

Changes in genome content generated via segregation of non-allelic homologs

Sanzhen Liu¹, Kai Ying², Cheng-Ting Yeh¹, Jinliang Yang², Ruth Swanson-Wagner^{2,†}, Wei Wu¹, Todd Richmond³, Daniel J. Gerhardt⁴, Jinsheng Lai⁵, Nathan Springer⁶, Dan Nettleton⁷, Jeffrey A. Jeddloh⁴ and Patrick S. Schnable^{1,*}

¹Department of Agronomy, 2035 Roy J. Carver Co-Lab, Iowa State University, Ames, IA 50011, USA,

²Interdepartmental Genetics, 2035 Roy J. Carver Co-Lab, Iowa State University, Ames, IA 50011, USA,

³Research Informatics, Roche NimbleGen, 500 South Rosa Road, Madison, WI 53705, USA,

⁴Development and Research, Roche NimbleGen, 500 South Rosa Road, Madison, WI 53705, USA,

⁵State Key Lab of Agrobiotechnology, China Agriculture University, Beijing, China,

⁶Department of Plant Biology, 250 Biological Science Center, St Paul, MN 55108, USA, and

⁷Department of Statistics, 2115 Snedecor, Iowa State University, Ames, IA 50011, USA

Received 18 May 2012; revised 15 June 2012; accepted 21 June 2012; published online 30 August 2012.

*For correspondence (e-mail schnable@iastate.edu).

[†]Present address: Monsanto, Leeuwenhoekweg 52, 2661 CZ Bergschenhoek, Netherlands
GenBank accession no. SRA036595.

SUMMARY

A careful analysis of two maize recombinant inbred lines (RILs) relative to their inbred parents revealed the presence of several hundred apparently *de novo* copy number variants (CNVs). These changes in genome content were validated via both PCR and whole exome-array capture-and-sequencing experiments. One hundred and eighty-five genomic regions, which overlap with 38 high-confidence genes, exhibited apparently *de novo* copy number variation (CNV) in these two RILs and in many instances the same apparently *de novo* CNV events were observed in multiple RILs. Further analyses revealed that these recurrent apparently *de novo* CNVs were caused by segregation of single-copy homologous sequences that are located in non-allelic positions in the two parental inbred lines. F₁ individuals derived from these inbred lines will be hemizygous for each of these non-allelic homologs but RIL genotypes will contain these sequences at zero, one or two genomic loci. Hence, the segregation of non-allelic homologs may contribute to transgressive segregation. Indeed, statistical associations between phenotypic quantitative trait loci and genomic losses were observed for two of 14 tested pairs of non-allelic homologs.

Keywords: copy number variation, comparative genomic hybridization, maize, non-allelic homologs, transgressive, recombinant inbred line.

INTRODUCTION

Genomes were once thought to be relatively stable structures having conserved gene order and gene content among individuals within a species. More recently, it has been demonstrated that the genomes of two individuals of the same species can exhibit substantial structural variation. For example, copy number variation (CNV), which refers to the number of copies of a particular sequence in different individuals of the same species, has been observed in many species (Iafate *et al.*, 2004; Sebat *et al.*, 2004; Fadista *et al.*, 2008, 2010; She *et al.*, 2008; Chen *et al.*, 2009; Daines *et al.*, 2009; Springer *et al.*, 2009; Conrad *et al.*, 2010; DeBolt, 2010; Hurwitz *et al.*, 2010). Geneticists are still exploring the ramifications of this intraspecific structural variation.

The maize lineage experienced a whole genome duplication approximately 5–12 million years ago (Ma), followed by gene fractionation which removed one member from some pairs of duplicated genes (Langham *et al.*, 2004; Woodhouse *et al.*, 2010). In addition transposons and perhaps other mechanisms have duplicated and transposed genomic sequences (including genes) throughout the genome (Benetzen, 2000; Jiang *et al.*, 2004; Lai *et al.*, 2005; Zhang and Peterson, 2005). In combination these processes have resulted in a large and complex genome. Maize exhibits high levels of intraspecific variation (Schnable *et al.*, 2009; Lai *et al.*, 2010), including thousands of CNVs and genic presence-absence variants (PAVs) (Springer *et al.*, 2009;

Belo *et al.*, 2010; Swanson-Wagner *et al.*, 2010). Maize is an excellent model for the study of quantitative trait variation. It exhibits prevalent transgressive segregation such that the recombinant offspring of two inbred varieties can exhibit phenotypes outside of the parental range.

In this report we document the existence of numerous non-allelic homologs in maize. These non-allelic homologs represent single-copy sequences that are present at different chromosomal locations in different individuals. A detailed analysis of CNVs was undertaken using several recombinant inbred lines (RILs) relative to their inbred parents. We detected more examples of apparently *de novo* CNV than expected and noted that several of these apparently *de novo* CNV were found in multiple RILs. Our investigations of these instances of recurrent apparently *de novo* CNV (Brunetti-Pierrri *et al.*, 2008; Fernandez *et al.*, 2010; Shinawi *et al.*, 2010; Neill *et al.*, 2011) revealed that they were in at least most instances the result of segregation of non-allelic homologs (SNH), which generated RILs that completely lack or have extra copies of a given sequence. Finally, we provide evidence that these changes in sequence content can contribute to phenotypic variation.

RESULTS

Chromosomal segments that exhibit non-parental-signals

Array-based comparative genomic hybridization (aCGH) experiments were conducted using genomic DNAs from two maize inbred lines, B73 and Mo17 (Springer *et al.*, 2009), and two of the inter-mated B73 × Mo17 (IBM) RILs (Lee *et al.*,

2002): M0022 and M0023 (Fu *et al.*, 2010) derived from these parental inbred lines. A careful analysis of the aCGH signals of single-copy (i.e. those mapped to single genomic positions in the B73 genome) probes revealed that 0.06% (1086/1 780 475 in RIL M0022) and 0.13% (2338/1 780 475 in RIL M0023) of probes yielded signals in the RILs that were statistically different from those of *both* parents (Figure 1 and Figures S1–S3 and Table S1 in Supporting information). We defined these probes as putative *de novo* CNV probes. Many of these probes exhibiting unique levels of hybridization signal in the RILs relative to the parental genotypes can be grouped into ‘segments’ using DNACopy software (Experimental Procedures). There are 67 chromosomal segments in RIL M0022, and 130 in RIL M0023 (Table 1 and Table S2) that represent putative *de novo* CNV events in the RILs. These include both gain and loss events, indicating copy number gain and loss, respectively. The average length of these CNV

Table 1 Numbers of putative *de novo* copy number variation (CNV) segments (genes in putative *de novo* CNV segments) derived from M0022 and M0023

	M0023 Gain	M0023 Loss	No gain/ loss in M0023	Total
M0022 Gain	2 (3)	0 (0)	2 (0)	4 (3)
M0022 Loss	3 (3)	7 (1)	53 (10)	63 (14)
No gain/loss in M0022	11 (2)	107 (19)	0 (0)	118 (21)
Total	16 (8)	114 (20)	55 (10)	185 (38)

