Article/Discoveries

The contribution of transposon insertion polymorphisms, structural variation and SNPs to the evolution of the melon genome

Walter Sanseverino[1†*], Elizabeth Hénaff[2!*], Cristina Vives[2], Sara Pinosio[3], William Burgos-Paz[2&], Michele Morgante[3], Sebastián E Ramos-Onsins[2#], Jordi Garcia-Mas[1#], Josep Maria Casacuberta[2#]

[1]Institut de Recerca i Tecnologia Agroalimentàries, Centre for Research in Agricultural Genomics CSIC-IRTA-UAB-UB, 08193 Barcelona, Spain

[2]Centre for Research in Agricultural Genomics CSIC-IRTA-UAB-UB, 08193 Barcelona, Spain

[3]Dipartimento di szience agrarie e ambientali, Università degli studi di Udine, Udine, Italy

[#]Correspondence: Jordi Garcia-Mas, jordi.garcia@irta.cat; Sebastián Ramos-Onsins, sebastian.ramos@cragenomica.es; Josep Casacuberta, josep.casacuberta@cragenomica.es.

[*]Both authors contributed equally to this work

[†]Present Address: Sequentia Biotech, Campus UAB, Edifici CRAG, Bellaterra, Cerdanyola del Vallès, 08193, Barcelona, Spain

[!]Present Address: Institute for Computational Biomedicine, Weill Cornell Medical College, New York, NY 11021, USA

[&]Present Address: Programa de Mejoramiento Genético, Universidad de Nariño. Ciudadela Universitaria Torobajo, Pasto, Colombia.

Abstract

The availability of extensive databases of crop genome sequences should allow analysis of crop variability at an unprecedented scale, which should have an important impact in plant breeding. However, up to now the analysis of genetic variability at the whole-genome scale has been mainly restricted to single nucleotide polymorphisms (SNPs). This is a strong limitation as structural variation and transposon insertion polymorphisms are frequent in plant species and have had an important mutational role in crop domestication and breeding. Here we present the first comprehensive analysis of melon genetic diversity, which includes a detailed analysis of SNPs, structural variation and transposon insertion polymorphisms.

The variability found among 7 melon varieties representing the species diversity and including wild accessions and highly-breed lines, is relatively high due in part to the marked divergence of some lineages. The diversity is distributed non-uniformly across the genome, being lower at the extremes of the chromosomes and higher in the pericentromeric regions, which is compatible with the effect of purifying selection and recombination forces over functional regions. Additionally, this variability is greatly reduced among elite varieties, probably due to selection during breeding. We have found some chromosomal regions showing a high differentiation of the elite varieties versus the rest, which could be considered as strongly-selected candidate regions. Our data also suggest that transposons and SV may be at the origin of an important fraction of the variability in melon, which highlights the importance of analyzing all types of genetic variability to understand crop genome evolution.

Introduction

Improving cultivars by breeding is essential to increase plant yield and ensure food security. While traditional breeding and marker-assisted breeding have been extremely successful in the past, the challenges agriculture has to face, which include the need to feed a growing human population, scarcity of land and water available for agriculture and climate change that will drastically modify growth conditions, impose an urgent need for improving these techniques (Godfray, et al. 2010). The availability of huge databases of crop genome sequences promises a new leap in plant breeding. However, in order to bridge the gap between genome sequences and trait improvement, there is a need to understand the links between genome variability and phenotypic variation. Re-sequencing crop varieties with interesting phenotypic traits to analyze their genomic variability promises to be an exceptional approach (Gebhardt 2013; Huang, et al. 2013; Myles 2013). The analysis of genome variability using re-sequencing data has been used to shed light on the domestication history of different crops including, e.g. maize (Hufford, et al. 2012; Jiao, et al. 2012), rice (Xu, et al. 2012), peach (Verde, et al. 2013), cucumber (Qi, et al. 2013), soybean (Lam, et al. 2010), watermelon (Guo, et al. 2013) and tomato (Lin, et al. 2014). Genomic regions with low genetic variability among domesticated varieties have been detected by these approaches, which likely highlight the genes that were selected for during domestication. In a similar way, the comparison of whole genome sequences of varieties with contrasting phenotypic traits can be used as a means to identify genes responsible for particular agronomic traits. Sequencing the genome of parental lines and high-resolution genotyping of recombinant inbred lines identified candidate genes for QTLs associated to the increased yield of hybrid rice varieties (Gao, et al. 2013). In summary, the analysis of genetic variability at the whole-genome level among varieties and cultivars should enable a new approach in plant breeding that will strengthen currently used strategies such as genome-wide association analyses (GWAS) as it has been recently proposed for rice (Huang, et al. 2013).

The vast majority of published studies dealing with genetic variability at the whole-genome level in plants have concentrated in single nucleotide polymorphisms (SNPs) as the main type of genetic variability (3KRGP 2014; Lin, et al. 2013; Qi, et al. 2013). However, genomes are rife with other types of variations, and many other types of sequence modifications are responsible for genetic variability relevant for plant genome evolution. Structural variation (SV), including copy number variation (CNV) and presence/absence variation (PAV), has been shown to be frequent in plant species (Saxena, et al. 2014). These SVs have traditionally been discovered using microarray-based methods, but the advent of next-generation sequencing technologies has made it possible to do so in a non-biased manner, although the methods to do so remain computationally challenging. A well analyzed example is maize, where two inbred lines may differ by more than 50% of the genome due in part to very frequent PAV of genes (Brunner, et al. 2005; Morgante, et al. 2007; Springer, et al. 2009). This variability in the genic component within individuals of the same species justifies the introduction of the concept of the pan-genome to refer to the ensemble of the core set of genes common to all individuals and the dispensable genome fraction specific to some of them. Interestingly, this high genome variability is enriched at loci associated with important traits (Chia, et al. 2012), and translates into transcriptome differences. Indeed, a recent transcriptome analysis of 503 maize inbred lines has shown that only the 16% of transcripts are present in all lines whereas the remaining 83% are expressed in subsets of the lines (Hirsch, et al. 2014). Therefore, this variability can be at the origin of phenotypic diversity of traits important for fitness and adaptation (Olsen and Wendel 2013; Hirsch, et al. 2014) and therefore of agricultural importance.

Transposons are at the origin of an important fraction of the CNV and PAV due to their capacity of mobilizing gene sequences within the genome (Morgante, et al. 2007). They can also contribute to genetic diversity in many other ways, the most important being the generation of transposon insertion polymorphisms. Indeed, transposon-related polymorphisms are at the origin of an important fraction of variability relevant for plant genome evolution both in the wild and in breeding processes (Lisch 2013; Olsen and Wendel 2013). For example, it has been shown that the critical increase in expression of the maize domestication gene *tb1* was the consequence of a transposon insertion in its promoter (Studer, et al. 2011). Similarly, accumulated evidence shows that transposon insertion polymorphisms are responsible for phenotypic variation in agronomically important traits such as the skin or flesh color of the orange, grape and peach (Kobayashi, et al. 2004; Butelli, et al. 2012; Falchi, et al. 2013). A recent survey of 60 genes related to plant domestication and breeding showed that 15% of them harbor transposable element insertions that have functional effects, which suggests that transposable elements have an important mutational role in domesticated plant genomes (Meyer and Purugganan 2013).

In addition to genetic variation, epigenetic variation has also been shown to be highly relevant for plant evolution and transposons can be major mediators of such variability (Lisch 2013; Pecinka, et al. 2013). Analyses of maize populations reveal that changes in DNA methylation are associated with changes in expression of some 300 genes, and that many of these differentially methylated regions are associated with transposons (Eichten, et al. 2013). Transposons are also at the origin of variations in the epigenetic state of genes responsible for important agronomic traits. For example, changes in sex determination in melon are due to the epigenetic silencing of a sex determination gene induced by an upstream transposon insertion (Martin, et al. 2009). It is thus highly relevant to study epigenetic variability in crops and to pay particular attention to transposon polymorphisms.

Melon (*Cucumis melo* L., 2*n*=2*x*=24) is an important vegetable crop of the Cucurbitaceae family that is highly appreciated for its fruit quality. It has been proposed that melon originated in Africa, although recent studies suggest a possible Asian origin (Sebastian, et al. 2010). It is a highly diverse species that has been classified in two subspecies, *melo* and *agrestis* (Jeffrey 1980), according to the pubescence of the female flower hypanthium although it has been shown that this classification does not completely agree with the molecular phylogeny (Stepansky, et al. 1999). Both subspecies have been further divided in several botanical groups, which include both edible and wild varieties (Pitrat 2008). Genetic diversity in melon has been studied using several types of molecular markers, such as restriction fragment length polymorphism (RFLP) (Silberstein, et al. 1999), random amplified polymorphic DNA (RAPD) (Stepansky, et al. 1999), amplified fragment length polymorphism (AFLP) (Garcia-Mas, et al. 2000), simple sequence repeat (SSR) (Monforte, et al. 2003) and SNPs (Esteras, et al. 2013). The reference genome sequence of melon is available for DHL92 (Garcia-Mas, et al. 2012), a double-haploid line derived from the cross between PI 161375 (SC) (*conomon* group, ssp. *agrestis*) and the 'Piel de sapo' line T111 (PS) (*inodorus* group, ssp. *melo*). The 375 Mb assembled melon genome contains 27,427 annotated genes, and 19.7 % of the sequence was shown to correspond to transposable elements (TEs) (Garcia-Ma, et al. 2012). However, this was a conservative annotation of the most recent TEs. A less-conservative annotation showed that up to 40% of melon genome is composed of TE-related sequences (unpublished).

Here we present the first analysis of melon genetic diversity at the whole-genome level using re-sequencing data from 7 melon accessions from diverse origins, which includes a detailed analysis of SNPs, structural variation (including PAV, CNV and inversions), and transposon insertion polymorphisms. To this end we used an array of already available and newly developed bioinformatic tools.

