Next-generation Sequencing and Bioinformatics for Plant Science

Edited by
Vijai Bhadauria

Caister Academic Press

Cataloguing Plant Genome Structural Variations

Xingtan Zhang[1,2], Xuequn Chen[1,2], Pingping Liang[1,2] and Haibao Tang[1,2]*

[1]Center for Genomics and Biotechnology, Fujian Provincial Key Laboratory of Haixia Applied Plant Systems Biology, Fujian Agriculture and Forestry University, Fuzhou, China.
[2]Haixia Institute of Science and Technology, Fujian Agriculture and Forestry University, Fuzhou, China. Email: zhangxt@fafu.edu.cn; chenxq@fafu.edu.cn; lppleiyu@gmail.com; and tanghaibao@gmail.com

*Correspondence: tanghaibao@gmail.com

https://doi.org/10.21775/9781910190654.12

Abstract

Structural variation (SV) is a type of genetic variation identified through the comparison of genome structures which often have direct and significant associations with phenotypic variations. Building on the next-generation sequencing (NGS) technologies, research on plant structural variations are gaining momentum and have revolutionized our view on the functional impact of the 'hidden' diversity that were largely understudied before. Herein, we first describe the current state of plant genomic SV research based on NGS and in particular focus on the biological insights gained from the large-scale identification of various types of plant SVs. Specific examples are chosen to demonstrate the genetic basis for phenotype diversity in model plant and major agricultural crops. Additionally, development of new genomic mapping technologies, including optical mapping and long read sequencing, as well as improved computational algorithms associated with these technologies have helped to pinpoint the exact nature and location of genomic SVs with much better resolution and precision. Future direction of plant research on SVs should focus on the population level to build a comprehensive catalogue of SVs, leading to full assessment of their impact on biological diversity.

Introduction

Structural variations (SVs) are a collection of complex genomic DNA mutations that differentiate among individuals in a certain population. In contrast to single nucleotide polymorphisms (SNPs) and short insertions and deletions (indels), SVs typically consist of DNA changes that are relatively long in size. The initial detection of SVs before the advent of the sequencing era is often based on detection of large scale chromosomal changes from karyotypic observation under microscope, including abnormal number of chromosomes such as aneuploidies (Jacobs *et al.*, 1959; Edwards *et al.*, 1960), chromosomal rearrangement (Bobrow *et al.*, 1971) and copy number variations (CNVs) (Bailey and Eichler 2006). Karyotypic mutations larger than 3 Mb in size can often be observed with *in situ* hybridization. Building on the next-generation sequencing (NGS) technology, we are moving towards a new phase of variant discovery that focus on identification of SVs with their boundaries mapped at single-base resolution. Extensive studies on the human genome have revealed that SVs play important roles in human health (Spielmann and Klopocki 2013; Weischenfeldt *et al.*, 2013). For example, Alzheimer's disease (AD) is a chronic neurodegenerative disease that usually badly affects the elder with symptoms such as problems with language, disorientation, mood swings and loss of motivation (Burns and Iliffe, 2009). Recent genetic research links the duplications of dosage-sensitive *APP* genes and Alzheimer's disease (Rovelet-Lecrux *et al.*, 2006). Similarly, higher *C4* copy number leads to more *C4* expression in the brain, contributing to a 1.4-fold risk to develop schizophrenia (Sekar *et al.*, 2016). These disease studies in human have indicated that abnormal copy number variation of certain genes significantly impact health and traits.

In addition to their leading role in causing a wide spectrum of human disease, structural variations are pervasive when comparing plant genomes and these changes are involved in many important biological processes (Saxena *et al.*, 2014; Żmieńko *et al.*, 2014). Structural variations represent an essential part in plant adaptive evolution and functional diversity. For example, re-sequencing of 50 rice accessions identified 21 genes with variable copy numbers that were annotated as potential resistance genes, suggesting that CNVs were involved in response to environmental cues and essential in their adaptability (Xu *et al.*, 2012). Additionally, presence or absence variations (PAVs) underlie the diversification in molecular functions that are shown to be beneficial in respective ecological niches. Sequencing of many individual ecotypes and cultivars in the population have vastly enriched our knowledge about the nature and impact of SVs over the past few years. Herein, we first describe the current state of plant SV research based on NGS as well as the biological implications of these genetic variations, categorized under several major types of structural variations. We then review several specific examples to show how SVs influence phenotypic diversity in plants in a case-study fashion. We then provide an overview of newly developed technologies, especially single molecular sequencing and genomic mapping that have become popular in recent years as powerful tools for the discovery of SVs with increasing resolution and accuracy. We conclude this review with several important directions in future SV studies in plants.

Large-scale detection of structural variations in plants

While different types of structural variations have been reviewed before, often with a human-centric view (Alkan *et al.*, 2011; Weischenfeldt *et al.*, 2013; Saxena *et al.*, 2014) – in this review we go through a list of SVs that represent the most abundant types in plant genomes, including copy number variations (CNVs), presence and absence variations (PAVs), mobile element insertions (MEIs) and homeologous exchanges (HEs).

Copy number variation

Copy number variations are variations of number of repeated sequences in the genome between individuals. Growing evidence support that CNVs are prevalent in plant genomes, which makes it the most extensively studied types of SVs in plants. CNV is defined as unbalanced changes in the genome structure and cover deletions (which is the same as copy number of zero) as well as duplications of > 50 bp in size (Girirajan *et al.*, 2011). Based on the data from the genomic hybridization (CGH) array and NGS technologies, a couple of plant CNVs (Table 11.1) were identified at genome-wide level (DeBolt, 2010; Swanson-Wagner *et al.*, 2010; Zheng *et al.*, 2011; Muñoz-Amatriaín *et al.*, 2013; Ping Yu *et al.*, 2013; Boocock *et al.*, 2015; Zhang *et al.*, 2015; Zhou *et al.*, 2015; Bai *et al.*, 2016; Cardone *et al.*, 2016; Hardigan *et al.*, 2016).

CNVs have been shown to play important roles in the biodiversity in plants, including adaptation to abiotic and biotic stress and yield-related traits that could be important targets for crop improvement. DeBolt performed CGH to detect CNVs in *Arabidopsis* grown under different temperature and stress regimes for the fifth-generation offspring from a common ancestor (DeBolt, 2010). Comparison between the siblings identified numerous CNVs, which account for approximately 0.38% of *Arabidopsis* annotated genes. The study found that many CNVs have effect on genome instability under environment with different abiotic inputs. Structural variations analysis from four soybean genotypes identified a list of genes overlapping with CNV regions. Functional annotation revealed that these genes were associated with disease resistance genes, suggesting that CNVs have contributed to the host defence against pathogens (McHale *et al.*, 2012). A recent report displayed that 30.2% of the potato genome were contained within CNV regions with nearly 30% of protein-coding genes affected. These CNVs are implicated in the expansions of stress-related gene families that vary between species and within species (Hardigan *et al.*, 2016). CNVs have