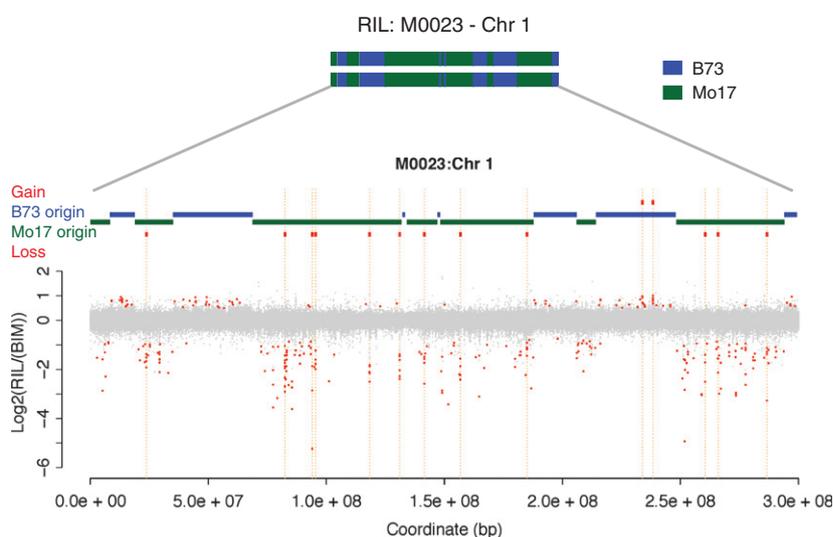


Figure 1. Distribution of putative *de novo* copy number variation (CNV) on chromosome 1.

The physical chromosomal position of each array-based comparative genomic hybridization (aCGH) probe is plotted versus the \log_2 ratio of its hybridization signals in recombinant inbred line M0023 and the appropriate parent (i.e. B73 or Mo17) [$\log_2(\text{RIL}/\text{BIM})$, see Experimental Procedures]. Probes that did or did not exhibit statistically significant signal losses or gains relative to *both* parents are highlighted in red and grey, respectively. Above the X–Y plot, the chromosomal regions from RIL M0023 that were derived from B73 and Mo17 based on genotyping experiments are color-coded in blue and green, respectively. These marker-based assignments of parental origin of chromosomal regions were inferred via segmentation of all aCGH probes that could be reliably classified as having B73-like or Mo17-like signals. Subsequently, a second segmentation was conducted to identify putative *de novo* CNV segments that are indicated in red.

segments is 2.8 kb with the longest being 53.4 kb (Figure S1). The majority (69%) of the 197 chromosomal segments do not exhibit substantial differences in aCGH signal between B73 and Mo17. Hence, these aCGH results indicate that although the corresponding probes and chromosomal segments do not exhibit CNV between the inbred parents, these sequences do exhibit CNV in the RILs. These putative *de novo* CNV comprise 0.009 and 0.012% of the M0022 and M0023 genomes, respectively.

To confirm and extend these results a NimbleGen whole exome-array was used to capture and sequence the genic portions of the chromosomal segments that exhibit non-parental hybridization levels (see Experimental Procedures). We separately ($n = 4$) captured genomic DNAs from the parental inbred lines (B73 and Mo17) and the two RILs (M0022 and M0023) using a previously published protocol (Haun *et al.*, 2010). Then 32–37 million 40-bp paired-end Illumina reads generated from each capture were aligned to the B73 reference genome (GenBank accession no. SRA036595). Reads that uniquely mapped to the 197 chromosomal segments identified via aCGH were counted for all four genotypes. Figure 2 compares the aCGH and exome-Seq results of two chromosomal segments in both RILs. For subsequent analyses we focused on the 61 segments for which at least 30 reads were obtained from the B73 capture. Of these, 54 and 7 segments had exhibited aCGH signal losses (copy losses) and gains (copy gains) in the RILs relative to B73, respectively (Figure S4). The aCGH and exome-Seq count data for these segments are highly correlated (correlation = 0.62). In the vast majority (45/54) of signal loss segments, at least 80% fewer reads were obtained from the affected RIL than from B73. Notably, no reads were obtained from the affected RIL in 24 (of 54) chromosomal segments that exhibited signal loss in the aCGH experiments, demonstrating that these segments are completely absent from the RIL genome. Consistently, more reads were recovered from the RILs than from B73 for all seven segments that exhibited signal gains in the RILs relative to B73, indicating that the RIL genomes contain more copies of these segments than do the parental genomes (Figure S4). Similar results were observed in comparisons to Mo17 (Figure S4). These exome-Seq results confirm the existence of apparently *de novo* CNV for many of the chromosomal segments in the RIL genotypes.

Apparently *de novo* copy number variation is the result of segregation of non-allelic homologs

There are a number of potential mechanisms for *de novo* CNV formation including non-allelic homologous recombination (NAHR), rearrangements in the absence of extended sequence similarity associated with DNA repair by non-homologous end-joining (NHEJ) or with microhomology-mediated break-induced replication (MMBIR), con-

traction or expansions of variable number tandem repeats (VNTRs) and mobile element insertions (MEI) (Mills *et al.*, 2011). Alternatively, apparently *de novo* CNV can be formed by segregation of single-copy sequences that are located in non-allelic positions. If two parental lines both contain a single copy of a sequence that is located at unlinked genomic positions, then the F_1 will be hemizygous for each of these copies and meiotic segregation will generate F_2 (or RIL) genotypes with zero, one or two copies of the sequence (Lu *et al.*, 2012). The relatively high rate of apparently *de novo* CNV observed in the RILs suggested that segregation of non-allelic homologs (SNH) might be responsible.

Several lines of evidence support this hypothesis. First, we noted that the locations of copy number gains and losses in the RILs exhibit dependence on the parental origins of the chromosomal segments containing these gains or losses. Those associated with significant signal losses in the RILs are embedded within Mo17-derived chromosomal regions, while 19/20 apparently *de novo* CNVs with significant signal gains are embedded within B73-derived regions (Figure 1). This would be expected if the apparently *de novo* CNV arose via SNH. Second, 12 chromosomal segments exhibited apparently *de novo* CNV in both RILs, resulting in 185 non-redundant segments (Table 1 and Table S2). SNH would be expected to yield a high rate of recurrent apparently *de novo* CNV. To further examine the degree to which the apparently *de novo* CNVs are shared among RILs, PCR primers were designed based on several chromosomal segments that exhibited signal loss in at least one of the RILs and used to amplify products in 300 IBM RILs. Every one of these segments was missing in multiple RILs (6–34%) (Table S3), providing evidence that these segments exhibit frequent apparently *de novo* CNV origin, consistent with SNH.

The third piece of evidence that supports the role of SNH in the generation of these apparently *de novo* CNVs is based on mapping of these sequences in B73 and Mo17 using mate pairs (see Experimental Procedures). Using conservative criteria 40 (M0022) and 68 (M0023) of the 197 segments exist in non-allelic positions in the B73 and Mo17 genomes (Tables S4 and S5). In most instances (5621/5709) the positions of non-allelic homologs inferred by aligning the Mo17 mate pair reads to the B73 reference genome do not correspond to expectations based on the whole genome duplication event (Krzywinski *et al.*, 2009; Schnable *et al.*, 2009), indicating that differential losses of genes in duplicated genomes is not the dominant mechanism underlying SNH. In contrast, only two (M0022) and six (M0023) of 197 random control segments mapped to non-allelic positions in the Mo17 genome (Figure 3, Table S4). Hence, consistent with the SNH model, more than half of the sequences that give rise to apparently *de novo* CNV are located in non-allelic positions in the B73 and Mo17 genomes.

