A genome-wide scan for correlated mutations detects macromolecular and chromatin loop interactions in *Arabidopsis thaliana*

Laura Perlaza-Jiménez & Dirk Walther
Max Planck Institute of Molecular Plant Physiology
Am Mühlenberg 1
14476 Potsdam-Golm
Germany
perlaza@mpimp-golm.mpg.de, walther@mpimp-golm.mpg.de
Phone: +49-0-331-5678 108

Abstract

The concept of exploiting correlated mutations has been used in the past to efficiently identify interactions between biological macromolecules. Its rationale lies in the preservation of interactions via compensatory mutations. With the massive increase of horizontal as well as vertical sequence information, approaches based on correlated mutations have regained considerable attention.

We analyzed a set of 10,707,430 single nucleotide polymorphisms detected in 1,135 accessions of the plant *Arabidopsis thaliana.* To measure their covariance and to reveal the global genome-wide sequence correlation structure of the Arabidopsis genome, the adjusted mutual information has been estimated for each possible pair of polymorphic sites. We developed a series of filtering steps to account for genetic linkage and/or lineage relations between Arabidopsis accessions, as well as transitive covariance as possible confounding factors. We show that upon appropriate filtering, correlated mutations prove indeed informative with regard to molecular interactions, and furthermore, appear to reflect on chromosomal loop interactions.

Our study demonstrates that the concept of correlated mutations can also be applied successfully to within-species sequence variation and establishes a promising approach to help unravel the complex molecular interactions in *A. thaliana* and other species with broad available sequence information.

Keywords: *Arabidopsis thaliana*, 1001genome project, SNPs, Evolution, Genome structure, Compensatory Mutations, Molecular Interactions, Adjusted Mutual Information

Introduction

At the molecular level, biological functions result from the sum of complex and dynamic interactions between individual molecules (e.g. protein-protein, miRNA-mRNAs, DNA-DNA, DNA-RNA). With the great variety of possible molecular interactions, the methods to experimentally detect and computationally predict them are also very diverse (e.g.(1–3)). Despite tremendous progress, obtaining a comprehensive view of interactions of all genome-encoded molecules and genomic elements remains highly challenging. However, independent of the type of interaction, there is one assumption in common. Molecules "communicate" between each other through structure-mediated physical interactions. Thus, the information that defines the structure of molecules, i.e. their sequences or molecular makeup, should contain information about their molecular interactions. Given the challenges associated with predicting structures and their interactions based on favorable – in the sense of low free energy - physical interactions, statistical approaches have been developed that harness the wealth of genome sequence information to inform on molecular interactions. Such statistical approaches exploit covariation at different sites detected in alignments of genomic/protein/RNA sequences (4–10). Their rationale relies on the concept of compensatory mutations. As one interaction partner is mutated, leading to a change of structure or interaction surface potential, natural selection will favor mutations in the cognate molecule that compensates for this change in order to preserve the interaction. Thus, correlated mutations can be taken as indicators of possible molecular interactions and can be detected by performing covariation analyses of sets of aligned sequences.

In the plant *Arabidopsis thaliana*, the application of the concept of covariance to deduce structural and functional interactions has been demonstrated successfully for prediction of structure-function relationships of the pentatricopeptide repeat proteins (PPR). *In silico* and experimental evidence has demonstrated that covariance in genomic sequences of orthologous protein-RNA pairs allow detecting protein-RNA recognition events (5). This covariance was identified by finding correlated mutations in the alignments between PPR binding factors and their RNA targets from different species (5–7, 11, 12). Correlated mutations have also been used successfully to identify DNA-binding domains and sites of transcription factors, such as MerR-family proteins (13). Extended to whole genomes, this concept was used in the Drosophila genus (14). In this study, investigators searched for long-range covariation clusters on a genomic scale based on genome-triple "fingerprints", finding that compensatory mutations suggest long range interactions between exons of mRNAs and also between noncoding RNAs.

Although the rationale of using correlated mutations to inform on molecular interactions seems very plausible and straightforward, it has also been shown that covarying sites do not

necessarily correspond to sites that are spatially interacting (4–10, 15) . Phylogenetic bias, transitivity effects, passenger mutations are some of the reasons why the detected covariance can be misleading (4). However, several approaches have been devised that circumvent these methodological problems. Maximum-Entropy probability models (MEPM) have been developed to predict the 3D-structure of proteins and RNA-molecules as well as protein-protein and protein-RNA interactions by using global statistical models of covariation over deep alignments of gene families (8–10). MEPMs were adapted for each case of interaction or genomic element, which allowed avoiding the transitivity effect and phylogenetic sampling bias depending on the genomic element of study (10). All these approaches are implemented for alignments of sequences from diverse species, avoiding, or at least reducing, any phylogenetic bias. By contrast, addressing compensatory mutations to detect functional associations in closely related genomes remains a challenge.

In this study, we set out to exploit the recently released genome sequence information of 1,135 genomes of *Arabidopsis thaliana* accessions (16). Thus, we focused on intra-species variation with mutations represented as single nucleotide polymorphisms (SNPs). This dataset contains around 10 million polymorphic sites, and all possible pairs were tested exhaustively for correlated mutations as a means to detect functional associations. To determine the molecules involved in these functional associations, the availability of detailed genome annotations is critical. In Arabidopsis, aabout 77% of the genes in the genome have been annotated (17), and other genomic elements such as pseudogenes, and mobile elements have also been identified. To detect covariance, we used as a metric the Adjusted Mutual Information (AMI) between pairs of polymorphic sites (18). When working with genomes from the same species, tight covariation can also emerge from factors other than compensatory selection such as linkage, lineage, and transitivity. We devised strategies to control for these confounding factors to enrich for correlated mutations pairs with functional meaning. We provide a genome-wide covariation statistic and report higher than expected events of correlated mutations in molecules that have been reported to interact (protein-protein, miRNA-mRNA) as well as present evidence for the validity of putative interactions. Most strikingly, we found that genomic sites reported to physically interact to form chromatin loops were enriched for correlated mutations as well.

Methods

Polymorphic site information

We used sequencing data and variant calls that is available as part of the Arabidopsis 1001 genome project including genomic sequences of 1,135 *A. thaliana* accessions (16). We

considered the nuclear genome only (chromosomes 1-5) and required polymorphic sites to be biallelic and without any insertions/deletions (indels) in any accession. In total, 10,707,430 SNP-sites were available for analysis. When computing pairwise correlations of SNP-sites, we required the respective minor allele frequency to be greater than 5% at both sites. Allele frequency was computed relative to the number of accessions with unambiguous base calls (i.e. A, C, G, T only and with no ambiguous calls, such as N, allowed) at both sites, respectively. In total, approximately 5.7E+13 pairs of SNP-sites were inspected for evidence of correlated mutations.

Adjusted Mutual Information (AMI)

To detect correlated mutations, the adjusted mutual information (AMI) was computed for every possible pair of polymorphic sites, using the analytical solution proposed by Romano et al., 2014 (18) (Eqs. 1-5). In essence, AMI measures the agreement of the clustering of accessions based on alleles at two different sites corrected for their chance agreement. Given a specific polymorphic site, not all accessions in the dataset may have an unambiguous nucleotide call at this site, instead showing Ns or other ambiguity codes. The lowest percentage of ambiguous base calls across all SNP-sites was determined as 4%. To detect covariance between two sites and without including ambiguous calls, the AMI was calculated for the intersection of accessions with definitive nucleotides called at both sites. We required this intersection to comprise at least 300 accessions. In total, approximately 5.7E+13 AMI-computations had to be performed, requiring efficient computing and program execution distributed across many CPUs. A C-program was implemented to perform the calculations in parallel on several cores per node using shared memory (C program available upon request).

$$(1) \qquad H_a = -\sum_{i=1}^{r} \frac{a_i}{N} \log \frac{a_i}{N},$$

$$(2) \qquad H_b = -\sum_{j=1}^{r} \frac{b_j}{N} \log \frac{b_j}{N},$$

$$(3) \qquad MI_{ab} = H_a + H_b - H_{ab},$$

$$(4) \qquad EMI_{ab} = \sum_{i,j}^{r} \sum_{n_{ij}} \frac{n_{ij}}{N} \log \frac{n_{ij} N}{a_i b_j} P(n_{ij}),$$

$$(5) \qquad AMI = \frac{MI_{ab} - EMI_{ab}}{\sqrt{H_a * H_b} - EMI_{ab}},$$

where *MI* is Mutual Information, *EMI* the expected MI, $H_{a/b}$ is the entropy of variable *a/b*, respectively, and H_{ab} is the joint entropy, *N* is the total length of the variable vector. In our case, a_i and b_j are the base counts or type $i=1..r$ *and* $j=1..r$ found at a particular SNP-sites *a* or *b*, respectively, *N* is the number of accessions with unambiguous base calls at both sites and *r* is the number of different alleles at a site (here, *r*=2, for biallelic SNP-sites), n_{ij} is the number of accessions clustered together based alleles *i* and *j* at the two sites *a* and *b*. The probability $P(n_{ij})$ is estimated from a hypergeometric distribution. AMI values range between 0, denoting random co-segregation of accessions, and 1, indicating perfect agreement of the allele-based partitioning at sites *a* and *b*.

Pairs of polymorphic sites that scored above an AMI of 0.9 were considered correlated mutations (*pair hits*). The threshold value of 0.9 was chosen as it was found to be significantly larger than the value of AMI that may arise by chance. This threshold was determined from an empirical permutation test, which was carried out using a subset of randomly selected 966,243 SNP-sites and vertically (across accessions) shuffled bases. The shuffling was repeated 1000 times. Then, computing all pairwise AMI values yielded a maximum AMI value of 0.88 with the bulk of AMI values at considerably smaller values (Supplementary Figure 1). Therefore, AMI >0.9 is a sensible threshold value. This sampling strategy to obtain a random reference distribution was chosen as the number of "N"-accessions is highly variable across different SNP-sites rendering analytical computations of the combinatory challenging.

With that many pairwise AMI computations for ~10Mill SNPs (5.7E+13), multiple testing appears to be a major concern at first. However, note that the probability of two SNP-sites randomly showing exactly correlated base changes is very small, and is, worst case 1.3E-25 for 300 non-N accessions and 5% (=15 accessions) minor allele frequency at both sites. In addition, with 1000 repeat runs of about ~1Mill shuffled SNP-sites, we performed about 10 times more random control simulations compared to actual comparisons. Thus, protecting against false correlations due to random chance can be safely controlled for. More critically, non-random effects leading to correlated mutations other than functional associations (e.g. via linkage) proved more challenging (see below).

Annotation of pair hits

Pair hits (pairs of SNP-sites with AMI >0.9) were classified with regard to molecular and functional annotation based on the following criteria. First, for both of the SNP-sites forming a pair, the types of *genomic elements* to which they map were identified. Genomic element refers to

the functional class of the encoded molecules (e.g. protein-coding or miRNA) at the respective sites or the type of local genomic region (e.g. intergenic, intron). A pair hit was then annotated as a particular *Class* based on the combination of the genomic element annotations associated with the respective individual SNP-sites comprising the pair hit (*e.g.*, Genomic Element (*A*): miRNA, Genomic Element (*B*): Protein Coding yields *Class* (*A,B*): miRNA-Protein Coding). Class annotations with reversed order were combined into one class. Second, the sequence ontology classification associated with every SNP-site was mapped as well as the corresponding gene identifiers (AGI IDs) assigned if the site was found within known genes. This information was obtained from the VCF file from the 1001 genomes project (available http://1001genomes.org/data/GMI-MPI/releases/v3.1/1001genomes_snpeff_v3.1/1001genomes_snp-short-indel_only_ACGTN_v3.1.vcf.snpeff.gz). Pairs sharing identical gene IDs and genome element classification were considered intra-molecular/element pairs, otherwise, they were considered inter-molecular/element pairs.

Accounting for linkage

SNP-sites at close genomic distances on the same chromosome will be linked because of the low probability of crossing-over between the two sites and the associated reshuffling of alleles. Linked SNP-sites will yield to high AMI values, which most likely do not reflect functional associations (but not strictly excluding a functional link either). Based on the observed distance intervals associated with pair hits, we set an effective distance cutoff of 10K bps to exclude linked sites (Figure 2).

Sequence Distance and Average Distance Diameter: Correction for Lineage

The possibility that a pair of SNP-sites can have high AMI due to phylogenetic bias (lineage effect) was taken into account by the following approach. For all 1,135 Arabidopsis accessions, their respective pairwise genome-sequence-based distances were calculated based on 200,000 randomly selected polymorphic sites. The sequence distances were calculated using the software Phylip (19). For each pair hit, each of the two biallelic polymorphic sites split the set of accessions into two clusters, one with allele *a*, the other with allele *b*. For each accession cluster, the *Average Distance Diameter* (*ADD*) was calculated. ADD is the average pairwise sequence distance in the cluster of accessions sharing the same allele at the respective site (Supplementary Figure 2). ADD allows measuring how different the accessions in the cluster are with regard to their genomic sequence. In a given pair hit, the set of accessions may be divided into clusters

because all accessions in each cluster originated from the same ancestor, and therefore, all its descendants have the same SNP mutation. Alternatively, the accessions in a cluster may be evolutionarily distant, but natural selection favors the specific allele observed at this site; i.e. positive selection preserved an allele in otherwise changing genomic sequences. When a cluster is explained by ancestry or phylogenetic relations, the ADD will be small because the sequences are related and, hence, similar. By contrast, when a cluster is explained by positive selective pressure, the same allele is preserved even for distant accessions. In the last case, this can happen possibly to compensate the cognate mutation, therefore the evolutionary sequence relations may be distant and the ADD high. Each pair hit is composed of two SNP-sites, each SNP-site contains two clusters (*a*, *b*), each cluster has an ADD value that expresses the sequence divergence of that cluster. Thus, for every pair hit there are four ADD values, one for each accession cluster (two clusters) of each SNP-site (two SNP-sites). Given that the pair hits have high covariance (AMI >0.9), both SNP-sites involved in a pair hit cause the same (or almost the same) clustering of accessions. Therefore, the values of ADD differ between clusters of accessions, but not between SNP-sites in the pair hit (Supplementary Figure 2). The values for the ADD can vary depending on the sequence distance inside the clusters and several combinations of ADD between the clusters are possible. For example, cluster *a* can have a small ADD, and cluster *b* as well. This means that both clusters are likely caused by ancestry relationships. Cluster *a* can have a small, and cluster *b* a large ADD (or vice versa), then in cluster *a* accessions are closely related, but in cluster *b* they are not, in which case cluster *b* is likely arose due to natural selection favoring that allele. Finally, both clusters can have large values of ADD showing a strong selective pressure for those SNPs in both clusters.

