Additional File 1:

**Figure S1** Distribution of DNA copy number in regions covered by CNV_610 and CNV_611 in 80 natural accessions of Arabidopsis (MPIcao2010 set).

**Figure S2**. A schematic map of Arabidopsis genes covered by AthMSH2-MLPA assay.

**Figure S3** Exemplar electropherograms of AthMSH2-MLPA assay results.

**Figure S4** Pairwise correlation of MLPA signals obtained with probes mlpAb-mlpaG in AthMSH2-MLPA genotyping assay of Arabidopsis populations.

**Figure S5** Nonhierarchical phylogenetic network of a subset of 154 accessions based on 20-kb regions flanking the CNVs and its relation to the genetic groups defined by 1001 Genomes Consortium.

**Figure S6** Linkage disequilibrium (LD) at genomic regions surrounding the investigated CNVs.

**Figure S7** Rate of missing calls at AT3G18530-AT3G18535 loci in pseudogene sequences of 1135 Arabidopsis accessions.

**Figure S8** The sequence composition of the left and right breakpoints in accessions with “del-2” and “dupl-2” genotypes.

**Figure S9** Sequence alignment of CNV breakpoints in accessions with “del-2” genotype.

**Figure S10** Sequence alignment of CNV breakpoints in accessions with simple “dupl-2” genotype.

**Figure S11** Sequence alignment of CNV breakpoints in accessions with “dupl-2” genotype harboring extended duplication, that involves also the 3’ flank of the right LCR.

**Figure S12** Optimization of genomic DNA template input for ddPCR.

**Figure S13** Optimization of primer annealing temperatures for ddPCR.

**Table S2** Sequences of MLPA probes.

**Table S3** Gene specific primers used for ddPCR assays.
Fig. S1 Distribution of DNA copy number in regions covered by CNV_610 and CNV_611 in 80 natural accessions of A. thaliana (MPICao2010 set). The source data are from [29]. DNA copy number is presented as relative to the reference genome, Col-0. The division according to the geographic origin of the accessions is identical to that proposed in the original paper.
Fig. S2 A schematic map of *A. thaliana* genes covered by AthMSH2-MLPA assay. Approximate positions of MLPA probes are indicated over the gene models. For genes with alternative splicing isoforms (*HDA15, BRC1, DCL1, AT4G21580*) the longest gene model is presented and the probes are localized in conserved exons. Blue rectangles represent exons (dark for coding sequence, light for untranslated regions); blue lines represent introns; arrows show the gene orientation on genomic DNA reference sequence. Green boxes indicate the overlap of CNV regions inferred by WGS data analysis [29]. Approximate positions of primer pairs used for ddPCR-based verification are indicated below the gene models.

**AT3G18520 (HDA15)**

**AT3G18524 (MSH2)**

**AT3G18530; AT3G18535**

**AT3G18550 (BRC1)**

**AT1G01040 (DCL1)**
Fig. S3 Exemplar electropherograms of AthMSH2-MLPA assay results. The data are representative for accessions with: A – no copy number changes detected; B - amplification of MSH2, AT3G18530 and AT3G18535 genes; C – deletion of AT3G18530 and AT3G18535 genes. Normalized peak height from the Gene Marker v.2.4.0 analysis are presented (note that the scale on electropherogram B is different from A and C). Probes’ IDs (see Table 1 for detailed description) are shown under the electropherograms. Colors indicate target genes: MSH2 (green) AT3G18530 (red) AT3G18535 (blue), HDA15 (brown), BRC1 (purple). All control probes are in black. DNA fragment sizes are indicated on the upper scale. The data presented on the Figure were obtained for Col-0 (A), Wal-HasB-4 (B) and Voeran-1 (C) accessions.
Fig. S4 Pairwise correlation of MLPA signals obtained with probes mlpA-B-mlpA-G in AthMSH2-MLPA genotyping assay of Arabidopsis populations. $R^2$ – the square of the Pearson linear regression coefficient