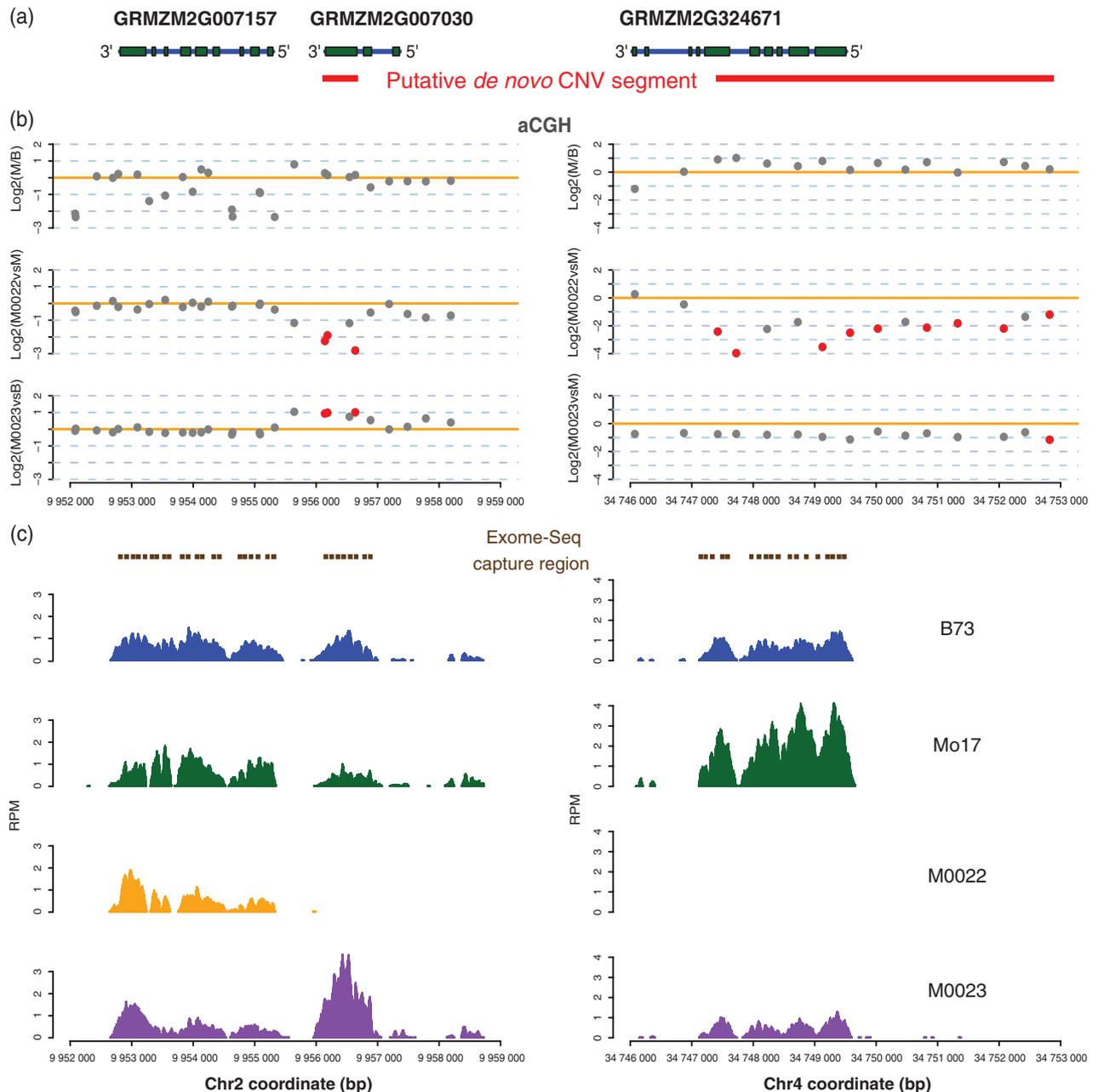


Figure 2. Analysis of two apparently *de novo* copy number variation (CNV) segments via array-based comparative genomic hybridization (aCGH) and exome-Seq. (a) Coding regions of three genes from two genomic regions are illustrated by green boxes. (b) The \log_2 of the ratios of normalized probe signals between different genotypes from aCGH are plotted versus the probes' physical positions. Probes that did or did not exhibit statistically significant signal losses or gains relative to both parents are highlighted in red and grey, respectively. Probes in red represent putative *de novo* CNV probes. Comparisons were conducted between the recombinant inbred lines and the parent that contributed the relevant chromosomal segment. (c) Read counts from the exome capture experiment at each nucleotide position were normalized to reads per million reads (RPM) for each genotype.

Impact of segregation of non-allelic homolog-derived copy number variation on phenotypic traits

The SNH model predicts that meiotic segregation will act on non-allelic homologs, resulting in novel complements of sequences (losses and gains) among progeny relative to parental haplotypes (Figure 4). Those SNH-derived CNVs

that involve genes would be particularly interesting because they could result in novel genic complements in progeny relative to parents.

The maize genome sequencing project defined a set of 32 540 high-quality gene annotations that is referred to as the Filtered Gene Set (FGS, ver. 4a.53). Prior to inclusion in the FGS, gene models were rigorously filtered to remove

