Chapter 13
CNV and Structural Variation in Plants: Prospects of NGS Approaches

Enrico Francia, Nicola Pecchioni, Alberto Policriti, and Simone Scalabrin

The present chapter focuses on copy number variants (CNVs). It firstly summarizes how CNVs are classified within structural variants (SVs), which are the mechanisms causing their onset, and to which extent they have been discovered in plant genomes. Moreover, as most of the CNVs reported so far overlap with protein-coding sequences and result in gains and losses of gene copies that might have a straight influence on gene/transcript dosage (Chia et al. 2012), particular attention is given to the role played by copy number variation (CNV) in the regulation of relevant adaptive traits, e.g. plant development, as well as resistance to abiotic stresses. A full range of structural variation could thus be detected from next-generation sequencing (NGS) data, including translocations, and CNVs (for a review, see Abel and Duncavage 2013). However, the complexity of plant genomes and the short read length obtained from NGS platforms pose new bioinformatic challenges associated with their detection. After the discussion about the computational issues, the array of available methods for CNV discovery from NGS data is reviewed. Notably, although numerous

E. Francia, Ph.D. • N. Pecchioni, Ph.D. (✉)
Department of Life Sciences, University of Modena and Reggio Emilia, via Amendola, 2, Reggio Emilia 42122, Italy

CGR – Center for Genome Research, University of Modena and Reggio Emilia, Via Campi, 287, Modena 41126, Italy
e-mail: nicola.pecchioni@unimore.it

A. Policriti, Ph.D.
Department of Mathematics and Computer Science, University of Udine, Udine 33100, Italy

IGA - Institute of Applied Genomics, Parco Scientifico e Tecnologico "L. Danieli", Udine 33100, Italy

S. Scalabrin, Ph.D.
Department of Mathematics and Computer Science, University of Udine, Udine 33100 , Italy

IGA Technology Services, Parco Scientifico e Tecnologico "L. Danieli", Udine 33100, Italy

G. Sablok et al. (eds.), *Advances in the Understanding of Biological Sciences Using Next Generation Sequencing (NGS) Approaches*,
DOI 10.1007/978-3-319-17157-9_13

software packages are available for NGS analysis, there is currently no single informatic method capable of identifying the full range of structural DNA variation, and multiple complementary tools are required for robust CNVs detection. Finally, future bioinformatic and applicative prospects for such genomic variants are discussed.

Copy Number Variation Is Part of Genome Structural Variation

Plant nuclear genomes display extensive variation in size, chromosome and gene number, and number of genome copies per nucleus (Kellogg and Bennetzen 2004). Such genomic variability can be present in many forms, including single nucleotide polymorphisms (SNPs), variable number of tandem repeats (VNTRs; e.g., mini- and microsatellites), presence/absence of transposable elements (e.g., retrotransposons and DNA transposons), and different forms of structural variation (SV) (Fig. 13.1). On the basis of their nature, SVs are classified in (1) chromosomal inversions when a segment of a chromosome is reversed end to end, (2) translocations in which rearrangements of parts of non-homologous chromosomes are involved, and (3) CNVs. Scherer (2007) masterly overviewed how descriptors of variation began in the realm of cytogenetics in the 1960s and in the 1970s, continued in the field of molecular genetics and, most recently, in that of cytogenomics, which bridges the gap for detection of genomic variants. Owing to Feuk et al. (2006), and as said in the introductory paragraph, SV should cover by definition the genomic variation that affect large DNA segments, ranging from 1 kb to several Mb ("submicroscopic" size). The designation of the category "1 kb to submicroscopic" is somewhat arbitrary at both ends, but is used for operational definition. In a broad sense, structural variation has been used to refer to genomic segments both smaller and larger than the narrower operational definition. CNVs are currently defined as unbalanced changes in the genome structure and represent a large category of genomic structural variation, which according to Alkan et al. (2011) should include by definition insertions (i.e., the addition of one or more base pairs into a DNA sequence), deletions (i.e., the loss of any number of nucleotides, from a single base to an entire piece of chromosome), tandem or interspersed duplications (i.e., any duplication of a region of DNA). According to these authors, also the single base INDELs should be ideally ascribed to CNVs. NGS, in conjunction with increasingly powerful bioinformatic tools, made possible the identification of polymorphic regions of >50 bp in size, traditionally defined as INDELs, that could be included among SVs (Alkan et al. 2011). In other reports (e.g., Springer et al. 2009), the definition of CNV is associated with that of presence-absence variation (PAV), that should include the insertions and deletions distinct from the typical CNVs; to which, in a restrictive view, should only be ascribed duplications. In the present chapter we prefer to embrace the wide-angle vision of CNVs, by including present-absent variants (PAVs) into this group of SVs. Then, in accordance with most literature reports, we also prefer to exclude from CNVs the structural variations <1 kb (Fig. 13.1). CNV sometimes exhibits strong associations with specific biological functions.

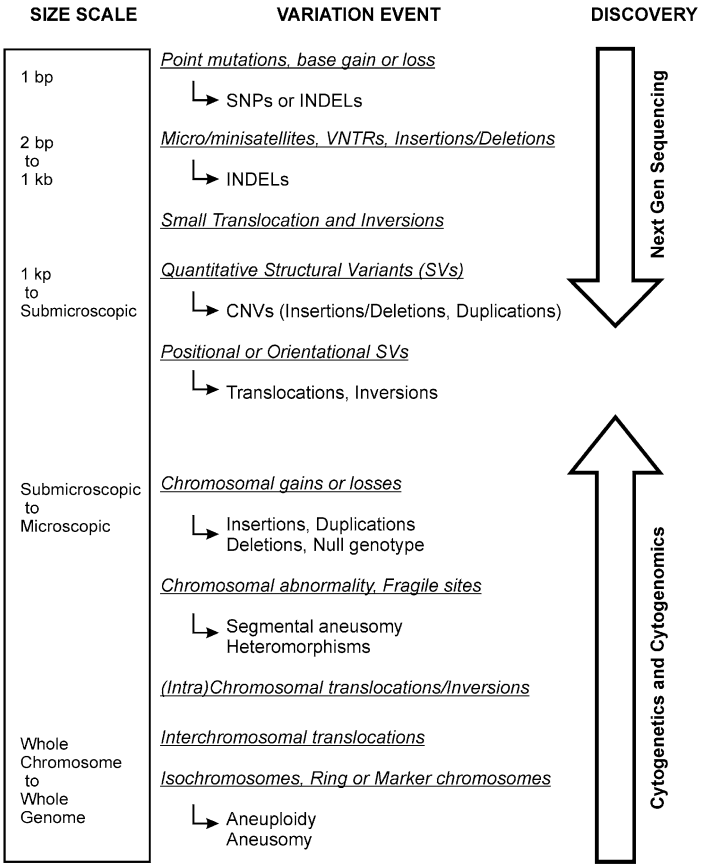

Fig. 13.1 General diagram of genomic structural variation (*SV*). Events ranging from single-base sequence variation to either whole chromosome or genome changes are *italics underlined* and ordered in the figure according to their physical size (*rectangle* to the left). Major categories of single nucleotide variants (*SNVs*) and structural variants (*SVs*) resulting from variation events are indicated by *solid arrows*; SVs <1 kb were excluded from CNVs in accordance with most literature records (see text for details). *Open arrows* show the general approaches applied for discovery the different variation events. Figure modified from Scherer et al. (2007)

Structural variation is therefore increasingly recognized, in humans as well as in other organisms, as a common feature and evolutionary force of genomes, where CNVs and associated gene dosage effects have been implicated in a number of trait/phenotypes (Girirajan et al. 2011; Cantsilieris and White 2013).