Reducing redundant pair hits: Correction for transitivity

AMI has been calculated for each pair of SNP-sites individually. As with any correlation measure, it may suffer from indirect correlations, also referred to as transitivity. When a SNP *A* is found to be correlated with both *B* and *C*, and based on a true mechanistic coupling, then *B* and *C* will appear correlated regardless whether or not they truly are associated mechanistically (Supplementary Figure 3). In our case, we correct for transitivity based on profile vectors at SNP-sites; i.e. the pattern of base changes across the different accessions at a given site. For example, if detected that two molecules harbor SNP-sites with identical profile vectors, we may postulate a connection between them as they are correlated. If however, the very same pattern links these molecules to yet another molecule (or many others), we penalize this SNP-site according to the transitivity-value (or clustering coefficient) computed by the function “transitivity” of the R

packages "igraph". Thus, the transitivity value allows us to identify correlations that are less likely to originate from indirect correlations by enriching for sites with low transitivity value and reducing those that are redundant (high transitivity value). For isolated pairs of SNP-sites; i.e. both sites do not appear correlated to any other site, transitivity cannot be computed and an arbitrary transitivity value of -0.1 was assigned to both sites. Note that transitivity relates to unique patterns in SNP-sites; i.e. if molecule *X* is found correlated to molecule *Y* based two sites of identical SNP-patterns, and molecule *Y* is found correlated to molecule *Z*, but based on another SNP-site with a different pattern, we will not conclude *X* and *Z* to be linked as well.

Testing for the effect of the introduced pair hit filtering steps using intergenic regions as reference

We tested for the effect of the introduced filtering steps, linkage, lineage, and transitivity, according to the following rationale. Given the set of actual SNP pair hits, we determined their associated characteristics with regard to annotation (pair hit classes as explained above, the genomic region/ molecule type). If the observed pair hits reflect true enrichment of particular classes, then a set of randomly obtained pair hits should show different characteristics.

The enrichment of particular classes was determined relative to pair hits that are unlikely to be functional. For the purpose of this study, pair hits in intergenic regions were assumed to be non-functional associations. Thus, the probability of finding pair hits in intergenic regions, with both SNP sites located in intergenic regions, was taken as the background probability.

The number of pair hits for a given interaction class was normalized by the number of possible comparisons amongst all SNP-sites in that class according to equations 6-10.

$$nComp_{xy} = n_x * n_y, \quad (6)$$

$$nComp_{xx} = \frac{((n_x)*(n_x-1))}{2}, \quad (7)$$

$$normFreq_{xy} = \frac{nPairHits_{xy}}{nComp_{xy}}, \quad (8)$$

$$Exp_{xy} = nComp_{xy} * normFreq_{intergenic::intergenic}, \quad (9)$$

$$E_{xy} = log\left(\frac{nPairHits_{xy}}{Exp_{xy}}\right), \quad (10)$$

where $nComp_{xy}$ is the number of comparisons (SNP-SNP pairs) between two genomic elements x and y; $nComp_{xx}$, the number of comparisons within the same genomic element x; $n_{x/y}$, are the numbers of SNPs in genomic element x or y; $nPairHits_{xy}$ is the number of pair hits between genomic elements x and y, $normFreq_{xy}$ is the normalized count of the number of pair hits ($nPairhits_{xy}$) over the number of comparisons ($nComp_{xy}$); $Exp_{xy,}$ is the expected number of hits between class x and y using pairs between intergenic regions-intergenic regions as the reference. Enrichment (or depletion) factors (E) of pair hits relative to intergenic regions are expressed as logarithmic ratios (natural log) of observed vs. expected number of pair hits. In equations 8-10, if $x=y$, pair hits between the same type of genomic elements are captured, with equation 7 serving to compute the number of possible pairs.

For all possible combinations of genomic elements, the obtained enrichment factors, E, were compared to those obtained from random re-sampling of SNP pairs. Here, only those SNP-sites, which were identified as part of pair hits, were considered, thereby eliminating SNP-sites from the original SNP-set that represent noise (regions that are difficult to sequence and, thus, rich in polymorphisms) or any other SNPs that are unlikely to represent any interaction. Furthermore, this selection renders the data manageable as for ~5.7E+13 pairs, the fraction of N-base calls needs to be tracked resulting in prohibitively large data files. This approach of re-sampling only from those SNP-sites that were part of at least one pair hit was applied throughout this study whenever re-sampling was performed or the total number of comparisons was to be estimated.

The random re-sampling aimed to destroy the covariance of the SNP-sites but to maintain the characteristics of the observed pair hits that can be explained by confounding factors (linkage, lineage, transitivity). In the case that this random resampling reflects the observed pair hits with regard to their genomic class characteristics, the correlated mutations are unlikely to reflect functional associations but just confounding factors. The random re-sampling was done by taking SNP-sites and repairing them randomly applying three conditions. (1) The distribution of the genomic distance between SNP-sites (number of bps between the SNP-sites) is the same for the observed pair hits and re-sampling. (2) The SNP-sites included in the resampling have the same ADD values threshold (see above), and (3) the same cluster coefficient values threshold. Thus, the confounding characteristics were kept, but the actual covariance was destroyed by randomly pairing up SNP-sites. Next, the actual enrichment profile with regard to pair hit class was compared to the randomized data via the Pearson linear correlation coefficient, r. We applied this procedure for the various filtering steps with an effective filtering becoming apparent by a low correlation coefficient r. High correlations would suggest no difference of observed vs.

randomized pair hit statistics, while low correlations would be indicative of differences, and therefore, reflect a signal in the observed data.

Annotation of known interactions, enrichment analysis, and semantic similarity of predicted protein-protein interactions

Several publicly available databases on molecular interactions were used to assign information to each pair hit, and to determine whether the polymorphic sites in a pair hit belong to a pair of molecules reported to interact. Protein-protein interactions (PPI) were identified using the AtPIN database (20), ANAP (21), and IntAct (22). Consolidated across all three resources, 1,927 pairs of PPI were used as a benchmark set after filtering the databases avoiding PPIs detected by predicting methods. Interactions between molecules not (yet) reported interacting, but with predicted interaction based on correlated mutations were tested for biological plausibility using the concept of semantic similarity. This method is based on gene ontology (GO) term annotation (23) and tests for involvement of molecules (proteins) in the same biological processes. The databases miRBase (24) and miRTarBase (25) were used to identify miRNA-mRNA-target interactions. To test for interactions at the chromatin level, we used the coordinates of chromatin loop formations at gene-resolution as reported in the study of Liu and collaborators in 2016 (26). SNP-sites located inside the coordinate intervals of DNA loops described to be in contact were considered as interacting polymorphic sites.

The enrichment of pair hits associated with interacting molecules (PPI, miRNA-mRNA, or loops) was tested by comparing the frequency of actual pair hits detected on molecules reported to interact to those obtained from random sampling. The random sampling was repeated 100 times and reported as a z-score statistic, with z-score computed as z-score=$(f_a-\langle f_r\rangle)/\sigma_r$, where f_a is the actually observed frequency, $\langle f_r\rangle$ the mean random frequency, and σ_r the associated standard deviation. This comparison was done first using all the originally obtained pair hits against 100 samples of the same set, and secondly, using the filtered set (for lineage, linkage, and transitivity) compared to 100 samples of the same set. As described above, in every random sample run, SNP sites were randomly paired up creating the same number of pairs as in the original set and preserving the original chromosomal distance distribution.

Gene co-expression/ GO-term enrichment analysis

Gene expression information were obtained from NASCArray (27) with reported expression values for 20,807 unique nuclear encoded genes probed across 5,295 hybridizations covering a diverse set of conditions and using the ATH1 Affymetrix Arabidopsis gene-chip. Raw values

were log-transformed and jointly normalized applying a quantile normalization using the "normalize.quantile" routine of the R package "preprocessCore". Pairwise gene co-expression was assessed by computing Pearson correlation coefficients of the normalized expression values across all hybridizations.

Gene-ontology (GO) information was obtained from TAIR (28, 29). Enrichment of specific GO-slim terms in one particular gene set compared to another set was assessed by performing Fisher's exact test of the respective term counts in the two sets to be compared. Associated p-values were tested for multiple GO-term testing based on the procedure introduced by Benjamini and Hochberg (30). Gene sets were not rendered unique; i.e. genes can be contained in the set engaging in positive and negative pairwise co-expression (with respectively differing gene partners) as well as, if a gene was reported in multiple pairings, its GO-term annotations were counted repeatedly effectively increasing their weights.

Results

From an all-against-all comparison of all SNP-sites passing the selection criteria and across all considered 1,135 *A. thaliana* accessions, 41,065,045 pairs of SNPs were detected at an Adjusted Mutual Information (AMI) level of above the employed threshold value of 0.9 and were considered significant correlations, henceforth referred to as *pair hits.* The total number of pair hits corresponds to $7*10^{-5}$% of all computed comparisons. Of all 10,707,430 polymorphic sites available for analysis, 1,329,228 (~12%) were part of at least one pair hit (Table 1). The distribution of individual SNP-sites in the five nuclear chromosomes of *A. thaliana* shows a specific segregation by genomic elements. There is a high SNP occurrence in protein-coding regions and transposable element genes. The SNP-density profile also reflects the chromosomal structure with high SNP densities near the centromeric regions, especially related to transposable element genes (Figure 1A). Most pair hits (84%) were found to be involved in inter-molecular interactions, with pair hits corresponding to protein-protein interactions being the most frequent correlated pair hit class (Figure 1B). However, upon normalization, we found that the most abundant pair hits relative to the corresponding number of comparisons were found on pairs of snoRNA-snoRNA, tRNA-tRNA, transposable element (TE)-TE, and TE-pseudogenes molecules (Figure 1C).

Linkage disequilibrium

Correlated mutations will also reveal linkage effects, the joint inheritance of proximal SNP-sites due to reduced probabilities of crossing over events, known also as linkage disequilibrium.

Indeed, a pronounced increased frequency of pair hits at close genomic distances extending to a distance of up to 10K bps is evident (Figure 2). Most of the pair hits were found in close proximity along chromosomes with 47% of all pair hits (including inter-chromosomal pairs) detected within the genomic distance range of up to 10K bps on the same chromosome. Linkage disequilibrium appears lost at a distance of approximately 10K bps (Figure 2). This agrees well with previous reports on linkage effects in *A. thaliana* (31). Thus, linkage is an important source of SNP-SNP correlations. It does not require any functional and compensatory origin, and hence, needs to be accounted for in order to reduce false positive associations. Therefore, when comparing the observed correlation structure to random data, the random sampling was designed to reproduce the observed linkage decay (see Methods).

Correlation structure of the Arabidopsis nuclear genome

We inspected the global correlation structure of SNP-sites in the nuclear genome of *A. thaliana*. A heat map representing the ratio of the number of observed pair hits and the number of pairs in the distance-biased resampling shows that, indeed, regions of increased correlations are evident (Figure 3). The ratio was calculated with the number of pair hits per genomic-region-bin (100K) over the number of pairs per genomic-region-bin (bin size 100K) in the distance-biased re-sampling. Both local clusters as well as long-range paired regions of elevated correlation can be identified. Local clusters primarily correspond to genomic regions around centromers. Long-range interactions also appear associated with these chromosomal segments. While this global picture does not provide any immediate functional insights, it does show that the actual correlation structure differs from random expectation. Note that the emerging clusters of elevated pair hit frequency do not reflect local SNP-density as the random re-sampling does consider local SNP density (see Methods).

Filtering out confounding factors

To identify pair hits that reflect functional associations via compensatory mechanisms, confounding factors need to be accounted for, such as linkage (Lk), lineage (Ln), and transitivity (Tr). These biases have also been discussed in other studies (4, 8–10, 32). Hence, we aimed to determine their impact on our results and developed strategies to eliminate pair hits that are predominantly affected by linkage, lineage, and the transitivity effect. To detect whether the signal in the correlated mutations is coming from functional associations, we compared actual pair hits to carefully composed randomizations. The generated random re-parings of SNP-sites were designed to follow the same linkage distribution and were subjected to the same lineage and

transitivity thresholds as applied to the actual SNP-pairs (see Methods). In testing, whether pair hits reflect functional associations as opposed to result from confounding factors, we pursued the following rationale: If observed pair hits are considerably different from random pair hits with respect to their genomic element type statistics (profiles as displayed in Figure 4), functional associations have likely been detected. By contrast, in cases of similar statistics of the genomic element types, the observed pair hits will predominantly be uninformative. Intergenic regions were assumed to be non-interacting and used as the background probability. For all pair hits, we found that interactions between RNAs (miRNA, tRNA, snoRNA, sRNA), transposable elements (TE), and pseudogenes have a probability of being correlated above what is expected based on the background probability, while many other classes occur less frequently (Figure 4A). Naively, we expected other types of interactions such as protein-protein interactions to be enriched relative to intergenic regions. Instead, they were found to be depleted. Furthermore, upon randomly re-sampling, a very similar enrichment profile was obtained as for the observed pair hits data (Pearson correlation coefficient, r=0.86, Figure 4A). Thus, when considering all pair hits, no strong evidence of functional relevance of correlated mutations was apparent.

To account for linkage as a driver of correlated mutations, all pair hits with distances of smaller than 10K bps were discarded based on the observed distance dependence of pair hits (Figure 2). This subset is referred to as *subset Lk*. Again, enrichment of genomic classes represented in the set was tested relative to intergenic regions as the reference (Figure 4B). There was no pronounced difference in the enrichment factor distribution between the subset Lk and the complete results using intergenic regions-reference. Furthermore, when comparing the subset Lk to its random sampling, pair hits still exhibit a high correlation level (r=0.79).

Using different accessions from the same plant species can be expected to confer a strong phylogenetic bias to our results as the different accessions are related by descent. Therefore, it is essential to discriminate, whether the detected covariance is functional or primarily reflects the phylogenetic history (lineage). To account for phylogenetic bias, we measured all pairwise sequence distances between accessions and used the Average Distance Diameter (ADD) to assess the level of genome sequence diversity. The ADD was calculated among the groups of accessions that correspond to clusters of accessions based on the particular allele at a SNP-site (see Methods, Supplementary Figure 2). The rationale for this approach is to use SNP-sites resulting in clusters of accessions that correspond to high sequence diversity, as this can be taken as evidence that natural selection may have favor a particular allele at a given site. Consequently, only pair hits with corresponding clusters of accessions displaying high diversity were considered. The top-

75% percentile of the ADD diversity was used as the threshold. This subset was denoted as *subset Lk+Ln*, comprising 34% of the all pair hits.

The distribution genomic element enrichment of subset Lk+Ln displayed a different distribution compared to the pair hits obtained without any filters (Figure 4C). The distribution of subset Lk+Ln also showed increased deviation relative to its random sampling, albeit still at a high overall level, r=0.71. Thus, accounting for confounding factors still appears incomplete.