Table 11.1 Selected examples of genome-wide CNV studies in plants

Species	CNVs identification	Methods	Implication	Reference
Arabidopsis	Identified intersibling CNVs in *Arabidopsis* grown under different temperature and SA treatment for the fifth-generation plants from one common ancestor	CGH	This study uncovered that CNVs contributed to genome instability and shaped genome diversity in *Arabidopsis*	DeBolt, 2010
Rice	Identification of CNVs from 20 Asian cultivated rice comprising 6 *indica*, 3 *aus*, 2 *rayada*, 2 aromatic, 3 tropical *japonica* and 4 temperate *japonica* cultivars. Found 2886 CNV regions in total	CGH	Some rice CNVs might arise independently within different groups and contribute to group differences	Yu *et al*., 2013
Rice	Identified 9916 deletions compared to the reference genome Nipponbare and 2806 genes were affected by CNVs	NGS	Experimentally validated 28 functional CNV genes including *OsMADS56*, *BPH14*, *OsDCL2b* and *OsMADS30*, implying that CNVs might contributed to phenotypic variations in rice	Bai *et al*., 2016
Maize	Identified 479 genes exhibiting higher copy number in some genotypes and 3410 genes that have either fewer copies or are missing in specific genomes	CGH	Although 10% of Maize annotated genes exhibit CNV/PAV events relative to the B73 reference genome, the majority of SVs were observed in both maize and teosinte genomes, suggesting these variants predate maize domestication	Swanson-Wagner *et al*., 2010
Cucumber	Identified 26,788 SVs based on deep resequencing of 115 diverse accessions	NGS	These SVs affect the coding regions of 1676 genes, some of which are associated with cucumber domestication. Genome-wide association analysis revealed that a CNV defined the *Female* locus	Zhang *et al*., 2015
Potato	Identified 219.8 (30.2%) of potato genomes with nearly 30% of annotated genes were affected by CNVs	NGS	This study showed that CNV was the major component of the significant genomic diversity of clonally propagated potato and CNV-affected genes highly enriched in functions related to adaption	Hardigan *et al*., 2016
Barley	CNVs in Barley were account for 14.9% of genome and affected 9.5% of coding genes from 14 different accessions	CGH	CNV-affected genes are overlapping R genes and might contribute to the molecular mechanism for adaptation to biotic and abiotic stress in barley	Muñoz-Amatriaín *et al*., 2013
Apple	876 CNV regions were detected from 30 domesticated apple accessions, which spanned 3.5% of the apple genome	NGS	Resistance genes were significantly enriched in CNV regions	Boocock *et al*., 2015
Sorghum	Identified 17,111 CNVs in two sweet and one grain sorghum inbred lines, affecting 2600 genes	NGS	CNVs were significantly enriched in genes encoding cellulose synthases, pectinesterases, GRAS transcription factors, BTB/POZ and auxin-responsive proteins, suggesting that these genes might be involved in differentiation	Zheng *et al*., 2011
Soybean	A total of 302 soybean accessions, including 62 wild soybeans (*G. soja*), 130 landraces and 110 improved cultivars	NGS	(1) 162 CNVs were potentially involved in sorghum domestication and improvement. (2) GWAS showed that CNVs contributed to important traits, such as hilum colour and plant height	Zhou *et al*., 2015
Soybean	Four soybean genotype were used to detect CNVs and PAVs	CGH	This study identified a full set of CNV-affected genes were associated with disease resistance genes, and provided insight into the mechanisms of soybean resistance gene evolution	McHale *et al*., 2012
Grapevine	Identified 746 CNVs from four grapevine varieties and more than 2000 genes were affected by CNVs	NGS and CGH	Genes with polymorphisms were related to auxin/hormone response, berry growth and development	Cardone *et al*., 2016

also conduced to population differentiation. A total of 2886 CNV regions were detected from 20 Asian cultivated rice comprising six *indica*, three *aus*, two *rayada*, two *aromatic*, three tropical *japonica* and four temperate *japonica* cultivars. Comparison of various CNV events observed in *ssp. indica* and ssp. *japonica* expounded that 57.8% of the CNVs were only present in either ssp. *indica* or ssp. *japonica* (Yu *et al.*, 2013), suggesting that CNVs arose independently within different groups and might contribute to population differentiation.

Presence or absence variation

Gene contents often vary among even closely related individuals or cultivars, the union of which is termed 'pan-genome' (Vernikos *et al.*, 2015). The gene set present in all individuals is often called the 'core genome'. In contrast to core genome, the 'dispensable genome' consists of the sequences present only in a subset of individuals. The dispensable genome often contains a large number of PAVs. It has been hypothesized that the core genome is critical in the fundamental pathways that control plant growth and development, while the dispensable genome played important roles in phenotypic variation and adaptive evolution.

A robust way to compare PAV variation in the population is to perform sequencing on a large collection of individuals. A couple of recent publications that utilized NGS methods to detect PAV genes with important functions (Table 11.2). The largest data collected thus far come from the '*Arabidopsis* 1001 Genomes' project (Alonso-Blanco *et al.*, 2016) and the 'Rice 3000 Genomes' project (Zhikang *et al.*, 2014). Multiple *Arabidopsis* genomes were sequenced in the *Arabidopsis* 1001 Genomes project to identify both the core and dispensable genomes (Weigel and Mott, 2009; Cao *et al.*, 2011; Gan *et al.*, 2011). Tan and colleagues surveyed 80 *Arabidopsis* accessions and reported 2407 genes that were not present in reference genome (Tan *et al.*, 2012). Functional investigation revealed that largest proportions of these genes were related to membranes and binding, suggesting that the associated PAVs were involved in pattern recognition in molecular function. In addition, 31% PAV genes showed uneven distribution in chromosomes, which might be resulted from unequal recombination as the likely genetic mechanism underlying the origin of PAVs. A pan-genome study of three rice accessions showed that PAV genes were common in the *Oryza* species and ~10% of protein-coding genes were genome specific (Schatz *et al.*, 2014). Another method based on metagenome assembly was used to perform large-scale identification of dispensable genomes based on low-coverage NGS data (Yao *et al.*, 2015). Re-sequencing data from 1483 cultivated rice accessions were divided into two subpopulations: *indica* and *japonica*. Unmapped reads from each set of individuals were *de novo* assembled separately, with *indica*-related accessions or *japonica*-related accessions merged to generate a metagenome assembly. By mapping the respective 'population' assemblies to reference genome, this approach was able to identify PAVs associated with agronomic traits and specific metabolite profiling (Yao *et al.*, 2015). A recent pan-genome study in soybean re-sequenced and *de novo* assembled seven wild soybeans (Li *et al.*, 2014). Comparisons among wild soybean (*G. soja*) genomes and the reference genome (*G. max*) revealed 2.3 to 3.9 Mb dispensable sequences in wild soybean. These PAV genes typically had significantly higher *dN/dS* ratios, indicating that dispensable genes had undergone weaker purifying selection and/or greater positive selection compared to the core genes that are more conserved (Li *et al.*, 2014).