Results and discussion

SNP identification from re-sequence data

Three and four melon accessions of the ssp. *melo* and the ssp. *agrestis* subspecies, respectively, were selected as representative of the main melon groups (Supplementary Table S1). Some of these accessions are parental lines for several melon mapping populations which have been extensively used for constructing genetic maps and mapping agronomically important traits. The DHL92 line, which is the genotype of the published melon reference genome (Garcia-Mas, et al. 2012) was also included in the analysis as a control. The homozygous DHL92 double-haploid line is derived from the cross between two varieties also included in this study, PI 161375 (Songwhan Charmi, spp. *agrestis*) (SC) and the "Piel de Sapo" T111 line (ssp. *melo*) (PS).

The seven melon varieties were re-sequenced using a paired-end approach, with libraries of 500 bp fragment length and 150 bp reads sequenced to an average of 19.45x depth and 80,3 % breadth coverage of the assembled genome for each line (Supplementary Table S2). In total we produced 273 million paired-end reads (63.9 Gb), which were mapped to the reference genome DHL92 v3.5 (Garcia-Mas, et al. 2012) and variants were called using the SUPERW pipeline (Supplementary Figure S1). A total of 4,556,377 SNPs and 718,832 short deletion and insertion polymorphisms (DIPs < 200 bp) were identified (Table 1). A high proportion of SNPs between PS and SC have been validated using the GoldenGate and Fluidigm platforms in the context of other projects (Argyris, et al. 2014).

Nucleotide diversity at the whole genome level

We used a total of 4,391,835 SNPs detected from 254,721,076 aligned positions, after excluding SNP positions with missing data in any of the lines, to perform a global variability analysis. Global nucleotide diversity (π_{tot}=0.0066) was among the highest reported in crop species (Qi, et al. 2013). Within the melon varieties analyzed, the improved lines (Elite) were the ones with the lowest diversity (π_{total_Elite}=0.0035), the cultivated landraces had an intermediate value ($\pi_{total\_Landrace}$=0.0052) and the wild ecotypes the most diverse (π_{total_Wild}=0.0094), which is concordant with comparisons of cultivated and wild varieties in other species (Qi, et al. 2013). Over the complete set of samples, the whole genome synonymous diversity (π_{syn}=0.0055) was lower than the total silent diversity (π_{sil}=0.0074). Although the variability at silent positions is assumed to be constrained on regulatory regions and on unknown functional positions (Mackay, et al. 2012), those are only a small fraction of the silent positions. In addition, synonymous positions can also be constrained due, for example, to codon usage preference (Ingvarsson, 2010). Non-synonymous diversity (π_{nsyn}=0.0015) was lower (4-fold) than the diversity at synonymous positions, as expected. The nucleotide diversity was distributed non-uniformly across the genome, being generally lower at the extremes of the chromosomes (towards the telomeres) and higher in the pericentromeric regions (Figure 1A).

The wild accessions Cabo Verde (CV) and Trigonus (TRI) showed the highest number of unique SNPs and DIPs (Table 1). A pairwise comparison between varieties showed a clear pattern of differentiation of the CV line versus the other six lines, CV having many more differences versus the rest of the lines that any other pairwise combination (Supplementary Table S3). This suggests that CV may have been isolated from the rest during a long period. Principal Component analysis also showed CV clearly separated from the rest in the first principal component, while TRI separated from the remaining five lines with the second principal component (Supplementary Figure S2).

Nucleotide divergence in melon lines compared to cucumber and melon population structure

Melon and cucumber lineages split 10.1 million years ago (Sebastian, et al. 2010), and cucumber is the closest relative to melon with an available genome sequence. For this reason, we used re-sequencing data from a breeding cucumber line, mapped to the melon reference, to analyse the divergence of melon versus cucumber (*Cucumis sativus* L.). This analysis was restricted to positions found within regions that can be aligned between the two species, totalling to 117,514,001 positions (46% of the total number of positions used for the variability analyses within the melon varieties), and detected 1,264,489 SNPs. The level of diversity per nucleotide for all the categories detected was relatively low (π_{tot}= 0.0041, π_{sil} = 0.0049, π_{syn} = 0.0045, π_{nsyn} = 0.0010), which was as expected as only conserved regions are included. The analysis showed a clear differentiation (the divergence per position corrected by HKY is K=0.034), indicating that it is suitable as outgroup, and can be used for polarizing melon ancestry from the derived variants.

The genealogical representation using the matrix of pairwise distances for the whole genome using cucumber as an outgroup (Figure 2A) shows that the ssp. *melo* forms a monophyletic group, while the *agrestis* group is a heterogeneous group, with Calcuta (CAL) closer to the *melo* clade and CV being the more divergent line. The topology shown is supported by an statistical analysis shown in Supplementary Figure S3. A recent study of 93 melon accessions, including 75 *melo* and 18 *agrestis* types, showed a strong population structure with different levels of admixture depending on the melon type, and identified two *melo* (*inodorus* and *cantalupensis*) and three *agrestis* (Indian *momordica*, African *agrestis* and Far-eastern *conomon*) subpopulations (Esteras, et al. 2013). Five of the 7 melon accessions used here were included in this study, and CAL was also located closer to the *melo* accessions than the rest of the *agrestis* accessions. A per chromosome phylogenetic reconstruction (Supplementary Figure S4) showed relatively consistent nodes among all chromosomes, being PS-VED, PS-VED-IRK-CAL and TRI-SC common groups across the genome. Note that the PS-VED-IRK node (ssp. *melon*), although common, is usually mixed with CAL lineages, suggesting a recent possible introgression or close recent ancestors. In an attempt to analyse in more detail the general history of these lineages, we calculated Neighbor-Joining trees using windows of 10Kb and 100Kb across the whole genome. We observed 5,480 different topologies when using windows of 10Kb (1,185 in windows of 100Kb) across the whole genome and from a total of 30,812 (3,169) windows (Supplementary Figure S4). Up to 120 (117 in 100Kb bins) different nodes (clusters of lineages) were observed. Note that the maximum number of topologies for 7 lineages using a rooted tree is 10,395 (Felsenstein, 1978). Although the percentage of support of each node at each subset tree seems quite low (no more than 30% in the best case), the tree based on the whole genome is highly robust because its branches are the most frequently observed (TRN=3, see the definition of this indicator in Materials and Methods) and their bootstrap support high. The large number of observed window topologies (gene trees) is expected for lineages from the same species because recombination frequently happened along the history of the species. On the other hand, the number of topologies differs significantly from being obtained by chance (Pvalue << 1e-6), indicating a consistent population structure in the melon species.

The role of selection at amino acid positions

Several plant species as *Helianthus annuus* (sunflower), *Populus tremula* or *Capsella grandiflora* show evidence of positive selection at non-synonymous positions (Fay

2011) while this was not the case in many others (e.g. crop species such as *Zea mays, Sorghum bicolor* and *Oryza sativa*). In order to test whether the melon genome is evolving under selection we performed a MacDonald-Kreitman test (MKT, (McDonald and Kreitman 1991)) that tests whether the ratio between variability and divergence is different over synonymous and non-synonymous positions by means of a Fisher exact test. We detected a very significant result of MKT (MKT=36.3, Pvalue=1.7E-9), suggesting that, indeed, the melon genome as a whole has evolved under selection. We calculated the proportion of non-synonymous positions affected by positive selection as $\alpha = 1 - (K_{syn}\ \pi_{nsyn}) / (\pi_{syn}\ K_{nsyn})$. The value of α was estimated as -0.24, indicating that negative selection predominates in the evolution at non-synonymous positions (Supplementary Table S4). α was very negative when calculated for only polymorphic singletons (α=-0.31), indicating that many of the low frequency amino-acid variants were negatively selected, but was also negative for the higher variant frequencies (α=-0.19 for variants at the highest frequency, that is, derived mutation at six samples (Messer and Petrov 2013)). This indicates a global negative effect of selection on non-synonymous positions throughout the history of melon.

The role of selection considering the patterns of association of genomic features with polymorphism and divergence

We sought to determine whether the evolution of the melon genome has been mainly subjected to positive, neutral or background selection. For this analysis, we assumed that genomic features such as recombination and gene density have structural patterns associated to the whole species and therefore, that there are not important differences in the gene density distribution neither in the recombination levels between different lineages nor between populations. Moreover, the distribution of the variability across the genome over different groups is rather similar, so we considered studying jointly the species sample. In melon, the recombination rate estimates (Argyris, et al. 2014) and the codon density content show an accused non-uniform distribution, as they concentrate in the distal parts of the chromosomes (Figure 1B-1C). We analyzed the distribution of silent (synonymous and non-coding) variability relative to coding region densities and to recombination rate and found that it is negatively associated with codon density (partial correlation R=-0.37, Pvalue << 1E-15) but not with the recombination rate (partial correlation R=-0.02, NS, Figure 3). This pattern does not fit with a strict neutral model of evolution, which predicts no association with these two features. This could be due to a reduction of variability in regions rich in coding regions (selectively constricted at non-synonymous positions) or a mutational bias, where high numbers of mutations were located at low codon density regions. We did not detect association of variability with the percentage of GCs when considering codon density (Supplementary Figure S5), suggesting that a mutation bias for this dinucleotide is not responsible for the patterns of variability observed. On the other hand, we did not observe a negative association of synonymous polymorphisms with non-synonymous divergence that would be expected for evolution under positive selection (recurrent selective sweeps) (Figure 4). Our data suggest an evolution under background selection. However, the expected positive correlation of neutral (silent) variability with recombination under this model is not observed (Supplementary 3B). Negative or no association (as it is our case) of neutral variability with recombination rate has already been described in other plant genomes and has been explained by a non-uniform distribution of coding regions that would generate a non-uniform distribution of selective mutations (Flowers, et al. 2012).