As the last explicitly considered confounding factor, we inspected possible effects from transitivity (Tr). Indirect covariances can result in overestimation of pair hits. Using pair hits with a smaller cluster coefficient, and hence, likely less affected by the transitivity effect, we aimed to reduce the number of redundant pairs. The 25% percentile of the lowest values of transitivity were considered for analysis (*subset Lk+Ln+Tr*, comprising pair hits with larger distance than 10K bps, the 75% highest ADD, and the 25% lowest transitivity values corresponding to 20% of the original set (8,549,704 pair hits)). The combination of the three filters yielded a distribution of the data that could not be reproduced by sampling. Sample and subset Lk+Ln+Tr have an associated correlation coefficient of r=0.16 (Figure 4D). Thus, following the rationale that actual profiles of genomic-class associations should not be reproduced by random sampling, we conclude that subset Lk+Ln+Tr may be enriched for functionally relevant correlations, and thus, may reflect molecular interactions. Interestingly, interactions such as tRNA-tRNA, protein-snoRNA, snoRNA-tRNA, miRNA-miRNA, miRNA-snoRNA, protein-tRNA, TE-TE, protein-protein (PPI), miRNA-protein, miRNA-miRNA, protein-tRNA, Protein-snoRNA were found overrepresented in this set relative to background (pairs of intergenic-intergenic regions). As these interaction classes appear biologically plausible, e.g. base-pair complementarity for RNA-mediated interactions or possible correlated amino acid changes in interacting proteins, or miRNA-target protein interactions reflecting the potential interaction of miRNA to the associated mRNA, the results obtained for subset Lk+Ln+Tr appear biologically reasonable. Likewise, pairings including pseudogenes were found at low levels, which can also be taken as evidence of the validity of our results as pseudogenes can be assumed non-functional. Note that the three introduced filtering steps were not applied sequentially, but were each based on the complete set and the intersection of those was taken. As the filtering reduced the number of pair hits, the correlation between actual and sampled pairs annotations may decrease purely due to increased stochasticity (smaller numbers). However, the number of considered pair hits after filtering remained large. Furthermore, the enrichment factors obtained for the various classes appear biologically plausible. This view is supported further when inspecting protein-protein and

miRNA-target gene interactions for the original and filtered set, where the filtered set yielded a larger fold increase of true interactions relative to random sampling (see below).

While the overall correlation structure after filtering (subset Lk+Ln+Tr) as assessed by the heatmap of all-against-all chromosomal regions (Figure 5A, 100K bins) appear similar to the unfiltered pair hit set (Figure 3), pronounced differences can also be observed. Regions of increased density of correlated SNPs appear more condensed to large, but local clusters. The respective cross-peaks indicative of possible non-local interactions are relatively depleted except for correlations between chromosomes 2 and 3, for which the two respective local clusters also appear correlated among each other. Evidently, as the heatmap presents the averaged overview, individual pair hits can still be present across all chromosomes and distance intervals. Furthermore, inter-chromosomal regions corresponding to possible correlations between centromeric regions, observed with high SNP frequency (Figure 5A), appear depleted for pair hits after filtering, indicating a reduced probability of true functional interactions via compensatory mutations.

Pronounced differences are apparent with regard to relative abundance of pair hits for particular genomic element classes. As described before, classes that appear biologically plausible have moved to higher rank positions compared to the unfiltered set (Figure 5B compared to Figure 1C). For example, protein-protein interactions are ranked seventh after filtering, but were ranked 24th in the original set.

Correlated mutations reflect known interactions

The subset Lk+Ln+Tr was tested for enrichment of known interactions between molecules. We focused on three types of interactions: protein-protein interactions (PPI), interactions between miRNAs and their target genes, and DNA-DNA interactions in chromatin loops.

Concerning PPIs, 4,020 pair hits were found between proteins reported to interact; i.e. each of the two correlated SNP-sites were found present on proteins reported to interact. This set contained 290 unique gene-gene pairs corresponding to 642 unique protein-protein pairs as some genes encode more than one protein variant via splice variation. When performing 100 runs of random re-pairings of the same set of polymorphic sites (subset Lk+Ln+Tr), counts of pair hits on interacting proteins were found with a mean of 1,053+/-80 (s.d.). By comparison, actual pair hits were enriched 3.8 fold (z-score=36.9, p=1.9E-298) for known PPI relative to random pairings, and thus, may reflect functional associations between proteins. Without filters (All pair hits) the difference between the actual number of PPIs found with correlated mutations and the PPI found in the sampling was less pronounced and increased by a factor of 1.75 only (6,049,457 pair hits,

z-score of 429.8, p=0). Thus, the introduced filtering steps proved effective in enriching for true interactions, but also significantly reduced the number of candidate associations, likely producing false negatives as well.

With regard to PPIs, the notion of compensatory mutations should have merit only in cases of nonsynonymous (i.e. amino acid changing) mutations. Synonymous (or silent) changes should not change the physics of interactions at all (but may affect other processes such as translation efficiency). Indeed, when examining SNP-SNP pairs on proteins known to interact, we found AMI-values associated with nonsynonymous-nonsynonymous changes were significantly larger (mean=0.021, median=0.0096, Wilcoxon test p-value= 3.6E-118) than those obtained for synonymous-synonymous pairs (mean=0.018, median=0.007) (Figure 6). While significant, the magnitude of difference is small, possibly explained by synonymous changes being in turn correlated with nearby functionally relevant nonsynonymous changes or by providing a suitable genomic context for fixing nonsynonymous changes.

Aside from protein-protein interactions, miRNA-protein-coding-mRNA are natural candidates molecular interaction types to harbor compensatory mutations, as for this type of molecular interaction, mutations directly translate into (almost) binary interaction decisions with base pairing possible or not. By contrast, mutations in proteins may either not result in any change of protein as they can be silent, and even if changing the amino acid, interactions between different amino acid residue types are not of binary character. Indeed, for miRNA-mRNA, we found 177 pair hits on 26 unique miRNA-mRNA molecule pairs corresponding to a 26.4 fold enrichment relative to random sampling (100 runs of random repairing, z-score=157.7, p=0). Again, when taking all pair hits; i.e. not including the filtering steps, the difference was less pronounced (fold=1.63, z-score=31.2, p=2.35E-214).

Finally, we probed for correlated mutations to inform on physical interactions at the DNA/chromosomal level. Here, we exploited the recently published Hi-C data on gene loop formations in Arabidopsis (26). In reported DNA-DNA contacts in gene loops, we found 60 pair hits (fold=2.8 relative to the mean obtained from 100 runs of random re-pairings, z-score=8.74, p=2.14E-18). In this test, the introduced distance filter of 10K bps eliminated most of the reported gene loop contacts. When including all pair hits (without any filtering), we detected 432,870 pair hits, 4.3 times more than the mean obtained from the 100 runs of random re-pairings of the same number of original pair hits (z-score=1,102, p=0). Note that, while there will be substantial linkage within 10K bps, and thus many pair hits (Figure 2), in the random repairing the distance-dependent frequency distribution was preserved to account for linkage. Thus, correlated mutations also reflect on chromosomal gene loop contacts.

Prediction of novel protein-protein interactions

As pair hits were indeed found to reflect molecular interactions between proteins, it can be assumed that also those pair hits that link two different proteins that are not reported to interact based on the currently available information, may in fact, interact. To test for this possibility and taken as an indirect evidence, we checked whether two proteins connected via at least one identified pair hit, but not reported to interact, are involved in the same biological process as captured by their GO-term annotation. Indeed, the semantic similarity of the putative 471,002 PPIs based on pair hits proved significantly higher semantic than obtained for a random sampling of protein pairs (Figure 7, Wilcoxon test p-value < 2.2e-16).

In addition to novel protein-protein interactions, Table 2 lists the pair counts of all other molecule types predicted to interact based on detected pair hits (filtered set) with one and, imposing a stricter filtering, five or more showing them as correlated. Actual molecule pairs associated with the latter are available as Supplementary File 1. Of note are the highlighted predictions for the types detected to be enriched relative to expectation (Figure 4D).

Co-expression of gene pairs linked via correlated mutations

In addition to direct physical interactions, functional links between two genes linked via correlated mutations could also become evident as correlated gene expression behavior. Exploiting a large compendium of gene-expression microarray data (see Methods), we indeed detected differences of the levels of gene-co-expression for gene pairs linked pair hits compared to gene pairs from a random pairing of SNPs (Figure 8). Here, and to allow for the effect of possible regulatory effects, we considered all SNP-locations within gene; i.e. including also non-CDSs sites such as introns or UTR. However, contrary to expectation, we did not find correlation levels to be shifted to larger positive values, but instead, a shift to larger absolute values. This result was even more pronounced when selecting only those gene pairs with five or more pair hits linking them (medians of absolute pairwise gene expression Pearson correlation coefficients of random gene pairs=0.086, median of gene pairs linked by at least one/five pair hit=0.091/0.097, p=7.7E-87/2.46E-169 (Wilcoxon test)), albeit the magnitude of difference was small. Thus, the expression of genes linked via pair hits is notably different from their random expectation.

To understand the unexpected increase of not only positive, but also negative correlation levels between genes with correlated mutations, we performed a GO-term (functional terms) enrichment analysis of gene sets with genes engaging in pairwise positive correlation (r> 0.2) and contrasted those to a set of genes with pairwise negative co-expression regulation (r< -0.2). In

line with expectations, genes engaging in positive pairwise gene co-expression and with at least 5 or more pair hits were enriched for the GO term "protein binding", p_{FDR}=1.8E-11 (other GO terms with p_{FDR}<0.05: "receptor binding or activity" p_{FDR}=5.8E-4; "nucleotide binding", p_{FDR}<2.6E-3). Thus, the notion of correlated mutations reflecting functional, and primarily physical, interactions seems validated as positive gene expression correlation is immediately consistent with this mode of interaction. By contrast, among genes with negative expression correlations, but also linked via at least 5 pair hits, no single GO term reached the same level of significance as observed for "protein binding" in the positive correlation set. Instead, GO terms "DNA or RNA binding", p_{FDR}=1.97E-3; "unknown molecular function", p_{FDR}=1.97E-3; "transferase activity", p_{FDR}=2.6E-3; "transporter activity", p_{FDR}=3.1E-3; "nucleic acid binding", p_{FDR}=1.2E-2; "other binding", p_{FDR}=1.2E-2, were obtained at borderline significance levels.

Discussion

The concept of detecting and exploiting correlated mutations as indications of physical or, in any way, functional interactions has been appealing from its inception (33–35). Its underlying rationale appears straightforward and convincing. Furthermore, the basis for its successful application - the availability of broad sequencing information - is growing rapidly. The power of this approach has also been demonstrated by its successful application in predicting structural aspects or proteins (e.g. (8, 11)) and RNA (32). At its core, the method requires evolution to have acted to select for compensatory mutations. Thus, sufficient evolutionary time needs to have elapsed in order for evolution to take effect. Therefore, previous studies on correlated mutation aimed to increase evolutionary distances between the considered genomes and included different species only.

Here, we set out to test and apply the concept of correlated mutations within the same species (*A. thaliana*). Evidently, the considered genomes will be more closely related with less evolutionary time separating the different Arabidopsis accessions, and therefore, the selective forces of evolution will have had less time to act. It has been estimated that *Arabidopsis thaliana* diverged from its ancestors around 5 MYA (36); i.e. relatively recently compared to the ~450 million years of land plant evolution (37). Thus, it was not obvious whether the concept of correlated mutations can be used successfully when applying it to within-species variation. Nonetheless, by implementing a series of filtering steps and by performing appropriate controls, our study has proven that correlated mutations can also reveal physical, or broadly functional, interactions between molecules or genomic regions even within a species. Correlated mutations were found to be enriched in molecule pairs known to interact. Thus, this confirmation also

suggests that this approach may also have predictive value. Our study also reveals the overall genome-wide correlation structure of the Arabidopsis genome and required implementing an efficient computation scheme rending the computation of around 10E+13 comparisons possible. Subsequently, we shall discuss methodological aspects as well as the biological implications of our results. By requiring a minimum of 5% minor allele, we effectively selected older polymorphisms and excluded polymorphisms found exclusively in a small number of accessions ("private" polymorphisms"). Thus, we enriched for sites on which selected may have taken effect.

Mutual information - the metric employed in this study – is a mathematically robust method to detect covariance between categorical variables (here, alleles). However, its power to detect functional associations has also been questioned due to different confounding factors (4, 8, 10). Two positions in an alignment can be covarying because they descended from the same ancestor in which two random mutations appeared (lineage) or because they are inherited jointly due to proximity on the chromosome (linkage) (31). In addition, methods for measuring covariance frequently fail to properly account for the transitivity effect, thereby overestimating the number of interactive pairs (8, 10). The transitivity effect is the result of indirect covariance. Additionally, sequencing errors could result in false and repetitive SNPs. This could lead to a substantial overestimation of the number of pair hits due to the transitivity effect, because all repetitions appear to be covarying with each other. To address these difficulties, we devised as series of filters. By maximizing genomic sequence distance within the splits of accessions introduced by the two alleles at a given site, we aimed to reduce the lineage effect. By penalizing sites with a high clustering coefficient, we diminished the influence of the transitivity effect. Linkage was avoided by only considering pairs at a distance of 10K bps or larger. In effect, the annotations characteristics of the actually detected pair hits differed significantly from its random control (Figure 4D), which can be taken as evidence that the filters have led to non-random effects. As, furthermore, the fold enrichment of known interactions detected by pair hits increased after filtering, we can also conclude that the effect of the introduced filtering steps was not only leading to non-random pair hits, but that those are enriched for true associations.

Our reasoning to conclude that we detected functional associations relied on comparisons to random controls. When generating random background pair distributions, we also aimed to avoid spurious associations by accounting for confounding factors. First, by maintaining the distance distribution between pairs in the random sampling, we tested if the enrichment was related to linkage, recombination patterns, gene conversion, chromosome structure or another factor related with the location of polymorphisms that could be influencing our findings. Without any filtering, these location-associated effects were indeed found to explain to large extent the

observed correlation structure (Figure 4A), and thus needs to be taken into account. However, randomly sampled pairs with low transitivity and high diversity in combination with maintaining the distance distribution showed little correlation with the real, filtered results. This suggests that the enrichment observed after filtering is neither related to the distribution of the polymorphic sites in the genome, nor to the influence of phylogenetic biases, nor to the transitivity originating from indirect covariation. Consequently, we feel supported in the conclusion that the observed covariance will reflect functional association. The recognition of linkage, lineage, and transitivity as strongly influencing factors is also important for the prediction of new interactions. However, these factors do not preclude true functional association, and thus, by filtering for them, truly functionally correlated sites may be lost upon filtering.