Mobile element insertion and deletion (MEI)

Transposable element (TE) insertions and deletions are also one of the most prevalent genetic variations in plants. Sequencing of 80 *Arabidopsis* accessions from eight populations explicated that ~80% of TEs were partially or completely absent from the genomes of at least one of the 80 individuals (Cao *et al.*, 2011). Crops with larger and more complex genomes typically contain a lot more TEs than the small genome of *Arabidopsis* or rice. Although first considered as selfish DNA or 'genome parasites', transposable elements have shown a dramatic impact on genome evolution and regulation of gene expression (Bennetzen and Wang, 2014). Accumulation of TEs, in particular LTR retrotransposons, were responsible for genome size variation (Vitte and Panaud, 2005; Hawkins *et al.*, 2006; Zedek *et al.*, 2010). For example, a study in genus *Eleocharis* revealed a positive correlation between *Ty1-copia* densities and genome size (Zedek *et al.*, 2010). El Baidouri and colleagues investigated 40 sequenced

Table 11.2 Selected examples of genome-wide PAV studies in plants

Species	PAVs identification	Methods	Implication	Reference
Arabidopsis	The authors investigated 80 re-sequenced *Arabidopsis* accessions and identified 2407 PAV genes	NGS	A large number of PAV genes were related to membranes, suggesting that these genes were involved in pattern recognition. In addition, 31% of PAV genes were unevenly distributed in chromosomes and showed a cluster pattern, suggesting that unequal recombination was a major mechanism for PAVs	Tan *et al.*, 2012
Soybean	*De novo* genome assembly of seven wild soybean identified 2.3–3.9 Mb of *G. soja*-specific dispensable genome	NGS	The PAV genes had significantly higher *dN/dS* ratios than the core genes, indicating that dispensable genes have undergone weaker purifying selection and/or greater positive selection than core genes. In addition, these PAV genes were related to receptor activity, structural molecule activity and antioxidant activity	Li *et al.*, 2014
Rice	A metagenome-like assembly strategy was used to identify rice dispensable genome from 1483 rice accessions	NGS	The new approach is powerful to identify dispensable genome using low-coverage NGS data. In addition, association mapping of rice dispensable genome found genes related to grain width and metabolic traits were located in dispensable genome, linking PAVs to important rice traits	Yao *et al.*, 2015
Potato	Re-sequencing of 12 potato monoploid and doubled monoploid potato clones identified ~7000 genes with PAVs	NGS	These dispensable genes were associated with limited transcription and/or a recent evolutionary history, with lower deletion frequency observed in genes conserved across angiosperms	Hardigan *et al.*, 2016
Maize	Transcriptome sequencing of seedlings from 503 maize uncovered 8681 representative transcript assemblies not present in the reference genome (B73)	NGS	This study investigated maize pan-transcriptome and showed that RNA-seq was also useful to uncover genetic variations and important phenotype related genes	Hirsch *et al.*, 2014

plant genomes and observed at least one case of horizontal TE transfer in 26 genomes and most of these TEs had remained functional after the transfer. Therefore, TE-driven genetic material exchange played an important role in genome evolution (El Baidouri *et al.*, 2014). It would be natural to expect that horizontal TE transfer might also be frequent between different individuals within species.

In addition to the impact on genome structure, TEs also have the capability to alter gene structure and expression. TE insertion could invade the space occupied by the protein-coding genes and disrupt the open reading frame (ORF), resulting in abnormal phenotypes (Chen *et al.*, 2005). TEs are also frequently observed to rewire existing regulatory networks by inserting into *cis*-regulatory elements. Study of DNA transposons *mPing* in rice genome showed that insertion of TEs have affected expression of 710 genes (Naito *et al.*, 2009). Once inserted into near-gene region, specific TE amplification could trigger the epigenetic silencing pathway through methylation or small RNAs (Hollister and Gaut, 2009; Joly-Lopez and Bureau, 2014), which regulated the gene expression in close proximity. Another study in soybean showed that a 5.7-kb transposon (*Tgm-Express I*) was inserted into intron 2 of the flavanone 3-hydroxylase gene (*F3H*), which converted purple flowers to pink (Zabala and Vodkin, 2005). Interestingly, the transposon consisted of five unrelated host gene fragments and the gene fragments were processed through a complex alternative splicing mechanism as added exons (Zabala and Vodkin, 2007).

Homeologous exchange (HE)

One type of structural variation has been relatively understudied until recently and is relatively unique to the plant genomes. In polyploid plant genomes, there has been sequence exchanges between homeologous chromosomes (often in different subgenomes in the same polyploid cell) leading to simultaneous gene deletions and amplification, or Homeologous Exchanges (HEs). For example, gene deletions and HEs between subgenomes in *Brassica napus* have led to the reduction of seed glucosinolate (GSL) content (Chalhoub *et al.*, 2014). Genes responsible for the flowering time (FLC) were greatly expanded from a single copy in *Arabidopsis* to nine copies in the oilseed genome with several expansion events due to HEs. The identified HEs also co-localize with known QTLs for vernalization requirements and flowering time. Interestingly, such HEs vary across different re-sequenced *B. napus* lines and even re-synthesized lines (Chalhoub *et al.*, 2014). Indeed, we can link much of the SVs that have occurred in *B. napus* to its unique adaptive and agronomic traits. Similarly, HEs occur frequently in the tetraploid cotton (*G. hirsutum*) (Li *et al.*, 2015). Re-sequencing of polyploid varieties or individuals could reveal additional genome restructuring events that might have been consciously or unconsciously selected for during the history of human cultivation and crop improvements.

Case studies: genetic basis for phenotype diversity contributed by plant SVs

Structural variations play important roles in phenotype diversity. In this part, we select a couple of examples that collectively demonstrated that how SVs influence a wide array of phenotypic variations, highlighting their critical role in the context of agricultural crops.

Sex determination

Cucumber is a major vegetable crop consumed worldwide and also serves as a model system for sex determination studies (Tanurdzic and Banks, 2004). Most cultivated cucumbers are monoecious, which contain male and female flowers on a single plant. In contrast, another type, gynoecious cucumber, bears only female flowers (Fig. 11.1A). Previous studies showed that gynoecy was associated with Mendelian locus *Female* (*F*-locus), which contained four protein-coding genes including aminocyclopropane-1-carboxylic acid synthase gene (*ACS1*), a truncated MYB transcription factor, a branched-chain amino acid aminotransferase and a gene with unknown function (Tanurdzic and Banks, 2004). Comparison between the gynoecious and monoecious cucumber genomes led to the discovery of a 30.2-kb duplicated segment that was significantly associated with gynoecy. This particular CNV event is the underlying genetic cause for the variation at the *Female* locus that has led to the development of gynoecy in cucumber (Zhang *et al.*, 2015).

Response to abiotic stress response

Algerian barley landrace Sahara showed superior boron-toxicity tolerance (Fig. 11.1B), when compared to intolerant genotypes (Sutton *et al.*, 2007). It has been reported that boron-tolerance is associated with the ability to maintain a low level of boron concentration in the shoot. Sutton and colleagues identified an efflux transporter (*Bot1*) as the candidate gene for the boron-toxicity tolerance. Comparison between Sahara and Clipper genomes revealed that Sahara contained 4x more copies of *Bot1* than Clipper, explaining their differential response to the toxic boron. Moreover, mRNA expression of this gene in Sahara were substantially higher than that in Clipper, consistent with the higher copy number of *Bot1*. Expression of the Sahara *Bot1* gene in yeast indicated that *Bot1* provided boron tolerance in yeast under high level treatment of H_3BO_3. These results, starting from the identification of the CNV at the *Bot1* locus to the functional validation, demonstrated that higher copy number of *Bot1* gene conferred boron tolerance in barley Sahara (Sutton *et al.*, 2007). Similarly, genic copy number variation that contributes to aluminium (Al) tolerance is found in maize (Maron *et al.*, 2013). Three tandem multidrug and toxic compound extrusion 1 (*MATE1*) copies were identified in a maize recombinant inbred line population. The expression of *MATE1* conferred a significant increase in Al tolerance and root citrate exudation in response to Al. Consequently, *MATE1* copy number was associated with its mRNA expression, which in turn resulted in superior Al tolerance in maize (Maron *et al.*, 2013).

Response to biotic stress

Soybean cyst nematode (SCN) is one of the most economically damaging pathogens for soybean.