The distribution of nucleotide divergence along the genome is non-uniform, being lower in pericentromeric regions and higher towards the telomeres (Supplementary Figure S6). The neutral model also predicts a positive association between variability and divergence. Our results show that there is a negative association of silent polymorphism with divergence (Kendall tau = -0.33, Supplementary Figure S7, see also Figure 1A and S6A). This association is also negative and very significant when considering codon density and recombination (Supplementary Figure S8A), and in contrast to the patterns of polymorphism versus divergence at synonymous (Kendall tau = +0.13) and non-synonymous (Kendall tau = +0.26) positions (Supplementary Figure S8B and C), suggesting that silent positions may not behave as neutral. This pattern is unusual, in rice a positive or null association between polymorphism and divergence was found (Flowers, et al. 2012). On the other hand, in regions of high coding density or near genes we observed no association of polymorphism with divergence (Supplementary Figure S9), which is similar to what has been previously reported in rice (Flowers, et al. 2012), where they used silent regions close to genes. The intriguing silent divergence pattern observed may be due (total or partially) to the difficulty to align regions with low gene density. The ratio of unaligned positions with the outgroup is highest in the pericentromeric regions (~80%) and lowest at rest of chromosomal arms (~40%).

Although this difference can be explained by the rapid expansion of TE elements across the melon genome after speciation (mostly at pericentromeric regions, see the following sections), only the 62 % of the cucumber genome was successfully aligned with melon, indicating that only the vicinity of the highest conserved regions were taken into account for this comparison.

Regions of interest for their singular patterns of variability and differentiation

As a first step towards identifying the regions of the melon genome that have been selected during its recent evolution we analysed the patterns of variability and divergence between different sets of varieties, including elite (VED and PS) versus NO-Elite (SC, TRI, CAL, IRK and CV) and *melo* (PS, VED, IRK) versus *agrestis* (CAL, SC, TRI, CV) subspecies.

We observed high divergence across the whole genome when comparing *melo* versus *agrestis* subspecies. High diversity at the *melo* group was located at chromosome 7 but also coincident with a peak in *agrestis* (Supplementary Figure S10). The comparison of elite (VED and PS) versus non-Elite varieties (SC, TRI, CAL, Irak - IRK-, and CV) shows that the total variability (π) is extremely reduced within the elite group in chromosomes 1 and 6 (Supplementary Figure S11) but very high at others (chromosomes 3 and 8). Of particular interest is a region in chromosome 1 showing almost no variability among the elite varieties while being high among the non-elite varieties (Supplementary Figure S11).

An analysis of the Fixation Index *Fst* (Hudson, et al. 1992) per chromosome showed clear differences when comparing the varieties belonging to the *melo* subspecies to the *agrestis* ones, and also when comparing the elite highly inbred lines to the rest (NO-Elite) (Supplementary Figure S12, Supplementary Table S5). The highest difference was found between the elite varieties and NO-Elite (mean Fst=0.191). These two elite lines (PS and VED) are the closest lines among those analyzed and therefore have the lowest variability (Supplementary Figure S11A) (π_{Elite}= 0.0035, $\pi_{\text{NO-Elite}}$=0.0070). The comparison of the *melo* and *agrestis* groups also showed that the melo group is more homogeneous (π_{melo}= 0.0041, π_{agrestis}= 0.0076).

In order to look for particular regions of differentiation, we also analysed the variability in in 500 Kb windows across the genome. The level of variability depends on the chromosome location, being generally higher at pericentromeric regions. In contrast, *Fst* is midly associated to codon density and it is not associated to recombination (Supplementary Figure S13), although heteroscedasticity is observed (that is, heterogeneity in the variance across the codon density parameter). In order to choose the more significant values we used a variable threshold value in relation to codon density (traced assuming two times the standard deviation calculated in bins of ten fragments, which is 99% in a Normal distribution), because their variance is different across this variable (Supplementary Figure S14). *Fst* values above the threshold curve may be associated to regions involved in the differentiation between these two groups. In the comparisons of *melo* versus *agrestis* subspecies, the more extreme population differentiation regions (Supplementary Figure S14B) were located at chromosomes 1 (positions 12e6 to 12.5e6, 19e6 to 20e6, 28.5e6 to 29.5e6), 3 (positions 21e6 to 21.5e6), 6 (positions 13e6 to 13.5e6 and from 17.5 to 18e6), 7 (positions 13e6 to 13.5e6), 8 (positions 1 to 0.5e6) and 11 (positions 11 to 11.5e6 and 21.5 to 22e6). For the comparison among elite and non-elite varieties, the outlier windows for *Fst*-index considering codon density are located at chromosome 1 (positions 3e6 to 3.5e6), 2 (positions 24e6 to 24.5e6), 3 (positions 21e6 to 21.5 e6), and 6 (positions 25e6 to 26e6).

Non-transposon structural variation

The most common methods used to discover SVs are based either on discordant mapping signatures of paired reads or by variations in read-depth (Alkan, et al. 2011). We developed an *in silico* workflow that implements these two approaches to identify SV in the seven melon lines with respect to the reference, and combined, have identified 3,609 SVs. SVs were classified as deletions or presence-absence variations PAV (n=2,541), inversions (n=620) and duplications or CNV (n=448) (Supplementary Table S6). The number and types of SV found are in general consistent with what has been reported in other plant species (Saxena, et al. 2014), although the use of different methods and parameters in different studies make them difficult to compare. A total of 902 genes are affected by a SV in at least one variety (Supplementary Table S7). Of these, 745 fell in deletions, 142 in tandem duplications and 15 in inverted regions. According to the public available melon Gene Ontology (GO) functional annotation (Garcia-Mas, et al. 2012), refined in this paper using Annotation with Human Readable Description (AHRD), fifty-three genes were related to agronomically relevant pathways, including disease resistance (29), cell-wall metabolism (10), aroma volatiles metabolism (9), sugar metabolism (4) and carotenoid biosynthesis (1).

The five largest deletions were found to range between 82 and 416 kb long (Supplementary Table S8). One of these large deletions is located in chromosome 5 and spans 146 kb in CV and 82 kb in SC, affecting 6 resistance genes (R-genes) of the NBS-LRR class (MELO3C04318-4324). This deletion is found in a 1,1 Mb region of chromosome 5 where the largest R-gene cluster in the melon genome is located, containing 23 R-genes (Garcia-Mas, et al. 2012; González, et al. 2014). The same deletion in CV and SC was also described in a recent presence/absence gene variability study using a subset of our melon lines (CV, SC, PS and IRK) and a different discovery pipeline (González, et al. 2013). Additional R-gene clusters are affected by SV, namely the MELO3C004289-4295 interval in the vicinity of the above-mentioned deletion in chromosome 5 (Supplementary Table S7, in yellow) and another in chromosome 1 (MELO3C023566-23578) (Supplementary Table S7, in yellow). A similarly high variability in resistance gene clusters has been reported recently in soybean using array hybridization and targeted re-sequencing (McHale, et al. 2012). Four out of the five large deletions described in Supplementary Table S8 were also detected in González et al (2013), confirming the accuracy of our SV discovery pipeline.

Transposon insertion polymorphisms

Mobile elements are an important source of the variability necessary for evolution (Lisch 2013; Olsen and Wendel 2013). In order to analyze the contribution of transposon insertions to the genotypic variability in melon, we first refined the transposon annotation we previously performed (Garcia-Mas, et al. 2012) with an additional search for Miniature Inverted Terminal-repeat Elements (MITEs) using Subotir (Henaff, et al. 2014) and MITE-Hunter (Han and Wessler 2010). MITEs are a particular type of transposons abundant and active in plant genomes and therefore it was important to include them in the annotation (Casacuberta and Santiago 2003). The previously performed annotation included MITEs related to other annotated class II elements (these are included within the different class II TEs superfamilies, see Table 2); with this additional search the annotation also includes MITEs non-related to the previously detected families of class II elements (these MITEs are classified as “other

MITEs", see Table 2). We used a combination of publicly available programs and tools newly developed in our laboratory to identify transposon deletions and insertions in the re-sequenced melon varieties with respect to the reference genome. Tools are routinely compared to others in the context of methods papers, sometimes with simulated data (Layer, et al. 2014), or real datasets that include gold standard variants from projects such as the 1000GP or GIAB (Rausch, et al. 2012) but the performance of these depends greatly on the sequencing depth and complexity of the reference. To date there is no reference dataset of gold standard variants in plants, and the 1000GP data used to benchmark the gold standard calls is much lower coverage and shorter reads than our dataset. In order to evaluate the performance of the programs to be used on our specific genome and with our sequencing characteristics, we took advantage of the unique possibility offered by the fact that we have resequencing data for the same line that was used to generate the reference sequence. We generated a simulated reference genome by deleting 1,871 transposons from the assembled DHL92 reference genome and inserting them in randomly chosen locations (see Supplementary material). These deletions and insertions can be used for benchmarking as they should be detected as insertions and deletions, respectively, in the same DHL92 line mapped to the modified reference, and is the most accurate measure of sensitivity and specificity as we model the complexity of the reference genome and the characteristic of our sequencing datasets. We evaluated BreakDancer (Chen, et al. 2009) and Pindel (Ye, et al. 2009) in their ability to detect deletions with respect to the reference. In our hands (see material and method section for details), Breakdancer has a sensitivity of 79 % and a positive predictive value (PPV) of less than 64 %. Pindel shows comparable sensitivity (76 %) yet higher PPV (85 %) (Supplementary Table S9). Therefore we chose to use Pindel to detect TE deletions with respect to the reference genome.

To detect insertions in the re-sequenced genomes with respect to the reference, we used a program recently developed in our laboratory which relies on discordant and soft-clipped read signatures to predict TE insertion loci, named Jitterbug (Hénaff et al, submitted). Using the previously described simulation, we turned the parameters to achieve 82.71 % sensitivity and 98.68 % PPV with our dataset. We then ran Jitterbug on the data for all 7 lines, and identified a total of 2,688 insertions, consisting in 2,056 polymorphic loci (corresponding to an insertion at the same locus in one or more lines). In order to confirm this high sensitivity and specificity values we analyzed by PCR 23 of the predicted polymorphic loci amongst the 7 lines, with primer sets designed to detect both the TE insertion and the reference allele (or empty locus). All the 23 polymorphic loci were confirmed by PCR (Supplementary Figure S15). However, while in 20 cases the genotype predicted for the 7 varieties was confirmed by PCR, in one case we detected the empty site for one of the varieties that was supposed to contain the insertion and in 3 additional cases we failed to amplify the region in some varieties that were predicted to not contain the insertion (Supplementary Figure S15). These discrepancies may be due to a lack of fixation of the insertion within the population, as the DNA used for sequencing and PCR analysis came from individuals different from those used for re-sequencing experiments. Indeed, although the elite PS and VED varieties are highly inbred and show a very high degree of homozygosity, the wild varieties and landraces may show a higher degree of heterozygosity and TE insertions may segregate in the population. A clear example is CAL which is heterozygous for the insertion in at least 4 of the polymorphic sites surveyed by PCR. The lack of amplification in PS of the empty site corresponding to the CM_4552 locus is probably the result of an insertion in PS that was not predicted by Jitterbug due to a low resequencing coverage in this variety at this particular location (not shown).