Diffusion of CNVs Within Genomes

Owing to the biomedical focus of most studies, at present the best data on CNVs come from human genome. Great interest in CNVs was stimulated by two initial papers of Iafrate et al. (2004) and Sebat et al. (2004) in the early 2000s. Both these

papers described large-scale copy number polymorphism in the human genome. CNV was found surprisingly common among humans. For example, an early study by McCarroll et al. (2008) has revealed that a sizable proportion—from 175 to 230 autosomal loci spanning approximately six megabases—of the human genome varies in copy number between two unrelated individuals. Considering only protein-coding genes, studies show that any two humans are likely to differ at CNVs completely encompassing approximately 105 genes. Interestingly, a considerable higher gene CNV was found in maize (Swanson-Wagner et al. 2010); however, the importance of this result cannot be overemphasized: any two individual genomes taken from nature, in any species, will have dozens to hundreds of differences in their total number of functional genes. Currently, it is estimated that common CNVs occur in approximately 10–12 % of the human reference genome (Conrad et al. 2010; Redon et al. 2006). In human genome CNVs are often detected in regions that contain protein-coding genes or important regulatory elements. CNVs may also affect gene regulation by position effects, and CNVs that partially overlap a gene sequence may disrupt the structure of the gene and impair its function (for a review see Zmieńko et al. 2014). In a comparison study between humans and chimpanzee, beside a conservation of many CNV regions between the two species, some of these regions appeared to be "hotspots" for the genesis of this kind of variation (Perry et al. 2006). CNVs in plants have not been so thoroughly studied, notwithstanding the significant number of diverse fully sequenced genomes since 2000. It is only in the last 5 years that CNVs have attracted the attention of plant biologists and geneticists, likely stimulated by the first findings of association to phenotypes in 2009 and 2010, and leading to the first estimates of the extent of CNV in plant genomes. Notably, in plant genetics, the individual organisms are mainly treated as representatives of one of the following sub-types: (a) cultivars (also named varieties), which are distinct, often intentionally bred subsets of a species that will behave uniformly and predictably when grown in the environment to which they are adapted or (b) accessions, which are collections of plant material from a particular location that are given unique identifiers. Accordingly, CNVs in plants are often recognized and discussed as polymorphisms distinguishing cultivars/accessions of one species rather than affecting individual plants (Chia et al. 2012; Cao et al. 2011; Xu et al. 2012). The crop plant in which CNVs have been primarily investigated, and for which exist the deepest knowledge is maize—*Zea mays* L. (Springer et al. 2009; Swanson-Wagner et al. 2010; Beló et al. 2010; Jiao et al. 2012). After the release of the complete genome sequence of the inbred line B73, the extremely high genomic diversity exhibited by maize has become accessible at a level of detail never had before. Several studies revealed extensive structural variation, including hundreds of CNVs and thousands of the cited PAVs. Basing on comparative genomic hybridization (CGH) Springer et al. (2009) and Beló et al. (2010) detected thousands of dispersed as well as clustered CNVs in the maize genome, between B73 and Mo17 inbred lines or among 13 inbreds compared to B73, respectively. Two main factors affected the estimation of the number of CNVs detected between different inbreds. First, the microarray platform used was primarily developed for gene expression with not uniform distribution of genes along the maize genome (e.g., with fewer probes in

the paracentromeric regions). Second, the majority of the probes were designed to be complementary to the B73 allele, and therefore sequences absent from B73 could not be detected. As a consequence, the number of CNVs identified was underestimated, especially with respect to small CNVs, as the methodology favors detection of large insertion–deletion variants. However, the high level of structural variation and differences in genome content observed in maize are unprecedented among higher eukaryotes. Lai et al. (2010) characterized genetic variation in the six elite strains most commonly used to make commercial hybrids. As already hypothesized by Springer et al. (2009), the authors discussed the potential roles of complementation of gene PAVs, CNVs, and other mutations in contributing to heterosis. Swanson-Wagner et al. (2010) analyzed structural variation between diverse maize inbreds and inbred wild teosinte lines, providing evidence for widespread genome content variation. Over 70 % of the CNV/PAV examples were identified in multiple genotypes, and the majority of events were observed in both maize and teosinte, suggesting that these variants predate domestication and that it seems not having been strong selection acting against them. Partially in contrast with this observation, Jiao et al. (2012) reported extensive CNVs occurring through the maize breeding history. By sequencing of 278 inbred lines from different periods of breeding history, including deep resequencing of 4 lines with known pedigree information, these authors could conclude that, even within identity-by-descent regions, extensive variation caused by SNPs, INDELs, together with CNVs occurred quite rapidly during breeding. In particular, 8.5 % of maize genes showed CNV among the four compared genomes, and an average CNV rate was calculated, although lower for maize compared to that described in humans (8.57×10^{-4} per gene per year vs. 1.2×10^{-2}) (Jiao et al. 2012). As a second important crop surveyed for CNVs, soybean reference genome of cultivar Williams 82 has been compared with introgressed regions from parent Kingwa by analyzing nucleotide and structural differences between Williams 82 individuals (Haun et al. 2011). The authors found that in soybean the impact of intracultivar genetic heterogeneity can be significant, with a high rate of structural and gene content variation and, as hypothesized in humans, the presence of conspicuous CNV hotspots. McHale et al. (2012) combined and compared two approaches for the evaluation of genome-wide structural and gene content variation among four soybean genotypes: microarray CGH and exome DNA capture and resequencing. As an interesting result of the analyses, the regions most enriched for SVs were gene-rich regions harboring clusters of multigene families. Only members of multigene families that are located within clusters tend to be associated with CNV regions. Among these multigene families, the most abundant were the nucleotide-binding and receptor-like classes, presumably important for plant defense against pathogens. In terms of CNV distribution, soybean showed relatively long chromosomal regions (and nearly entire chromosomes) that exhibit virtually no SV among genotypes, interspersed with pockets of high SV ranging from several kb to greater than 10 Mb in length. By resequencing and comparing two sweet and one grain sorghum (*Sorghum bicolor* L.) inbred lines to the reference accession BTx623, Zheng et al. (2011) came to similar result. Along with INDELs PAVs and SNPs, more than 17,000 CNVs (>2 kb in length) were retrieved. While the majority of the

large-effect structural variations resided in genes containing LRR, PPR repeats and in disease resistance R genes, annotation analysis showed that 2,600 genes had 3,234 CNVs, and 32 genes had CNVs in all three sorghum lines (Zheng et al. 2011). The first catalog of CNVs in a diploid Triticeae species has been reported by Muñoz-Amatriaín et al. (2013) for the barley (*Hordeum vulgare* L.) crop. The authors developed a CGH array covering approximately 50 Mb of repeat-masked sequence of the reference cv. Morex and compared via genomic hybridization a collection of 14 genotypes including eight cultivars and six wild barleys. Almost 15 % of all the sequences considered were affected by CNV and more than 60 % were found in two or more genotypes. As already observed in the maize genome (Springer et al. 2009; Swanson-Wagner et al. 2010; Beló et al. 2010) CNVs in barley are enriched near to chromosome ends, apart in one chromosome (4H), that showed the lowest frequency of CNVs. CNV affects 9.5 % of the coding sequences represented on the array and, similarly to what observed in soybean, the genes affected by CNV are enriched for sequences annotated as disease resistance proteins and protein kinases. The list of agriculturally relevant species surveyed for presence of CNVs extends at least to allotetraploid wheat (Saintenac et al. 2011), rice (Yu et al. 2013), and tomato (Causse et al. 2013). A significant presence of such SVs has been verified consistently in all the three species. By a sequence capture assay restricted to 3.5 Mb exon regions, for a total of 3,497 genes of tetraploid wheat compared between cultivar Langdon and a wild emmer accession, Saintenac et al. (2011) found 85 CNV targets; among these, 77 variants were due to an elevated number of copies in the Langdon genome and only 8 variants resulted from copy increase in the wild emmer genome. In the rice CGH study, Yu et al. (2013) identified 2.69 % of rice genome interested by CN variable regions (CNVRs), overlapping 1,321 genes, these significantly enriched for cell death, protein phosphorylation, and defense response, as already observed in soybean and barley. The 1,686 putative CNV regions identified in tomato impacted a total of 1,235 genes, with significant differences between the eight resequenced genotypes, and cell death process genes represented in significant excess (Causse et al. 2013).