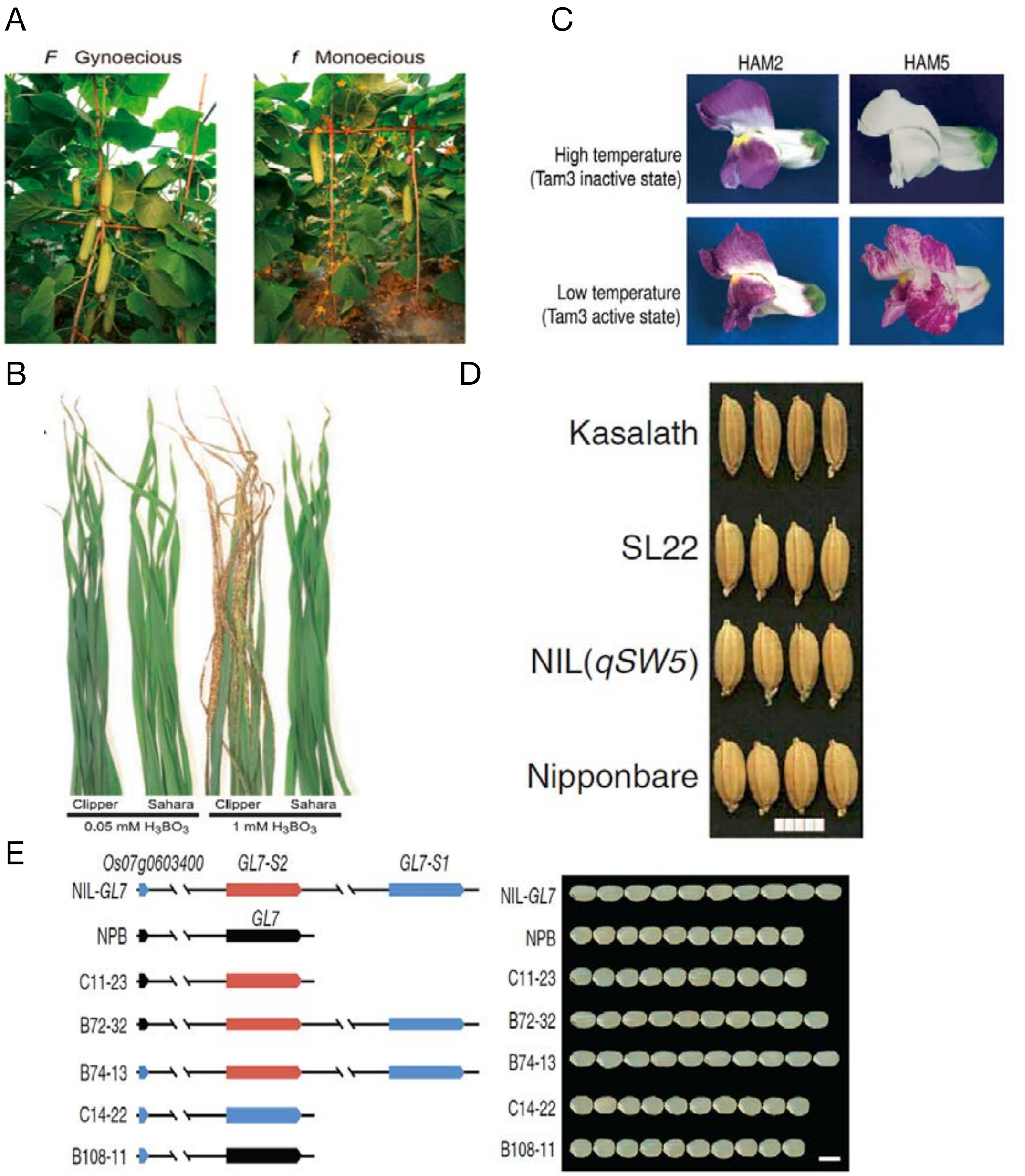

Figure 11.1 Functional impact of structural variations in plant genomes. (A) Phenotypes of gynoecious (WI1983G) and monoecious (WI1983GM) cucumbers. Monoecious cucumbers contain both male and female flowers on a single plant, while gynoecious cucumbers contain 30.2-kb duplication in their genomes that lead to the development of female flowers only (Zhang *et al*., 2015). (B) Boron-toxicity symptoms in leaf blades of boron-intolerant (Clipper) and boron-tolerant (Sahara) barley plants. Compared with intolerant genotypes, Sahara contains four times more copies of *Bot1*, which encodes an efflux transporter and confers tolerance to boron-toxicity (Sutton *et al*., 2007). (C) Different flower pigmentation patterns in *Antirrhinum majus* lines HAM2 and HAM5. HAM2 only contains one copy of transposon *Tam3* in the upstream of *niv* gene, which is responsible for pale-red petal colour. In contrast, HAM5 contains two copies of *Tam3* in the upstream and downstream of *niv* gene, respectively. The paired *Tam3* inhibits the expression of *niv* and results in white petal colour (Uchiyama *et al*., 2013). (D) Phenotypic variation among rice Kasalath, SL22, NIL (*qSW5*) and Nipponbare accessions. Nipponbare contains a 1,212-bp deletion in the *qSW5* region and shows increased grain size (Shomura *et al*., 2008). (E) Rice grain length diversity is associated with copy number variations in *GL7* locus (Wang *et al*., 2015). Images were based on the respective studies (Sutton *et al*., 2007; Shomura *et al*., 2008; Uchiyama *et al*., 2013; Wang *et al*., 2015; Zhang *et al*., 2015), and reproduced with permission.

QTL analysis located that *Rhg1* regions consistently show a large contribution to SCN resistance in SCN-resistant cultivated soybean (Concibido *et al.*, 2004). *Rhg1* locus is a 31-kb segment which encodes multiple gene products. In susceptible varieties, only one copy of this 31-kb segment was observed. In contrast, ten tandem copies were identified in SCN-resistant soybeans. Overexpression of the cluster genes in *Rhg1* regions conferred enhanced SCN resistant, therefore confirming that copy number variation of *Rhg1* is associated with SCN resistance in soybean (Cook *et al.*, 2012).

Flower colour

Flowers attract much attention due to their diversity in plants, and also for their critical role in agricultural production. A recent study shows that transposon is one of the major factors that contributes to the floral pigmentation and flower colour in *Antirrhinum* (Clegg and Durbin, 2003; Wei and Cao, 2016). The *nivea* (*niv*) gene encodes chalcone synthase (*CHS*) and is responsible for flower pigmentation (Sommer *et al.*, 1985). Expression of *niv* gene resulted in pale-red petal colour; while suppression of *niv* mRNA level lead to white colour. In HAM5 accession, *Tam3* transposons are inserted into both upstream and downstream of *niv* gene, and the corresponding inserted sequences paired and formed a loop structure under high temperature, thus preventing the binding of correlated transcription factor. Therefore, the transposon insertion inhibited the expression of *niv* gene, resulting in white petal colour. In contrast, only one *Tam3* transposon was observed in the promoter of *niv* gene in HAM2 line. The absence of the loop structure maintained an active *niv* expression and led to pale-red petal colour (Fig. 11.1C) (Uchiyama *et al.*, 2013).