Using Pindel and Jitterbug to call deletions and insertions respectively in the re-sequenced varieties with respect to the reference we detected 2,735 polymorphic TE insertions. Two thirds of these polymorphisms were caused by retrotransposons, DNA transposons insertion/deletions roughly accounting for the remaining 1/3 (we were not able to categorize 3.3 % of the polymorphic sites due to their complex nature) (Table 2). Among retrotransposon-related polymorphisms, most of them were contributed by *copia* (23 %) and *gypsy* (24 %) elements, while most of the DNA transposon-related polymorphisms were contributed by MITEs (Table 2). A small number of TE families were responsible for the large part of the observed polymorphisms. Indeed, 9 families, which represent less than 4 % of the TE copies annotated in the melon genome, account for more than 1/3 of the polymorphic TE insertions (Table 3). Judging from their sequence similarity, these families contain relatively young elements (not shown), which is consistent with their recent activity during melon domestication and breeding. It is interesting to note that as much as 60 % of the polymorphic TEs are present in only one variety, which is also consistent with a recent TE activity (Supplementary Table S10). We used the TE insertions shared by more than one variety to construct a dendrogram of the phylogenetic relationships of the 7 melon varieties. We used a NJ approach to obtain a dendrogram as the non-constant evolutionary rate of TEs is not consistent with the UPGMA approach. Indeed, it is generally accepted that the activity of TEs, and in particular that of LTR-retrotransposons and MITEs, is not constant over time, with burst of transposition being followed by periods of relatively low activity (Lu, et al. 2012; El Baidouri and Panaud 2013). The dendrogram obtained with the TE data shows a pattern consistent with the dendrogram based on SNPs, although the relative length of the branches are very different (much longer at external nodes), possibly by the faster and non-constant rate of evolution of TEs (Figure 2B).

The annotation and analysis of the transposons present in the melon reference genome, and the comparison of these data with the transposon content in cucumber, showed that transposons have been very active during melon recent evolution (Garcia-Mas, et al. 2012), transposing and amplifying to a greater extent in melon compared to cucumber. The results presented here confirm the recent transposon activity in melon genome and suggest that transposons may be at the origin of an important fraction of the variability in this species.

Transposon density usually shows a non-random distribution along plant chromosomes, with TEs concentrating in the pericentromeric regions where gene density is lower (Arabidopsis Genome 2000; International Rice Genome Sequencing 2005; Paterson, et al. 2009; Schnable, et al. 2009; Schmutz, et al. 2010; Tian, et al. 2012; Tomato Genome Consortium 2012; Choulet, et al. 2014). In plants these regions frequently show a low recombination rate and it has been proposed that this may also explain the higher concentration of TEs in these regions as they may be more difficult to eliminate by selection (Gaut, et al. 2007). The distribution of fixed (annotated in the reference and non-polymorphic) TEs in melon is non-random, with higher density in large regions flanking the centromeres (Garcia-Mas, et al. 2012) (Figure 1D). We confirmed that fixed TEs show an inversely correlated distribution with respect to gene density as well as to recombination rate, and we tested which of these two features influences TE distribution. Our results show that the density of fixed TEs is strongly negatively associated with codon density (as seen in partial correlations, Supplementary Figure S16) indicating an accumulation of fixed TEs in non-functional regions. On the contrary, the frequency of fixed TEs shows no association with recombination rate (Supplementary Figure S16).

The frequency of polymorphic TEs across the genome is more homogeneously distributed (Supplementary Figures S17-S18) than that of SNPs. However, when using the theta estimate of Zeng (Zeng et al. 2006), which weights more the high-frequency variants, it can be seen that there is a slight bias to pericentromeric regions (Supplementary Figures S17-S18). This pattern suggests the effect of selection on TE distribution, possibly eliminating those elements that affect functional regions and allowing their fixation in regions with less functional constraints. Alternatively, this could also be the consequence of an insertion preference within pericentromeric regions of some TE families. Indeed, it has been previously shown that some TE families, in particular among retrotransposons, target pericentromeric regions for insertion and accumulate almost exclusively within these regions (Peterson-Burch, et al. 2004; Du, et al. 2010; Sharma and Presting 2014).

We found generally lower TE diversity in elite varieties when compared with the rest (Supplementary Figure S11B). Indeed, there are 613 polymorphic TE insertions between VED and PS, whereas there are 2,092 polymorphic sites among non-elite varieties, and there is no combination of varieties showing a level of polymorphisms comparable or smaller than that of the two elite varieties (p=0). This low TE diversity is particularly clear in some chromosomes and chromosomal regions such as portions of chromosomes 1 and 6. For example, a 11 Mb region (positions $5x10^6$ to $16x10^6$) in chromosome 1 shows no TE polymorphisms between elite varieties and 99 TE between the non-elite. This number is significantly lower (p=0) than any other region for the elite varieties, and not significantly different (p=0.99) for the non-elite varieties. Interestingly these regions also show very low nucleotide diversity (Supplementary Figure S11A) which suggests that these regions may have been fixed during breeding. An analysis of the 306 genes found in this region that have an associated GO term (of the 532 genes present in this region) shows enrichment in GO terms related to cellulose biosynthesis (p value=2.2×10^{-5}). Although cellulose synthesis and breakdown may be related with fruit softening and this is an important trait for melon breeding, more work is needed to evaluate the biological significance of this finding.

The analysis of the *Fst* index (Supplementary Figure S19) measuring differences between elite and NO-Elite showed several peaks of high differentiation, also suggesting that some TE insertions may have been fixed during the breeding process of these two elite lines.

In general, there is an important fraction of the polymorphic TE insertions located in genic regions, suggesting a potential impact on genes. Indeed, more than 22% of the 2,735 polymorphic TEs are located within an annotated gene, and an additional 7.8% are located within 500 nt of a gene. Among TEs located within genes, 361 are within exons and 250 in introns and untranslated regions (Table 4). This dataset should allow in the future to evaluate the impact of transposons on the evolution of the melon genome. Of particular interest is the fact that we identified 165 TE insertions in coding regions that are polymorphic between the two elite lines VED and PS. These two elite lines are closely related phylogenetically (Figure 2A), but still they show important differences in key agronomical traits such as fruit shape, flesh color, ripening behavior, sugar content, and aromas. The possibility that one of these TE insertions within genes may have altered one of these characteristics is highly likely. In fact, an analysis of the genes showing transposon insertion polymorphisms shows that an important fraction of them are related to the development of reproductive structures, hormone signalling and sugar metabolism, suggesting that transposon insertions may have modified some of the metabolic pathways or regulatory networks that underlie these important agronomic traits.

Conclusions

We present here a pioneer work in plants consisting of a comprehensive analysis of variability in a crop species, from SNPs to large SV, including transposons insertion polymorphisms, and which includes new tools and bioinformatic pipelines to integrate these analyses. Our benchmarking of these algorithms using the resequence of the reference genome is a novel approach and ensures an accurate estimation of the specificity and sensitivity of the results for our specific dataset. In particular, our assessment of Jitterbug shows that it is a very specific new tool for TE insertion polymorphisms identification.

The variability found among 7 melon varieties that represent the extant diversity of the species and include wild accessions and breeding lines, is relatively high due in part to the structure of the species. The nucleotide diversity is distributed non-uniformly across the genome, being generally lower at the extremes of the chromosomes, coinciding with gene-rich and high-recombination regions, and higher in the pericentromeric regions, where gene density and recombination rate are low and there is a higher accumulation of TEs. However, this variability is greatly reduced among elite varieties, probably due to the selection during breeding. We have found some chromosomal regions that show a high differentiation of the elite varieties versus the rest of the varieties analyzed, which could be considered as regions that suffered strong selection. Interestingly, some of these regions also show a high differentiation between elite and non-elite varieties with respect to polymorphic TE insertions suggesting that these regions may have been fixed during breeding. Our data also suggest that transposons may be at the origin of an important fraction of the variability in melon, in addition to the variation due to SNPs and SVs. As much as 60 % of the polymorphic TEs are present in only one variety, suggesting that there have been an important transposon activity very recently during melon evolution.

Additionally, a total of 902 genes were shown to be affected by a SV in at least one variety, with a significant enrichment in regions harbouring R-gene clusters. We found that the largest R-gene cluster in the melon genome, located in chromosome 5 and comprising 23 genes, has been partially deleted in some of the accessions.

As re-sequencing costs decrease, the analysis of large data sets of varieties that represent the extant variability of crop species is becoming feasible. However, up to now these analyses have been restricted to SNPs and, in few cases to large SV. The approach presented here describes a comprehensive analysis of variability including SNPs, SVs and also TE insertion polymorphisms, and implements the in-depth variability analysis that can be used to detect genomic regions involved in domestication and selection when the re-sequence of a wide collection of melon germplasm is available.