Mechanisms Leading to Variation in Number of Copies

As a general rule, alteration in copy number involves change in the structure of the chromosomes such that two formerly separated DNA sequences are joined together. Several mechanisms have been postulated to explain the formation and then the variation in number of copies of CNVs (Hastings et al. 2009a). However, the mechanisms of all structural changes that involve chromosomal DNA are substantially the same, and occur by two general mechanisms: homologous recombination (HR) and non-homologous recombination (NHR). HR is a complex process whereby DNA segments that share significant sequence homology are exchanged. This definition entails the requirement for broad DNA sequence identity; however, in yeast it is thought that as little as 30 bp are sufficient (Haber 2000). In plants, a few hundred

base pairs can engage the HR machinery (Puchta and Hohn 1991), but it is still unclear whether there is a lower limit, nor what is the dependence on the type of partners (Lieberman-Lazarovich and Levy 2011). Sequence microhomology (i.e., very few bases of identity) or no homology are instead the basic events for NHR. Although HR provides vital repair mechanisms, meiosis requires crossing over and a possible side effect of this requirement is the rather high frequency of CNVs produced—according to the estimates reported by Lupski (2007). According to this author, such frequency ranges from 10^{-6} to 10^{-4} copy number changes per gamete. Several mechanisms are based on HR for repairing DNA breaks and gaps; among these, the best studied is called double-stranded break (DSB)-induced recombination. Owing to previous research done in *Saccharomyces cerevisiae* one of DSB repair models (namely, synthesis-dependent strand annealing—SDSA), which does not generate crossovers, could produce variations in copy number when the DNA template contains direct repeats (for a review see, Pâques and Haber 1999). A more important HR mechanism is the non-allelic homologous recombination (NAHR), between DNA segments on the same chromosome and of high similarity, but that are not alleles. NAHR usually involves low-copy repeats (LCRs)—DNA segments larger than 1 kb that are generated during ancient duplication events. Depending on the LCR location, NAHR can lead to intrachromatid, interchromatid, or interchromosomal rearrangements. The type of rearrangement depends on LCR orientation: the repeats may be direct, opposite or mixed. The orientation determines whether NAHR leads to the deletion, reciprocal duplication, or inversion of the DNA segment flanked by the LCRs. In maize, some transposon elements have been shown capable of directly inducing tandem sequence duplications, and let to hypothesize that this activity has contributed to the evolution of the maize genome (Zhang et al. 2013). Besides repairing two-ended DSBs, HR can repair collapsed or broken replication forks in a process called break-induced replication (BIR). Several authors discussed the possible involvement of BIR in a microhomology-mediated mechanism of copy number change (Hastings et al. 2009b). Finally, a minor HR player in the formation of CNVs is a DSB mechanism known as single-strand annealing (SSA). In yeast, SSA has been found responsible for deletions of up to a few tens of kb (Pâques and Haber 1999), while in plants SSA can lead to efficient sequence deletions between direct repeats and this might, for example, explain the accumulation of single long terminal repeats of retroelements in cereal genomes (Puchta 2005).

Concerning NHR, other mechanisms of DSB repair either do not require homology or need very short micro-homologies for DNA repair: non-homologous end joining (NHEJ), microhomology-mediated end joining (MMEJ), and breakage–fusion–bridge cycle (Puchta 2005). All these phenomena increase the probability of genetic changes such as CNV. Another potential non-homologous mechanism is fork stalling and template switching (FoSTeS). FoSTeS is caused by DNA replication errors when replication forks stall, in a manner in which the 3′ primer end of a DNA strand can change templates to an ssDNA template in a nearby replication fork (Lee et al. 2007). FoSTeS events may generate insertions, deletions, or more complex rearrangements such as CNVs (Lee et al. 2007). However, Hastings et al.

(2009b) proposed a new model—microhomology-mediated break-induced replication (MMBIR)—that in addition to the events included in FoSTeS could also lead to translocations. Interestingly, MMBIR supports the hypothesis of an increase in the frequency of CNVs produced when cells are under stress. This observation from molecular evidences is consistent with the intriguing hypothesis of an adaptive evolutionary value of CNV, when organisms are challenged by environmental stresses (see below). Such switch from high-fidelity to error-prone DSB repair in stress conditions seems common from bacteria to plants (DeBolt 2010). Finally, it must be underlined that CNVs are not randomly distributed in genomes, but tend to be clustered in CNV hotspots (Perry et al. 2006), in regions of complex genomic architecture. There is therefore ample evidence that specific features of chromosomal architecture are also involved in CNV generation, and this entails that multiple genomic features can affect the probability of CNV occurrence.

Do CNVs Have a Biological Meaning?

Association to Phenotypes

Since the 1980s it is known that the human genome contains apart from single base and short repeat polymorphisms another abundant source of variation, involving deletions, insertions, duplications, and complex rearrangements. Nevertheless, the first evidence of a phenotypic role of CNVs has come with the elucidation of the etiology of Charcot–Marie–Tooth neuropathy type 1A, due to gene duplication rather than to point mutations (Lupski et al. 1991). Since 1991 and until 2006, with the scientific world fully dedicated to the exploitation of SNP-associated traits, only a small number of pioneer studies advanced knowledge of CNV impacts on human diseases, before the systematic characterization of Redon et al. (2006). They identified CNVs covering approximately 12 % of the human genome, and hypothesized potential alterations of gene dosage, gene disruption or perturbed regulation of their expression, even at long-range distances. After the first global searches aimed to discover and catalog these structural variations in the human and mouse genome, an array of different experiments mostly performed as case–control studies allowed to characterize an increasing number of CNV-associated phenotypes (diseases) in humans, such as Crohn's disease (McCarroll et al. 2008). Diskin et al. (2009) demonstrated, in a disease for which SNP variations are known to influence susceptibility, that CNV at 1q21.1 is associated with neuroblastoma and implicates a novel gene in early tumorigenesis. Sometimes, genetic risk factors have been missed because association studies have sought risk-associated SNPs, while ignoring structural variation causing gene copy number changes. This is the case of CNVs associated with colorectal adenoma recurrence (Laukaitis et al. 2010). But it is in particular with many developmental neuropsychiatric disorders that rare CNVs have unprecedented levels of statistical association. These CNV-associated disorders include schizophrenia, autism spectrum disorders, intellectual disability, and attention

deficit hyperactivity disorder (ADHD); however, as CNVs often include multiple genes, causal genes responsible for CNV-associated diagnoses and traits are still poorly understood (Hiroi et al. 2013). Among these associations, the 16p11.2 copy variant phenotype of neurocognitive defects was found to be driven by the KCTD13 gene dosage changes within the CNV region encompassing 29 genes (Golzio et al. 2012). As regards other investigated traits, finally and interestingly, severe obesity and being underweight could be mirror extreme phenotypes of the same CNV at 16p11.2 locus, respectively, associated with a large (600 kb) deletion vs. a duplication of the region (Jacquemont et al. 2011).

The intuitive scientific question whether CNVs can modify gene expression is a key issue for their association to phenotypes where differential gene expression plays a role. The majority of experiments found out that not only variations in gene copy numbers can modify gene expression in carrier genotypes, but importantly they can also significantly influence expression time courses. In a global survey in humans, Stranger et al. (2007) observed that CNVs captured a significant percentage of the total genetic variation in gene expression, 17.7 %, although lower than the remaining part attributed to SNPs (83.6 %). In a study throughout mouse development, Chaignat et al. (2011) observed that CNV genes are significantly enriched within transcripts showing variable time courses between mice strains; thus, modifications of the copy number of a gene may alter not only gene expression, but also potentially alter timing of its expression. Henrichsen et al. (2009) found that not only expression of human genes within CNVs tend to correlate with copy number changes, but also that CNVs can influence the expression of close genes, with an effect extending in the vicinity up to a distance of 0.5 Mb; moreover, they can also have a global influence on transcriptome. An intriguing effect on gene expression has been shown by a promoter competition between copy number variant α-globin genes and the NME4 gene, located 300 kb apart from the α-globin cluster, for which the deletion of two α-globin genes is unlocking higher NME4 expression by a regulator (Lower et al. 2009).