Grain size

Grain size is one of the most agronomically important trait for cereal domestication. Recent reports demonstrated that both CNV and PAV contributed to rice grain dimension as well as increased yields (Shomura *et al.*, 2008; Wang *et al.*, 2015). In an F_2 cross population between Nipponbare (*japonica*) and Kasalath (*indica*) cultivars, Izawa and his team identified several QTLs responsible for grain width. One of the QTLs, *qSW5* explained 38.5% of variation in the F_2 population (Shomura *et al.*, 2008). Compared to Kasalath, Nipponbare has increased grain size and its genome contained a specific 1212-bp deletion that appeared to be associated with grain width. Two rice lines were tested in confirm this association – one line with substitution of Kasalath chromosome 5 in a Nipponbare genetic background (SL22) and another nearly isogenic line (NIL) that both contained around 90 kb of Kasalath fragments of the *qSW5* region in a Nipponbare background. Both rice lines, in which *qSW5* region was present, showed relatively reduced rice grain size compared to Nipponbare (Fig. 11.1D). For grain length, another rice study revealed that tandem duplications of 17.1-kb segment containing *GL7* locus contributed to the increase in grain length and improvement of appearance quality (Wang *et al.*, 2015). Rice lines (NIL-GL7, B72–32 and B74–13) harbour multiple copies of *GL7* as well as increased grain width. In contrast, rice lines (NPB, C11–23, C14–22 and B108–11) that only harbour one copy of *GL7* displayed relatively narrow grain (Fig. 11.1E). *GL7* encodes a homologous protein to the LONGIFOLIA proteins, which is known to regulate longitudinal cell elongation in *Arabidopsis*. These two studies – with PAV controlling grain width and CNV controlling grain length – provide classical examples that specific SVs were probably selected for during the breeding history of rice.

New experimental and computational approaches to detect SVs

Structural variations attracted much attention in the genomics field in the past decade and a number of methods have been developed on the basis of large amounts of NGS data. Most of these efforts focus on the careful analyses of alignment of short reads – such as read depth, split reads (reads that map to multiple distinct locations), and discordant read pairs (paired-end reads that show an abnormal distance between the two ends) (Zhao *et al.* 2013). Many of these alignment-based approaches have been extensively reviewed elsewhere (Escaramís *et al.*, 2015; Tattini *et al.*, 2015). We instead focus on the more recent genome mapping and sequencing technologies, including optical mapping and PacBio sequencing that show promise in the discovery of SVs that were intrinsically difficult to do with only short reads.

Optical mapping

Optical mapping is an approach that combines DNA tagging with imaging to produce an ordered map of restriction sites or sequence motifs from a single linearized DNA molecule (Chaney *et al.*, 2016). There are two commercial vendors with slightly different variation of conceptually similar technologies – OpGen and BioNano (Tang *et al.*, 2015). We will briefly review the OpGen method here. Long DNA molecules are linearized and stretched first, followed by restriction enzyme digestion. Images of the resulting digested fragments are then captured with a camera, allowing a precise measurement of the distance between restriction sites. The resulting restriction map can be viewed as unique 'barcodes' for different regions in the genome (Lam *et al.*, 2012). Since optical mapping has the ability to barcode long DNA single molecules, this technology was initially used for the improvement of *de novo* genome assemblies (Tang *et al.*, 2015). Naturally, optical mapping is also a very powerful tool to detect large structural variations among individuals in a population. For example, Mak and colleagues generated genome maps for a trio family with long, fluorescently labelled DNA molecules using the BioNano Irys platform (Mak *et al.*, 2016). Comparison of these maps between closely related individuals and the reference human genome led to a comprehensive catalogue of relatively large structural variations (> 5 kb in size), including insertion and deletion, inversion and CNVs. Although so far very few studies plant structural variations were studied using optical mapping, there have been emerging proposals and pilot studies to map many individuals using optical mapping related technologies (Chaney *et al.*, 2016).

PacBio single-molecule sequencing

Although the second-generation sequencing technologies are widely used in genomics, including *de novo* genome assembly and detection of small-scale genetic variations, their apparent disadvantages, namely the GC bias and short read length, make them unsuitable for calling variations that are larger scale or more complex in nature (Rhoads and Au, 2015). Repetitive sequences in complex genomes can often lead to multiple, ambiguous mapping of short reads that misidentifies structural variations. The alternative PacBio sequencing technology offers long, single-molecule sequencing and produces much more robust data for *de novo* assembly and the characterization of structural variations. Chaisson and colleagues sequenced and analysed a haploid human genome (CHM1) using PacBio platform (Chaisson *et al.*, 2015). Comparative analysis of the PacBio assembled genome and human GRCh37 reference genome resolved the complete sequences of 26,079 euchromatic structural variants at the base-pair level, most of which were not previously published (Chaisson *et al.*, 2015). Combining Illumina and PacBio technologies, Korbinian and colleagues generated *Arabidopsis thaliana* Landsberg erecta (Ler) assembly at chromosome-level (Dong *et al.*, 2016). Whole-genome comparison of the Ler assembly against the TAIR10 reference genome identified a number of SVs, including transpositions, inversions, rearranged regions as well as PAVs. In addition, the authors highlighted one case of 1.2 Mb inversion and postulated that this inversion might have reduced the level of meiotic recombination that could lead to the genetically isolated haplotypes in the worldwide population of *A. thaliana* (Dong *et al.*, 2016).

Algorithmic advancement to detect structural variations

It is important to point out that the increased power and sensitivity achieved with the PacBio technology to detect SVs would be impossible without the algorithmic advances in the analyses of read alignments. Ritz *et al.* proposed a new algorithm, MultiBreak-SV, that tolerated high error rates of PacBio reads and employed a probabilistic approach to consider all possible solutions for break-reads alignments at the same time (Ritz *et al.*, 2014). Another mapping-based SV detection method (PB Honey) proposed two strategies, interrupted long-read mapping (PBHoney-Tails) and intra-read discordance (PBHoney-Spots) (English *et al.*, 2014). PBHoney-Tails first extracts and re-aligns unmapped tails from mapped long reads. By analysing piece-alignments with similarly mapped tails based on their respective location and orientation, one can then annotate each cluster as either a deletion, insertion, or translocation and predict breakpoints using the average interrupted position of each read. By leveraging the experimentally determined 15% per-base error rate, PBHoney-Spots was able to identify discordant 'spots' within the reference where the error rate was higher than expected (intra-read discordance).

Due to the overlapping goals of genome assembly and SV discovery, software used to assess the quality of genome assemblies could also be useful to discover SVs. For instance, Assemblytics (Nattestad and Schatz, 2016) was able to identify six classes of variants based on distinct alignment signatures between related genome assemblies.

Future perspective

With recent technological and algorithmic advancements, much of the previously hidden genetic variations could be discovered across different plant taxa and crop varieties. Future studies of SVs should focus on the integration of various studies and databases, and associate the SVs with specific traits and phenotypes, in order to gain a comprehensive understanding of the genetic underpinnings of a rich set of biological functions.

Population analysis on a grand scale

Population analysis of SNPs have been extensively studied in plants over past two decades since SNP is perhaps the easiest form of genomic variation to compare across individuals. However, most SV studies have been only limited to a small number of individuals and their population genetics is very much in its infancy. Recent CNV research in human sequenced 236 individuals, representing 125 distinct human populations and observed that duplications exhibit drastically different genetic and selective signatures between populations. For example, comparative analysis of CNVs at the population level found that large duplications that introgressed from the extinct Denisova lineage were exclusively present in Oceanic populations (Sudmant *et al.*, 2015). In plants, population analysis of 302 soybean re-sequencing lines showed that many CNVs could be targets of artificial selection in addition to the SNPs. Using a metric called relative frequency difference (RFD) to prioritize CNVs, the authors identified 162 potentially CNVs that are potential targets for human selection during domestication and improvement. Almost all of the discovered CNVs overlap with known domestication-related QTL regions (Zhou *et al.*, 2015).