Materials and Methods

Plant material

Seven melon accessions were used in this study, three from the ssp. *melo* and four from the ssp. *agrestis* (Supplementary Table S1). The ssp. *melo* accessions were the “Piel de sapo” line T111 (PS) (*inodorus* group), the cantaloupe type Védrantais (VED) (*cantaloupensis* group) and the C-1012 cultivar (IRK) (*dudaim* group). The ssp. *agrestis* accessions were PI 161375 (Songwhan charmi, SC) (*conomon* group), PI 124112

(Calcuta, CAL) (*momordica* group) and the accessions Ames-24297, previously classified as *Cucumis trigonus* (TRI), and C-386 from Cabo Verde (CV). According to morphological and agronomic data, CV/TRI, SC/CAL/IRK and PS/VED may be considered wild, landrace and elite lines, respectively. DHL92, a doubled-haploid line obtained from the cross between PS and SC, and which was used to obtain the reference genome sequence of melon (Garcia-Mas, et al. 2012) was also included in the analysis as a control. Seeds from the eight accessions were planted in trays and plants were grown under the same greenhouse conditions as previously described (Eduardo, et al. 2005). For library construction, young leaf samples were used for DNA extraction (Garcia-Mas, et al. 2001), mixing leaves of 5 plants per accession except for CV and IRK, where a single plant was used. DNA for PCR analysis was extracted from different individuals than the ones used for library preparation.

Re-sequence analysis and SNP calling

Paired-end libraries with an average insert size of 500 bp were produced and sequenced with the Illumina Genome Analyser IIx technology at the Centre Nacional d'Anàlisi Genòmica (CNAG, Barcelona) (Supplementary Table S2). On average, more than 30 million paired reads were obtained for each melon accession with a read length of 150 bp. Re-sequencing data of DHL92, PS, SC, IRK and CV has been already described in previous works (Garcia-Mas, et al. 2012) (González, et al. 2013).

An in-house pipeline called SUPERW (Simply Unified Pair-End Read Workflow) was developed to create a dynamic and fast tool to analyze the variation data produced from the re-sequencing experiments (Supplementary Figure S1, Supplementary material). The SUPER pipeline and the filtering script SUPERRA were developed and used with the melon re-sequencing data. The melon reference genome used was v3.5 (melon_genome_ pseudomolecules_V3.5), available at http://www.melonomics.net. The parameters used were i) for the filtering and mapping step a read PRHED quality >25, a minimum length of 35 bp, removing all the Illumina adaptors and a mapping quality of PHRED >10, ii) for the variation calling step (SNPs and DIPs) a genotype quality >= 20, a locus quality >30 and a minimum depth coverage of 5 reads for both small and large variations, iii) only DIPs up to 200 bp were kept, iv) at each variable locus, the allele frequency (AF) of the variant was calculated as the ratio of reads supporting a homozygous or heterozygous state. Variants were filtered to keep those with an allele frequency (AF) >0.75. Several benchmarks were used to reach an optimal quality for the data produced. The re-sequenced line DHL92, the same variety used to assemble the melon reference genome, was used as control to remove all the variants caused by 454 sequencing errors (Supplementary Table S11). The re-sequence of DHL92 enabled us to determine the false discovery rate of called SNPs as 2.05 x 10E-5 per bp. The same approach was used to calculate the false discovery rate between the reference genome and one of its parental lines, in a region inherited from SC in chr12, being 2.66 x 10E-5. After the quality filtering, only homozygous sites were considered.

Calling large structural variants

Both pair-end and depth of coverage approaches were used to predict large structural variations (SV). The paired-end approach was used to calculate the SV from 200 bp up to 25 Kb while depth of coverage was used to identify SV larger than 25 Kb. The results of the two algorithms were merged in order to create a non-redundant set of large SVs. The pair-read approach of Pindel v2.4 (Ye, et al. 2009) was used to call variations up to 25 Kb. Alignment files created with *bwa sampe* and *bwa aln* (bwa v 0.7.0) without removing multiple mapped reads were used to extract deletions, duplications, small insertions and inversions considering the difference between the pair-end distance and mapped distance. All SVs were filtered for a depth of coverage to require at least 5x. Moreover specific filters were added to each SV: i) should have a PHRED quality of 20, ii) inversions should be longer than 200 bp, iii) deletions should have both forward and reverse supporting reads and iv) SVs of different types that fell in same genomic region (conflict SVs) were removed. In addition, for each SV the genes that have suffered a variation were extracted. All genes in regions affected by SVs were tallied and functionally annotated using a tool to assign Automated Assignment of Human Readable Descriptions, (AHRD v2, https://github.com/groupschoof/AHRD). AHRD is able to select descriptions that are concise and informative, using BLAST hits taken from searches against Uniprot/trEMBL, Uniprot/Swissprot and TAIR10.

Population genetic analysis

Estimates of nucleotide variability were calculated using Achaz equations (Achaz 2009) for Watterson, Tajima's, Fu and Li, Fay and Wu and Zeng's theta with the folded and unfolded frequency spectrum, if necessary. Patterns of variability were inferred from calculation of the following neutrality tests: Tajima's D (Tajima 1989), Fu and Li's D and F (Fu and Li 1993), Fay and Wu's H (Fay and Wu 2000; Zeng, et al. 2006). Population differentiation - *Fst* (Hudson, et al. 1992) - was calculated between chosen groups. An Illumina Genome Analyser IIx genome re-sequence of a cucumber inbreed line, kindly obtained from Semillas Fitó SA, was used for the divergence studies. Fixed variants and nucleotide divergence was calculated between *C. melo* and *C. sativus* using the number of differences from the total positions. Divergence was corrected for multiple mutations using the HKY model (Hasegawa, et al. 1985). *mstatspop* was used to calculate all these statistics (made available by the authors at http://bioinformatics.cragenomica.es/numgenomics/people/sebas/software/software.html).

The groups considered in this study were *C. melo,* ssp. *melo* (PS, VED, IRK), *C. melo* ssp. *agrestis* (CAL, SC, TRI, CV), 'Elite' (PS and VED), 'NO-Elite' (TRI, CV, IRK, CAL, SC), “Landrace” (IRK, CAL, SC), 'Domesticated' (PS, VED, IRK, CAL, SC) and 'Wild' (TRI, CV).

The analyses were performed on total, silent, synonymous and non-synonymous positions using the GTF annotation file, considering the whole genome but also calculating separately the statistics by windows of 50 Kb, 100 Kb and 500 Kb. In order to analyse a possible influence of read depth on the measured variability and differentiation the correlation of the mean read depth in windows of 50Kb and 500Kb and the levels of variability and differentiation among populations was calculated. No correlation was found between read depth and differentiation. A very low correlation (-0.104, Pvalue=4.57e-35) was observed between read depth and variability, and no differences were observed when introducing read depth as a dependent factor in our analyses.

A table with the presence/absence of all annotated Transposon Elements (TE) was used to calculate the number and the frequency of these elements on the whole and on each desired window. Estimates of diversity (Watterson, Tajima, Fu and Li, Zeng) were calculated with folded and with unfolded frequency spectrum for each window of size 50 Kb, 100 Kb and 500 Kb. Smaller windows showed higher variance in the studied statistics and were not used in global analysis. The frequency spectra and the Tajima's D test were also calculated to study the patterns of TE variability. Finally, population differentiation was also calculated between different groups or populations.

Kendall rank association values and their probabilities were calculated with the R-environment (http://www.rproject.org) to estimate the association between any two variables. Similarly to Cai et al. (2009), we calculated the mean of the variable located on the y-axis on 100 separated bins for the variable located on the x-axis. These values were plotted in red together with the whole values.

Partial correlation analyses were calculated assuming a normal distribution and comparing the residuals of the variables in relation to a third variable to account. Correlation and signification was obtained using the Pearson method.

The methods used for constructing dendrograms, performing Principal Component Analysis, estimating the proportion of non-synonymous positions under selection and the analysis using recombination estimates are detailed at Supplementary material.

Transposable Element Insertion polymorphism analysis

Quality of reads was assessed with FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Reads were filtered using SGA (https://github.com/jts/sga) with *preprocess* (-q 10 -m 50 --permute-ambiguous –pe-mode=1), then *index* (-d 2000000) and *correct* with default parameters. Reads were corrected and trimmed using SGA in order to maintain read pairs intact (as opposed to filtering performed for SNP analysis). Filtered reads were mapped to the assembled reference genome DHL92 v3.5 (Garcia-Mas, et al. 2012) using BWA v 0.7.0 (Li and Durbin 2009) with *aln* (-n 6 -o 1 -e 1) , and *sampe* with –s default parameter. SAMtools v 0.1.19 was used (Li, et al. 2009) to sort and index all *bam* files.

Pindel v2.4 (Ye, et al. 2009) and Breakdancer v1.2.6 (Chen, et al. 2009) were used to detect deletions in the melon lines with respect to the reference genome. Pindel was run with a maximum range index of 5 and an anchor quality of 35. In order to decrease computational time, reporting of both inversions and duplications were disabled. Breakdancer was used with default parameters. Both sets of predictions were merged to remove redundancies, and an additional filtering step was applied to select those greater than 200 bp and less than 25 Kb. Of these, those overlapping annotated transposons were retained for further analysis.

Analysis to detect transposon insertions was performed with Jitterbug (Hénaff et al, submitted), using a value of 35 for the minimal mapping quality of the reads (-q 35). Jitterbug was parallelized using a bin size of 1 million, and insertions were filtered using the companion scripts supplied and the default parameters calculated on the fly by Jitterbug. The sensitivity and specificity of Pindel, Breakdancer and Jitterbug was assessed on a simulated dataset where a subset of annotated TEs we shuffled to random positions. The elements were selected randomly over the set of annotated elements in order to get a distribution in size of the elements (excluding annotated fragments smaller than 200bp). These were cut-and-paste into random locations in the genome (excluding regions within 100bp of already annotated TEs and Ns). The script used to perform this simulation is available at https://sourceforge.net/projects/kitchen-drawer/files/sim_SV.py/download. The reads from DHL92 line were then mapped to the modified references and we assessed the ability for Pindel and Breakdancer to recover elements inserted into the reference (and absent from that locus in the sample, so detected as deletions) and that of Jitterbug to detect elements deleted from the reference (and seen as insertions in the sample). Sensitivity was calculated as the number of True Positives (TP) over the number of simulated variants, and PPV was calculated as the number of TP over the total number of predictions.