Also in plants, SV has been hypothesized to be a driving force behind phenotypic variation (Chia et al. 2012). First clear associations to phenotypes in plants follow at distance the discoveries made in human genetics, with the first report in barley dating 2007. The boron-toxicity tolerant cultivar Sahara contains about four times as many *Bot1* boron transporter gene copies compared to intolerant genotypes, and produces significantly more *Bot1* transcripts. *Bot1* transcript levels identified in barley tissues are consistent with an avoidance strategy, by limiting the net entry of boron into the root and by disposing boron from leaves via hydathode guttation (Sutton et al. 2007). A very similar genetic strategy has been observed recently in maize for tolerance to Aluminum, i.e. to acidic soils. The expansion in MATE1 (multidrug and toxic compound extrusion 1) copy number is associated with higher MATE1 expression, which in turn results in superior Al tolerance; the three MATE1 copies in the rare tolerant genotypes (all containing three copies) are identical and are part of a tandem triplication, absent in the vast majority of susceptible accessions that carry a single copy of the transporter gene (Maron et al. 2013). The frost-tolerant barley cultivar Nure contains tandem segmental duplications through the

CBF2A-CBF4B genomic region of the CBF gene cluster on chromosome 5H, that differentiate freeze-tolerant from sensitive genotypes, which carry single copies of those genes. The higher copy number of CBF genes is associated with higher gene expression in tolerant genotype Nure of the transcription factors under short days (Knox et al. 2010). Although observed for an effector gene, at the end of the final response cascade to cold, CNV of Y_2K_4 dehydrin in *Medicago* has been hypothesized as a duplication of dehydrin genes in cold-tolerant cultivated alfalfa genotypes (Castonguay et al. 2013). In a review of Oh et al. (2012) about tolerance of plants to extreme conditions, gene duplication is indicated as one of three possible strategies to cope with extreme abiotic stress conditions. Among the examples reported to support the hypothesis, HKT1, a plasma membrane Na^+/K^+ transporter considered to be a genetic determinant of salt tolerance, exists as tandem duplicated copies in two salt-tolerant *Thellungiella* species. As a second example, the duplication of NHX8 homologs, known to encode a putative Li^+ transporter in *A. thaliana*, leads to a constitutively higher expression in *Thellungiella parvula* than in *A. thaliana*, and this in turn might be responsible for the apparently enhanced tolerance of *T. parvula* to high Li^+ in its natural habitat. In maize, a recent genome-wide SNP screen of 103 diverse maize and teosinte lines (Chia et al. 2012) suggests a correlation between genomic regions containing structural variation – detected as read-depth variants (RDVs) in genome resequencing – and QTLs for agronomic traits. As an interesting example, genomic regions containing QTLs for leaf architecture and resistance to northern and southern leaf blight are enriched for RDVs. This suggests a potential role for CNV/PAV in generating phenotypic variation for these agronomic traits. Schnable and Springer (2013) hypothesize a generic role for gene CNV to help explaining heterosis. In fact, complementation of allelic variation, as well as complementation of variation in gene content and expression patterns, is likely to be important contributors to this trait of paramount importance in maize. CNV/PAV has been reported to be differentially represented among genes categorized as being involved in stress and stimulus response, perhaps in part because this category includes some large gene families (e.g., NBS-LRR genes). This pattern is detectable on a genome-wide scale in maize (Chia et al. 2012), rice (Xu et al. 2012) and in other plants. An interesting example of multiple resistance genes acting by means of a structural variation is *Rhg1* nematode resistance QTL in soybean. Cook et al. (2012) demonstrated how this resistance is governed by a peculiar CNV of multiple genes. Ten tandem copies of the 31-kilobase segment identifying the *Rhg1* locus are present in an rhg1-b resistant haplotype vs. one copy per haploid genome in susceptible varieties. In this multigene segment, overexpression of the individual genes was ineffective, but overexpression of the genes together conferred enhanced soil cyst nematode resistance. Hence, SCN resistance mediated by the soybean quantitative trait locus *Rhg1* is conferred by CNV that increases the expression of a set of dissimilar genes in a repeated multigene segment.

Regulation of plant development is the last group of plant phenotypic traits that are being increasingly associated with CN variations. In barley, Nitcher et al. (2013) demonstrate that the *HvFT1* (FLOWERING LOCUS T homolog, corresponding to the VRN-H3 locus) allele present in the barley accession BGS213 and associated

with a dominant spring growth habit, carries at least four identical copies of *HvFT1*, whereas most barley varieties harbor a single copy. The increased copy number is associated with earlier transcriptional up-regulation of *HvFT1*, thus giving further support to the hypothesis made in humans that CNV is not only leading to differences in gene expression, but also to differences in expression time course. In wheat, two key regulators of flowering in response to light and temperature have been found to be ruled by CNV associated with altered gene expression. Alleles with an increased copy number of photoperiod response gene *Ppd-B1* confer an early flowering day neutral phenotype and have arisen independently at least twice. At the same time, plants with an increased copy number of vernalization requirement gene *Vrn-A1* have an increased requirement for vernalization so that longer periods of cold are required to potentiate flowering (Díaz et al. 2012). The results shed new light on regulation of flowering in wheat, and intriguingly suggest that CNV plays a significant role in wheat and plant adaptation.

Evolutionary and Adaptive Value of CNVs

As stated by Schrider and Hahn (2010) and by Kondrashov (2012), although it might be too early to tell whether or not a substantial fraction of gene copies have initially achieved fixation in eukaryotes by positive selection for increased dosage, nevertheless enough examples have accumulated in the literature to strongly suggest an adaptive value for such genetic variation. As a consequence of this, a complete understanding of the molecular basis for adaptive natural selection must necessarily include the study of copy number variation. One of the clearest examples supporting such hypothesis comes from budding yeast (Stambuk et al. 2009). In five industrially important *S. cerevisiae* strains responsible for the production of fuel ethanol from sugarcane, there have been found significant amplifications of the telomeric SNO and SNZ genes, which are involved in the biosynthesis of vitamins B6 (pyridoxine) and B1 (thiamin), and confer the ability to grow more efficiently under the repressing effects of thiamin, especially with high sugar concentrations. These genetic changes have likely been adaptive and selected for in that specific industrial environment. Similar effects of the feeding environment on CNV were observed in wood decaying fungi, where CNV was observed in members of the detoxification pathways belonging to multigenic families such as the cytochrome P450 monooxygenases and the glutathione transferases, as an adaptive strategy allowing these basidiomycetes to deal with the plethora of potential toxic compounds resulting at least partly from wood degradation (Morel et al. 2013). In humans, the AMY1 α-amylase gene, which encodes a protein catalyzing starch degradation constitutes an interesting example. It has been found a gene copy number three times higher in humans compared to chimpanzees, and higher expression levels of salivary amylase protein, suggesting that humans were favored in the gene dosage due to an increase of starch consumption in their evolutionary history (Perry et al. 2007). As pointed out by Bailey et al. (2008), in a global survey of human copy

number genes, many examples of gene CNVs described within the human population due to their association with phenotype and disease, also before the NGS era, can be postulated to have played important roles in human adaptation to changing environmental conditions and infectious pathogens.