More and better GWAS studies

Building on the ever-increasing catalogue of SVs, it is natural to extend the statistical tests of Genome-Wide Association Studies (GWAS) to SVs. It has been suggested that many SVs could well account for the 'missing heritability' (Manolio *et al.*, 2009). In 2008, McCarroll already proposed extension of GWAS to CNVs since SNP-based GWAS only explain a modest fraction (2–15%) of heritable variation for certain disease risk (McCarroll, 2008). With more re-sequencing data, it has become more feasible to extend the gene–trait association across all genetic variants discovered – from simple SNPs to the more complex SVs. In a pilot study, Zhou and colleagues performed GWAS tests using CNV data identified by re-sequencing 302 soybeans and uncovered genes with varying copy numbers to be associated with cyst nematode resistance and hilum colour (Zhou *et al.*, 2015). GWAS with SVs was also performed in cucumber and seven SV loci were identified to be significantly related with the tuberculate fruit trait, with one matching locus from a previous study (Zhang *et al.*, 2015). While still rudimentary and relatively small-scale in terms of sample size, these studies did show promising applications of GWAS tests that are directly performed on SVs. Better SV-related GWAS analyses should become routinely performed at the population level for all re-sequencing projects, with the ultimate goal to identify much more gene–trait associations than previously suggested by SNPs alone in plants.

Acknowledgements

We thank the Fujian provincial government for a Fujian '100 Talent Plan' award to HT. This work was supported by the National Key Research and Development Program of China (2016YFD0100305).

References

1001 Genomes Consortium. Electronic address: magnus.nordborg@gmi.oeaw.ac.at. (2016). 1,135 Genomes Reveal the Global Pattern of Polymorphism in *Arabidopsis thaliana*. Cell *166*, 481–491. http://dx.doi.org/10.1016/j.cell.2016.05.063

Alkan, C., Coe, B.P., and Eichler, E.E. (2011). Genome structural variation discovery and genotyping. Nat. Rev. Genet. *12*, 363–376. http://dx.doi.org/10.1038/nrg2958

Bai, Z., Chen, J., Liao, Y., Wang, M., Liu, R., Ge, S., Wing, R.A., and Chen, M. (2016). The impact and origin of copy number variations in the *Oryza* species. BMC Genomics *17*, 261. http://dx.doi.org/10.1186/s12864-016-2589-2

Bailey, J.A., and Eichler, E.E. (2006). Primate segmental duplications: crucibles of evolution, diversity and disease. Nat. Rev. Genet. *7*, 552–564.

Bennetzen, J.L., and Wang, H. (2014). The contributions of transposable elements to the structure, function, and evolution of plant genomes. Annu. Rev. Plant Biol. *65*, 505–530. http://dx.doi.org/10.1146/annurev-arplant-050213-035811

Bobrow, M., Joness, L.F., and Clarke, G. (1971). A complex chromosomal rearrangement with formation of a ring 4. J. Med. Genet. *8*, 235–239.

Boocock, J., Chagné, D., Merriman, T.R., and Black, M.A. (2015). The distribution and impact of common copy-number variation in the genome of the domesticated apple, Malus x domestica Borkh. BMC Genomics *16*, 848. http://dx.doi.org/10.1186/s12864-015-2096-x

Burns, A., and Iliffe, S. (2009). Alzheimer's disease. BMJ *338*, b158. http://dx.doi.org/10.1136/bmj.b158

Cao, J., Schneeberger, K., Ossowski, S., Günther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O., Lippert, C., *et al.* (2011). Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. Nat. Genet. *43*, 956–963. http://dx.doi.org/10.1038/ng.911

Cardone, M.F., D'Addabbo, P., Alkan, C., Bergamini, C., Catacchio, C.R., Anaclerio, F., Chiatante, G., Marra, A., Giannuzzi, G., Perniola, R., *et al.* (2016). Inter-varietal structural variation in grapevine genomes. Plant J. *88*, 648–661. http://dx.doi.org/10.1111/tpj.13274

Chaisson, M.J., Huddleston, J., Dennis, M.Y., Sudmant, P.H., Malig, M., Hormozdiari, F., Antonacci, F., Surti, U., Sandstrom, R., Boitano, M., *et al.* (2015). Resolving the complexity of the human genome using single-molecule sequencing. Nature *517*, 608–611. http://dx.doi.org/10.1038/nature13907

Chalhoub, B., Denoeud, F., Liu, S., Parkin, I.A., Tang, H., Wang, X., Chiquet, J., Beleram, H., Tong, C., Samans, B., *et al.* (2014). Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. Science *345*, 950–953.

Chaney, L., Sharp, A.R., Evans, C.R., and Udall, J.A. (2016). Genome mapping in plant comparative genomics. Trends Plant Sci. *21*, 770–780. http://dx.doi.org/10.1016/j.tplants.2016.05.004

Chen, J.M., Stenson, P.D., Cooper, D.N., and Ferec, C. (2005). A systematic analysis of LINE-1 endonuclease-dependent retrotranspositional events causing human genetic disease. Hum. Genet. *117*, 411–427.

Clegg, M.T., and Durbin, M.L. (2003). Tracing floral adaptations from ecology to molecules. Nat. Rev. Genet. *4*, 206–215. http://dx.doi.org/10.1038/nrg1023

Concibido, V.C., Diers, B.W., and Arelli, P.R. (2004). A decade of QTL mapping for cyst nematode resistance in soybean. Crop Sci. *44*, 1121–1131.

Cook, D.E., Lee, T.G., Guo, X., Melito, S., Wang, K., Bayless, A.M., Wang, J., Hughes, T.J., Willis, D.K., Clemente, T.E., *et al.* (2012). Copy number variation of multiple genes at Rhg1 mediates nematode resistance in soybean. Science *338*, 1206–1209. http://dx.doi.org/10.1126/science.1228746

DeBolt, S. (2010). Copy number variation shapes genome diversity in *Arabidopsis* over immediate family generational scales. Genome Biology and Evolution *2*, 441–453.

Dong, J., Feng, Y., Kumar, D., Zhang, W., Zhu, T., Luo, M.C., and Messing, J. (2016). Analysis of tandem gene copies in maize chromosomal regions reconstructed from long sequence reads. Proc. Natl. Acad. Sci. U.S.A. *113*, 7949–7956. http://dx.doi.org/10.1073/pnas.1608775113

Edwards, J.H., Harnden, D.G., Cameron, A.H., Crosse, V.M., and Wolff, O.H. (1960). A new trisomic syndrome. Lancet *1*, 787–790.