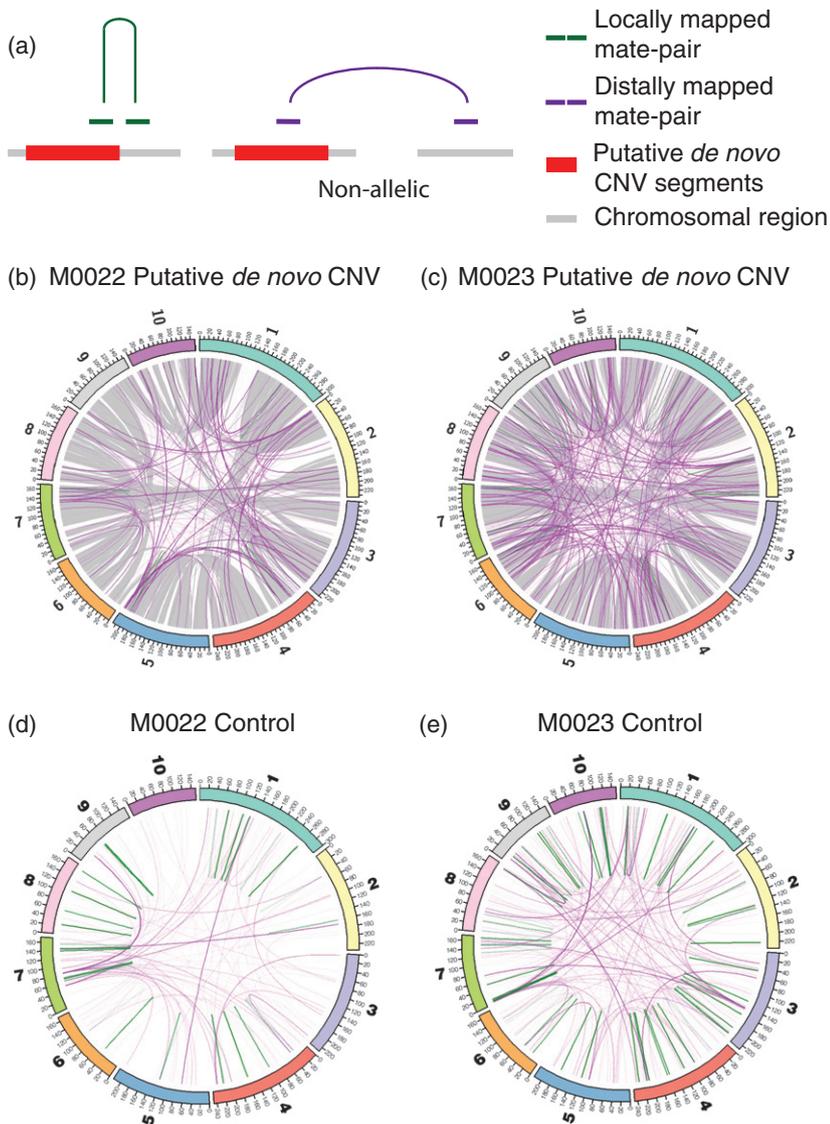


Figure 3. Determination of the allelic status in the B73 and Mo17 genomes of sequences that give rise to apparently *de novo* copy number variation (CNV).

(a) Mo17 mate-pair reads with 5-kb of insertion were mapped separately to the B73 reference genome. Pairs with at least one read uniquely mapped to the apparently *de novo* CNV segments (red horizontal bars) and the other read were uniquely mapped to the reference genome were categorized to two types, locally mapped mate pair (green horizontal bars) and distally mapped mate pair (purple horizontal bars). A locally mapped mate pair linked by a green line represents mate-pair reads that were mapped within 1 Mb, while distally mapped mate pair reads that were mapped at >1 Mb distance or on different chromosomes. These rules were applied to draw figures in the panels (b–e).

(b–e) Mate-pair reads associated with the apparently *de novo* CNV segments and control segments from the indicated genotypes are color-coded as shown in (a). Each line connects mapped locations of two Mo17 reads of each mate pair, at least one of which was uniquely mapped to the apparently *de novo* CNV segments of M0022 (b) and the apparently *de novo* CNV segments of M0023 (c), randomly selected segments simulating M0022 apparently *de novo* CNV segments (d) and randomly selected segments simulating M0023 apparently *de novo* CNV segments (e). The number of read pairs clustered was intensity-coded as shown in (a). The transparency factor for each line is 0.05. In (b) and (c) the homeologous sites of duplicated blocks derived from the ancestral whole genome duplication in the B73 reference genome are shown in grey.

gene fragments and sequences with similarity to transposons. Thirty-five of the observed cases of SNH-derived CNVs overlap (partially or completely) with 38 of the high-quality gene models in the FGS (Table S6). The RNA-Seq data from apices (GenBank accession no. SRA036595; see Experimental Procedures) provided evidence of expression for 24 (63%) of these genes that are affected by SNH-derived CNV (Table S6). In addition, 27 (71%) of these genes have homologs in sorghum or rice, indicating phylogenetic conservation. In combination, these lines of evidence suggest that many of the genes affected by SNH-derived CNV are probably functional.

To test whether changes in gene complement caused by SNH-derived CNV have phenotypic consequences, we collected data on a number of phenotypic traits from the ~300 IBM RILs discussed above. Each of the 14 assayed cases of SNH-derived CNV fully or partially overlaps at least one of

the high-confidence genes in the FGS. We then compared the average phenotypic trait values of RILs that did or did not experience gene loss via SNH. After controlling for multiple testing (see Experimental Procedures), losses of two of the 14 tested chromosomal intervals were significantly associated with phenotypic variation. Chromosomal interval M0022_seg30 is significantly associated with reduced cob diameter (adjusted P -value = 0.03) and kernel row number (adjusted P -value = 0.01). Similarly chromosomal interval M0022_seg15/M0023_seg22 (which includes a putative peptidyl–prolyl *cis–trans* isomerase gene) is associated with increased tiller number (adjusted P -value = 0.01) (Table S7).

DISCUSSION

De novo CNV has been hypothesized to arise via transposon-, recombination- and replication-mediated mechanisms

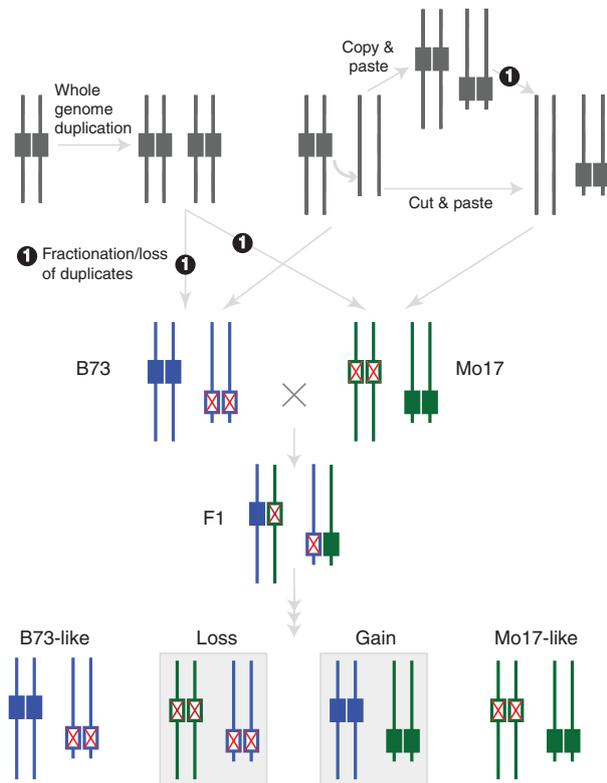


Figure 4. The segregation of non-allelic homologs model for the origin of recurrent apparently *de novo* copy number variation (CNV).

Identical or nearly identical DNA sequences are located on different chromosomes in the B73 and Mo17 genomes (non-allelic homologs). B73- and Mo17-derived chromosomes are indicated by blue and green, respectively. Filled boxes and open boxes containing a red 'X' designate the presence and absence of a non-allelic homolog, respectively. This model is consistent with our finding that most recurrent apparently *de novo* CNVs exhibit losses in ~25% (two loci) or ~12.5% (three loci) of the recombinant inbred lines. It is also consistent with our finding that all of the 165 losses are embedded in Mo17-derived segments and that almost all (19/20) of the copy number gains are embedded in B73-derived segments (Figure 1). This model predicts that copy number losses and gains should occur at equal frequencies. We hypothesize that losses exceed gains due to ascertainment bias (i.e. copy number losses are more easily detected than copy number gains).