In plants, the common observed association between abiotic and biotic stress tolerant phenotypes and gene CNV is coupled to the observation in Arabidopsis (DeBolt 2010), although common to all organisms (Hastings et al. 2009a; Freeman et al. 2006), that CNVs form at a faster rate than other types of mutation. A striking example of such a faster rate is the generation of significant numbers of CNVs in Arabidopsis lineages after only five generations under low and high temperature and chemical (salycilic acid spray) stresses, with positive selection for fecundity, while genotypes deriving from the same mother plants by selfing did not display any differences in CNV when growing under normal conditions (DeBolt 2010). Boyko and Kovalchuk (2011), from their previous experiments about signaling in plant–pathogen interactions, hypothesize the generation in plants infected with a compatible pathogen of a systemic recombination signal (SRS) that precedes the spread of pathogens and results in an increase of the somatic and meiotic recombination frequency. Although yet to be fully validated, the hypothesis is an intriguing further support to a wide environmental adaptive role for the origin of SVs. In a very interesting review about genetic variation in extremophile plants (adapted to extreme environmental conditions), Oh et al. (2012) argue that there is little overall evidence that polyploidy itself is a major evolutionary driving force leading to extremophiles, while tandem duplications seem to have a more important role in shaping genomes for stress adaptations. The evolutionary meaning of local gene duplications could be in fact viewed also in comparison with polyploidy, common in plants, and for a long time considered as a main evolutionary driver in these organisms. In humans, Makino and McLysaght (2010) observe that duplicated genes deriving from two ancestral WGD (whole genome duplication; i.e. ohnologs) have rarely experienced subsequent small-scale duplication (SSD), are refractory to CNV, are dosage-balanced and preferably retained in human populations; by contrast, genes that have experienced SSD are more likely to also display CNV and dosage unbalance. Similar observations in plants took Birchler (2012) to conclude that different fates can be observed for duplicate genes depending on whole genome or segmental duplication. Following polyploidy formation, members of macromolecular complexes persist in the evolutionary lineage longer than random genes, while a complementary pattern is found for segmental duplications in that there is an underrepresentation of members of macromolecular complexes.

What written about adaptive value of CNVs is mostly referred to examples of copy number variable genes, and the majority of validated phenotypes present in the literature refer to these cases. However, the case of the relatively large structural variation at *Rhg1* locus in soybean suggests that also other more complex copy unbalances in higher organisms, if at similar faster mutation rates, can be included within the same evolutionary meaning.

NGS Approaches and Bioinformatic Tools for CNV Detection

In this section, we are going to discuss some bioinformatics issues involved in the discovery and classification of SV, with special emphasis on CNV in plants. We consider these issues trying to answer the following questions:

1. general: Given the mathematical definitions of the problems we want to solve, what are the main computational bottlenecks to face and what kind of limits can we put to the (abstractly obtainable) answers?
2. practical: On the grounds of the given definitions, which ones among the concrete solutions proposed in the literature—and to what extent—reach the potential frontiers of implementable tools?
3. technological: Is the interplay among proposed definitions, computational problems, and available (or foreseeable) technologies for data production, going to change significantly the landscape in the (near) future?

We will see that the search of variations among genomes of different organisms of the same species is a challenging subject, as a result of the difficulties involved in answering each one of the above questions. The problem is mathematically elusive, as a precise definition is either quickly unrealistic or impossible to satisfy; practical solutions proposed are often difficult to judge or classify, because of the large amount of specific and rapidly changing sets of heuristics implemented. It is often not clear how the technological changes that we expect will take place in the near future, will modify the amounts and the kind of data that soon will be available for analysis. Nevertheless, the bioinformatics aspects involved in the field make the challenges exciting, as it is clear that only a coordinated effort towards a clear specification and a compilation of realistic needs can result in the design of a new generation of useful tools.

The Computational Problem

From a computational point of view, we begin by attempting a classification of SV and CNV. Any classification must assume the existence of a reference genome G for the organism under study. The reference can be either the first (or most reliable) available sequenced genome for the species, or a core genome resulting as a common factor of previous analyses. The first class of objects (SV) is usually defined as the collection of sub-sequences σ that may or may not appear in G. In such terms, SVs include *any* possible variation to search and classify. Among SVs we can isolate CNVs as those sub-sequences γ whose characteristic feature is presence/absence together with the number of their occurrences. As we pointed out, any sensible definition is to be given with respect to a genome sequence to be considered our fixed reference system. This is true even in cases in which an "official" reference is not available and a comparative study between two or more individuals is carried out: in these cases, the reference is fixed on-the-fly but is, however, present. Hence, for example, presence in the individual under study and absence in the reference corresponds to a number of occurrences equal to one against a number of

occurrences equal to zero (infinite ratio). In general, when a γ occurs in the reference we can talk about the rate of its occurrence. For both SVs and CNVs, the definition should be further refined by (at least) specifying:

- the (lower) limits in length for σ's and γ's, thereby introducing a finer classification on both categories;
- the number and kind of allowed alignment errors, while establishing presence/absence or evaluating the number of occurrences.

The above classification cannot be rigid: two shorter sub-sequences cannot be considered equal by the same percentage of errors (mismatch, insertion/deletions of characters) employed for significantly longer ones. Moreover, even though CNVs do change the total length of the genome, a detection based on a variation of the total length is not of any practical use. On the ground of the grid defined above, we can then finally enter within a more functional analysis of the sub-sequence considered. Each σ or γ can be classified on special patterns defining its encoding, compositional, or otherwise syntactically characterizing feature.

NGS and the Main Techniques of CNV Discovery

Historically, two general categories of methods were used to detect CNVs and regions with overlapping CNVs (CNVRs): array-based comparative genome hybridization (CGH) and reference genome-based NGS. The first ("hybridization-based mapping") followed the observation that any region duplicated or deleted in an individual sample will show an excess or deficit, respectively, of DNA that is highly similar to that region relative to the reference genome. These methods were therefore aimed at detecting these localized differences in relative DNA content. The second category of methods (sequencing) does not detect the duplications and deletions directly, but instead detects length differences in the size of captured fragments from a sample relative to the reference genome. Fragments that appear too large must contain insertions or duplications, while those that are too small must contain deletions. Other methods, such as quantitative PCR (D'haene et al. 2010) and fluorescent in situ hybridization (Cook et al. 2012), can be used to verify CNVs but they are generally not useful for the discovery process. The current approach for CNV discovery uses NGS high-throughput DNA sequencing technology. This approach has been proven effective for the discovery and mapping of SVs at nucleotide resolution in plants, animals and humans (Cao et al. 2011; Daines et al. 2009; Yoon et al. 2009; Mills et al. 2011; Bickhart et al. 2012).

A Classification of NGS Technologies

The previously used array-based methods could still provide a cost-effective mean for CNV discovery but they suffer of low throughput and low resolution of breakpoints, in the best cases hundreds of bp (Conrad et al. 2010; Park et al. 2010).

Precise characterization of breakpoints, which may capture the signature of potential mutational mechanisms, is crucial for designing robust genotyping assays and assessing the functional content of detected CNVs (Li and Olivier 2013). Moreover, these methods are limited to sequence present in the reference assembly used to design the probes and they cannot neither identify balanced structural variations nor specify the location of a duplication (Alkan et al. 2011). In order to overcome the above problems sequencing has been used in the last years. Initially only Sanger sequencing (Tuzun et al. 2005; Korbel et al. 2007; Kidd et al. 2008; Korbel et al. 2009) was used, then also Second (Bentley et al. 2008; Hormozdiari et al. 2009; Campbell et al. 2008) and Third Generation Technologies (Maron et al. 2013) were exploited. Sanger sequences are about 1 kb long with nearly perfect accuracy and can be produced only at very low throughput and high costs. Second Generation sequences, e.g. Illumina sequences, are much shorter, 100–150 bp for HiSeq machines and 250–300 bp for MiSeq instruments, of good accuracy with only 1 % erroneous bases and throughput increase of orders of magnitude with several Gb produced daily at very limited cost. Finally, Third Generation, single molecule-derived, sequences, e.g. PacBio sequences, are a few kb long, still of limited accuracy with more than 10 % erroneous bases, but with dozens of Mb produced daily at limited costs. Therefore, apart from timing and costs, Second and Third Generation sequences mainly differ on read length and accuracy, and throughput. These factors highly influence the kind of methods to be used to tackle the problem of CNV detection.