El Baidouri, M., Carpentier, M.C., Cooke, R., Gao, D., Lasserre, E., Llauro, C., Mirouze, M., Picault, N., Jackson, S.A., and Panaud, O. (2014). Widespread and frequent horizontal transfers of transposable elements in plants. Genome Res. *24*, 831–838. http://dx.doi.org/10.1101/gr.164400.113

English, A.C., Salerno, W.J., and Reid, J.G. (2014). PBHoney: identifying genomic variants via long-read discordance and interrupted mapping. BMC Bioinf. *15*, 180. http://dx.doi.org/10.1186/1471-2105-15-180

Escaramís, G., Docampo, E., and Rabionet, R. (2015). A decade of structural variants: description, history and methods to detect structural variation. Brief. Funct. Genomics. *14*, 305–314. http://dx.doi.org/10.1093/bfgp/elv014

Gan, X., Stegle, O., Behr, J., Steffen, J.G., Drewe, P., Hildebrand, K.L., Lyngsoe, R., Schultheiss, S.J., Osborne, E.J., Sreedharan, V.T., *et al.* (2011). Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. Nature *477*, 419–423. http://dx.doi.org/10.1038/nature10414

Girirajan, S., Campbell, C.D., and Eichler, E.E. (2011). Human copy number variation and complex genetic disease. Annu. Rev. Genet. *45*, 203–226. http://dx.doi.org/10.1146/annurev-genet-102209-163544

Sommer, H., Carpenter, R., Harrison, B.J., and Saedler, H. (1985). The transposable element Tam3 of *Antirrhinum majus* generates a novel type of sequence alterations upon excision. Mol. Gen. Genet. *199*, 225–231.

Hardigan, M.A., *et al.* (2016). 'Genome reduction uncovers a large dispensable genome and adaptive role for copy number variation in asexually propagated Solanum tuberosum.' Plant Cell

Hawkins, J.S., Kim, H., Nason, J.D., Wing, R.A., and Wendel, J.F. (2006). Differential lineage-specific amplification of transposable elements is responsible for genome size variation in Gossypium. Genome Res. *16*, 1252–1261.

Hirsch, C.N., Foerster, J.M., Johnson, J.M., Sekhon, R.S., Muttoni, G., Vaillancourt, B., Peñagaricano, F., Lindquist, E., Pedraza, M.A., Barry, K., *et al.* (2014). Insights into the maize pan-genome and pan-transcriptome. Plant Cell *26*, 121–135. http://dx.doi.org/10.1105/tpc.113.119982

Hollister, J.D., and Gaut, B.S. (2009). Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. Genome Res. *19*, 1419–1428. http://dx.doi.org/10.1101/gr.091678.109

Jacobs, P.A., Baikie, A.G., Court Brown, W.M., and Strong, J.A. (1959). The somatic chromosomes in mongolism. Lancet *1*, 710.

Joly-Lopez, Z., and Bureau, T.E. (2014). Diversity and evolution of transposable elements in *Arabidopsis*. Chromosome. Res. *22*, 203–216. http://dx.doi.org/10.1007/s10577-014-9418-8

Lam, E.T., Hastie, A., Lin, C., Ehrlich, D., Das, S.K., Austin, M.D., Deshpande, P., Cao, H., Nagarajan, N., Xiao, M., *et al.* (2012). Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. Nat. Biotechnol. *30*, 771–776.

Li, F., Fan, G., Lu, C., Xiao, G., Zou, C., Kohel, R.J., Ma, Z., Shang, H., Ma, X., Wu, J., *et al.* (2015). Genome sequence of cultivated Upland cotton (*Gossypium hirsutum* TM-1) provides insights into genome evolution. Nat. Biotechnol. *33*, 524–530. http://dx.doi.org/10.1038/nbt.3208

Li, Y.H., Zhou, G., Ma, J., Jiang, W., Jin, L.G., Zhang, Z., Guo, Y., Zhang, J., Sui, Y., Zheng, L., *et al.* (2014). De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. Nat. Biotechnol. *32*, 1045–1052. http://dx.doi.org/10.1038/nbt.2979

Mak, A.C., Lai, Y.Y., Lam, E.T., Kwok, T.P., Leung, A.K., Poon, A., Mostovoy, Y., Hastie, A.R., Stedman, W., Anantharaman, T., *et al.* (2016). Genome-wide structural variation detection by genome mapping on nanochannel arrays. Genetics *202*, 351–362. http://dx.doi.org/10.1534/genetics.115.183483

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., *et al.* (2009). Finding the missing heritability of complex diseases. Nature *461*, 747–753. http://dx.doi.org/10.1038/nature08494

Maron, L.G., Guimarães, C.T., Kirst, M., Albert, P.S., Birchler, J.A., Bradbury, P.J., Buckler, E.S., Coluccio, A.E., Danilova, T.V., Kudrna, D., *et al.* (2013). Aluminum tolerance in maize is associated with higher MATE1 gene copy number. Proc. Natl. Acad. Sci. U.S.A. *110*, 5241–5246. http://dx.doi.org/10.1073/pnas.1220766110

McCarroll, S.A. (2008). Extending genome-wide association studies to copy-number variation. Hum. Mol. Genet. *17*, R135–42. http://dx.doi.org/10.1093/hmg/ddn282

McHale, L.K., Haun, W.J., Xu, W.W., Bhaskar, P.B., Anderson, J.E., Hyten, D.L., Gerhardt, D.J., Jeddeloh, J.A., and Stupar, R.M. (2012). Structural variants in the soybean genome localize to clusters of biotic stress-response genes. Plant Physiol. *159*, 1295–1308. http://dx.doi.org/10.1104/pp.112.194605

Muñoz-Amatriaín, M., Eichten, S.R., Wicker, T., Richmond, T.A., Mascher, M., Steuernagel, B., Scholz, U., Ariyadasa, R., Spannagl, M., Nussbaumer, T., *et al.* (2013). Distribution, functional impact, and origin mechanisms of copy number variation in the barley genome. Genome Biol. *14*, R58. http://dx.doi.org/10.1186/gb-2013-14-6-r58

Naito, K., Zhang, F., Tsukiyama, T., Saito, H., Hancock, C.N., Richardson, A.O., Okumoto, Y., Tanisaka, T., and Wessler, S.R. (2009). Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. Nature *461*, 1130–1134. http://dx.doi.org/10.1038/nature08479

Nattestad, M., and Schatz, M.C. (2016). Assemblytics: a web analytics tool for the detection of variants from an assembly. Bioinformatics 32, 3021–3023. http://dx.doi.org/10.1093/bioinformatics/btw369

Yu, P., Wang, C.H., Xu, Q., Feng, Y., Yuan, X.P., Yu, H.Y., Wang, Y.P., Tang, S.X., and Wei, X.H. (2013). Genome-wide copy number variations in Oryza sativa L. BMC Genomics *14*, 649. http://dx.doi.org/10.1186/1471-2164-14-649

Rhoads, A., and Au, K.F. (2015). PacBio sequencing and its applications. genomics. proteomics. Bioinformatics. *13*, 278–289. http://dx.doi.org/10.1016/j.gpb.2015.08.002

Ritz, A., Bashir, A., Sindi, S., Hsu, D., Hajirasouliha, I., and Raphael, B.J. (2014). Characterization of structural variants with single molecule and hybrid sequencing approaches. Bioinformatics *30*, 3458–3466. http://dx.doi.org/10.1093/bioinformatics/btu714

Rovelet-Lecrux, A., Hannequin, D., Raux, G., Le Meur, N., Laquerriere, A., Vital, A., Dumachin, Cl, Feuillette, S., Brice, A., Vercelletto, M., *et al.* (2006). APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. Nat. Genet. *38*, 24–26.

Saxena, R.K., Edwards, D., and Varshney, R.K. (2014). Structural variations in plant genomes. Brief. Funct. Genomics. *13*, 296–307. http://dx.doi.org/10.1093/bfgp/elu016

Schatz, M.C., Maron, L.G., Stein, J.C., Hernandez Wences, A., Gurtowski, J., Biggers, E., Lee, H., Kramer, M., Antoniou, E., Ghiban, E., *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.