|          | MSH2   | AT3G18530 | AT3G18535 |
|----------|--------|-----------|-----------|
| mlpA     | 1.000  | 0.988     | 0.472     |
| mlpB     | 1.000  | 0.532     | 0.507     |
| mlpC     | 1.000  | 0.957     | 0.968     |
| mlpD     | 1.000  | 0.994     | 0.991     |
| mlpE     |        | 1.000     | 0.982     |
| mlpF     |        |           | 1.000     |
| mlpG     |        |           |           |
Fig. S5 Nonhierarchical phylogenetic network of a subset of 154 accessions (including Col-0) based on 20-kb regions flanking the CNVs and its relation to the genetic groups defined by 1001 Genomes Consortium. The Network was constructed with the NeighborNet algorithm, based on bi-allelic SNPs of at least 10% frequency, surrounding the CNV region from both sides (20-kb flanks). Each accession is marked with a symbol shape indicating its CNV pattern and with a color indicating the genetic group it belongs to, according to a recent population study of 1135 Arabidopsis accessions [44]. No clear evolutionary splits between the accessions harboring distinct copy number of MSH2, AT3G18530 and AT3G18535 genes could be observed (see main text for details).
Fig. 56 Linkage disequilibrium (LD) at genomic regions surrounding the investigated CNV. LD plot shows correlation between the pairs of bi-allelic SNPs in the 20-kb DNA regions which flank MSH2-AT3G18530-AT3G18535 genes, in 154 accessions (including Col-0) as well as the correlation between the SNPs and the CNV patterns. The SNPs are mapped to their physical positions along the chromosome 3 (top) with black lines. Solid green arrows indicate genes; the yellow rectangle highlights MSH2, AT3G18530 and AT3G18535 genes. CNV genotypes were incorporated into the analysis as TRUE or FALSE values (i.e. the accession harbors/not harbors the indicated genotype, respectively; “dupl-1” genotype was excluded from this analysis). Although LD blocks were detected in the analyzed region from each side of CNV, we observed no correlation between any CNV pattern and any SNP ($R^2 < 0.3$, data highlighted by the red frame).
Fig. S7 Rate of missing calls at AT3G18530-AT3G18535 loci in pseudogenome sequences of 1135 Arabidopsis accessions.
Fig. S8 The sequence composition of the left and right breakpoints in accessions with “del-2” and “dupl-2” genotypes. A – “basic” genotype. The 9-nt sequence (TGAACAAGG) localized at the 3’ end of the left LCR shows homology (1 mismatch) to the sequence located downstream of the right LCR (TGAATAAGG); B – the “del-2” genotype; B – the simple “dupl-2” genotype, observed in Nie 1-2, Star-8, Tamm-2, Lecho-1 and Rubezhnoe-1 accessions; D – the “dupl-2” genotype with extended duplication region, observed in La-0, Etna-2 and Di-G accessions, with the indication of additional, microhomology-based sequence template switching between the 9-nt sequences (the recombined sequence is underlined). The sequence alignments of the regions marked by dashed-line boxes are presented in Supplementary Figs. S9-S11.
Fig. S9 Sequence alignment of CNV breakpoints in accessions with “del-2” genotype. The sequences are aligned along with the LCRs. Left LCR (Chr3:6372413..6373650) is presented with its 38 nt of its 5’ flanking region (highlighted in pink) and right LCR (Chr3:6377368..6378605) is presented with its with 114 nt of its 3’ flanking region (highlighted in teal). The colors of LCR flanking regions match those in Supplementary Fig. S8. Positions distinguishing both LCRs are highlighted by yellow boxes.
**Fig. S10** Sequence alignment of CNV breakpoints in accessions with simple “dupl-2” genotype. The sequences are aligned along with the LCRs. Left LCR (Chr3:6372413..6373650) is presented with its 3’ flanking region (highlighted in purple) and right LCR (Chr3:6377368..6378605) is presented with its 5’ flanking region (highlighted in green). The colors of LCR flanking regions match those in Supplementary Fig. S8. Positions distinguishing both LCRs are highlighted by yellow boxes.
On next pages:

**Fig. S11** Sequence alignment of CNV breakpoints in accessions with “dupl-2” genotype harboring extended duplication, that involves also the 3’ flank of the right LCR. The sequences are aligned along with the LCRs. Left LCR (Chr3:6372413..6373650) is presented with its 3’ flanking region (highlighted in purple) while right LCR (Chr3:6377368..6378605) is presented with both its with 5’ flanking region (highlighted in green) and 3’ flanking region (highlighted in teal). Positions distinguishing both LCRs are highlighted by yellow boxes, except for the last differentiating position, which is highlighted in orange. This position is within the microhomology region, which mediated the template switching in Di-G Etna-2 and La-0 accessions. The colors of LCR flanking regions match those in Supplementary Fig. S8.
Fig. S12 Optimization of genomic DNA template input for ddPCR. A – Each column represents a single well of ~18,000 droplets generated from 20-µl reaction mix containing indicated amount of Col-0 genomic DNA and a single set of primers, targeting DCL1 gene. B – The linearity of the dynamic range of ddPCR assays presented in (A).
Fig. S13 Optimization of primer annealing temperatures for ddPCR. A – Separation of positive and negative droplets at a range of primer annealing temperatures (56-60 °C). Each column (separated by yellow lines) represents a single well of ~18,000 droplets generated from 20-μl reaction mix containing 1 ng of Col-0 genomic DNA and a single set of primers, targeting DCL1, HDA15, MSH2, AT3G18530, AT3G18535 or BRC1 gene (as indicated on the bottom of panel B). B - The target concentrations detected in the assays presented in (A). The error bars indicate the Poisson 95% confidence intervals.
Table S2. Sequences of MLPA probes. Each probe consists of two half probes design to hybridize to adjacent positions of the target genomic DNA. Each half probe consists of a target-specific sequence (red), universal primer sequence (blue) and (optional) stuffer sequence (black).