NGS Technologies vs. Computational Techniques

Most of the current algorithms for SVs detection are modeled on computational methods that were initially developed to analyze Sanger sequences (Tuzun et al. 2005; Korbel et al. 2007; Kidd et al. 2008; Korbel et al. 2009). So far, NGS based methods to detect SVs can be categorized into five different strategies: paired-end mapping (PEM), split read mapping (SRM), depth of coverage (DOC), de novo assembly (DNA), and a combination of the above approaches (COMBI). PEM was historically the first method based on sequencing used to discover SVs (Tuzun et al. 2005). It assesses the span and orientation of paired reads detecting discordant pairs whose orientation is not as expected or distance is significantly different from the predetermined average insert size. PEM-based tools applied to NGS datasets usually search for clusters of such signatures (Medvedev et al. 2009). Multiple evidences are required to strengthen the signal of usually short NGS reads. Gathering information from multiple evidences is called clustering and it can be divided into hard and soft clustering. In hard clustering reads that map to multiple locations are discarded in order to avoid false positive SVs due to repetitive regions. In soft clustering (Hormozdiari et al. 2009), instead, in order to improve sensitivity, such reads are not discarded and assigned to a single cluster. PEM-based tools can be used to detect effectively deletions, short insertions, inversions, translocations, and

duplications at almost single base pair resolution. Insertions size is limited to library insert size unless multiple evidences are used, e.g. clusters of single reads with only one read of the pair of a fragment can safely be positioned in the genome may hint the insertion of a repetitive element (Platzer et al. 2012; Fiston-Lavier et al. 2011). PEM methods based either on hard or soft clustering suffer, respectively, of sensitivity and specificity in repetitive and low-complexity regions. SRM methods are based on gapped alignment of a single read to the reference genome and can be used to determine SV breakpoints down to base pair resolution. If a read does not align entirely, then a gapped alignment is applied.

SRM was first applied to long Sanger reads (Mills et al. 2006) but later SRM methods were developed also for the NGS technologies with some modifications: (1) given the high coverage of NGS experiments, clusters of split reads are requested as proper signature, (2) given the short length of NGS sequences, split reads usually tend to map to multiple locations of a genome. To overcome this problem the mapping of their mate is used as a reliable anchor, severely limiting the search space for the split read, (3) elongate single reads producing overlapping paired read libraries that can be merged into a single longer read, (4) complement PEM methods providing putative SVs with breakpoints not determined at base pair resolution. In general, SRM methods heavily rely on the length of reads, still a problem for Second Generation Sequences, and are not applicable to repetitive or low-complexity regions. SRM-based tools can be used to detect effectively deletions and very short insertions at base pair resolution. The limitation on insertions is given by the read length itself. DOC methods are still based on read alignment but unlike PEM and SRM methods, they mainly care on DOC and less on single base resolution. Their main assumption is that the number of mapping reads follows a Poisson distribution and regions deleted or duplicated will have less or more reads assigned to them, respectively. DOC-based tools can be classified in at least two categories: single sample and multi-sample. In the first case, average read depth is estimated using mathematical models and then regions that depart from it are discovered. In the second case, a sample is used as control and all other samples are compared to its coverage rather than to an average read depth. Therefore, in the first case copy numbers are absolute numbers while in the second case they are relative to the control sample. In general, DOC methods follow a four-step procedure composed of: (1) independently mapping reads of each sample towards a reference genome, (2) normalize coverage along a sliding window where read depth of a single window is computed according to the number of reads mapped in it (normalization basically serves to correct potential biases in read depths mainly caused by GC content and repetitive regions), (3) estimation of copy number, either absolute or relative, along the sliding window in order to determine possible gain or loss, (4) segmentation, merging adjacent genomic regions with a similar copy number using statistical models. Sliding windows can be computed in a variety of ways as with a fixed width or with a predefined amount of reads mapping within it. DOC methods can detect CNVs with respect to what is present in the reference genome, therefore novel insertions and inversions cannot be detected. The detection reveals the copy number but not the location of possible new copies. The breakpoint resolution is very poor and is on the level of several kb.

Methods based on DNA differ from previously described methods as they do not rely on a first step of read alignment toward a reference genome but directly use the reads to assemble them into contigs that are later compared to a reference genome in order to discover discrepancies. Comparison is usually performed through sequence alignment to the reference genome. An alternative approach is proposed by the software Cortex (Iqbal et al. 2012; Leggett et al. 2013), designed to directly discover CNVs among multi-samples: as most assemblers it is based on de Bruijn graphs with the exception that nodes and edges are marked in different colors to differentiate different samples. Unfortunately, although a range of assemblers have been developed (Simpson et al. 2009; Gnerre et al. 2011; Luo et al. 2012; Simpson and Durbin 2012; Zimin et al. 2013), given the short length of NGS sequences, DNA is still challenging and the accuracy of contigs produced is unsatisfactory especially in repetitive regions that are often a great source of variations. An emerging branch in the field is the assembly of limited regions, e.g. exomes or fosmid clones, which should lead to improved assemblies and consequently improved CNV detection, though at the cost of restricting to limited portions of the genome.

Although methods based on the four previously described categories (PEM, SRM, DOC, and DNA) have been greatly improved and a wide number of tools have been recently developed, none is able to reliably detect SVs, either in terms of sensitivity or specificity. Each has different strengths and weaknesses in detection, depending on the kind of variant or the sequence at the studied *locus*. To overcome the implicit limitations of individual methods, often operating in a complementary manner, it is possible to implement approaches (COMBI) including the different methods and therefore improve the detection performance and reduce the number of false positives. While PEM and SRM methods are related to each other, the other methods, DOC and DNA represent complementary methodologies that could benefit from each other. In some cases, they can detect identical events, perhaps with different strength and precision, while in other cases they can detect very independent events that cannot be discovered by all methodologies.

Future Perspectives

Examples of association of SVs to agronomically relevant phenotypes can be found in a recent review on putative dispensable regions of plant genomes (Marroni et al. 2014). The repertoire of functional and evolutionary consequences of SVs is expanding, but a comprehensive map of all causative SVs is still far from complete. The advent of NGS technologies highly improved the detection rate of SVs even if such technologies are affected by two main drawbacks: difficulty with reliably mapping short reads to DNA repeats (Treangen and Salzberg 2012) and platform-specific biases, which result in lower read coverage of some parts of the genome (for example, GC-rich regions) (Dohm et al. 2008). Compared to detection of SVs using a single tool, the combination of different software has proved effective in overcoming the main drawbacks of NGS technologies and in improving SVs prediction accuracy. In addition, the use of libraries with different characteristics has been

proved effective in the detection of SVs. Moreover, longer reads may greatly improve the specificity of reads mapping and consequently SVs detection. In this context, third generation sequencing (TGS) provide reads as long as few kb and could solve most of the problems of shorter reads, in particular in presence of repetitive regions source of most misalignments. Currently, its main problem is the accuracy of base calls, much lower than most of previous sequencing technologies. A final comment on the availability of proper infrastructures for SVs detection is needed. The huge quantity of NGS data requires a large hardware infrastructure to handle it in terms of both disk space and computational resources. Comprehensive databases of already discovered SVs could highly improve the detection but also the evaluation of putative newly discovered SVs. Although already available for human genetics the Database of Genomic Variants (http://dgv.tcag.ca/dgv/app/home; http://www.ncbi.nlm.nih.gov/dbvar/; http://www.ebi.ac.uk/dgva/), the need for an SV database for the plant kingdom could be seriously considered by bioinformatics institutes in the near future.

On the applicative, plant breeding side, we should consider whether and how CNVs will be effectively used for genomic-assisted selection. A relevant starting consideration is that for too much time this kind of variation has been excluded from genetic association studies. Not only in plants, but also in humans, sometimes genetic risk factors have been missed because association studies have sought risk-associated SNPs, while ignoring structural variation causing gene copy number changes, as reported for colorectal adenoma recurrence (Laukaitis et al. 2010). To avoid this, as anticipated by Stranger et al. (2007) and by Beckmann et al. (2007) for humans, the interrogation of the genomes for both types of variants (SNPs and CNVs) in association studies may be an effective way to elucidate the causes of complex phenotypes in humans, animals, and plants.