A subset of the predicted TE insertion polymorphisms were analyzed by PCR. PCR products were obtained in a final volume of 20 µl containing 40 ng genomic DNA, 300 µM dNTPs, 20 µM for each primer and 2 units / 20 µl of LongAmp® Taq DNA Polymerase (New England BioLabs). Primer pairs were designed to be 20-26 bp long for PCR amplification using Primer3 software (Untergasser, et al. 2012). The oligonucleotides used are listed in Supplementary Table S12. Half of the PCR products were separated on a 1% agarose gel and stained with ethidium bromide for checking the PCR amplification. Fragment sizes were estimated with the 1 Kb DNA ladder (Biotools).

Statistical analysis of low variation in elite varieties

To test the significance of the lesser number of PM sites in the elite varieties, we performed a label permutation test: taking two varieties at random and counting the number of variable sites, over 5,000 iterations. The number of such combinations with a number of variable sites less than 613 (the number of variable sites between elite varieties) is 0, thus the pvalue is 0 / 5000 = 0 (definitely significant).

Statistical analysis of the low variation region on Chr1

The region in chr 1 from 5Mb to 16Mb (11Mb of sequence) has 0 PM sites between elite varieties, and 99 PM sites between non-elite varieties. To test whether this low variability in the elite varieties is significantly different from the rest of the genome, we did a resampling test: we divided the genome into 1Mb windows, selected 11 random windows and counted the number of PM sites between elite varieties. This resampling was performed 10,000 times, and out of these, 0 times were 0 PM sites over the 11 windows accounting for 11Mb of sequence. The pvalue is 0 / 10000 = 0, meaning that the variability between elite varieties is significantly lower in that region than elsewhere.

We did the same for the non-elite varieties and found that 9,936 times out of 10,000 resamplings, there were 99 PM sites or less in the sampled 11Mb. This corresponds to a pvalue of 0.99 which means that the number of PM site in non-elite varieties is not significantly lower in this window.

Data access

The SUPERW and SUPERRA tools are available and documented at Sourceforge (https://sourceforge.net/projects/superw/). Illumina paired-end sequences have been deposited in the European Nucleotide Archive SRA and are available at the URL http://www.ebi.ac.uk/ena/data/view/PRJEB7636.

Author's contributions

JM-C and JG-M conceived the study and designed the experiments. WS, SP and EH analyzed the raw sequence data and discovered SNPs, SV and TEs, respectively. WB-P performed the validation of the SNPs, the Principal Component Analysis and calculated the Codon and Gene Densities and the percentages of GCs. SER-O performed the population evolutionary analyses of variability and divergence for nucleotide positions and for the frequency of transposon regions, including the statistical correlation analyses and their interpretation. CV validated TE polymorphisms. JG-M, JMC, MM and SER-O drafted the manuscript. All authors read and approved the final version of the manuscript.

Acknowledgements

We thank Jordi Morata for his help dating retrotransposon insertions. This work was supported by Ministerio de Ciencia e Innovación (grant BFU2009-11932 to JMC and grant AGL2012-40130-C02-01 to JG-M); Ministerio de Economía y Competitividad (grant AGL2013-43244-R to JMC, grant AGL2013-41834-R to SER-O and grant AGL2009-12698-C02-01 to JG-M) and Fundación Genoma España to JG-M.

References

3KRGP. 2014. The 3,000 rice genomes project. GigaScience 3:7.

Achaz G. 2009. Frequency spectrum neutrality tests: one for all and all for one. Genetics 183:249-258.

Alkan C, Coe BP, Eichler EE. 2011. Genome structural variation discovery and genotyping. Nature reviews. Genetics 12:363-376.

Arabidopsis Genome I. 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408:796-815.

Argyris JM, Ruiz-Herrera A, Madriz Masis P, Sanseverino W, Morata J, Pujol M, Ramos-Onsins SE, Garcia-Mas J. 2014. Use of targeted SNP selection for an improved anchoring of the melon (*Cucumis melo* L.) scaffold genome assembly. BMC Genomics (in press).

Brunner S, Fengler K, Morgante M, Tingey S, Rafalski A. 2005. Evolution of DNA sequence nonhomologies among maize inbreds. Plant Cell 17:343-360.

Butelli E, Licciardello C, Zhang Y, Liu J, Mackay S, Bailey P, Reforgiato-Recupero G, Martin C. 2012. Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges. The Plant cell 24:1242-1255.

Cai JJ, Macpherson JM, Sella G, Petrov DA. 2009. Pervasive hitchhiking at coding and regulatory sites in humans. PLoS Genet 5:e1000336.

Casacuberta JM, Santiago N. 2003. Plant LTR-retrotransposons and MITEs: control of transposition and impact on the evolution of plant genes and genomes. Gene. 311:1-11.

Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang QY, Locke DP, et al. 2009. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nature methods 6:677-U676.

Chia JM, Song C, Bradbury PJ, Costich D, de Leon N, Doebley J, Elshire RJ, Gaut B, Geller L, Glaubitz JC, et al. 2012. Maize HapMap2 identifies extant variation from a genome in flux. Nature Genetics 44:803-U238.

Choulet F, Alberti A, Theil S, Glover N, Barbe V, Daron J, Pingault L, Sourdille P, Couloux A, Paux E, et al. 2014. Structural and functional partitioning of bread wheat chromosome 3B. Science (New York, N.Y.) 345:1249721.

Du J, Tian Z, Hans CS, Laten HM, Cannon SB, Jackson SA, Shoemaker RC, Ma J. 2010. Evolutionary conservation, diversity and specificity of LTR-retrotransposons in flowering plants: insights from genome-wide analysis and multi-specific comparison. The Plant journal : for cell and molecular biology 63:584-598.

Eduardo I, Arus P, Monforte AJ. 2005. Development of a genomic library of near isogenic lines (NILs) in melon (*Cucumis melo* L.) from the exotic accession PI161375. Theor Appl Genet 112:139-148.

Eichten SR, Briskine R, Song J, Li Q, Swanson-Wagner R, Hermanson PJ, Waters AJ, Starr E, West PT, Tiffin P, et al. 2013. Epigenetic and Genetic Influences on DNA Methylation Variation in Maize Populations. The Plant Cell Online.

El Baidouri M, Panaud O. 2013. Comparative genomic paleontology across plant kingdom reveals the dynamics of TE-driven genome evolution. Genome Biol Evol 5:954-965.

Esteras C, Formisano G, Roig C, Diaz A, Blanca J, Garcia-Mas J, Gomez-Guillamon ML, Lopez-Sese AI, Lazaro A, Monforte AJ, et al. 2013. SNP genotyping in melons: genetic variation, population structure, and linkage disequilibrium. Theor Appl Genet 126:1285-1303.

Falchi R, Vendramin E, Zanon L, Scalabrin S, Cipriani G, Verde I, Vizzotto G, Morgante M. 2013. Three distinct mutational mechanisms acting on a single gene underpin the origin of yellow flesh in peach. The Plant Journal 76:175-187.

Fay JC. 2011. Weighing the evidence for adaptation at the molecular level. Trends in Genetics 27:343-349.

Fay JC, Wu CI. 2000. Hitchhiking under positive Darwinian selection. Genetics 155:1405-1413.

Felsenstein, J. 1978. The number of Evolutionary Trees. Systematic Zoology. 27: 27-33.

Flowers JM, Molina J, Rubinstein S, Huang P, Schaal BA, Purugganan MD. 2012. Natural selection in gene-dense regions shapes the genomic pattern of polymorphism in wild and domesticated rice. Mol Biol Evol 29:675-687.

Fu YX, Li WH. 1993. Statistical tests of neutrality of mutations. Genetics 133:693-709.

Gao Z-Y, Zhao S-C, He W-M, Guo L-B, Peng Y-L, Wang J-J, Guo X-S, Zhang X-M, Rao Y-C, Zhang C, et al. 2013. Dissecting yield-associated loci in super hybrid rice by resequencing recombinant inbred lines and improving parental genome sequences. Proceedings of the National Academy of Sciences 110:14492-14497.

Garcia-Mas J, Benjak A, Sanseverino W, Bourgeois M, Mir G, Gonzalez VM, Henaff E, Camara F, Cozzuto L, Lowy E, et al. 2012. The genome of melon (*Cucumis melo* L.). Proc Natl Acad Sci USA 109:11872-11877.

Garcia-Mas J, Oliver M, Gomez-Paniagua H, de Vicente MC. 2000. Comparing AFLP, RAPD and RFLP markers for measuring genetic diversity in melon. Theoretical and Applied Genetics 101:860-864.

Garcia-Mas J, van Leeuwen H, Monfort A, Carmen de Vicente M, Puigdomènech P, Arús P. 2001. Cloning and mapping of resistance gene homologues in melon. Plant Science 161:165-172.

Gaut BS, Wright SI, Rizzon C, Dvorak J, Anderson LK. 2007. Recombination: an underappreciated factor in the evolution of plant genomes. Nat Rev Genet 8:77-84.

Gebhardt C. 2013. Bridging the gap between genome analysis and precision breeding in potato. Trends in Genetics 29:248-256.

Godfray HC, Beddington JR, Crute IR, Haddad L, Lawrence D, Muir JF, Pretty J, Robinson S, Thomas SM, Toulmin C. 2010. Food security: the challenge of feeding 9 billion people. Science 327:812-818.

González VM, Aventín N, Centeno E, Puigdomènech P. 2013. High presence/absence gene variability in defense-related gene clusters of Cucumis melo. BMC Genomics 14:782.

González VM, Aventín N, Centeno E, Puigdomènech P. 2014. Interspecific and intraspecific gene variability in a 1-Mb region containing the highest density of NBS-LRR genes found in the melon genome. BMC Genomics. 15:1131.

Guo S, Zhang J, Sun H, Salse J, Lucas WJ, Zhang H, Zheng Y, Mao L, Ren Y, Wang Z, et al. 2013. The draft genome of watermelon (*Citrullus lanatus*) and resequencing of 20 diverse accessions. Nat Genet 45:51-58.

Han Y, Wessler SR. 2010. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. Nucleic Acids Research 38:e199.

Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22:160-174.