Natural selection that favors specific adaptations to specific environmental conditions represents another origin for covariance, also referred to as purifying selection (38, 39). In essence, environment acts as an independent factor producing indirect correlations between SNP-sites that are themselves not directly functionally linked. For example, accessions in hot climates may show specific mutations increasing their fitness relative to those in colder habitats. Those sites would be detected as correlated, but they are caused by a third factor, the environment. Furthermore, the environment may not only introduce independent changes, but also affects the fixation of compensatory mutations. For example, it has been shown for a yeast model of a particular human disease that environmental conditions influence the fixation of compensatory mutations (40). However, it has been reported that purifying selection in coding-regions in plants is virtually absent (41). This would suggest that the extent of correlation due to environmental factors and purifying selection may be small in Arabidopsis. However, it needs to be cautioned that seemingly neutral mutations may reflect unrecognized adaptations. Furthermore, it has been found that at least 5% of noncoding regions in *A. thaliana* are under purifying selection in the family of Brassicaceae (42). Therefore, indirectly correlated mutations induced by external factors cannot be ruled out to contribute to the set of pair hits reported in this study.

When using closely related genomes as done in this study, detecting correlated mutations necessarily becomes challenging for reasons of direct ancestry and linkage. However, it also bears the advantage of allowing direct vertical alignment comparisons across the genomes without necessitating establishing homology between different molecules in different species. Yet, large deletions/ insertions and rearrangements may happen also within species and may constitute a source of error of our study (43, 44).

Epistasis plays an important role in the fixation of compensatory mutations and is pivotal for molecular evolution (40, 45). This phenomenon allows the fixation of deleterious mutations if

compensated by other mutations, or generally by the genetic background of the organism. Therefore, in essence, our effort to identify correlated mutations can also be seen as yielding epistatic events and reveals the extent of this mode of evolution. Epistasis may also explain the unexpectedly small difference between the AMI of nonsynonymous-nonsynonymous and synonymous-synonymous pair hits in known PPI (Figure 6). While not changing the amino acid, synonymous changes may provide the conducive background for a nonsynonymous change via an as yet unidentified mechanism.

As we were able to demonstrate that pair hits are enriched on molecules known to interact, those that are found on molecules not (yet) reported to interact may, in fact, suggest that they may be functionally linked and may form physical contacts (Table 2). A large proportion of those putative interactions are of the type protein-protein interactions, which, of course, also reflects the genomic sequence space protein coding genes occupy in the Arabidopsis genome (Figure 1A). The predicted interactions await experimental confirmation or support by other data. Note that our method allows detecting correlated mutations that have functional meaning and reflect functional association, however, given that compensatory mechanisms are evolutionary phenomena not necessarily all functional associations present correlated mutations. In other words, where you find correlated mutations (controlled for confounding factors) you have a functional association, but not always that you have a functional association you find a correlated mutation in the involved molecules.

Pairs of tRNA-tRNA genes were found enriched for correlated mutations compared with intergenic regions (Figure 4). This may originate from tRNA-tRNA interactions in ribosomes as described by Smith and Yarus in 1989 (46). The detected compensatory mutations may reflect contacts between the anticodon loops of the P-site and A-site of the tRNAs. Other RNAs such as miRNA and snoRNAs were also found enriched interacting in comparison with intergenic regions. miRNA-miRNA correlations may perhaps be explained due to competitive mechanisms (47). Some RNA transcripts are able to communicate and co-regulate other RNAs by competing with its shared RNA regulator molecule, typically a miRNA (47–49). These are referred to as competing endogenous RNAs (ceRNAs) and the term “target mimicry” is often also used to describe this phenomenon in plants. ceRNAs can be two or more mRNAs with the same miRNA-response element (MRE) competing for the same miRNA, or ncRNA-mRNA competing for a miRNA (50). snoRNA in Arabidopsis are often polycistronic being transcribed from a single promoter and are also found to be dicistronic with tRNAs (51). This last combination suggests that those snoRNA and tRNA share regulatory mechanisms, transcriptional proteins and are

evolutionary linked (52). Even though it is still obscure why these two types of molecules rely on the same transcriptional mechanisms, a joint functional role appears possible.

Of note also is the high number of novel interactions involving transposable elements (TEs). While the argument of SNP-density and covered genomic space of TE-genes holds as well, in the light of the reported functional roles of TEs, correlated mutations appear plausible as well. For example, it was shown that they act as gene expression regulators via small interfering RNAs (53) or by introducing novel cis-regulatory elements (54); i.e. different genomic sites are linked up, possibly reflected by correlated mutations.

Interestingly, pair hits were found enriched in 3D-genome gene loop contact regions suggesting a link between 3D-genome organization and the sequence correlation structure. This may indicate that the integrity of such contacts is preserved in evolution via compensatory mutations. Thus, our study suggests a link between the 3D-structural organization of the Arabidopsis genome and its sequence correlation patterns.

The continued and expanded sequencing efforts have yielded large SNP collections not only in Arabidopsis. The potential of these data to enable functional analyses has been demonstrated before, for example, specifically in *Arabidopsis thaliana* in the context of identifying novel cis-regulatory elements (55). Our study demonstrates the power of exploiting SNP information to unravel the global correlation structure of the Arabidopsis genome and to discern molecular interactions.

Conclusions

By developing a series of filtering steps and random controls, our study was successful in applying the concept of correlated mutations to the identification of functional, and primarily direct, physical interactions in closely related genomes within a single species by exploiting single nucleotide polymorphisms. This study succeeded in performing a genome-wide detection of the sequence-level correlation structure of a plant genome. As known physical interactions of molecules have been observed to be enriched for correlated mutations, correlated mutations on molecules not yet reported to interact or be functionally related may also have predictive value. With the rapidly increasing sequencing information, the approaches developed in this study may find broad application in other species.

Acknowledgements

We wish to expressly thank Marco Ehlert (Max Planck of Colloids and Interfaces, Potsdam-Golm, Germany) for the critical help implementing an efficient computation strategy making the necessary massive computations possible. Further computing support was provided by Andreas Donath from Max Planck for Molecular Plant Physiology (Potsdam-Golm, Germany) and additional computational resources kindly made available by Alessandra Buonanno (Max Planck Institute for Gravitational Physics, Albert Einstein Institute, Potsdam-Golm, Germany), and the Max Planck Computer and Data Facility.

Declarations

Authors' contributions

DW conceived the study. LPJ performed all computations except for the co-expression analysis performed by DW. Both authors devised the strategies and methods of statistical testing, analyzed and interpreted the results, and wrote the manuscript.

Competing interests

The authors declare that they have no competing interest associated with this study.

Availability of data and materials

All information of data resources and methods necessary to replicate this study has been disclosed.

References

1. Zhang,Y., Gao,P. and Yuan,J.S. (2010) Plant protein-protein interaction network and interactome. *Curr. Genomics*, **11**, 40–46.
2. Axtell,M.J. (2013) Classification and Comparison of Small RNAs from Plants. *Annu. Rev. Plant Biol.*, **64**, 137–159.
3. Panwar,B., Arora,A. and Raghava,G.P. (2014) Prediction and classification of ncRNAs using structural information. *BMC Genomics*, **15**, 127.
4. Lapedes,A.S., Giraud,B.G., Liu,L. and Stormo,G.D. (1999) Correlated Mutations in Models of Protein Sequences: Phylogenetic and Structural Effects. *IMS Lect. Notes-Monograph Ser.*, **33**, 236–256.
5. Barkan,A., Rojas,M., Fujii,S., Yap,A., Chong,Y.S., Bond,C.S. and Small,I. (2012) A Combinatorial Amino Acid Code for RNA Recognition by Pentatricopeptide Repeat Proteins. *PLoS Genet.*, **8**, 4–11.
6. Okuda,K., Hammani,K., Tanz,S.K., Peng,L., Fukao,Y., Myouga,F., Motohashi,R., Shinozaki,K., Small,I. and Shikanai,T. (2010) The pentatricopeptide repeat protein OTP82 is required for RNA editing of plastid ndhB and ndhG transcripts. *Plant J.*, **61**, 339–349.
7. Kindgren,P., Yap,A., Bond,C.S. and Small,I. (2015) Predictable Alteration of Sequence Recognition by RNA Editing Factors from Arabidopsis. *Plant Cell Online*, **27**, tpc.114.134189.
8. Marks,D.S., Colwell,L.J., Sheridan,R., Hopf,T.A., Pagnani,A., Zecchina,R. and Sander,C. (2011) Protein 3D structure computed from evolutionary sequence variation. *PLoS One*, **6**.
9. Hopf,T.A., Schärfe,C.P.I., Rodrigues,J.P.G.L.M., Green,A.G., Kohlbacher,O., Sander,C., Bonvin,A.M.J.J. and Marks,D.S. (2014) Sequence co-evolution gives 3D contacts and structures of protein complexes. *Elife*, **3**, e03430.
10. Stein,R.R., Marks,D.S. and Sander,C. (2015) Inferring Pairwise Interactions from Biological Data Using Maximum-Entropy Probability Models. *PLoS Comput. Biol.*, **11**, 1–22.
11. Barkan,A. and Small,I. (2014) Pentatricopeptide repeat proteins in plants. *Annu. Rev. Plant Biol.*, **65**, 415–42.
12. Okuda,K., Shoki,H., Arai,M., Shikanai,T., Small,I., Nakamura,T., Barkan,A. and Small,I. (2014) Quantitative analysis of motifs contributing to the interaction between PLS-subfamily members and their target RNA sequences in plastid RNA editing. *Annu. Rev. Plant Biol.*, **65**, 870–882.
13. Korostelev,Y.D., Zharov,I.A., Mironov,A.A., Rakhmaininova,A.B. and Gelfand,M.S. (2016) Identification of position-specific correlations between DNA-Binding domains and their

binding sites. application to the merr family of transcription factors. *PLoS One*, **11**, 1–23.

14. Bindewald,E. and Shapiro,B. a (2013) Computational detection of abundant long-range nucleotide covariation in Drosophila genomes. *RNA*, 10.1261/rna.037630.112.

15. Van Dijk,A.D.J. and Van Ham,R.C.H.J. (2010) Conserved and variable correlated mutations in the plant MADS protein network. *BMC Genomics*, **11**.

16. Alonso-Blanco,C., Andrade,J., Becker,C., Bemm,F., Bergelson,J., Borgwardt,K.M., Cao,J., Chae,E., Dezwaan,T.M., Ding,W., *et al.* (2016) 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell*, **166**, 481–491.

17. Lamesch,P., Berardini,T.Z., Li,D., Swarbreck,D., Wilks,C., Sasidharan,R., Muller,R., Dreher,K., Alexander,D.L., Garcia-Hernandez,M., *et al.* (2012) The Arabidopsis Information Resource (TAIR): Improved gene annotation and new tools. *Nucleic Acids Res.*, **40**, 1202–1210.

18. Romano,S., Bailey,J., Vinh,N.X. and Verspoor,K. (2014) Standardized Mutual Information for Clustering Comparisons: One Step Further in Adjustment for Chance. *Proc. 31st Int. Conf. Mach. Learn.*, **32**.

19. Tuimala,J. (2006) A primer to phylogenetic analysis using the PHYLIP package.

20. Brandão,M.M., Dantas,L.L. and Silva-Filho,M.C. (2009) AtPIN: Arabidopsis thaliana Protein Interaction Network. *BMC Bioinformatics*, **10**, 454.

21. Wang,C., Marshall, a., Zhang,D. and Wilson,Z. a. (2012) ANAP: An Integrated Knowledge Base for Arabidopsis Protein Interaction Network Analysis. *Plant Physiol.*, **158**, 1523–1533.

22. Orchard,S., Ammari,M., Aranda,B., Breuza,L., Briganti,L., Broackes-Carter,F., Campbell,N.H., Chavali,G., Chen,C., Del-Toro,N., *et al.* (2014) The MIntAct project - IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.*, **42**, 358–363.

23. Yu,G., Li,F., Qin,Y., Bo,X., Wu,Y. and Wang,S. (2010) GOSemSim: An R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics*, **26**, 976–978.

24. Kozomara,A. and Griffiths-Jones,S. (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.*, **42**, D68-73.

25. Chou,C.H., Chang,N.W., Shrestha,S., Hsu,S. Da, Lin,Y.L., Lee,W.H., Yang,C.D., Hong,H.C., Wei,T.Y., Tu,S.J., *et al.* (2016) miRTarBase 2016: Updates to the experimentally validated miRNA-target interactions database. *Nucleic Acids Res.*, **44**, D239–D247.

26. Liu,C., Wang,C., Wang,G., Becker,C., Zaidem,M. and Weigel,D. (2016) Genome-wide analysis of chromatin packing in *Arabidopsis thaliana* at single-gene resolution. *Genome*

Res., 10.1101/gr.204032.116.

27. Craigon,D.J., James,N., Okyere,J., Higgins,J., Jotham,J. and May,S. (2004) NASCArrays: a repository for microarray data generated by NASC's transcriptomics service. *Nucleic Acids Res.*, **32**, 575D–577.

28. Rhee,S.Y., Beavis,W., Berardini,T.Z., Chen,G., Dixon,D., Doyle,A., Garcia-Hernandez,M., Huala,E., Lander,G., Montoya,M., *et al.* (2003) The Arabidopsis Information Resource (TAIR): A model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Res.*, **31**, 224–228.

29. Berardini,T.Z., Reiser,L., Li,D., Mezheritsky,Y., Muller,R., Strait,E. and Huala,E. (2015) The arabidopsis information resource: Making and mining the 'gold standard' annotated reference plant genome. *Genesis*, **53**, 474–485.

30. Benjamini,Y. and Hochberg,Y. (1995) Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing Author (s): Yoav Benjamini and Yosef Hochberg Source : Journal of the Royal Statistical Society . Series B (Methodological), Vol . 57 , No . 1 Published by : *J. R. Stat. Soc. Ser. B*, **57**, 289–300.

31. Kim,S., Plagnol,V., Hu,T.T., Toomajian,C., Clark,R.M., Ossowski,S., Ecker,J.R., Weigel,D. and Nordborg,M. (2007) Recombination and linkage disequilibrium in Arabidopsis thaliana. *Nat. Genet.*, **39**, 1151–1155.

32. Weinreb,C., Riesselman,A.J., Ingraham,J.B., Gross,T., Sander,C. and Marks,D.S. (2016) 3D RNA and Functional Interactions from Evolutionary Couplings. *Cell*, **165**, 963–975.

33. Göbel,U., Sander,C., Schneider,R. and Valencia,A. (1994) Correlated mutations and residue contacts in proteins. *Proteins Struct. Funct. Genet.*, **18**, 309–317.

34. Halabi,N., Rivoire,O., Leibler,S. and Ranganathan,R. (2009) Protein Sectors: Evolutionary Units of Three-Dimensional Structure. *Cell*, **138**, 774–786.

35. Pazos,F., Helmer-Citterich,M., Ausiello,G. and Valencia, a (1997) Correlated mutations contain information about protein-protein interaction. *J. Mol. Biol.*, **271**, 511–523.