| PROBE | GENE   | TARGET-SPECIFIC REGION | LEFT HALF PROBE | RIGHT HALF PROBE |
|-------|--------|------------------------|-----------------|------------------|
| mlpA  | HDA15  | chr3:636365-636368     | GGGTTCCCTAAAGGTTGGA<actacta>GGCTTCACAATAATGCGGCGTTG | CTGCAATTAGTAGCACAAGCAGCAGGaatctacTCTAGATGGATCTTGCTGGCGC |
| mlpB  | MSH2   | chr3:636880-636885     | GGGTTCCCTAAAGGTTGGA<gcctactacta>GCCTTCACAATAATGCAGCCGTTGCACT | TGGAGAATTTCACACAAACAGCTTGTatgtactaatggcaaaactatcTCTAGAT TGGATCTTGCTGGCGC |
| mlpC  | AT3G18530 | chr3:637046-637051   | GGGTTCCCTAAAGGTTGGA<gcctactacta>GTGCATTAGTAGCACAAGCAGG | AACCACCTGATGAACTTCCAAGGTAaaactatcTCTAGATTGGATCTTGCTGGCGC |
| mlpD  | AT3G18535 | chr3:637386-637391 | GGGTTCCCTAAAGGTTGGA<gcctactacta>GGTGGTTTTGATGATAC | ACATGCAACACCATGAAAGAACAGGCTaatgtactaatggcaaaactatcTCTAGAT |
| mlpE  | AT3G18535 | chr3:637519-637524 | GGGTTCCCTAAAGGTTGGA<gcctactacta>GGTGGGTTTTGATGATAC | AGATTTGGAATCCTTGCTGGCGC |
| mlpF  | AT3G18535 | chr3:637594-637600 | GGGTTCCCTAAAGGTTGGA<gcctactacta>GGTGGGTTTTGATGATAC | ATCTAGATTGGATCTTGCTGGCGC |
| mlpG  | AT3G18535 | chr3:637645-637650 | GGGTTCCCTAAAGGTTGGA<gcctactacta>GGTGGGTTTTGATGATAC | CCGGTTGTCACAACATAGAAACAAGATGTCTAGATGGATCTTGCTGGCGC |
| mlpH  | BRC1   | chr3:637386-637391     | GGGTTCCCTAAAGGTTGGA<gcctactacta>GGTGGGTTTTGATGATAC | ACATGCAACACCATGAAAGAACAGGCTaatgtactaatggcaaaactatcTCTAGAT |
| ctrl1 | DCL1   | chr1:23416-31120       | GGGTTCCCTAAAGGTTGGA<gcctactacta>GGTGGGTTTTGATGATAC | CATCTGTCAGATCTTGCTGGGAACCACCTGATGAACTTCCAAGGTAaaactatcTCTAGATGGATCTTGCTGGCGC |
| ctrl2 | AT4G21580 | chr4:11475719-11477768 | GGGTTCCCTAAAGGTTGGA<gcctactacta>GGTGGGTTTTGATGATAC | GGTGTTGCTACTGTCATGATGGGCaatctacTCTAGATGGATCTTGCTGGCGC |
| ctrl3 | APG10  | chr2:15195538          | GGGTTCCCTAAAGGTTGGA<gcctactacta>GGTGGGTTTTGATGATAC | GGTAGGCAACTCAGACTTGGATATGGAACCACCTGATGAACTTCCAAGGTAaaactatcTCTAGATGGATCTTGCTGGCGC |
| ctrl4 | PFD5   | chr5:784781-784784     | GGGTTCCCTAAAGGTTGGA<gcctactacta>GGTGGGTTTTGATGATAC | GACAGGCAACAGTTAACCAGTGAatgtactaatggcaaaactatcTCTAGATGGATCTTGCTGGCGC |
| ctrl5 | PS2    | chr1:2746548-27465522  | GGGTTCCCTAAAGGTTGGA<gcctactacta>GGTGGGTTTTGATGATAC | GCTAATTCCGAGATGTTCAACAGATGGATGCGGGAATGGAatgtactaatggcaaaactatcTCTAGATGGATCTTGCTGGCGC |
Table S3. Gene specific primers used for ddPCR assays.

| GENE       | primer | sequence                  |
|------------|--------|---------------------------|
| HDA15      | dd01   | 5’- GTGCTGATTGTGGACTGGGTA-3’ |
|            | dd02   | 5’- GGACATCCTGCAACCGTAGTA-3’ |
| MSH2       | dd03   | 5’- GACACTCCACGCAGCTAACA-3’ |
|            | dd04   | 5’- ATCTTTGCCCCTGTAGGAGC-3’ |
| AT3G18530  | dd05   | 5’- TGCTGTGACTGCAGGGA-3’   |
|            | dd06   | 5’- GTGAGCTTTTGTCTCA-3’     |
| AT3G18535  | dd07   | 5’- CGACGGCTTCCTGTTGTTT-3’  |
|            | dd08   | 5’- TGGATGCTTTTCTGAGC-3’    |
| BRC1       | dd09   | 5’- CACTGAGCCCTCGGAACTA-3’  |
|            | dd10   | 5’- ACGATCTCGTGCCTTTTGG-3’  |
| DCL1       | dd11   | 5’- TCTCAGCGCGTTCTAAG-3’    |
|            | dd12   | 5’- TGTACCAGGCACAGCATCAA-3’ |