(Hastings *et al.*, 2009; Springer *et al.*, 2009; Conrad *et al.*, 2010; Innan and Kondrashov, 2010; Stankiewicz and Lupski, 2010; Mills *et al.*, 2011). The association between the distributions of gains and losses of apparently *de novo* CNV observed in this study and the parental origins of the surrounding chromosomal segments (Figure 1) is inconsistent with transposon-mediated mechanisms acting during the several generations required to produce the RILs. Further, the high rates of recurrence of apparently *de novo* CNV are inconsistent with recombination- and replication-driven mechanisms because these mechanisms are reported to generate losses and amplifications at much lower rates (van Ommen, 2005; Yandeu-Nelson *et al.*, 2006; Lupski, 2007; Turner *et al.*, 2008). In contrast, SNH-derived CNV is not the

result of active rearrangements of DNA but is instead the result of meiotic segregation acting upon transposed gene copies and, in some cases, fractionation events following whole genome duplication. Collectively, our observations suggest that SNH results in CNV for hundreds of maize loci.

The maize genome is a product of an ancient tetraploidization event and now consists of two 'subgenomes' (Schnable *et al.*, 2011) having different properties, including gene expression levels. Intra-chromosomal recombination events can result in the loss of the copy of a pair of homologs from one subgenome (Woodhouse *et al.*, 2010). Although the two subgenomes exhibit different rates of genes loss (Schnable *et al.*, 2011), this process and others such as transposon-mediated gene duplication/transposition (Jiang *et al.*, 2004; Lai *et al.*, 2005) have generated numerous non-allelic homologs (Eichten *et al.*, 2011). We have demonstrated that meiotic segregation of these non-allelic homologs generates CNV affecting hundreds of loci in the progeny of a single cross (Figure 4). The 185 detected SNH-derived CNV affect 38 high-confidence genes. Considering the stringent criteria used in this study, this frequency is likely to be an underestimate.

The SNH model exhibits similarities to the 'reciprocal gene loss model' first proposed by Lynch and Force (2000) to explain interspecific genomic incompatibility. This model proposed that the loss of different copies of duplicated genes in different populations would lead to gene loss in gametes from F₁ individuals. This process has been demonstrated in crosses among three yeast species (Scannell *et al.*, 2006) and between two fish species (Semon and Wolfe, 2007). It has also been shown to affect single genes in several intraspecific studies, including *Drosophila* (Masly *et al.*, 2006) and *Arabidopsis* (Bikard *et al.*, 2009). The SNH model differs from the reciprocal gene loss model in that it occurs intraspecifically, can generate copy number gains and can act on non-allelic homologs generated via various mechanisms. The SNH model would be expected to generate CNV in any species that contains non-allelic homologs and undergoes meiotic segregation. In maize we believe some of the non-allelic homologs arise via fractionation, but the mechanism outlined in Figure 4 can occur regardless of the mechanism by which the non-allelic homologs were originally generated. For example, with only minor modifications this mechanism could also generate CNVs in a species that contains non-allelic homologs generated via the transposition of single-copy genes (Vlad *et al.*, 2010).

Phenotypic effects of segregation of non-allelic homolog-derived copy number variation and presence-absence variants

Although CNV has previously been associated with genetic disorders in humans (Stankiewicz and Lupski, 2010), this report provides evidence that the segregation of CNV via the SNH model can also contribute to the phenotypic

variation present in crop species such as maize. This model may also shed light on transgressive segregation, i.e. the appearance of progeny from a bi-parental cross whose phenotypic values exceed those of their parents (Rieseberg *et al.*, 1999).

The findings reported here have significant implications for the large-scale efforts under way to identify the genetic determinants of phenotypic variation in humans, model and agricultural species. This is because the genetic determinants of phenotypic variation arising via the SNH model will not be detected via traditional single marker association studies or QTL analyses. Indeed, the synergistic effects from multiple unlinked genomic loci are likely to lower the power of such traditional one-dimensional analyses. Multiple-dimensional scans that consider the synergistic effects of multiple markers on phenotypes can overcome this limitation of traditional genetic mapping approaches. Although computationally intensive, such studies are now tractable (Koesterke *et al.*, 2011). Another concern is that if PAVs are not in linkage disequilibrium (LD) with nearby genetic markers such as SNPs, the power of association studies that rely on such markers will be reduced. It will be interesting to determine whether the direct genotyping of PAVs via CGH-based genotyping (Fu *et al.*, 2010) or genotyping-by-sequencing approaches (Fogu *et al.*, 2007; Huang *et al.*, 2009; Lai *et al.*, 2010; Andolfatto *et al.*, 2011; Elshire *et al.*, 2011) will uncover at least a fraction of the 'missing heritability' observed in genome-wide association studies (Kump *et al.*, 2010).

EXPERIMENTAL PROCEDURES

Genetic stocks

Two maize inbred lines, B73 and Mo17, and two RILs, M0022 and M0023, were extracted from the IBM Syn4 population (Lee *et al.*, 2002). The RILs used in this study were from the F7–9 generation.

Identification of putative *de novo* copy number variation probes

The aCGH experiments (GEO: GSE16938), data processing and statistical analyses were performed as described previously (Springer *et al.*, 2009; Fu *et al.*, 2010). Contrasts were performed between B73 versus Mo17, B73 versus RIL and Mo17 versus RIL. A *P*-value was determined for each probe from each contrast. To account for multiple testing, *P*-values were converted to *q*-values (Benjamini and Hochberg, 1995). Probes with significantly higher or lower signals in the RILs when compared with both B73 and Mo17 were termed putative *de novo* CNV probes. Signal loss probes were called using the criteria of *q*-value (RIL versus B73) <0.0001, *q*-value (RIL versus Mo17) <0.0001, $\log_2(\text{signal ratios of RIL/B73}) < 0$ and $\log_2(\text{signal ratios of RIL/Mo17}) < 0$; signal gain probes were called using the criteria of *q*-value (RIL versus B73) <0.001, *q*-value (RIL versus Mo17) <0.001, $\log_2(\text{signal ratios of RIL/B73}) > 0$ and $\log_2(\text{signal ratios of RIL/Mo17}) > 0$. Different *q*-value cutoffs were used to identify signal loss probes and signal gain probes because aCGH technology has a greater power to detect copy number losses than copy number gains (Altshuler *et al.*, 2010).

Determination of appropriate parental control for each probe in array-based comparative genomic hybridization analyses of recombinant inbred lines

For each aCGH probe, the \log_2 ratios of the hybridization signals of the RIL versus B73 [$\log_2(\text{RIL/B})$] and separately versus Mo17 [$\log_2(\text{RIL/M})$] were calculated. A \log_2 ratio that is greater (or smaller) than 0 indicates that a given probe yields a stronger (weaker) signal in the RIL than in a particular parental inbred. For each aCGH probe, the smaller of absolute value of $\log_2(\text{RIL/B})$ and the absolute value of $\log_2(\text{RIL/M})$ was used to identify the presumptive parental origin (B73 or Mo17) in the RIL of the chromosomal segment from which each probe was designed. For each probe this parental hybridization value [$\log_2(\text{RIL/BIM})$] was used for the calculations plotted in Figure 1.