Henaff E, Vives C, Desvoyes B, Chaurasia A, Payet J, Gutierrez C, Casacuberta JM. 2014. Extensive amplification of the E2F transcription factor binding sites by transposons during evolution of Brassica species. The Plant journal. 77:852-862.

Hirsch CN, Foerster JM, Johnson JM, Sekhon RS, Muttoni G, Vaillancourt B, Peñagaricano F, Lindquist E, Pedraza MA, Barry K, et al. 2014. Insights into the Maize Pan-Genome and Pan-Transcriptome. The Plant Cell Online.

Huang XH, Lu TT, Han B. 2013. Resequencing rice genomes: an emerging new era of rice genomics. Trends in Genetics 29:225-232.

Hudson RR, Slatkin M, Maddison WP. 1992. Estimation of levels of gene flow from DNA sequence data. Genetics 132:583-589.

Hufford MB, Xu X, van Heerwaarden J, Pyhajarvi T, Chia JM, Cartwright RA, Elshire RJ, Glaubitz JC, Guill KE, Kaeppler SM, et al. 2012. Comparative population genomics of maize domestication and improvement. Nature Genetics 44:808-U118.

International Rice Genome Sequencing P. 2005. The map-based sequence of the rice genome. Nature 436:793-800.

Ingvarsson, PK. 2010. Natural selection on synonymous and nonsynonymous mutations shapes patterns of polymorphism in Populus tremula. Mol Biol Evol. 27: 650-660.

Jeffrey C. 1980. A review of the Cucurbitaceae. Botanical Journal of the Linnean Society 81:233-247.

Jiao YP, Zhao HN, Ren LH, Song WB, Zeng B, Guo JJ, Wang BB, Liu ZP, Chen J, Li W, et al. 2012. Genome-wide genetic changes during modern breeding of maize. Nature Genetics 44:812-U124.

Kobayashi S, Goto-Yamamoto N, Hirochika H. 2004. Retrotransposon-Induced Mutations in Grape Skin Color. Science 304:982-.

Lam HM, Xu X, Liu X, Chen W, Yang G, Wong FL, Li MW, He W, Qin N, Wang B, et al. 2010. Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nat Genet 42:1053-1059.

Layer RM, Chiang C, Quinlan AR, Hall IM. 2014. LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. 15:R84.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754-1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078-2079.

Lin T, Zhu G, Zhang J, Xu X, Yu Q, Zheng Z, Zhang Z, Lun Y, Li S, Wang X, et al. 2014.

Genomic analyses provide insights into the history of tomato breeding. Nature Genetics doi:10.1038/ng.3117.

Lisch D. 2013. How important are transposons for plant evolution? Nature reviews. Genetics 14:49-61.

Lu C, Chen JJ, Zhang Y, Hu Q, Su WQ, Kuang HH. 2012. Miniature Inverted-Repeat Transposable Elements (MITEs) Have Been Accumulated through Amplification Bursts and Play Important Roles in Gene Expression and Species Diversity in Oryza sativa. Molecular Biology and Evolution 29:1005-1017.

Mackay TF, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, Casillas S, Han Y, Magwire MM, Cridland JM, Richardson MF, Anholt RR, Barrón M, Bess C, Blankenburg KP, Carbone MA, Castellano D, Chaboub L, Duncan L, Harris Z, Javaid M, Jayaseelan JC, Jhangiani SN, Jordan KW, Lara F, Lawrence F, Lee SL, Librado P, Linheiro RS, Lyman RF, Mackey AJ, Munidasa M, Muzny DM, Nazareth L, Newsham I, Perales L, Pu LL, Qu C, Ràmia M, Reid JG, Rollmann SM, Rozas J, Saada N, Turlapati L, Worley KC, Wu YQ, Yamamoto A, Zhu Y, Bergman CM, Thornton KR, Mittelman D, Gibbs RA. 2012. The Drosophila melanogaster Genetic Reference Panel. Nature 482:173-178.

Martin A, Troadec C, Boualem A, Rajab M, Fernandez R, Morin H, Pitrat M, Dogimont C, Bendahmane A. 2009. A transposon-induced epigenetic change leads to sex determination in melon. Nature 461:1135-1138.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. Nature 351:652-654.

McHale LK, Haun WJ, Xu WW, Bhaskar PB, Anderson JE, Hyten DL, Gerhardt DJ, Jeddeloh JA, Stupar RM. 2012. Structural variants in the soybean genome localize to clusters of biotic stress-response genes. Plant physiology 159:1295-1308.

Messer PW, Petrov DA. 2013. Frequent adaptation and the McDonald-Kreitman test. Proc Natl Acad Sci U S A 110:8615-8620.

Meyer RS, Purugganan MD. 2013. Evolution of crop species: genetics of domestication and diversification. Nat Rev Genet 14:840-852.

Monforte AJ, Garcia-Mas J, Arus P. 2003. Genetic variability in melon based on microsatellite variation. Plant Breeding 122:153-157.

Morgante M, De Paoli E, Radovic S. 2007. Transposable elements and the plant pan-genomes. Current Opinion in Plant Biology 10:149-155.

Myles S. 2013. Improving fruit and wine: what does genomics have to offer? Trends in Genetics 29:190-196.

Olsen KM, Wendel JF. 2013. A Bountiful Harvest: Genomic Insights into Crop Domestication Phenotypes. Annual Review of Plant Biology 64:47-70.

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, et al. 2009. The Sorghum bicolor genome and the diversification of grasses. Nature 457:551-556.

Pecinka A, Abdelsamad A, Vu GT. 2013. Hidden genetic nature of epigenetic natural variation in plants. Trends in Plant Science.

Peterson-Burch BD, Nettleton D, Voytas DF. 2004. Genomic neighborhoods for Arabidopsis retrotransposons: a role for targeted integration in the distribution of the Metaviridae. Genome biology 5.

Pitrat M. 2008. Melon (*Cucumis melo* L.). In: Prohens J, Nuez F, editors. Handbook of Crop Breeding Vol I: Vegetables. New York: Springer. p. 283–315.

Qi J, Liu X, Shen D, Miao H, Xie B, Li X, Zeng P, Wang S, Shang Y, Gu X, et al. 2013. A genomic variation map provides insights into the genetic basis of cucumber domestication and diversity. Nat Genet 45:1510-1515.

Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. 2012. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 28:i333-i339.

Saxena RK, Edwards D, Varshney RK. 2014. Structural variations in plant genomes. Briefings in functional genomics. 13:296-307

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, et al. 2010. Genome sequence of the palaeopolyploid soybean. Nature 463:178-183.

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, et al. 2009. The B73 maize genome: complexity, diversity, and dynamics. Science 326:1112-1115.

Sebastian P, Schaefer H, Telford IR, Renner SS. 2010. Cucumber (*Cucumis sativus*) and melon (*C. melo*) have numerous wild relatives in Asia and Australia, and the sister species of melon is from Australia. Proc Natl Acad Sci U S A 107:14269-14273.

Sharma A, Presting GG. 2014. Evolution of centromeric retrotransposons in grasses. Genome biology and evolution 6:1335-1352.

Silberstein L, Kovalski I, Huang R, Anagnostou K, Jahn M, Perl-Treves R. 1999. Molecular variation in melon (*Cucumis melo* L.) as revealed by RFLP and RAPD markers. Scientia Horticulturae.

Springer NM, Ying K, Fu Y, Ji T, Yeh C-T, Jia Y, Wu W, Richmond T, Kitzman J, Rosenbaum H, et al. 2009. Maize Inbreds Exhibit High Levels of Copy Number Variation (CNV) and Presence/Absence Variation (PAV) in Genome Content. PLoS Genet 5:e1000734.

Stepansky A, Kovalski I, Perl-Treves R. 1999. Intraspecific classification of melons (*Cucumis melo* L.) in view of their phenotypic and molecular variation. Plant Systematics and Evolution 217:313-332.

Studer A, Zhao Q, Ross-Ibarra J, Doebley J. 2011. Identification of a functional transposon insertion in the maize domestication gene tb1. Nature Genetics 43:1160-U1164.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585-595.

Tian Z, Zhao M, She M, Du J, Cannon SB, Liu X, Xu X, Qi X, Li M-WW, Lam H-MM, et al. 2012. Genome-wide characterization of nonreference transposons reveals evolutionary propensities of transposons in soybean. The Plant Cell 24:4422-4436.

Tomato Genome Consortium. 2012. The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485:635-641.

Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG. 2012. Primer3--new capabilities and interfaces. Nucleic Acids Research 40.

Verde I, Abbott AG, Scalabrin S, Jung S, Shu SQ, Marroni F, Zhebentyayeva T, Dettori MT, Grimwood J, Cattonaro F, et al. 2013. The high-quality draft genome of peach (Prunus persica) identifies unique patterns of genetic diversity, domestication and genome evolution. Nature Genetics 45:487-U447.

Xu X, Liu X, Ge S, Jensen JD, Hu FY, Li X, Dong Y, Gutenkunst RN, Fang L, Huang L, et al. 2012. Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nature biotechnology 30:105-U157.

Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. 2009. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics (Oxford, England) 25:2865-2871.

Zeng K, Fu Y-XX, Shi S, Wu C-II. 2006. Statistical tests for detecting positive selection by utilizing high-frequency variants. Genetics 174:1431-1439.

Tables

SNPs and DIPs	CV	IRK	PS	SC	TRI	CAL	VED	TOTAL
SNPs	2,189,790	1,110,612	835,481	679,942	1,281,217	1,439,763	1,331,441	4,556,377
DIPs	208,431	132,123	98,959	81,756	165,699	173,250	145,460	718,832
Unique SNPs	842,870	59,513	41,365	93,199	221,946	85,345	68,306	
Unique DIPs	69,884	6,194	3,549	8,860	23,348	9,553	6,174	

Table 1. SNPs and DIPs between the 7 melon lines and the DHL92 reference genome. Unique variants are those unique to that given line.