36. Koch,M. a and Matschinger,M. (2007) Evolution and genetic differentiation among relatives of Arabidopsis thaliana. *Proc. Natl. Acad. Sci. U. S. A.*, **104**, 6272–6277.

37. Wellman,C.H., Osterloff,P.L. and Mohiuddin,U. (2003) Fragments of the earliest land plants. *Nature*, **425**, 282–285.

38. Zhu,Y., Qian,W. and Hua,J. (2010) Temperature modulates plant defense responses through NB-LRR proteins. *PLoS Pathog.*, **6**, e1000844.

39. Wang,Y., Zhang,Y., Wang,Z., Zhang,X. and Yang,S. (2013) A missense mutation in CHS1, a TIR-NB protein, induces chilling sensitivity in Arabidopsis. *Plant J.*, **75**, 553–565.

40. Filteau,M., Hamel,V., Pouliot,M.-C., Gagnon-Arsenault,I., Dube, a. K. and Landry,C.R. (2015) Evolutionary rescue by compensatory mutations is constrained by genomic and environmental backgrounds. *Mol. Syst. Biol.*, **11**, 832–832.

41. Hough,J., Williamson,R.J. and Wright,S.I. (2013) Patterns of selection in plant genomes. *Annu. Rev. Ecol. Evol. Syst.*, **44**, 31–49.

42. Haudry,A., Platts,A.E., Vello,E., Hoen,D.R., Leclercq,M., Williamson,R.J., Forczek,E., Joly-Lopez,Z., Steffen,J.G., Hazzouri,K.M., *et al.* (2013) An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. *Nat. Genet.*, **45**, 891–8.

43. Long,Q., Rabanal,F.A., Meng,D., Huber,C.D., Farlow,A., Platzer,A., Zhang,Q., Vilhjálmsson,B.J., Korte,A., Nizhynska,V., *et al.* (2013) Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden. *Nat. Genet.*, **45**, 884–890.

44. Imprialou,M., Kahles,A., Steffen,J.G., Osborne,E.J., Gan,X., Lempe,J., Bhomra,A., Belfield,E., Visscher,A., Greenhalgh,R., *et al.* (2017) Genomic rearrangements in Arabidopsis considered as quantitative traits. *Genetics*, **205**, 1425–1441.

45. Breen,M.S., Kemena,C., Vlasov,P.K., Notredame,C. and Kondrashov,F.A. (2012) Epistasis as the primary factor in molecular evolution. *Nature*, **490**, 535–538.

46. Smith,D. and Yarus,M. (1989) tRNA-tRNA interactions within cellular ribosomes. *Proc. Natl. Acad. Sci.*, **86**, 4397–4401.

47. Salmena,L., Poliseno,L., Tay,Y., Kats,L. and Pandolfi,P.P. (2011) A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell*, **146**, 353–8.

48. Ala,U., Karreth,F.A., Bosia,C., Pagnani,A., Taulli,R., Léopold,V. and Tay,Y. (2013) Integrated transcriptional and competitive endogenous RNA networks are cross-regulated in permissive molecular environments. *Proc. Natl. Acad. Sci. U. S. A.*, **110**, 7154–7159.

49. Hansen,T.B., Jensen,T.I., Clausen,B.H., Bramsen,J.B., Finsen,B., Damgaard,C.K. and Kjems,J. (2013) Natural RNA circles function as efficient microRNA sponges. *Nature*, **495**, 384–8.

50. Tay,Y., Rinn,J. and Pandolfi,P.P. (2014) The multilayered complexity of ceRNA crosstalk and competition. *Nature*, **505**, 344–52.

51. Kruszka,K., Barneche,F., Guyot,R., Ailhas,J., Meneau,I., Schiffer,S., Marchfelder,A. and Echeverría,M. (2003) Plant dicistronic tRNA-snoRNA genes: A new mode of expression of the small nucleolar RNAs processed by RNase Z. *EMBO J.*, **22**, 621–632.

52. Qu,G., Kruszka,K., Plewka,P., Yang,S.-Y., Chiou,T.-J., Jarmolowski,A., Szweykowska-Kulinska,Z., Echeverria,M. and Karlowski,W.M. (2015) Promoter-based identification of

novel non-coding RNAs reveals the presence of dicistronic snoRNA-miRNA genes in Arabidopsis thaliana. *BMC Genomics*, **16**, 1009.

53. McCue,A.D. and Slotkin,R.K. (2012) Transposable element small RNAs as regulators of gene expression. *Trends Genet.*, **28**, 616–623.

54. Makarevitch,I., Waters,A.J., West,P.T., Stitzer,M., Hirsch,C.N., Ross-Ibarra,J. and Springer,N.M. (2015) Transposable Elements Contribute to Activation of Maize Genes in Response to Abiotic Stress. *PLoS Genet.*, **11**.

55. Korkuc,P., Schippers,J.H.M. and Walther,D. (2014) Characterization and Identification of cis-Regulatory Elements in Arabidopsis Based on Single-Nucleotide Polymorphism Information. *Plant Physiol.*, **164**, 181–200.

Figure legends

Figure 1. Summary of overall SNP and pair hits data and annotation characteristics. (A) Density of polymorphic sites along the genome as represented by the chromosome structure and molecule/genomic element type. (B) Percentage of pair hit annotation classes, counted as relative number of pair hits (left) and relative number of paired molecules (right). (C) Normalized abundance of pair hits representing the fraction of all significant SNP-SNP correlations carrying a particular type of class relative to all possible comparisons of any SNP-SNP site of this class. Numbers above the normalized frequency bars denote absolute counts of pair hits.

Figure 2. Frequency distribution of genomic distances in base-pairs (bp) of SNP-sites found correlated (pair hits) at bin width of 100 bps.

Figure 3. Heatmap of the natural logarithm of the ratio of number of observed pairs and the number of sampled pairs, counts per bins of 100K bps.

Figure 4. Enrichment (Eqs. 6-10) of pair hits in genomic classes using as reference intergenic regions-intergenic region correlations and compared to its resampling. (A) All pair hits. (B) *Subset Lk*, pair hits with more than 10K bps distance between them. (C) *Subset Lk + Ln*, pair hits with more than 10K bps distance between them and with associated values of average distance diameter (ADD) larger than the 75% percentile of ADD values. (D) *Subset Lk+Ln+T*r, pair hits with more than 10K distance between them, with values of ADD larger than 75% percentile of sequence distance, and with values of transitivity smaller than the 25% percentile of the cluster coefficient. Pearson correlation coefficients, r, of actual vs. random (sample) enrichment factors are given for each analysis. Wr correspond to correlation coefficients weighted according to the number of pairs entering the correlation. For comparison, raw enrichment values as reported in (A) are repeated in all graphs (light blue bars).

Figure 5. Summary of pair hit results obtained for subset Lk+Ln+Tr. (A) Heatmap of the natural logarithm of the ratio of number of observed pairs in the filtered data and the number of randomly sampled pairs, counts per bins of 100K bps. (B) Normalized abundance of pair hits representing the fraction of all significant SNP-SNP correlations carrying a particular type of genomic element class relative to all possible comparisons of any SNP-SNP site of this class. Numbers above the normalized frequency bars denote absolute counts of pair hits.

Figure 6. AMI-value distribution of SNP-SNP pairs on interacting proteins resulting either in nonsynonymous changes on both interacting proteins (mean=0.021, median=0.0096) or in synonymous changes (mean=0.018, median=0.0070), Wilcoxon-test p-value 3.63E-118.

Figure 7. Boxplot of the semantic similarity of proteins predicted to interact based on a pair hit linking them compared to a set of protein pairs formed randomly. In this statistic, only those protein pairs were considered that are not reported to interact using the available PPI information (Wilcoxon test p-value < 2.2e-16).

Figure 8. Pairwise gene expression correlation of transcripts encoded by gene pairs linked via correlated mutations (pair hits) and as measured by raw (left) and absolute (right graph) Pearson correlation coefficient, r. Considered gene-pair sets were linked by at least one (294,479 gene pairs with available gene expression information for both genes) or five pair hits (56,718 pairs) and compared to a set of randomized pairings (172,443 gene pairs) of the same set of SNP-sites. All SNP-locations on genes were considered; i.e. also including non-CDS regions such as introns or UTR.

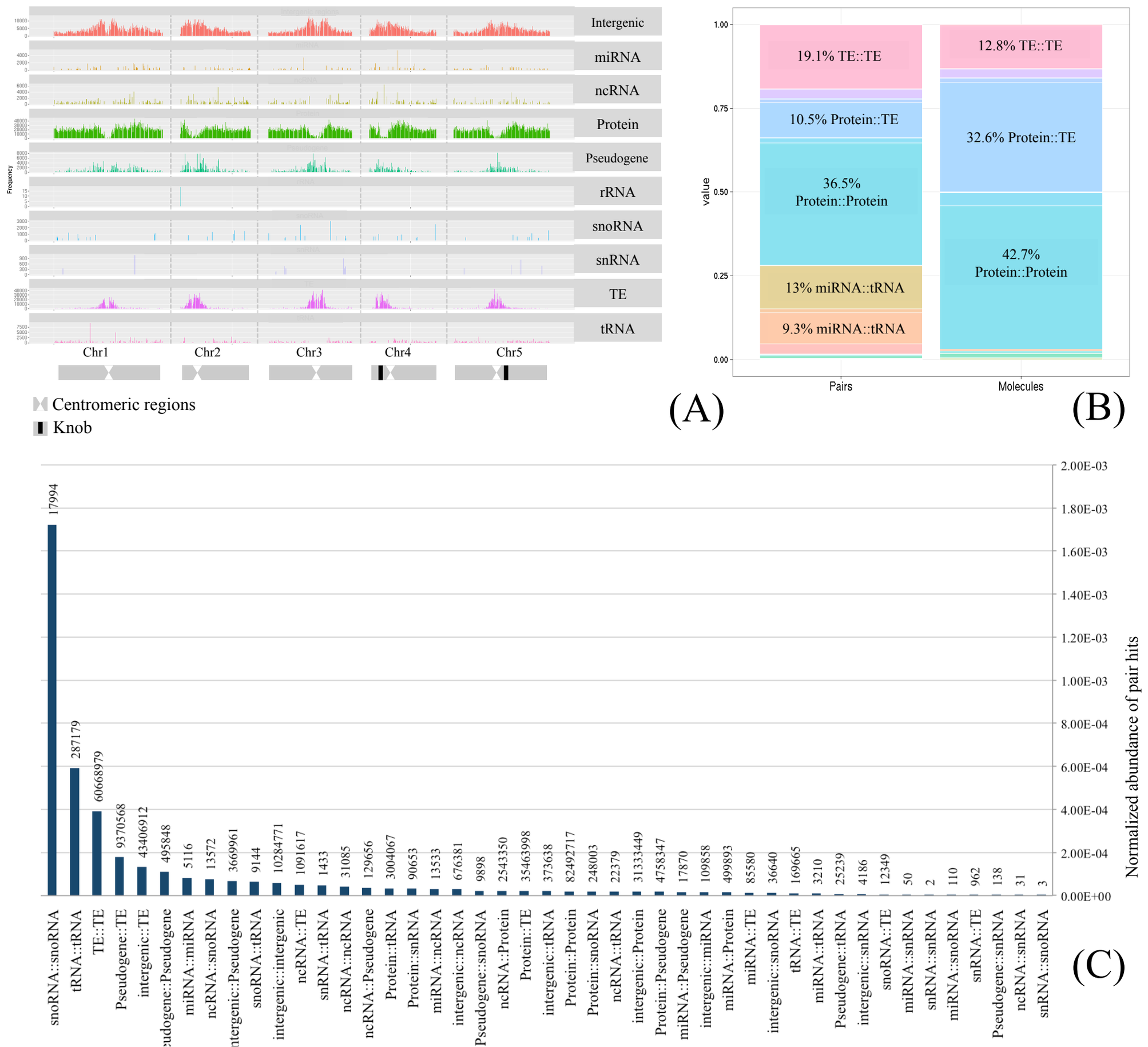
Tables

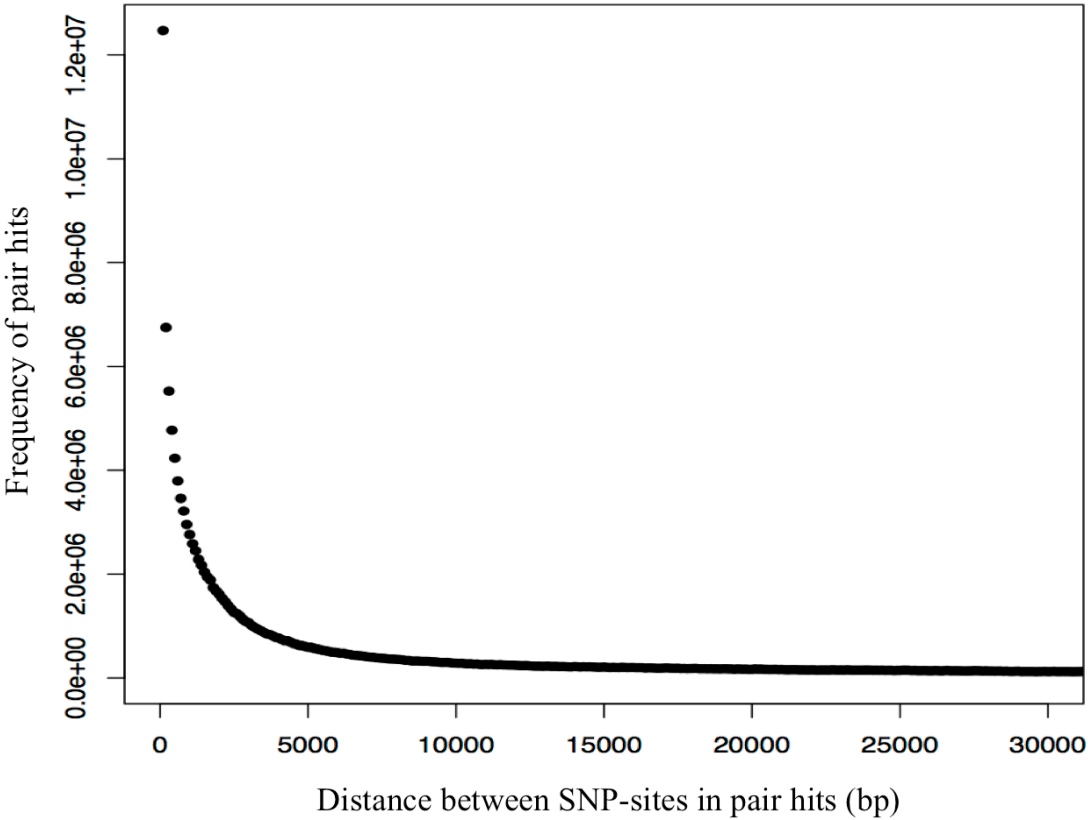
Chr	L, Length of Chr (bp)	N, Number of SNPs on Chr	N/L [%]	NP, Number of SNPs in at least in one pair hit	NP/N [%]
1	30427671	2597567	8.54%	310017	12%
2	19698289	1868168	9.48%	230652	12%
3	23459830	2193866	9.35%	251132	11%
4	18585056	1766544	9.51%	221331	13%
5	26975502	2281284	8.46%	316096	14%
Total 1-5	119146348	10707429	8.99%	1329228	12%