Segmentation of putative *de novo* copy number variation probes to identify putative *de novo* copy number variation segments

The putative *de novo* CNV probes were converted to 1 (copy gain compared with B73) or -1 (copy loss compared with B73). All other probes were assigned a value of 0. The converted binary data were subjected to segmentation via DNACopy (Olshen *et al.*, 2004) using the parameters: alpha = 0.01, nperm = 10 000, p.method = 'perm', eta = 0.01, min.width = 3. Putative *de novo* CNV segments were required to contain at least three putative *de novo* CNV probes and a median absolute deviation (MAD) equal to 0.

Recombinant inbred line segmentation to distinguish the origin of regions from either B73 or Mo17

Probes that distinguished B73 and Mo17 were identified using the following criteria: *q*-value (Mo17 versus B73) <0.0001 and $\log_2(\text{Mo17/B73}) < (-1)$. These probes were treated as genetic markers to genotype the RILs, most of which were grouped into Mo17-type [*q*-value (RIL versus B73) <0.001 and *q*-value (RIL versus Mo17) >0.1] or B73-type [*q*-value (RIL versus B73) >0.1 and *q*-value (RIL versus Mo17) <0.001].

To avoid the partition of a chromosomal region exhibiting the same origin into multiple segments, the array CGH genotyping results were converted to binary data (B73-type = 1; Mo17-type = 0) and merged to perform segmentation as described above. Segments smaller than 200 kb and having a mean segment value between 0.1 and 0.9 were removed from further analysis. By so doing, we excluded segments that were ambiguously assigned as being B73-type or Mo17-type and small segments that might represent mis-assemblies in the reference genome.

Exome-Seq

Gene annotation information was downloaded from <http://www.maizesequence.org>. This annotation set was defined as the entire set of evidence-based genes (predicted by Gramene GeneBuilder), complemented by a set of Fgenesh models. Pseudogenes, TE-encoded genes and low-confidence models were filtered out to produce the final annotation set. Coding sequence coordinates were extracted from this annotation (251 067 regions; 55.5 Mbp) and consolidated into non-overlapping regions (152 529 regions; 37.9 Mbp). These regions were padded to a minimum region size of 100 bp, and again consolidated into non-overlapping regions (151 929 regions; 39.7 Mbp). Final coordinates were offset by 35 bp to account for capture probe length overhang at the end of each region. Variable length capture probes (50- to 100-mers)

were selected by tiling through each region at an average spacing of 48 bp (measured from 5' start to 5' start). Repeat-masking was done by generating a histogram of all 15-mers in the maize genome and removing probes with an average 15-mer frequency greater than 100. Probe uniqueness was assessed using SSAHA (<http://www.sanger.ac.uk/resources/software/ssaha/>), using a minimum match size of 26. No more than three matches were allowed for capture probes. The final design covers 131 469 of the 151 929 original regions, and has a total capture space of 55.4 Mbp. The Sequence Capture Developer 2.1M feature array design 100224_ZmB73_public_exome_cap_HX3 is available for purchase from Roche NimbleGen (<http://www.nimblegen.com/>).

Analysis of Exome-Seq data

Novoalign 2.05.31 (<http://www.novocraft.com/>) was used to map all reads (40 bp) to reference genomes, including the B73 reference genome (B73ref_v1), the maize mitochondrial genome (GenBank accession no. AY506529.1) and the maize chloroplast genome (GenBank accession no. X86563.2). Reads that were uniquely mapped with two or fewer mismatches (insertions and deletions was counted as mismatches) were used for further analysis. Read counts of each putative *de novo* CNV segment were adjusted by the addition of 1 to avoid zero values and were then used to calculate $\log_2(\text{RIL}/\text{B73})$ and $\log_2(\text{RIL}/\text{Mo17})$.

RNA-Seq

The RNA was extracted from a pool of three to six apices of 14-day-old seedlings of the inbred lines B73 and Mo17 using the Qiagen RNeasy Plant Mini Kit (catalog no. 74903, <http://www.qiagen.com/default.aspx>). RNA-Seq was conducted using an Illumina GAIIX instrument at the Iowa State University DNA facility following an Illumina protocol (mRNA-seq Sample Preparation Guide; <http://www.illumina.com/>).

Genotyping of a subset of the apparently *de novo* copy number variation segments on the recombinant inbred line population

Primers were designed on a subset of the apparently *de novo* CNV segments that fully or partially overlap with transcription- or annotation-supported genes to directly genotype >300 IBM recombinant inbred lines. A primer pair was designed for each selected segment using the B73 reference sequence. Primers were aligned to JGI 454 Mo17 reads to ensure that primers work for Mo17. This PCR program consisted of 94°C for 10 min; 35 cycles of 94°C for 30 sec, 60°C for 45 sec, 72°C for 1.5 min, and a final extension at 72°C for 10 min in a 20- μ l volume. Three existing genetic markers of the IBM genetic map (Liu *et al.*, 2009), IDP525, IDP7957 and IDP7324, were used as the PCR control primers. The IDP525 marker has been used to genotype the full set of RILs previously (Liu *et al.*, 2009). The RILs with poor or no PCR amplification of any IDP markers or inconsistent IDP525 PCR results between the previous scores and the re-genotyping scores were not used for further analysis.

Phenotyping recombinant inbred lines

The following traits were collected from 291 IBM RILs: seedling dry weight, average kernel weight, cob diameter, cob length, cob weight, kernel count, kernel row number, total kernel weight, tiller number, ferulic acid (FA), *p*-coumaric acid (PCA), brace node number (BN), node number above primary ear (NA), node number below primary ear (NB) and total node number (NT). Four replicates of seedling dry weight trait data and three replications of other traits

data were collected. Least square means for each genotype were estimated using the SAS code for the traits except for tiller number. Mean of tiller number for each genotype was calculated among replicates.

Phenotypic associations of segregation of non-allelic homolog-derived copy number variants

We used *t*-tests to test the null hypothesis that no association exists between segmental copy loss and each of the phenotypic traits. For each tested segment, RILs were divided into two groups based on the results of PCR-based genotyping, i.e. RILs with and without the expected PCR bands. The *t*-test assuming equal variation between two groups was conducted for each trait, generating a *P*-value. A permutation test was implemented to account for multiple testing for each segmental copy loss. In each permutation of a given segment, the genotyping scores of the segment were randomly shuffled among the RILs. A similar *t*-test was performed on the shuffled genotyping data for each trait and a *P*-value was obtained for each trait. The smallest *P*-value was then determined among the multiple *P*-values from the multiple trait tests. This procedure was repeated 1000 times. One thousand *P*-values were obtained for each of the tested segments. The original *P*-values of the multiple traits computed from the observed data of a given segment were compared with the 1000 *P*-values from the permutation test of this segment to generate adjusted *P*-values.

