# Supplementary information for:Plant pan-genomes are highly vulnerable to methodological considerations

## Supplementary note 1 – analysis of DN+|MTP- nonreference pan-genes in *A. thaliana*

We investigated the methodological causes for the occurrence of DN+|MTP- pan-genes by mapping their transcript sequences to the reference genome and to the novel genomic sequences detected by the MTP approach. Based on the mapping results, DN+|MTP- pan-genes were classifying as:

1. Completely mapped to the reference genome
2. Partially mapped to the reference genome
3. Mapped to the MTP novel sequences

(also see Supplementary Figure S1).

If a DN+|MTP- transcript can be completely mapped to the reference genome it means that it is coded in a genomic region that is highly homologous to the reference and therefore not detected as novel by the MTP approach and not processed in the MTP annotation step. The transcript may map to a region marked as intergenic by the reference annotation, which can result from either gene emergence from previously noncoding regions in one of the annotated ecotypes, over-annotation by the DN pipeline, or under-annotation of the reference. Owing to the high quality of the reference annotation, the first two options appear more plausible. Alternatively, a DN+|MTP- transcript may successfully map to a reference genomic region coding for a reference gene, but translate into a highly different protein sequence. This can occur when the genomic sequence in one of the ecotypes contains a mutation that substantially modifies the protein sequence (e.g. frame-shift or stop codon loss/gain). We found that 1,794 DN+|MTP- (40% of the total DN+|MTP- pool) transcripts could be completely mapped to the reference genome (query coverage > 95%). Out of these, 38% overlap with reference genes (coordinate overlap of gene features > 50%), whereas 62% map to reference intergenic regions.

In other cases, a DN+|MTP- transcript only partially maps to the reference sequence. Such cases indicate that the gene is annotated in a region containing a combination of reference-homologous and non-reference-homologous sequences. If the stretches of nonhomologous sequences shorter than a certain cutoff (300 bp in this case), then they will be ignored in the MTP annotation process. We found that 2,476 (55%) DN+|MTP- transcripts partially mapped to the reference genome (20% < query coverage < 95%).

Finally, we found 144 (3%) DN+|MTP- transcripts that can be completely mapped to a novel sequence identified by the MTP approach. These are likely the result of the more limited genomic context in which annotation is performed in the MTP approach.

Together these results suggest that the higher number of nonreference genes detected by the DN approach stems from: (1) its ability to predict genes in reference sequences; and (2) its ability to analyze partially novel genomic sequences.

## Supplementary note 2 – analysis of DN-|MTP+ nonreference pan-genes in *A. thaliana*

All genomic sequences annotated in the MTP approach are extracted from genome assemblies, and as such are necessarily also annotated by the DN approach. Therefore, the existence of 574 DN-|MTP+ nonreference genes might be surprising at first, and points to methodological biases in both construction approaches (Supplementary Figure S1).

To investigate these methodological biases, we first mapped protein sequences of DN-|MTP+ pan-genes to those of DN+|MTP- pan-genes. 60 proteins (10%) show high similarity to DN+|MTP- proteins (percent identity > 95%), however most are in fact truncated versions of those DN+|MTP- genes. Such cases likely occur when novel genomic regions flanked by reference-homologous regions are annotated by the DN procedure.

We performed another analysis, this time mapping DN-|MTP+ protein sequences to protein sequences derived directly from whole genome annotations of the DN pan-genome, before orthology clustering and selection of representative genes (see Figure 1). 381 of the DN-|MTP+ (66%) show very high similarity (percent identity > 95% and length difference < 10%) to annotated DN proteins. This indicates that the corresponding nonreference genes were in fact detected by the DN approach, however they were not selected as the representative of the orthology cluster to which they belong, for example because this cluster contains a reference gene.

The remaining 24% may result from over-annotation by the MTP approach.