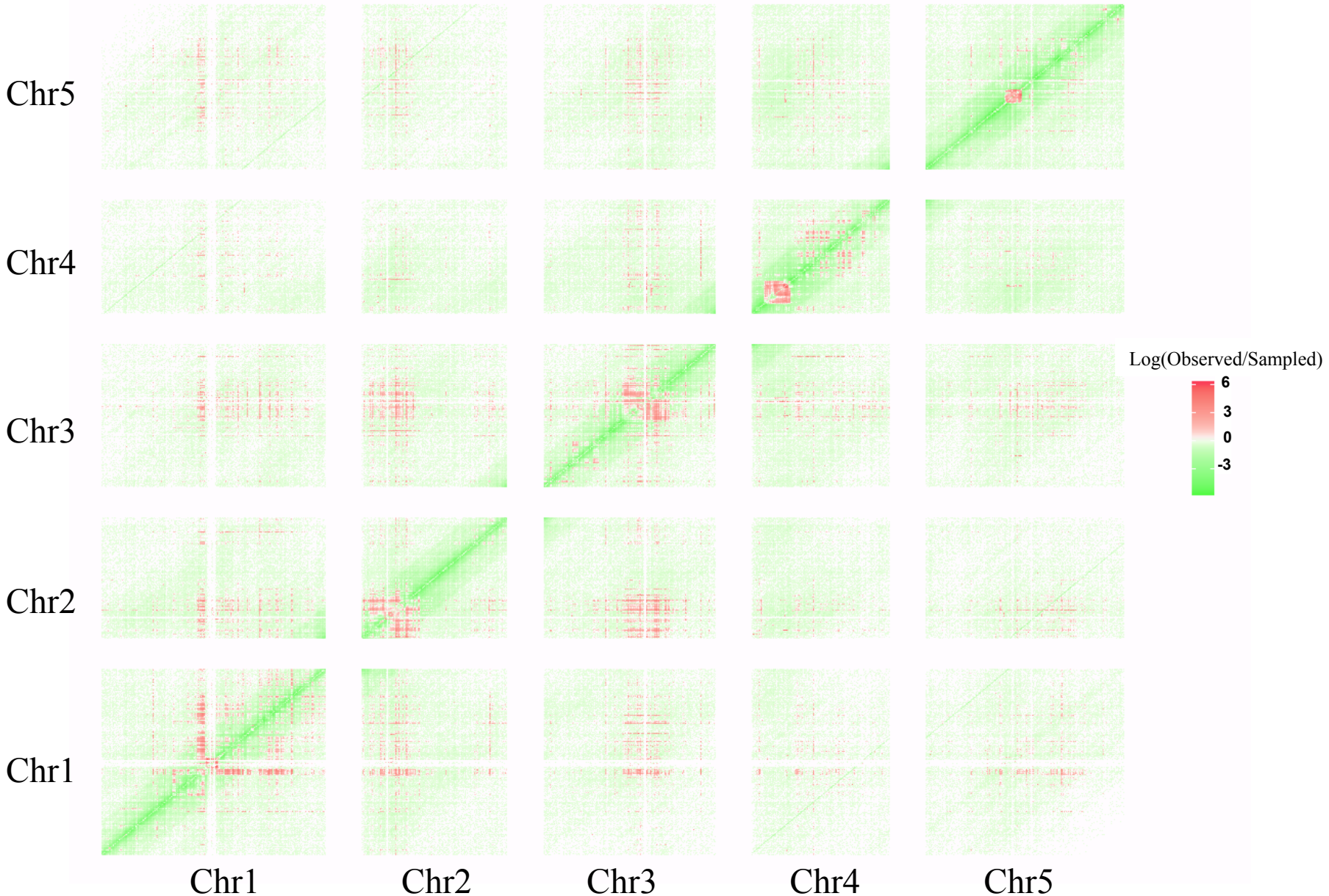
Table 1. Summary of the numbers of SNPs per chromosome (Chr) and their relation to the size of chromosomes and the number of SNPs reported in pair hits.

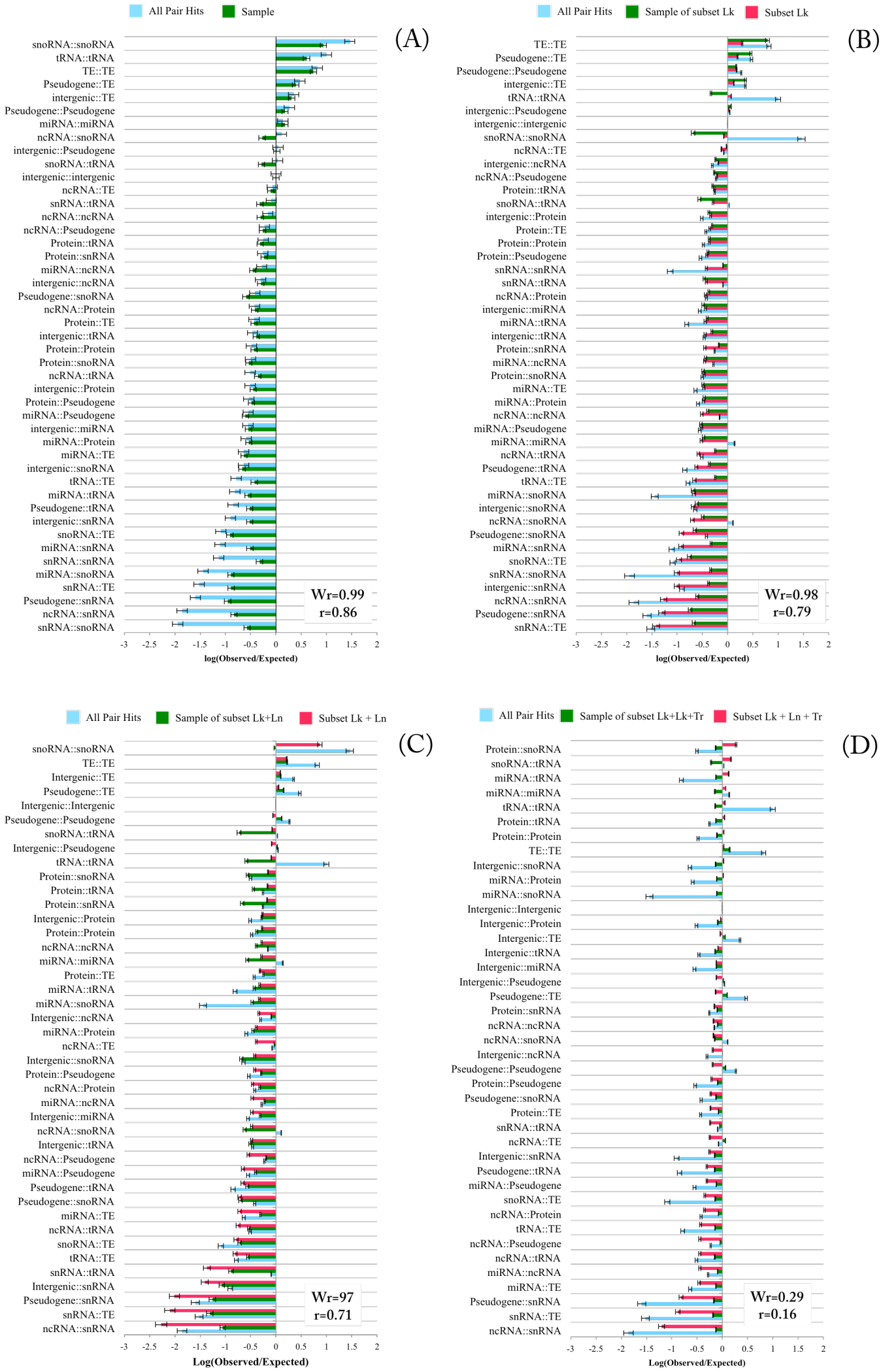
Class	At least 1 pair hit	5 of more pair hits
Protein::Protein	**471002**	**70079**
Protein::TE	391524	53514
TE::TE	**170575**	**38862**
ncRNA::Protein	12848	1234
Protein::tRNA	**11805**	**1110**
ncRNA::TE	7059	1158
miRNA::Protein	**4705**	**367**
tRNA::TE	4046	307
Protein::snoRNA	**2519**	**195**
miRNA::TE	1706	84
snoRNA::TE	742	33
Protein::snRNA	210	12
ncRNA::tRNA	155	4
ncRNA::ncRNA	103	10
tRNA::tRNA	**86**	**1**
miRNA::tRNA	**69**	**3**
miRNA::ncRNA	58	1
snRNA::TE	45	1
ncRNA::snoRNA	32	0
snoRNA::tRNA	**25**	**0**
miRNA::miRNA	**15**	**0**
miRNA::snoRNA	**5**	**0**
snRNA::tRNA	3	0
ncRNA::snRNA	1	0

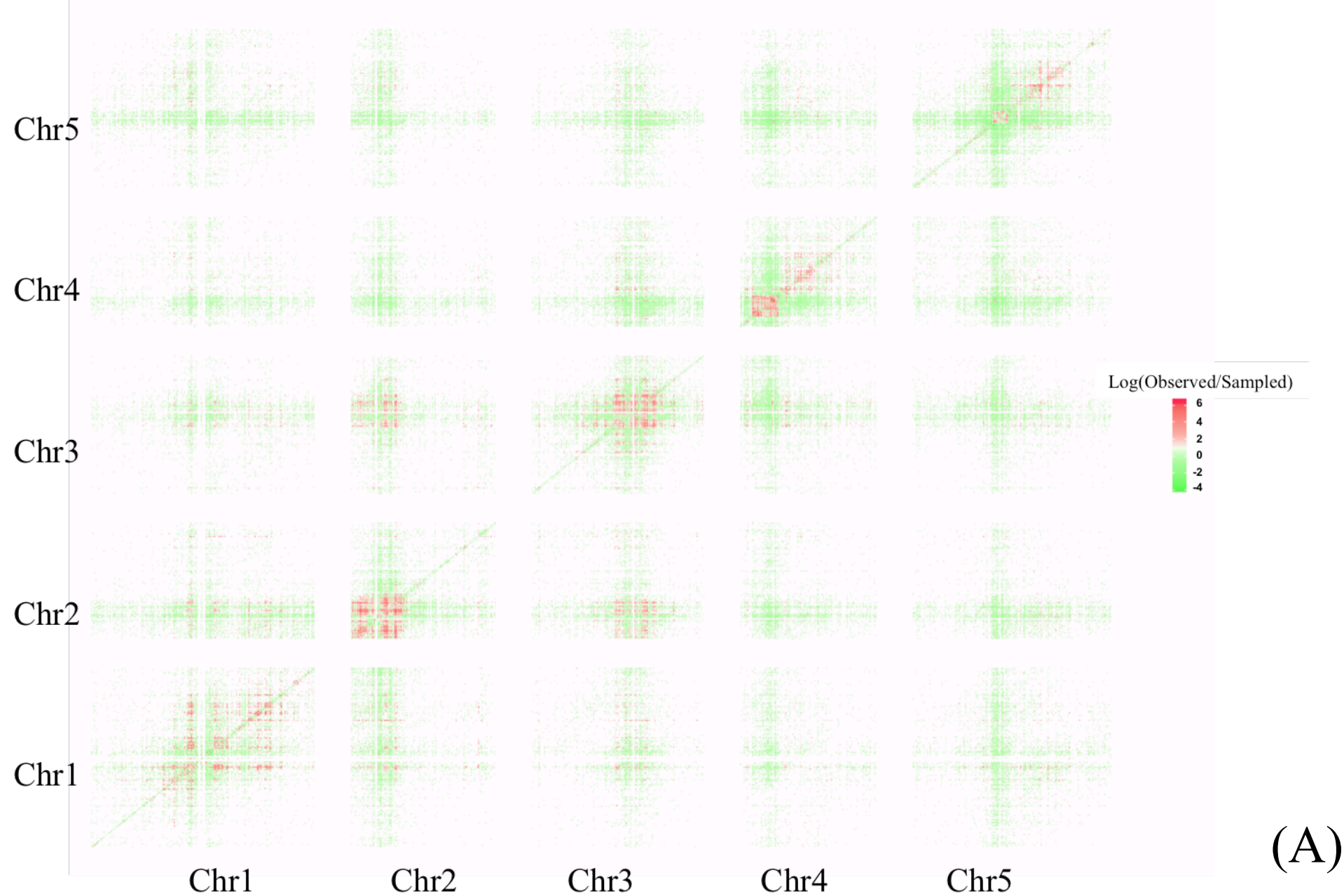
Table 2. Predicted interactions of molecules based on the presence of at least one/ five pair hit(s) revealing them as correlated. Highlighted bold are the molecule type pairs identified as enriched for pair hits in the genome-wide screen and after Filtering (Figure 4D) (Transfer RNA (tRNAs), small nuclear RNA (snRNA), Small nucleolar RNAs (snoRNAs), non-coding RNA (ncRNA), microRNA (miRNA), protein coding sequences, Transposable Elements (TE)). Actual molecule pairs predicted to interact based on five or more pair hits are available as Supplementary File 1.