Sekar, A., Bialas, A.R., de Rivera, H., Davis, A., Hammond, T.R., Kamitaki, N., Tooley, K., Presumey, J., Baum, M., Van Doren, V., *et al.* (2016). Schizophrenia risk from complex variation of complement component 4. Nature *530*, 177–183. http://dx.doi.org/10.1038/nature16549

Tan, S., Zhong, Y., Hou, H., Yang, S., and Tian, D. (2012). Variation of presence/absence genes among *Arabidopsis* populations. BMC Evol. Biol. *12*, 86. http://dx.doi.org/10.1186/1471-2148-12-86

Shomura, A., Izawa, T., Ebana, K., Ebitani, T., Kanegae, H., Konishi, S., and Yano, M. (2008). Deletion in a gene associated with grain size increased yields during rice domestication. Nat. Genet. *40*, 1023–1028. http://dx.doi.org/10.1038/ng.169

Spielmann, M., and Klopocki, E. (2013). CNVs of noncoding cis-regulatory elements in human disease. Curr. Opin. Genet. Dev. *23*, 249–256. http://dx.doi.org/10.1016/j.gde.2013.02.013

Sudmant, P.H., Mallick, S., Nelson, B.J., Hormozdiari, F., Krumm, N., Huddleston, J., Coe, B.P., Baker, C., Nordenfelt, S., Bamshad, M., *et al.* (2015). Global diversity, population stratification, and selection of human copy-number variation. Science *349*, aab3761. http://dx.doi.org/10.1126/science.aab3761

Sutton, T., Baumann, U., Hayes, J., Collins, N.C., Shi, B.J., Schnurbusch, T., Hay, A., Mayo, G., Pallotta, M., Tester, M., *et al.* (2007). Boron-toxicity tolerance in barley arising from efflux transporter amplification. Science *318*, 1446–1449.

Swanson-Wagner, R.A., Eichten, S.R., Kumari, S., Tiffin, P., Stein, J.C., Ware, D., and Springer, N.M. (2010). Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. Genome Res. *20*, 1689–1699. http://dx.doi.org/10.1101/gr.109165.110

Tang, H., Lyons, E., and Town, C.D. (2015). Optical mapping in plant comparative genomics. GigaScience *4*, 3. http://dx.doi.org/10.1186/s13742-015-0044-y

Tanurdzic, M., and Banks, J.A. (2004). Sex-determining mechanisms in land plants. Plant Cell *16*, S61–71. http://dx.doi.org/10.1105/tpc.016667

Tattini, L., D'Aurizio, R., and Magi, A. (2015). Detection of genomic structural variants from next-generation sequencing data. Front. Bioeng. Biotechnol. *3*, 92. http://dx.doi.org/10.3389/fbioe.2015.00092

Uchiyama, T., Hiura, S., Ebinuma, I., Senda, M., Mikami, T., Martin, C., and Kishima, Y. (2013). A pair of transposons coordinately suppresses gene expression, independent of pathways mediated by siRNA in Antirrhinum. New Phytol. *197*, 431–440. http://dx.doi.org/10.1111/nph.12041

Vernikos, G., Medini, D., Riley, D.R., and Tettelin, H. (2015). Ten years of pan-genome analyses. Curr. Opin. Microbiol. *23*, 148–154. http://dx.doi.org/10.1016/j.mib.2014.11.016

Vitte, C., and Panaud, O. (2005). LTR retrotransposons and flowering plant genome size: emergence of the increase/decrease model. Cytogenet. Genome. Res. *110*, 91–107.

Wang, Y., Xiong, G., Hu, J., Jiang, L., Yu, H., Xu, J., Fang, Y., Zeng, L., Xu, E., Xu, J., *et al.* (2015). Copy number variation at the GL7 locus contributes to grain size diversity in rice. Nat. Genet. *47*, 944–948. http://dx.doi.org/10.1038/ng.3346

Wei, L., and Cao, X. (2016). The effect of transposable elements on phenotypic variation: insights from plants to humans. Sci. China Life Sci. *59*, 24–37. http://dx.doi.org/10.1007/s11427-015-4993-2

Weigel, D., and Mott, R. (2009). The 1001 genomes project for *Arabidopsis thaliana*. Genome Biol. *10*, 107. http://dx.doi.org/10.1186/gb-2009-10-5-107

Weischenfeldt, J., Symmons, O., Spitz, F., and Korbel, J.O. (2013). Phenotypic impact of genomic structural variation: insights from and for human disease. Nat. Rev. Genet. *14*, 125–138. http://dx.doi.org/10.1038/nrg3373

Xu, X., Liu, X., Ge, S., Jensen, J.D., Hu, F., Li, X., Dong, Y., Gutenkunst, R.N., Fang, L., Huang, L., *et al.* (2011). Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nat. Biotechnol. *30*, 105–111. http://dx.doi.org/10.1038/nbt.2050

Yao, W., Li, G., Zhao, H., Wang, G., Lian, X., and Xie, W. (2015). Exploring the rice dispensable genome using a metagenome-like assembly strategy. Genome Biol. *16*, 187. http://dx.doi.org/10.1186/s13059-015-0757-3

Zabala, G., and Vodkin, L. (2007). Novel exon combinations generated by alternative splicing of gene fragments mobilized by a CACTA transposon in Glycine max. BMC Plant Biol. *7*, 38.

Zabala, G., and Vodkin, L.O. (2005). The wp mutation of Glycine max carries a gene-fragment-rich transposon of the CACTA superfamily. Plant Cell *17*, 2619–2632.

Zedek, F., Smerda, J., Smarda, P., and Bureš, P. (2010). Correlated evolution of LTR retrotransposons and genome size in the genus Eleocharis. BMC Plant Biol. *10*, 265. http://dx.doi.org/10.1186/1471-2229-10-265

Zhang, Z., et al. (2015). 'Genome-Wide Mapping of Structural Variations Reveals a Copy Number Variant That Determines Reproductive Morphology in Cucumber.' Plant Cell **27**(6): 1595–1604.

Zhao, M., Wang, Q., Wang, Q., Jia, P., and Zhao, Z. (2013). Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. BMC Bioinf. *14* (Suppl. 11), S1. http://dx.doi.org/10.1186/1471-2105-14-S11-S1

Zheng, L.Y., Guo, X.S., He, B., Sun, L.J., Peng, Y., Dong, S.S., Liu, T.F., Jiang, S., Ramachandran, S., Liu, C.M., *et al.* (2011). Genome-wide patterns of genetic variation in sweet and grain sorghum (Sorghum bicolor). Genome Biol. *12*, R114. http://dx.doi.org/10.1186/gb-2011-12-11-r114

Zhikang, L., Fu, B.Y., Gao, Y.M., Wang, W.S., Xu, J.L., Zhang, F., Zhao, X.Q., Zhao, T.Q., Zheng, T.Q., Zhou, Y.L., *et al.* (2014). The 3,000 rice genomes project. Gigascience *3*, 7.

Zhou, Z., Jiang, Y., Wang, Z., Gou, Z. Lyu, J., Li, W., Yu, Y., Shu, L., Zhao, Y., Ma, Y., *et al.* (2015). Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nat. Biotechnol. *33*, 408–414.

Żmieńko, A., Samelak, A., Kozłowski, P., and Figlerowicz, M. (2014). Copy number polymorphism in plant genomes. Theor. Appl. Genet. *127*, 1–18. http://dx.doi.org/10.1007/s00122-013-2177-7