Physical mapping of Mo17 sequences involved in apparently *de novo* copy number variation

To determine the chromosomal location of paired reads, the mate-pair cluster mapping method was used. First, the individual reads from 5-kb Mo17 mate pairs (one lane from an Illumina/Solexa GA-IIX instrument) were separately mapped to the B73 reference genome. A mate pair was used for further analyses only if one and only one member of that mate pair uniquely mapped to an apparently *de novo* CNV segment (or within the 500 bp beyond either end of the segment because the actual endpoints of loss or duplication are not known with certainty) in the B73 reference genome. For each mate pair of this type, if the other member mapped within 1 Mb of the same apparently *de novo* CNV segment, the pair was categorized as having similar locations in the B73 and Mo17 genomes ('locally mapped'). If the mapped location of the other member of a mate pair was >1 Mb away or on a different chromosome, the mate pair was categorized as having different locations in the B73 and Mo17 genomes ('distally mapped'). The identification of multiple, independent distally mapped mate pairs clustered on a chromosomal region was considered evidence that the corresponding chromosomal segments in B73 and Mo17 are non-allelic. A 'cluster' of mate-pair reads was defined as consisting of two or fewer independent non-stacked reads that mapped within <100 kb of each other and that covered >80 bp of the CNV segment. Two sets of segments (one from M00022 and one from M0023) were randomly selected as the controls. For each RIL, the number of randomly selected segments from each chromosome was equal to the number of identified apparently *de novo* CNV segments on that chromosome in the corresponding RIL.

ACKNOWLEDGEMENTS

We thank Stephen Moldovan, Ho Man Tang and Marianne Smith for technical assistance and Tracy Millard and Dr Thomas J. Albert for sequencing and project support. We thank Roche NimbleGen for providing support for this study through the donation of reagents, and team-member time. The co-authors TR, DJG and JAJ recognize

a competing interest in this publication as employees of Roche NimbleGen, Inc. This research was also supported by a grant from the National Science Foundation to PSS (IOS-1027527) and data generated as part of NSF grant IOS-0820610 (Mike Scanlon, PI).

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Figure S1. Potential *de novo* copy number variation (CNV) probes and sizes of apparently *de novo* CNV segments.

Figure S2. Distribution of array-based comparative genomic hybridization probes on all 10 chromosomes of the recombinant inbred line M0022.

Figure S3. Distribution of apparently *de novo* copy number variation segments on nine chromosomes for the recombinant inbred line M0023.

Figure S4. Validation of apparently *de novo* copy number variation segments using exome-Seq data.

Table S1. Detailed information on array-based comparative genomic hybridization probes.

Table S2. Apparently *de novo* copy number variation segments in the recombinant inbred lines.

Table S3. Detection by PCR of copy number loss in some apparently *de novo* copy number variation segments.

Table S4. Mapping Mo17 mate-pair reads associated with apparently *de novo* copy number variation segments on the B73 reference genome.

Table S5. Mapping result of Mo17 5-kb mate-pair reads derived from apparently *de novo* copy number variation segments to the B73 reference genome.

Table S6. Genes in apparently *de novo* copy number variation segments.

Table S7. Genotype/phenotype associations between apparently *de novo* copy number variation segments and traits.

Please note: As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.

REFERENCES

Altshuler, D.M., Gibbs, R.A., Peltonen, L. *et al.* (2010) Integrating common and rare genetic variation in diverse human populations. *Nature*, **467**, 52–58.

Andolfatto, P., Davison, D., Erezylmaz, D., Hu, T.T., Mast, J., Sunayama-Morita, T. and Stern, D.L. (2011) Multiplexed shotgun genotyping for rapid and efficient genetic mapping. *Genome Res.* **21**, 610–617.

Belo, A., Beatty, M.K., Hondred, D., Fengler, K.A., Li, B. and Rafalski, A. (2010) Allelic genome structural variations in maize detected by array comparative genome hybridization. *Theor. Appl. Genet.* **120**, 355–367.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statistical Society, Series B*, **57**, 289–300.

Bennetzen, J.L. (2000) Transposable element contributions to plant gene and genome evolution. *Plant Mol. Biol.* **42**, 251–269.

Bikard, D., Patel, D., Le Mette, C., Giorgi, V., Camilleri, C., Bennett, M.J. and Loudet, O. (2009) Divergent evolution of duplicate genes leads to genetic incompatibilities within *A. thaliana*. *Science*, **323**, 623–626.

Brunetti-Pierri, N., Berg, J.S., Scaglia, F. *et al.* (2008) Recurrent reciprocal 1q21.1 deletions and duplications associated with microcephaly or macrocephaly and developmental and behavioral abnormalities. *Nat. Genet.* **40**, 1466–1471.

Chen, W.K., Swartz, J.D., Rush, L.J. and Alvarez, C.E. (2009) Mapping DNA structural variation in dogs. *Genome Res.* **19**, 500–509.

Conrad, D.F., Pinto, D., Redon, R. *et al.* (2010) Origins and functional impact of copy number variation in the human genome. *Nature*, **464**, 704–712.

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

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

Eichten, S.R., Foerster, J.M., de Leon, N., Kai, Y., Yeh, C.T., Liu, S., Jeddeloh, J.A., Schnable, P.S., Kaeppler, S.M. and Springer, N.M. (2011) B73-Mo17 near-isogenic lines demonstrate dispersed structural variation in maize. *Plant Physiol.* **156**, 1679–1690.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S. and Mitchell, S.E. (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One*, **6**, e19379.

Fadista, J., Nygaard, M., Holm, L.E., Thomsen, B. and Bendixen, C. (2008) A snapshot of CNVs in the pig genome. *PLoS One*, **3**, e3916.

Fadista, J., Thomsen, B., Holm, L.E. and Bendixen, C. (2010) Copy number variation in the bovine genome. *BMC Genomics*, **11**, 284.

Fernandez, B.A., Roberts, W., Chung, B. *et al.* (2010) Phenotypic spectrum associated with *de novo* and inherited deletions and duplications at 16p11.2 in individuals ascertained for diagnosis of autism spectrum disorder. *J. Med. Genet.* **47**, 195–203.

Fogu, G., Bandiera, P., Cambosu, F., Carta, A.R., Pilo, L., Serra, G., Soro, G., Tondi, M., Tusacchi, G. and Montella, A. (2007) Pure partial trisomy of 6p12.1-p22.1 secondary to a familial 12/6 insertion in two malformed babies. *Eur. J. Med. Genet.* **50**, 103–111.

Fu, Y., Springer, N.M., Ying, K. *et al.* (2010) High-resolution genotyping via whole genome hybridizations to microarrays containing long oligonucleotide probes. *PLoS One*, **5**, e14178.

Hastings, P.J., Ira, G. and Lupski, J.R. (2009) A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLoS Genet.* **5**, e1000327.

Haun, W.J., Hyten, D.L., Xu, W.W. *et al.* (2010) The composition and origins of genomic variation among individuals of the soybean reference cultivar Williams 82. *Plant Physiol.* **155**, 645–655.

Huang, X., Feng, Q., Qian, Q. *et al.* (2009) High-throughput genotyping by whole-genome resequencing. *Genome Res.* **19**, 1068–1076.

Hurwitz, B.L., Kudrna, D., Yu, Y., Sebastian, A., Zuccolo, A., Jackson, S.A., Ware, D., Wing, R.A. and Stein, L. (2010) Rice structural variation: a comparative analysis of structural variation between rice and three of its closest relatives in the genus *Oryza*. *Plant J.* **63**, 990–1003.