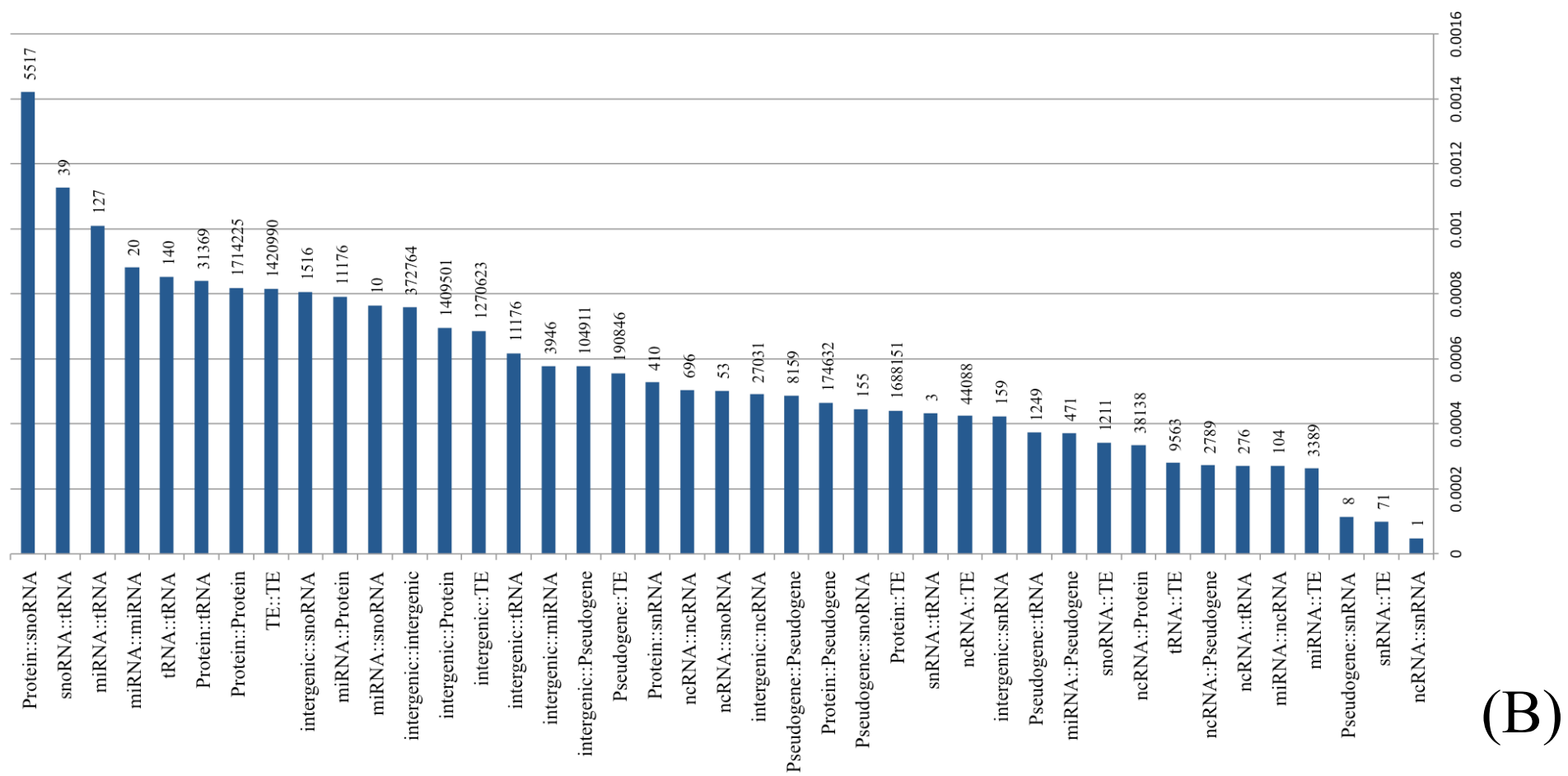

Intergenic
miRNA
ncRNA
Protein
Pseudogene
rRNA
snoRNA
snRNA
TE
tRNA
Frequency
Chr1
Chr2
Chr3
Chr4
Chr5
Centromeric regions
Knob
(A)
19.1% TE::TE
10.5% Protein::TE
36.5% Protein::Protein
13% miRNA::tRNA
9.3% miRNA::tRNA
12.8% TE::TE
32.6% Protein::TE
42.7% Protein::Protein
value
Pairs
Molecules
(B)
17994
287179
60668979
9370568
43406912
495848
5116
13572
3669961
9144
10284771
1091617
1433
31085
129656
3004067
90653
13533
676381
9898
2543350
35463998
373638
82492717
248003
22379
31333449
4758347
17870
109858
499893
85580
36640
169665
3210
25239
4186
12349
50
2
110
962
138
31
3
snoRNA::snoRNA
tRNA::tRNA
TE::TE
Pseudogene::TE
intergenic::TE
Pseudogene::Pseudogene
miRNA::miRNA
ncRNA::snoRNA
intergenic::Pseudogene
snoRNA::tRNA
intergenic::intergenic
ncRNA::TE
snRNA::tRNA
ncRNA::ncRNA
ncRNA::Pseudogene
Protein::tRNA
Protein::snRNA
miRNA::ncRNA
intergenic::ncRNA
Pseudogene::snoRNA
ncRNA::Protein
Protein::TE
intergenic::tRNA
Protein::Protein
Protein::snoRNA
ncRNA::tRNA
intergenic::Protein
Protein::Pseudogene
miRNA::Pseudogene
intergenic::miRNA
miRNA::Protein
miRNA::TE
intergenic::snoRNA
tRNA::TE
miRNA::tRNA
Pseudogene::tRNA
intergenic::snRNA
snoRNA::TE
miRNA::snRNA
snRNA::snRNA
miRNA::snoRNA
snRNA::TE
Pseudogene::snRNA
ncRNA::snRNA
snRNA::snoRNA
Normalized abundance of pair hits
(C)

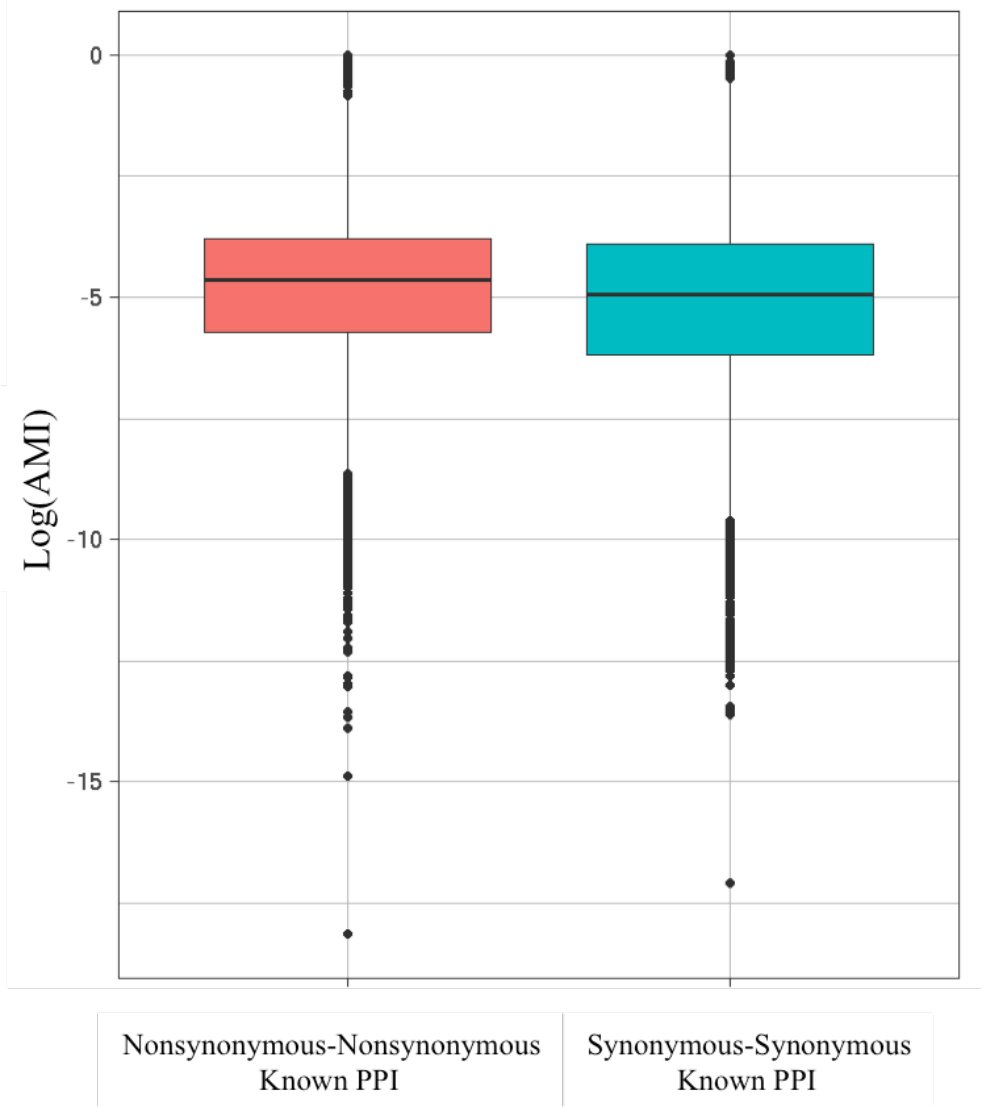

Frequency of pair hits
0.0e+00
2.0e+06
4.0e+06
6.0e+06
8.0e+06
1.0e+07
1.2e+07
0
5000
10000
15000
20000
25000
30000
Distance between SNP-sites in pair hits (bp)