Acknowledgements This work was partially supported by the FROSTMAP project of the Fondazione Cassa di Risparmio di Modena and by IGA Technology Services.

References

Abel HJ, Duncavage EJ (2013) Detection of structural DNA variation from next generation sequencing data: a review of informatic approaches. Cancer Genet 206(12):432–440

Alkan C, Coe BP, Eichler EE (2011) Genome structural variation discovery and genotyping. Nat Rev Genet 12(5):363–376

Bailey JA, Kidd JM, Eichler EE (2008) Human copy number polymorphic genes. Cytogenet Genome Res 123(1–4):234–243

Beckmann JS, Estivill X, Antonarakis SE (2007) Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability. Nat Rev Genet 8(8):639–646

Beló A, Beatty MK, Hondred D, Fengler KA, Li B, Rafalski A (2010) Allelic genome structural variations in maize detected by array comparative genome hybridization. Theor Appl Genet 120(2):355–367

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG et al (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456(7218):53–59

Bickhart DM, Hou Y, Schroeder SG, Alkan C, Cardone MF, Matukumalli LK et al (2012) Copy number variation of individual cattle genomes using next-generation sequencing. Genome Res 22(4):778–790

Birchler JA (2012) Insights from paleogenomic and population studies into the consequences of dosage sensitive gene expression in plants. Curr Opin Plant Biol 15(5):544–548

Boyko A, Kovalchuk I (2011) Genetic and epigenetic effects of plant-pathogen interactions: an evolutionary perspective. Mol Plant 4(6):1014–1023

Campbell PJ, Stephens PJ, Pleasance ED, O'Meara S, Li H, Santarius T et al (2008) Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat Genet 40(6):722–729

Cantsilieris S, White SJ (2013) Correlating multiallelic copy number polymorphisms with disease susceptibility. Hum Mutat 34(1):1–13

Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J et al (2011) Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. Nat Genet 43(10):956–963

Castonguay Y, Dubé M-P, Cloutier J, Bertrand A, Michaud R, Laberge S (2013) Molecular physiology and breeding at the crossroads of cold hardiness improvement. Physiol Plant 147(1): 64–74

Causse M, Desplat N, Pascual L, Le Paslier M-C, Sauvage C, Bauchet G et al (2013) Whole genome resequencing in tomato reveals variation associated with introgression and breeding events. BMC Genomics 14:791

Chaignat E, Yahya-Graison EA, Henrichsen CN, Chrast J, Schütz F, Pradervand S et al (2011) Copy number variation modifies expression time courses. Genome Res 21(1):106–113

Chia J-M, Song C, Bradbury PJ, Costich D, de Leon N, Doebley J et al (2012) Maize HapMap2 identifies extant variation from a genome in flux. Nat Genet 44(7):803–807

Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y et al (2010) Origins and functional impact of copy number variation in the human genome. Nature 464(7289):704–712

Cook DE, Lee TG, Guo X, Melito S, Wang K, Bayless AM et al (2012) Copy number variation of multiple genes at Rhg1 mediates nematode resistance in soybean. Science 338(6111): 1206–1209

D'haene B, Vandesompele J, Hellemans J (2010) Accurate and objective copy number profiling using real-time quantitative PCR. Methods 50(4):262–270

Daines B, Wang H, Li Y, Han Y, Gibbs R, Chen R (2009) High-throughput multiplex sequencing to discover copy number variants in Drosophila. Genetics 182(4):935–941

DeBolt S (2010) Copy number variation shapes genome diversity in Arabidopsis over immediate family generational scales. Genome Biol Evol 2:441–453

Díaz A, Zikhali M, Turner AS, Isaac P, Laurie DA (2012) Copy number variation affecting the Photoperiod-B1 and Vernalization-A1 genes is associated with altered flowering time in wheat (*Triticum aestivum*). PLoS One 7(3):e33234

Diskin SJ, Hou C, Glessner JT, Attiyeh EF, Laudenslager M, Bosse K et al (2009) Copy number variation at 1q21.1 associated with neuroblastoma. Nature 459(7249):987–991

Dohm JC, Lottaz C, Borodina T, Himmelbauer H (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res 36(16):e105

Feuk L, Carson AR, Scherer SW (2006) Structural variation in the human genome. Nat Rev Genet 7(2):85–97

Fiston-Lavier A-S, Carrigan M, Petrov DA, González J (2011) T-lex: a program for fast and accurate assessment of transposable element presence using next-generation sequencing data. Nucleic Acids Res 39(6):e36

Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM et al (2006) Copy number variation: new insights in genome diversity. Genome Res 16(8):949–961

Girirajan S, Campbell CD, Eichler EE (2011) Human copy number variation and complex genetic disease. Annu Rev Genet 45:203–226

Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ et al (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci U S A 108(4):1513–1518

Golzio C, Willer J, Talkowski ME, Oh EC, Taniguchi Y, Jacquemont S et al (2012) KCTD13 is a major driver of mirrored neuroanatomical phenotypes of the 16p11.2 copy number variant. Nature 485(7398):363–367

Haber JE (2000) Partners and pathways repairing a double-strand break. Trends Genet 16(6): 259–264

Hastings PJ, Lupski JR, Rosenberg SM, Ira G (2009a) Mechanisms of change in gene copy number. Nat Rev Genet 10(8):551–564

Hastings PJ, Ira G, Lupski JR (2009b) A microhomology-mediated break-induced replication model for the origin of human copy number variation. PLoS Genet 5(1):e1000327

Haun WJ, Hyten DL, Xu WW, Gerhardt DJ, Albert TJ, Richmond T et al (2011) The composition and origins of genomic variation among individuals of the soybean reference cultivar Williams 82. Plant Physiol 155(2):645–655

Henrichsen CN, Vinckenbosch N, Zöllner S, Chaignat E, Pradervand S, Schütz F et al (2009) Segmental copy number variation shapes tissue transcriptomes. Nat Genet 41(4):424–429

Hiroi N, Takahashi T, Hishimoto A, Izumi T, Boku S, Hiramoto T (2013) Copy number variation at 22q11.2: from rare variants to common mechanisms of developmental neuropsychiatric disorders. Mol Psychiatry 18(11):1153–1165

Hormozdiari F, Alkan C, Eichler EE, Sahinalp SC (2009) Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. Genome Res 19(7):1270–1278

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y et al (2004) Detection of large-scale variation in the human genome. Nat Genet 36(9):949–951

Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G (2012) De novo assembly and genotyping of variants using colored de Bruijn graphs. Nat Genet 44(2):226–232

Jacquemont S, Reymond A, Zufferey F, Harewood L, Walters RG, Kutalik Z et al (2011) Mirror extreme BMI phenotypes associated with gene dosage at the chromosome 16p11.2 locus. Nature 478(7367):97–102

Jiao Y, Zhao H, Ren L, Song W, Zeng B, Guo J et al (2012) Genome-wide genetic changes during modern breeding of maize. Nat Genet 44(7):812–815

Kellogg EA, Bennetzen JL (2004) The evolution of nuclear genome structure in seed plants. Am J Bot 91(10):1709–1725

Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T et al (2008) Mapping and sequencing of structural variation from eight human genomes. Nature 453(7191):56–64

Knox AK, Dhillon T, Cheng H, Tondelli A, Pecchioni N, Stockinger EJ (2010) CBF gene copy number variation at Frost Resistance-2 is associated with levels of freezing tolerance in temperate-climate cereals. Theor Appl Genet 121(1):21–35

Kondrashov FA (2012) Gene duplication as a mechanism of genomic adaptation to a changing environment. Proc Biol Sci 279(1749):5048–5057

Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF et al (2007) Paired-end mapping reveals extensive structural variation in the human genome. Science 318(5849):420–426

Korbel JO, Abyzov A, Mu XJ, Carriero N, Cayting P, Zhang Z et al (2009) PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. Genome Biol 10(2):R23

Lai J, Li R, Xu X, Jin W, Xu M, Zhao H et al (2010) Genome-wide patterns of genetic variation among elite maize inbred lines. Nat Genet 42(11):1027–1030