## Supplementary note 3 – *A. thaliana* genome assembly and contamination analysis

During the analysis of genome assemblies, we noticed that the assembly of the ecotype Ler shows considerably lower N50 values and larger assembly size compared to other ecotypes in most-data sets (Figure 3A and Supplementary Table S4). We suspected that this may be the result of biological sequencing contamination, specifically with aphid DNA. To test this, we generated a data base of whole genome sequences of five Hemiptera species (*Cimex lectularius, Rhodnius prolixus, Bemisia tabaci, Acyrthosiphon pisum,* and *Trialeurodes vaporariorum*), obtained from ENSEMBL Metazoa Release 52. All Ler contigs from the 50× data set which could not be reliably mapped to the *A. thaliana* reference genome during Panoramic's reference-guided assembly step were searched for within the Hemiptera DB, using Blastn. As expected, we found that Ler contigs with high similarity to Hemiptera genomes (% identity > 70) accounted for 52,762,315 bp out of the total 60,862,882 nonreference sequences (87%), indicating that the majority of nonreference genome sequences derives from aphid DNA sequencing contamination, assembled into contigs.

We tested the effect of this contamination on gene prediction results by mapping the transcript sequences of nonreference genes from the DN 50× pan-genome to the Hemiptera DB, using Minimap2. None of the transcripts showed significant similarity to Hemiptera genomes, indicating no gene models of aphid origin were incorporated into the pan-genome.

## Supplementary note 4 – analysis of gene presence-absence detection thresholds in the Map-to-pan approach

The last step of the MTP pan-genome construction procedure (Figure 1) is aimed at determining the presence of all reference and nonreference pan-genes in each of the examined accessions to create the gene PAV matrix. This is achieved through mapping of short sequencing reads to the genomic reference and nonreference pan-genome sequences and analysis of coverage patterns for each gene. This usually requires the choice of values for two parameters:

1. The depth threshold – the minimal number of mapped reads required to label a given position in the gene sequence "covered".
2. The coverage threshold – the minimal percentage of the gene sequence labeled as "covered" required to assign the gene as present in a given accession.

Different pan-genome studies have arbitrarily used different thresholds, without reporting the effects of the value choice on the obtained pan-genome. We therefore analyzed the effect of these two parameters to gain a general understanding on how they impact gene presence-absence detection by the MTP approach, and the inferred pan-genome composition.

We began by examining the effect of the depth threshold parameter on the percentage of gene coverage across all pan-genes of a single *A. thaliana* ecotype (An-1), using 50× sequencing data. As expected, using depth thresholds of 1, 5, 10, 15, 20, 30, 40, and 50 reads resulted in considerably different distributions of gene coverage, with lower threshold values leading to high gene coverage values (Supplementary Figure S5A). Notably, intermediate depth thresholds (20,30,40) resulted in higher variation among genes compared to more extreme values (1-15 and 50). Next, we computed the percentage of genes determined as present in the examined ecotype out of the total 27,960 pan-genes, using different values of the depth and coverage thresholds (Supplementary Figure S5B). Using lenient depth thresholds between 1 and 15 resulted in gene presence percentage > 90% for any coverage threshold < 95%. Using a depth threshold of 20 and coverage threshold of 75% resulted in a gene presence percentage similar to the one observed using the DN approach (87%). Depth thresholds of 30 or higher resulted in unrealistically low gene presence percentages for almost any coverage threshold.

We further examined the effect of the threshold parameters on two measures related to the pan-genome composition (using the 50× pan-genome): the percentage of core pan-genes (present in all eight accessions; Supplementary Figure S5C) and the overall gene occupancy (see the main text for details; Supplementary Figure S5D). The observed patterns were similar to those described above, with low depth and coverage threshold values resulting in high percentage of core genes and gene occupancies. The percentage of core genes detected by the DN approach (73%) was observed with depth and coverage thresholds considerably more stringent than commonly used (e.g. depth threshold 15 and coverage threshold 55%, or depth threshold 10 and coverage threshold 80%).

In most studies which applied the MTP approach (including this one), the depth threshold was specified as an absolute number of reads. However, the choice of this parameter value should obviously be based on the amount of available sequencing data, and thus may be specified as a fraction of the mean sequencing depth. We therefore analyzed the effect of the mean sequencing depth when using various depth fraction thresholds. For the purpose of this analysis, we fixed the coverage threshold at 50% and computed the mean sequencing depth against the percentage of gene presence in a single ecotype (An-1), using different depth fraction thresholds (0.1, 0.3, 0.5, 0.7, 0.9; Supplementary Figure S5E). The sequencing depth appears to have only a modest effect on the gene presence-absence inferences, indicating that setting the threshold as a fraction of the mean depth rather than as an absolute number is an effective way to perform this type of analysis.