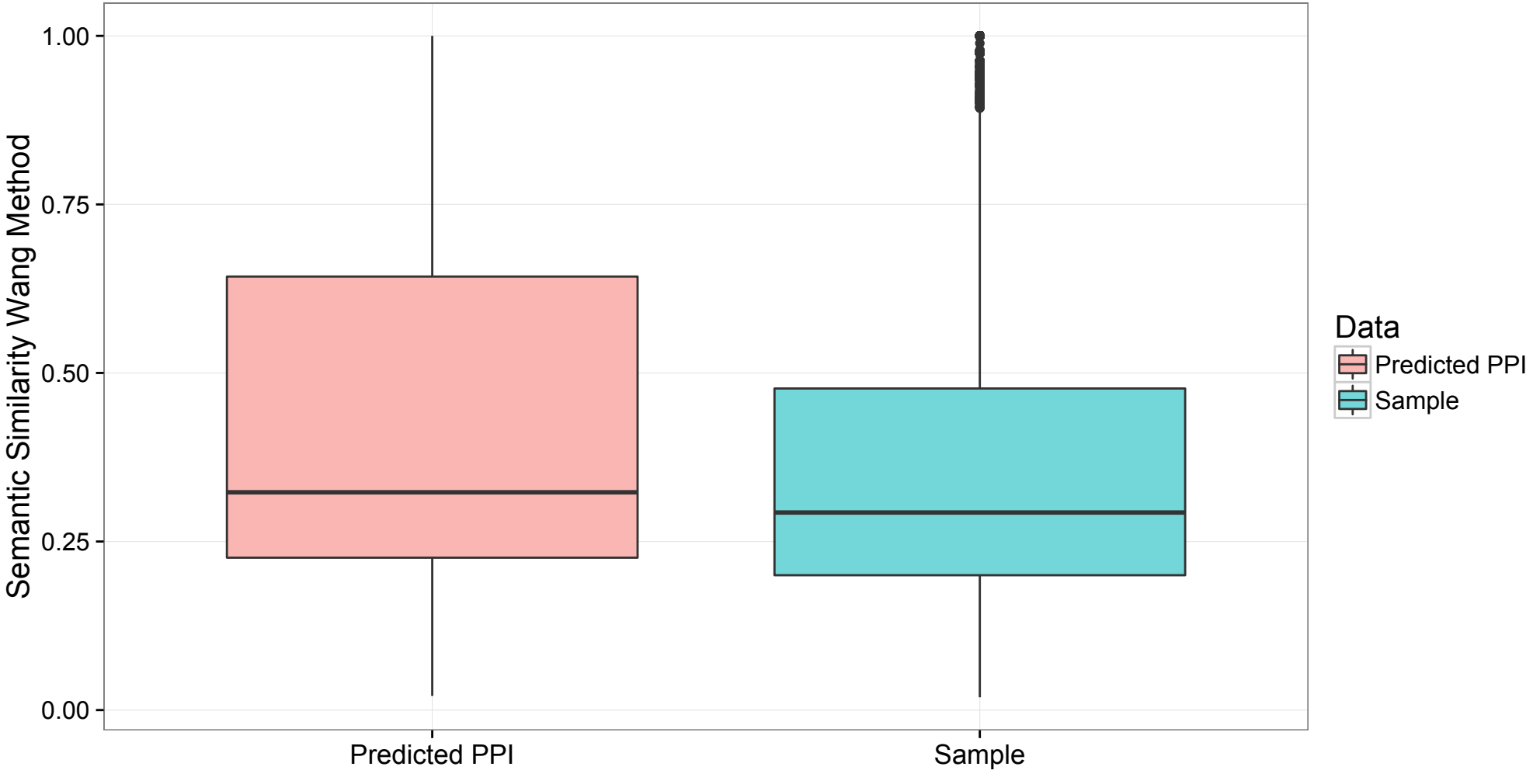

Chr5
Chr4
Chr3
Chr2
Chr1
Chr1
Chr2
Chr3
Chr4
Chr5
Log(Observed/Sampled)
6
3
0
-3

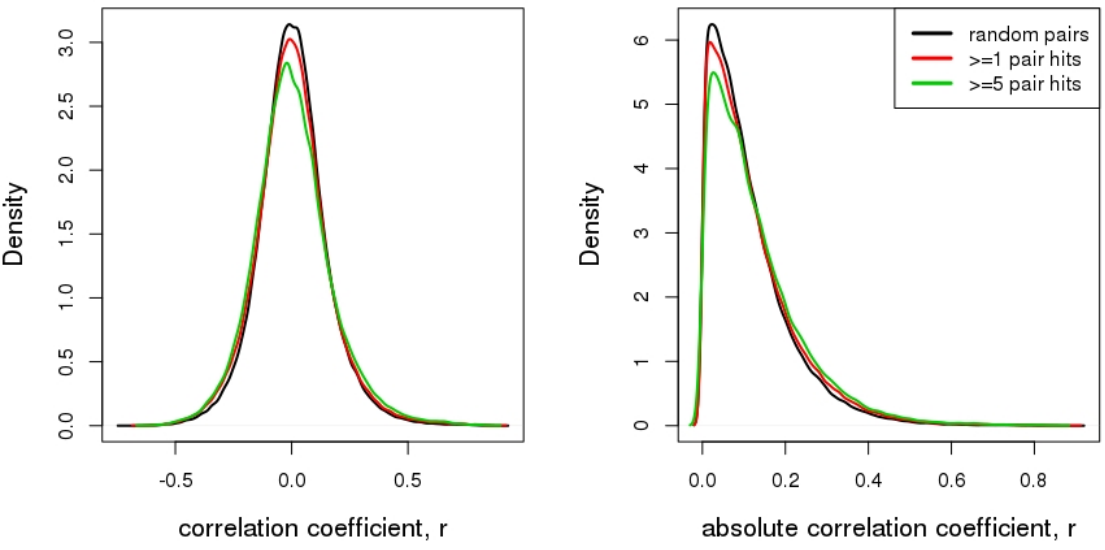

All Pair Hits
Sample
(A)
snoRNA::snoRNA
tRNA::tRNA
TE::TE
Pseudogene::TE
intergenic::TE
Pseudogene::Pseudogene
miRNA::miRNA
ncRNA::snoRNA
intergenic::Pseudogene
snoRNA::tRNA
intergenic::intergenic
ncRNA::TE
snRNA::tRNA
ncRNA::ncRNA
ncRNA::Pseudogene
Protein::tRNA
Protein::snRNA
miRNA::ncRNA
intergenic::ncRNA
Pseudogene::snoRNA
ncRNA::Protein
Protein::TE
intergenic::tRNA
Protein::Protein
Protein::snoRNA
ncRNA::tRNA
intergenic::Protein
Protein::Pseudogene
miRNA::Pseudogene
intergenic::miRNA
miRNA::Protein
miRNA::TE
intergenic::snoRNA
tRNA::TE
miRNA::tRNA
Pseudogene::tRNA
intergenic::snRNA
snoRNA::TE
miRNA::snRNA
snRNA::snRNA
miRNA::snoRNA
snRNA::TE
Pseudogene::snRNA
ncRNA::snRNA
snRNA::snoRNA
Wr=0.99
r=0.86
-3 -2.5 -2 -1.5 -1 -0.5 0 0.5 1 1.5 2
log(Observed/Expected)
All Pair Hits
Sample of subset Lk
Subset Lk
(B)
TE::TE
Pseudogene::TE
Pseudogene::Pseudogene
intergenic::TE
tRNA::tRNA
intergenic::Pseudogene
intergenic::intergenic
snoRNA::snoRNA
ncRNA::TE
intergenic::ncRNA
ncRNA::Pseudogene
Protein::tRNA
snoRNA::tRNA
intergenic::Protein
Protein::TE
Protein::Protein
Protein::Pseudogene
snRNA::snRNA
snRNA::tRNA
ncRNA::Protein
intergenic::miRNA
miRNA::tRNA
intergenic::tRNA
Protein::snRNA
miRNA::ncRNA
Protein::snoRNA
miRNA::TE
miRNA::Protein
ncRNA::ncRNA
miRNA::Pseudogene
miRNA::miRNA
ncRNA::tRNA
Pseudogene::tRNA
tRNA::TE
miRNA::snoRNA
intergenic::snoRNA
ncRNA::snoRNA
Pseudogene::snoRNA
miRNA::snRNA
snoRNA::TE
snRNA::snoRNA
intergenic::snRNA
ncRNA::snRNA
Pseudogene::snRNA
snRNA::TE
Wr=0.98
r=0.79
-3 -2.5 -2 -1.5 -1 -0.5 0 0.5 1 1.5 2
log(Observed/Expected)
All Pair Hits
Sample of subset Lk+Ln
Subset Lk + Ln
(C)
snoRNA::snoRNA
TE::TE
Intergenic::TE
Pseudogene::TE
Intergenic::Intergenic
Pseudogene::Pseudogene
snoRNA::tRNA
Intergenic::Pseudogene
tRNA::tRNA
Protein::snoRNA
Protein::tRNA
Protein::snRNA
Intergenic::Protein
Protein::Protein
ncRNA::ncRNA
miRNA::miRNA
Protein::TE
miRNA::tRNA
miRNA::snoRNA
Intergenic::ncRNA
miRNA::Protein
ncRNA::TE
Intergenic::snoRNA
Protein::Pseudogene
ncRNA::Protein
miRNA::ncRNA
Intergenic::miRNA
ncRNA::snoRNA
Intergenic::tRNA
ncRNA::Pseudogene
miRNA::Pseudogene
Pseudogene::tRNA
Pseudogene::snoRNA
miRNA::TE
ncRNA::tRNA
snoRNA::TE
tRNA::TE
snRNA::tRNA
Intergenic::snRNA
Pseudogene::snRNA
snRNA::TE
ncRNA::snRNA
Wr=97
r=0.71
-3 -2.5 -2 -1.5 -1 -0.5 0 0.5 1 1.5 2
Log(Observed/Expected)
All Pair Hits
Sample of subset Lk+Lk+Tr
Subset Lk + Ln + Tr
(D)
Protein::snoRNA
snoRNA::tRNA
miRNA::tRNA
miRNA::miRNA
tRNA::tRNA
Protein::tRNA
Protein::Protein
TE::TE
Intergenic::snoRNA
miRNA::Protein
miRNA::snoRNA
Intergenic::Intergenic
Intergenic::Protein
Intergenic::TE
Intergenic::tRNA
Intergenic::miRNA
Intergenic::Pseudogene
Pseudogene::TE
Protein::snRNA
ncRNA::ncRNA
ncRNA::snoRNA
Intergenic::ncRNA
Pseudogene::Pseudogene
Protein::Pseudogene
Pseudogene::snoRNA
Protein::TE
snRNA::tRNA
ncRNA::TE
Intergenic::snRNA
Pseudogene::tRNA
miRNA::Pseudogene
snoRNA::TE
ncRNA::Protein
tRNA::TE
ncRNA::Pseudogene
ncRNA::tRNA
miRNA::ncRNA
miRNA::TE
Pseudogene::snRNA
snRNA::TE
ncRNA::snRNA
Wr=0.29
r=0.16
-3 -2.5 -2 -1.5 -1 -0.5 0 0.5 1 1.5 2
Log(Observed/Expected)


Chr5
Chr4
Chr3
Chr2
Chr1
Chr1
Chr2
Chr3
Chr4
Chr5
Log(Observed/Sampled)
6
4
2
0
-2
-4
(A)


0
0.0002
0.0004
0.0006
0.0008
0.001
0.0012
0.0014
0.0016
Protein::snoRNA 5517
snoRNA::tRNA 39
miRNA::tRNA 127
miRNA::miRNA 20
tRNA::tRNA 140
Protein::tRNA 31369
Protein::Protein 1714225
TE::TE 1420990
intergenic::snoRNA 1516
miRNA::Protein 11176
miRNA::snoRNA 10
intergenic::intergenic 372764
intergenic::Protein 1409501
intergenic::TE 1270623
intergenic::tRNA 11176
intergenic::miRNA 3946
intergenic::Pseudogene 104911
Pseudogene::TE 190846
Protein::snRNA 410
ncRNA::ncRNA 696
ncRNA::snoRNA 53
intergenic::ncRNA 27031
Pseudogene::Pseudogene 8159
Protein::Pseudogene 174632
Pseudogene::snoRNA 155
Protein::TE 1688151
snRNA::tRNA 3
ncRNA::TE 44088
intergenic::snRNA 159
Pseudogene::tRNA 1249
miRNA::Pseudogene 471
snoRNA::TE 1211
ncRNA::Protein 38138
tRNA::TE 9563
ncRNA::Pseudogene 2789
ncRNA::tRNA 276
miRNA::ncRNA 104
miRNA::TE 3389
Pseudogene::snRNA 8
snRNA::TE 71
ncRNA::snRNA 1
(B)


0
-5
-10
-15
Log(AMI)
Nonsynonymous-Nonsynonymous Known PPI
Synonymous-Synonymous Known PPI

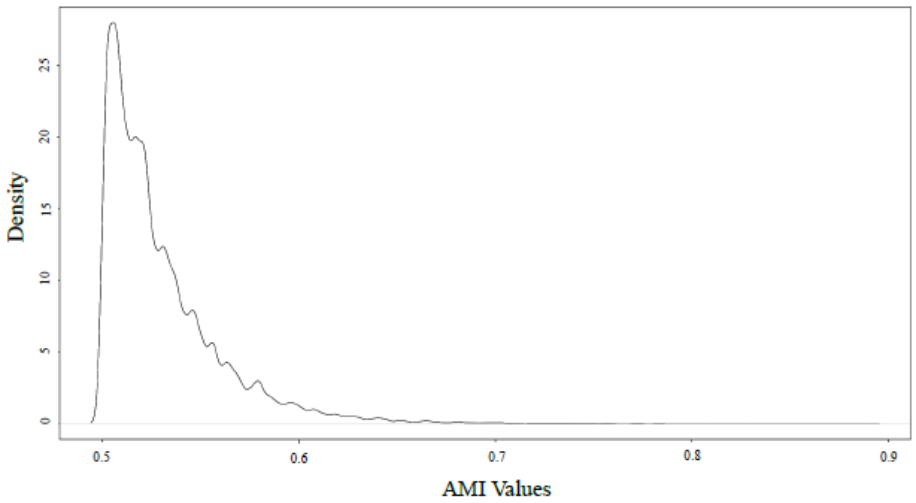

Semantic Similarity Wang Method
1.00
0.75
0.50
0.25
0.00
Predicted PPI
Sample
Data
Predicted PPI
Sample


Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
-0.5
0.0
0.5
correlation coefficient, r
Density
0
1
2
3
4
5
6
0.0
0.2
0.4
0.6
0.8
absolute correlation coefficient, r
random pairs
>=1 pair hits
>=5 pair hits

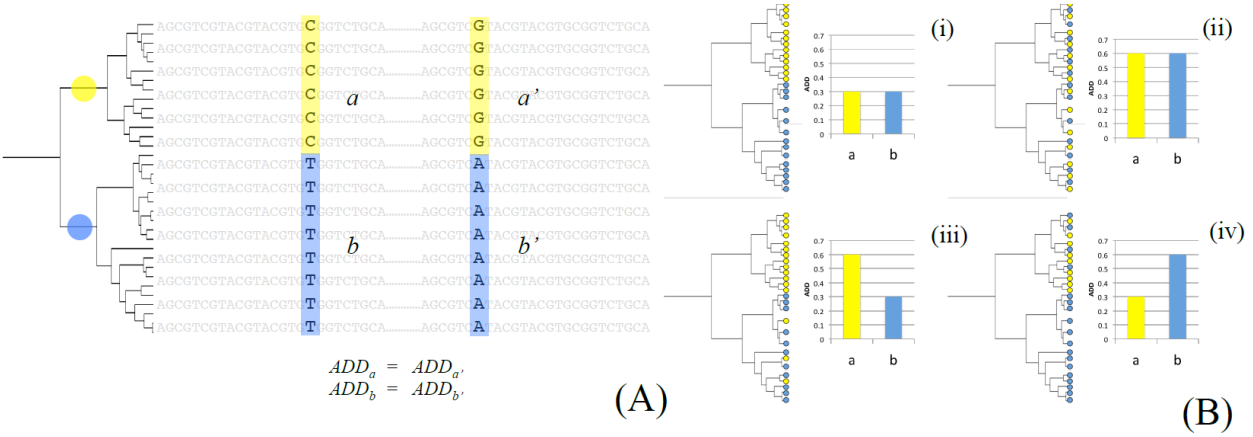
A genome-wide scan for correlated mutations detects macromolecular and chromatin loop interactions in *Arabidopsis thaliana*

Laura Perlaza-Jiménez & Dirk Walther

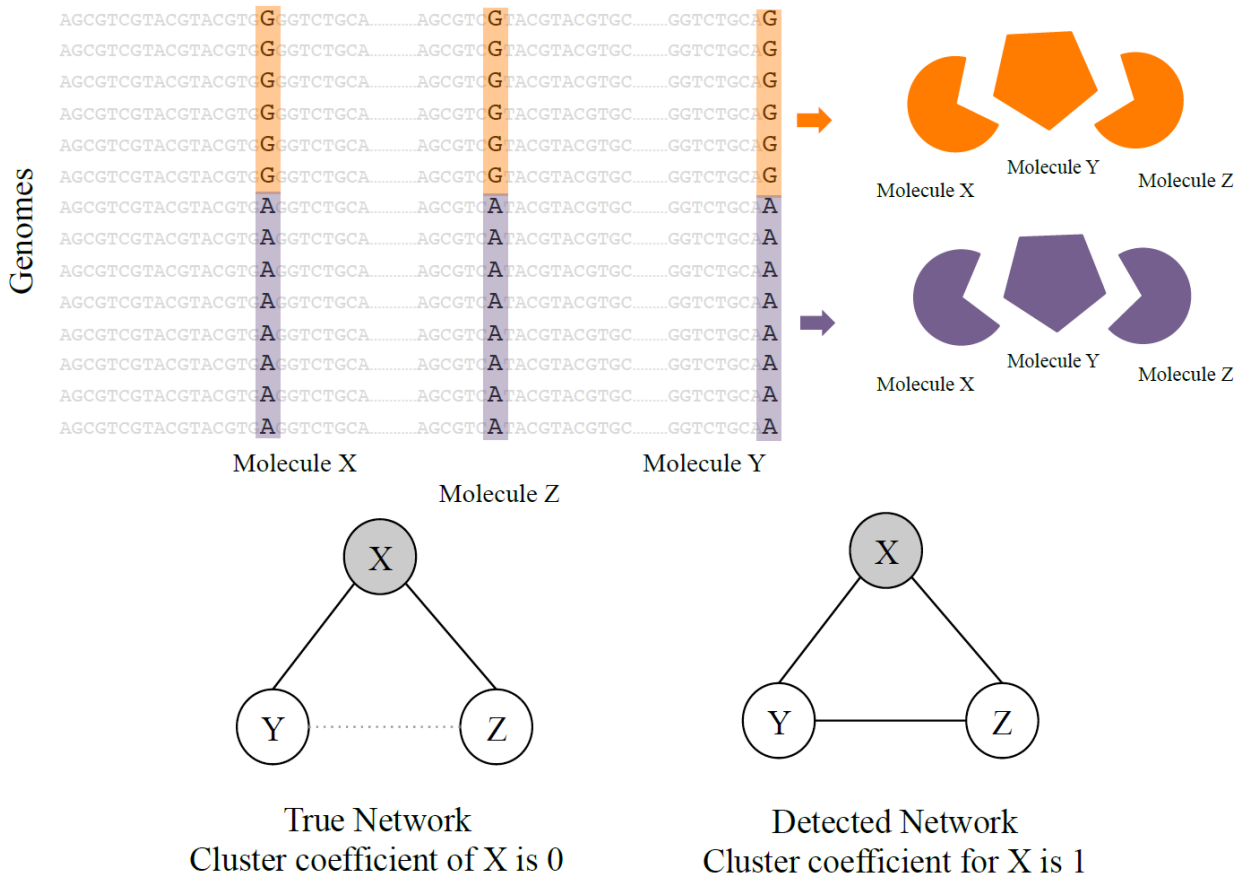
Supplementary Figures



Supplementary Figure 1. Density of the AMI values obtained by permutations of the SNP data.



Supplementary Figure 2. (A) The average distance diameter (*ADD)* is the variability determined in a cluster of accessions created by the alleles found at a polymorphic site and based on the pairwise sequence distances in a cluster. Both polymorphic sites in a pair hit share the values for the *ADD* for each cluster because they form the same clusters (ADD_a=$ADD_{a'}$). (B) Potential behavior of the ADD values and its corresponding phylogenetic representation. When the accessions (genomes) that compose a cluster at a polymorphic site are closely related, as is shown in cluster *a* and *b* (B(i)), the *ADD* is small. By contrast, when the accessions are not closely related, the *ADD* is large (B(ii-iv)).



Supplementary Figure 3. Illustration of transitivity (or cluster coefficient). In this example, the true functional network does not involve an interaction between molecules Y and Z. Their correlation is indirect via molecule X. However, based on the observed SNP patterns, AMI would falsely report Y and Z to be linked as well. Filtering for low transitivity values will reduce the number of such false correlations.