Iafraite, A.J., Feuk, L., Rivera, M.N., Listewnik, M.L., Donahoe, P.K., Qi, Y., Scherer, S.W. and Lee, C. (2004) Detection of large-scale variation in the human genome. *Nat. Genet.* **36**, 949–951.

Innan, H. and Kondrashov, F. (2010) The evolution of gene duplications: classifying and distinguishing between models. *Nat. Rev. Genet.* **11**, 97–108.

Jiang, N., Bao, Z., Zhang, X., Eddy, S.R. and Wessler, S.R. (2004) Pack-MULE transposable elements mediate gene evolution in plants. *Nature*, **431**, 569–573.

Koesterke, L., Stanzione, D., Vaughn, M., Welch, S.M., Kusnierczyk, W., Wang, J., Yeh, C., Nettleton, D. and Schnable, P.S. (2011) An efficient and scalable implementation of SNP-pair interaction testing for genetic association studies. In *IEEE International Workshop on High Performance Computational Biology*. Anchorage, Alaska.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J. and Marra, M.A. (2009) Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645.

Kump, K.L., Bradbury, P.J., Wisser, R.J. *et al.* (2010) Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. *Nat. Genet.* **43**, 163–168.

Lai, J., Li, Y., Messing, J. and Dooner, H.K. (2005) Gene movement by Helitron transposons contributes to the haplotype variability of maize. *Proc. Natl Acad. Sci. USA*, **102**, 9068–9073.

Lai, J., Li, R., Xu, X. *et al.* (2010) Genome-wide patterns of genetic variation among elite maize inbred lines. *Nat. Genet.* **42**, 1027–1030.

Langham, R.J., Walsh, J., Dunn, M., Ko, C., Goff, S.A. and Freeling, M. (2004) Genomic duplication, fractionation and the origin of regulatory novelty. *Genetics*, **166**, 935–945.

- Lee, M., Sharopova, N., Beavis, W.D., Grant, D., Katt, M., Blair, D. and Hallauer, A. (2002) Expanding the genetic map of maize with the intermated B73 x Mo17 (IBM) population. *Plant Mol. Biol.* **48**, 453–461.
- Liu, S., Yeh, C.T., Ji, T., Ying, K., Wu, H., Tang, H.M., Fu, Y., Nettleton, D. and Schnable, P.S. (2009) Mu transposon insertion sites and meiotic recombination events co-localize with epigenetic marks for open chromatin across the maize genome. *PLoS Genet.* **5**, e1000733.
- Liu, P., Han, X., Qi, J., Yang, J., Wijeratne, A.J., Li, T. and Ma, H. (2012) Analysis of Arabidopsis genome-wide variations before and after meiosis and meiotic recombination by resequencing Landsberg erecta and all four products of a single meiosis. *Genome Res.* **22**, 508–518.
- Lupski, J.R. (2007) Genomic rearrangements and sporadic disease. *Nat. Genet.* **39**, S43–S47.
- Lynch, M. and Force, A. (2000) The origin of interspecific genomic incompatibility via gene duplication. *Am. Nat.* **156**, 590–605.
- Masly, J.P., Jones, C.D., Noor, M.A., Locke, J. and Orr, H.A. (2006) Gene transposition as a cause of hybrid sterility in *Drosophila*. *Science*, **313**, 1448–1450.
- Mills, R.E., Walter, K., Stewart, C. *et al.* (2011) Mapping copy number variation by population-scale genome sequencing. *Nature*, **470**, 59–65.
- Neill, N.J., Ballif, B.C., Lamb, A.N., Parikh, S., Ravnán, J.B., Schultz, R.A., Torchia, B.S., Rosenfeld, J.A. and Shaffer, L.G. (2011) Recurrence, submicroscopic complexity, and potential clinical relevance of copy gains detected by array CGH that are shown to be unbalanced insertions by FISH. *Genome Res.* **21**, 535–544.
- Olshen, A.B., Venkatraman, E.S., Lucito, R. and Wigler, M. (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**, 557–572.
- van Ommen, G.J. (2005) Frequency of new copy number variation in humans. *Nat. Genet.* **37**, 333–334.
- Rieseberg, L.H., Archer, M.A. and Wayne, R.K. (1999) Transgressive segregation, adaptation and speciation. *Heredity*, **83**, 363–372.
- Scannell, D.R., Byrne, K.P., Gordon, J.L., Wong, S. and Wolfe, K.H. (2006) Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature*, **440**, 341–345.
- Schnable, P.S., Ware, D., Fulton, R.S. *et al.* (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science*, **326**, 1112–1115.
- Schnable, J.C., Springer, N.M. and Freeling, M. (2011) Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc. Natl Acad. Sci. USA*, **108**, 4069–4074.
- Sebat, J., Lakshmi, B., Troge, J. *et al.* (2004) Large-scale copy number polymorphism in the human genome. *Science*, **305**, 525–528.
- Semon, M. and Wolfe, K.H. (2007) Reciprocal gene loss between Tetraodon and zebrafish after whole genome duplication in their ancestor. *Trends Genet.* **23**, 108–112.
- She, X., Cheng, Z., Zollner, S., Church, D.M. and Eichler, E.E. (2008) Mouse segmental duplication and copy number variation. *Nat. Genet.* **40**, 909–914.
- Shinawi, M., Liu, P., Kang, S.H. *et al.* (2010) Recurrent reciprocal 16p11.2 rearrangements associated with global developmental delay, behavioural problems, dysmorphism, epilepsy, and abnormal head size. *J. Med. Genet.* **47**, 332–341.
- Springer, N.M., Ying, K., Fu, Y. *et al.* (2009) Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS Genet.* **5**, e1000734.
- Stankiewicz, P. and Lupski, J.R. (2010) Structural variation in the human genome and its role in disease. *Annu. Rev. Med.* **61**, 437–455.
- 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.* **168**, 9–1699.
- Turner, D.J., Miretti, M., Rajan, D., Fiegler, H., Carter, N.P., Blayney, M.L., Beck, S. and Hurles, M.E. (2008) Germline rates of de novo meiotic deletions and duplications causing several genomic disorders. *Nat. Genet.* **40**, 90–95.
- Vlad, D., Rappaport, F., Simon, M. and Loudet, O. (2010) Gene transposition causing natural variation for growth in *Arabidopsis thaliana*. *PLoS Genet.* **6**, e1000945.
- Woodhouse, M.R., Schnable, J.C., Pedersen, B.S., Lyons, E., Lisch, D., Subramaniam, S. and Freeling, M. (2010) Following tetraploidy in maize, a short deletion mechanism removed genes preferentially from one of the two homologs. *PLoS Biol.* **8**, e1000409.
- Yandea-Nelson, M.D., Xia, Y., Li, J., Neuffer, M.G. and Schnable, P.S. (2006) Unequal sister chromatid and homolog recombination at a tandem duplication of the A1 locus in maize. *Genetics*, **173**, 2211–2226.
- Zhang, J. and Peterson, T. (2005) A segmental deletion series generated by sister-chromatid transposition of Ac transposable elements in maize. *Genetics*, **171**, 333–344.