Laukaitis CM, Thompson P, Martinez ME, Gerner EW (2010) Identifying gene copy number variants associated with colorectal adenoma recurrence. Genome Biol 11(Suppl 1):24

Lee JA, Carvalho CMB, Lupski JR (2007) A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. Cell 131(7):1235–1247

Leggett RM, Ramirez-Gonzalez RH, Verweij W, Kawashima CG, Iqbal Z, Jones JDG et al (2013) Identifying and classifying trait linked polymorphisms in non-reference species by walking coloured de Bruijn graphs. PLoS One 8(3):e60058

Li W, Olivier M (2013) Current analysis platforms and methods for detecting copy number variation. Physiol Genomics 45(1):1–16

Lieberman-Lazarovich M, Levy AA (2011) Homologous recombination in plants: an antireview. Methods Mol Biol 701:51–65
Lower KM, Hughes JR, De Gobbi M, Henderson S, Viprakasit V, Fisher C et al (2009) Adventitious changes in long-range gene expression caused by polymorphic structural variation and promoter competition. Proc Natl Acad Sci U S A 106(51):21771–21776
Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J et al (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience 1(1):18
Lupski JR (2007) Genomic rearrangements and sporadic disease. Nat Genet 39(7 Suppl): S43–S47
Lupski JR, de Oca-Luna RM, Slaugenhaupt S, Pentao L, Guzzetta V, Trask BJ et al (1991) DNA duplication associated with Charcot-Marie-Tooth disease type 1A. Cell 66(2):219–232
Makino T, McLysaght A (2010) Ohnologs in the human genome are dosage balanced and frequently associated with disease. Proc Natl Acad Sci U S A 107(20):9270–9274
Maron LG, Guimarães CT, Kirst M, Albert PS, Birchler JA, Bradbury PJ et al (2013) Aluminum tolerance in maize is associated with higher MATE1 gene copy number. Proc Natl Acad Sci U S A 110(13):5241–5246
Marroni F, Pinosio S, Morgante M (2014) Structural variation and genome complexity: is dispensable really dispensable? Curr Opin Plant Biol 18C:31–36
McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A et al (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat Genet 40(10):1166–1174
McHale LK, Haun WJ, Xu WW, Bhaskar PB, Anderson JE, Hyten DL et al (2012) Structural variants in the soybean genome localize to clusters of biotic stress-response genes. Plant Physiol 159(4):1295–1308
Medvedev P, Stanciu M, Brudno M (2009) Computational methods for discovering structural variation with next-generation sequencing. Nat Methods 6(11 Suppl):S13–S20
Mills RE, Luttig CT, Larkins CE, Beauchamp A, Tsui C, Pittard WS et al (2006) An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Res 16(9): 1182–1190
Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C et al (2011) Mapping copy number variation by population-scale genome sequencing. Nature 470(7332):59–65
Morel M, Meux E, Mathieu Y, Thuillier A, Chibani K, Harvengt L et al (2013) Xenomic networks variability and adaptation traits in wood decaying fungi. Microb Biotechnol 6(3):248–263
Muñoz-Amatriaín M, Eichten SR, Wicker T, Richmond TA, Mascher M, Steuernagel B et al (2013) Distribution, functional impact, and origin mechanisms of copy number variation in the barley genome. Genome Biol 14(6):R58
Nitcher R, Distelfeld A, Tan C, Yan L, Dubcovsky J (2013) Increased copy number at the HvFT1 locus is associated with accelerated flowering time in barley. Mol Genet Genomics 288(5–6):261–275
Oh D-H, Dassanayake M, Bohnert HJ, Cheeseman JM (2012) Life at the extreme: lessons from the genome. Genome Biol 13(3):241
Pâques F, Haber JE (1999) Multiple pathways of recombination induced by double-strand breaks in *Saccharomyces cerevisiae*. Microbiol Mol Biol Rev 63(2):349–404
Park H, Kim J-I, Ju YS, Gokcumen O, Mills RE, Kim S et al (2010) Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. Nat Genet 42(5):400–405
Perry GH, Tchinda J, McGrath SD, Zhang J, Picker SR, Cáceres AM et al (2006) Hotspots for copy number variation in chimpanzees and humans. Proc Natl Acad Sci U S A 103(21): 8006–8011
Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R et al (2007) Diet and the evolution of human amylase gene copy number variation. Nat Genet 39(10):1256–1260
Platzer A, Nizhynska V, Long Q (2012) TE-Locate: a tool to locate and group transposable element occurrences using paired-end next-generation sequencing data. Biology 1(2):395–410

Puchta H (2005) The repair of double-strand breaks in plants: mechanisms and consequences for genome evolution. J Exp Bot 56(409):1–14

Puchta H, Hohn B (1991) A transient assay in plant cells reveals a positive correlation between extrachromosomal recombination rates and length of homologous overlap. Nucleic Acids Res 19(10):2693–2700

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD et al (2006) Global variation in copy number in the human genome. Nature 444(7118):444–454

Saintenac C, Jiang D, Akhunov ED (2011) Targeted analysis of nucleotide and copy number variation by exon capture in allotetraploid wheat genome. Genome Biol 12(9):R88

Scherer SW, Lee C, Birney E, Altshuler DM, Eichler EE, Carter NP et al (2007) Challenges and standards in integrating surveys of structural variation. Nat Genet 39(7 Suppl):S7–S15

Schnable PS, Springer NM (2013) Progress toward understanding heterosis in crop plants. Annu Rev Plant Biol 64:71–88

Schrider DR, Hahn MW (2010) Gene copy-number polymorphism in nature. Proc R Soc B Biol Sci 277(1698):3213–3221

Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P et al (2004) Large-scale copy number polymorphism in the human genome. Science 305(5683):525–528

Simpson JT, Durbin R (2012) Efficient de novo assembly of large genomes using compressed data structures. Genome Res 22(3):549–556

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I (2009) ABySS: a parallel assembler for short read sequence data. Genome Res 19(6):1117–1123

Springer NM, Ying K, Fu Y, Ji T, Yeh C-T, Jia 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(11):e1000734

Stambuk BU, Dunn B, Alves SL Jr, Duval EH, Sherlock G (2009) Industrial fuel ethanol yeasts contain adaptive copy number changes in genes involved in vitamin B1 and B6 biosynthesis. Genome Res 19(12):2271–2278

Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N et al (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315(5813): 848–853

Sutton T, Baumann U, Hayes J, Collins NC, Shi B-J, Schnurbusch T et al (2007) Boron-toxicity tolerance in barley arising from efflux transporter amplification. Science 318(5855): 1446–1449

Swanson-Wagner RA, Eichten SR, Kumari S, Tiffin P, Stein JC, Ware D et al (2010) Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. Genome Res 20(12):1689–1699

Treangen TJ, Salzberg SL (2012) Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet 13(1):36–46

Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM et al (2005) Fine-scale structural variation of the human genome. Nat Genet 37(7):727–732

Xu X, Liu X, Ge S, Jensen JD, Hu F, Li X et al (2012) Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nat Biotechnol 30(1):105–111

Yoon S, Xuan Z, Makarov V, Ye K, Sebat J (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res 19(9):1586–1592

Yu P, Wang C-H, Xu Q, Feng Y, Yuan X-P, Yu H-Y et al (2013) Genome-wide copy number variations in *Oryza sativa* L. BMC Genomics 14:649

Zhang J, Zuo T, Peterson T (2013) Generation of tandem direct duplications by reversed-ends transposition of maize ac elements. PLoS Genet 9(8):e1003691

Zheng L-Y, Guo X-S, He B, Sun L-J, Peng Y, Dong S-S et al (2011) Genome-wide patterns of genetic variation in sweet and grain sorghum (Sorghum bicolor). Genome Biol 12(11):R114

Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA (2013) The MaSuRCA genome assembler. Bioinformatics 29(21):2669–2677

Zmieńko A, Samelak A, Kozłowski P, Figlerowicz M (2014) Copy number polymorphism in plant genomes. Theor Appl Genet 127:1–18