Superfamily	# PM	% PM	# copies in the genome	% copies in the genome
gypsy	661	24.17	28,174	23.68
copia	635	23.22	17,346	14.58
non LTR retrotransposon	10	0.37	129	0.11
retrotransposon fragment	469	17.15	42,733	35.91
Total retrotransposons	1,775	64.90	88,382	74.27
CACTA*	93	3.40	6,944	5.84
hAT*	11	0.40	677	0.57
MULE*	239	8.74	9,497	7.98
Mariner*	3	0.11	609	0.51
Other MITEs	355	12.98	2,175	1.83
helitron	4	0.15	746	0.63
PIF*	147	5.37	3,094	2.60
DNA TE fragment	29	1.06	6,874	5.78
Total DNA TE	881	32.21	30,616	25.73
uncategorized	79	2.89		
TOTAL	2,735	100.00	118,998	100

* including short elements that could be MITEs

Table 2. Breakdown of TE families for TE-related polymorphic loci (PM).

Family	Superfamily	PM sites	%	Annotated	%
CM_MITE_2617	CACTA (MITE)	224	8.19	700	0.59
M_MULE_10	MULE	187	6.84	682	0.57

CM_gypsy_116	gypsy	120	4.39	177	0.15
CM_PIF_6	PIF	111	4.06	881	0.74
MELON_MITEs_1_43749	PIF (MITE)	110	4.02	117	0.1
CM_copia_96 *	copia	70	2.56	1695	1.42
CM_copia_45	copia	59	2.16	184	0.15
CM_copia_70 *	copia	59	2.16	39	0.03
CM_gypsy_137	gypsy	51	1.86	107	0.1
Total		991	36.23		3.85

* complex families composed of nested insertions

Table 3. Most of the polymorphisms were caused by the mobilization of a small number of transposon families.

	Total PM sites	**PM sites < 500 bp**	**%**	**PM sites in genes**	**%**	**PM sites in exons**	**%**
All lines	2,735	826	30.20	611	22.34	361	13.20
agrestis* vs *melo	31	4	12.90	4	12.90	3	9.68
elite vs others	69	13	18.84	12	17.39	8	11.59
PS vs VED	671	231	34.43	165	24.59	105	15.65

Table 4. TE insertions located in genic regions.

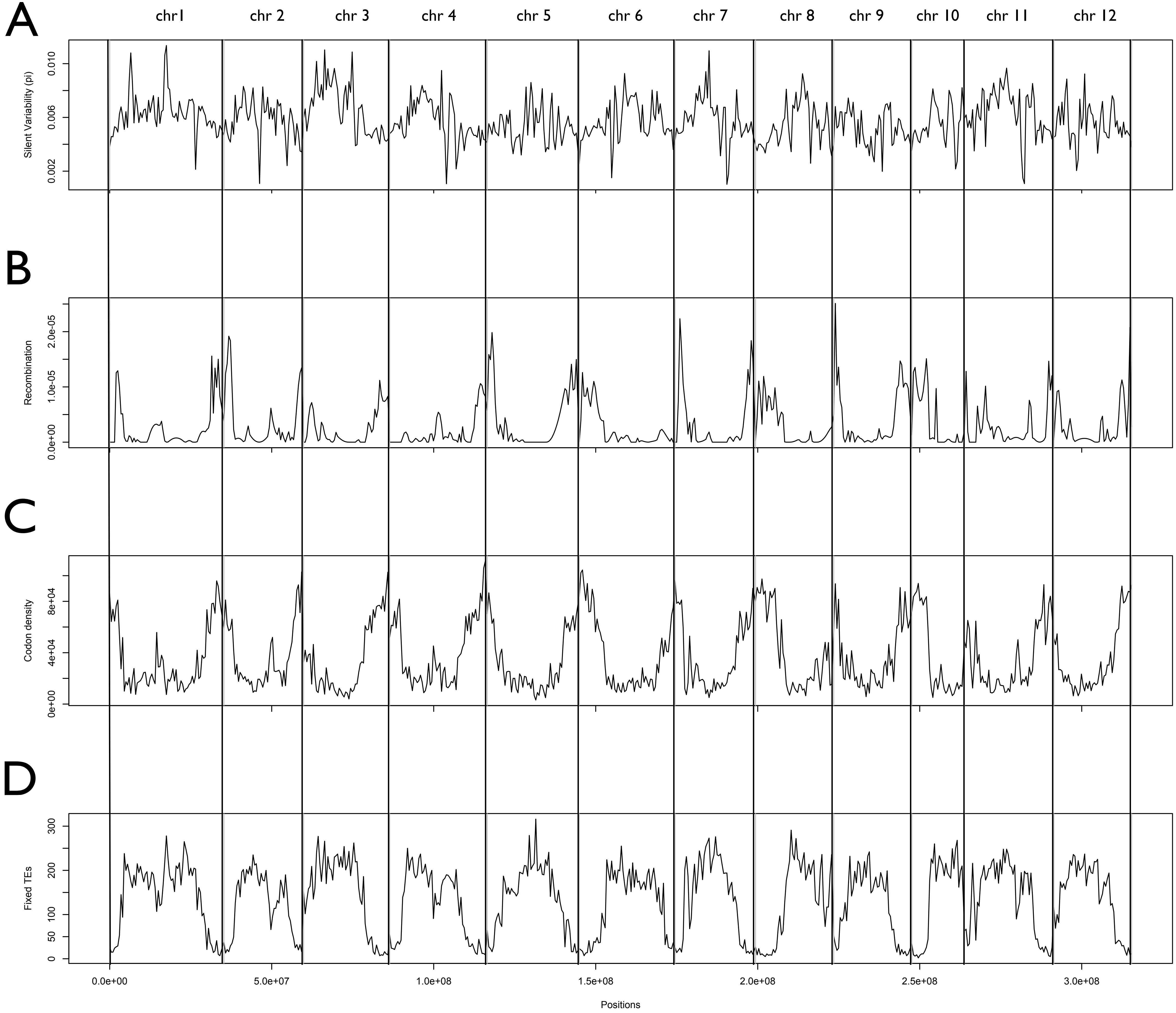
Legends to the Figures

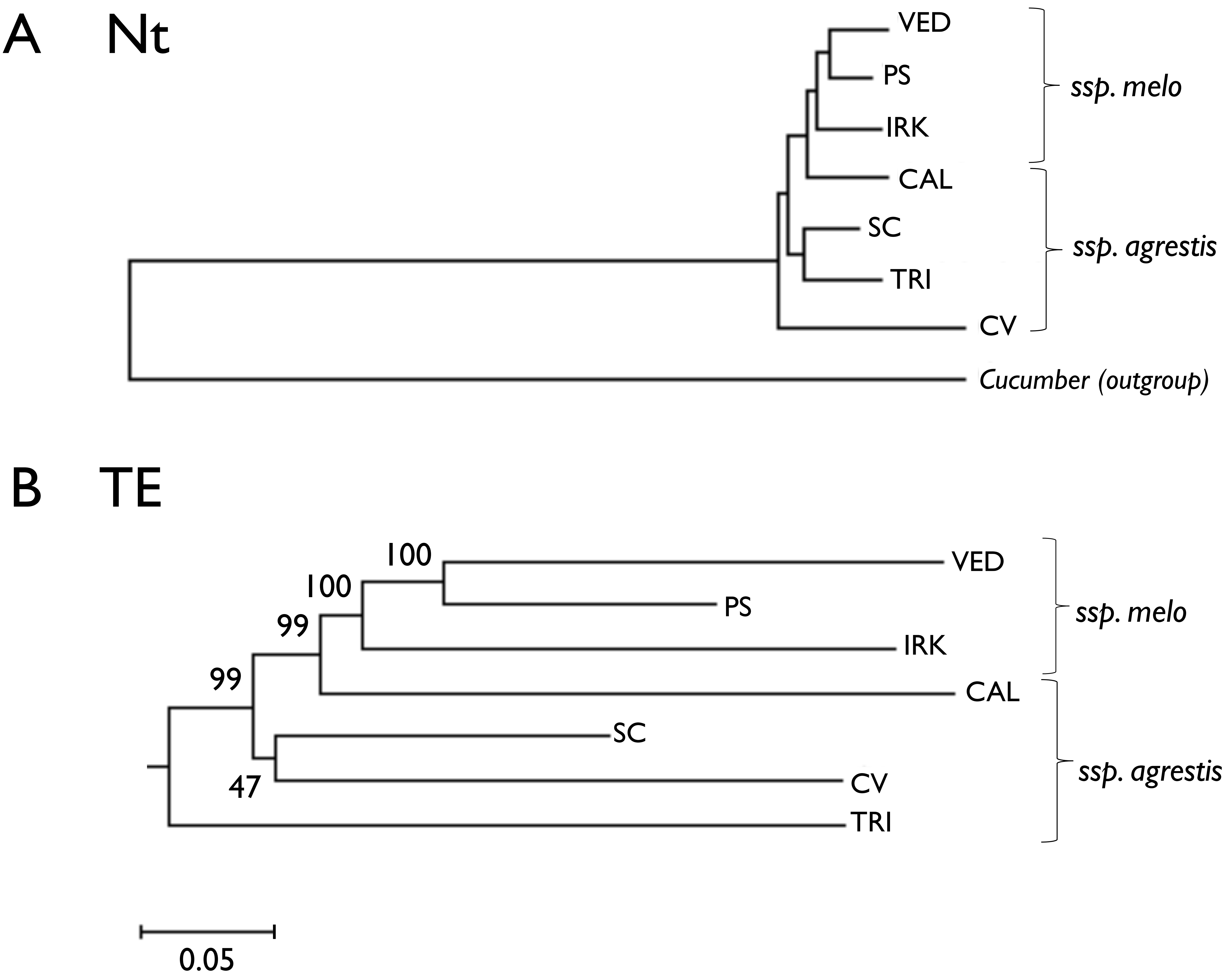
Figure 1. A. Nucleotide diversity of melon (π) across the genome in windows of 500 Kb. B. Distribution of the recombination rate across the genome in windows of 500 Kb. C. Distribution of codon density across the genome in windows of 500 Kb. D. Number of fixed TE across the genome in windows of 500 Kb.

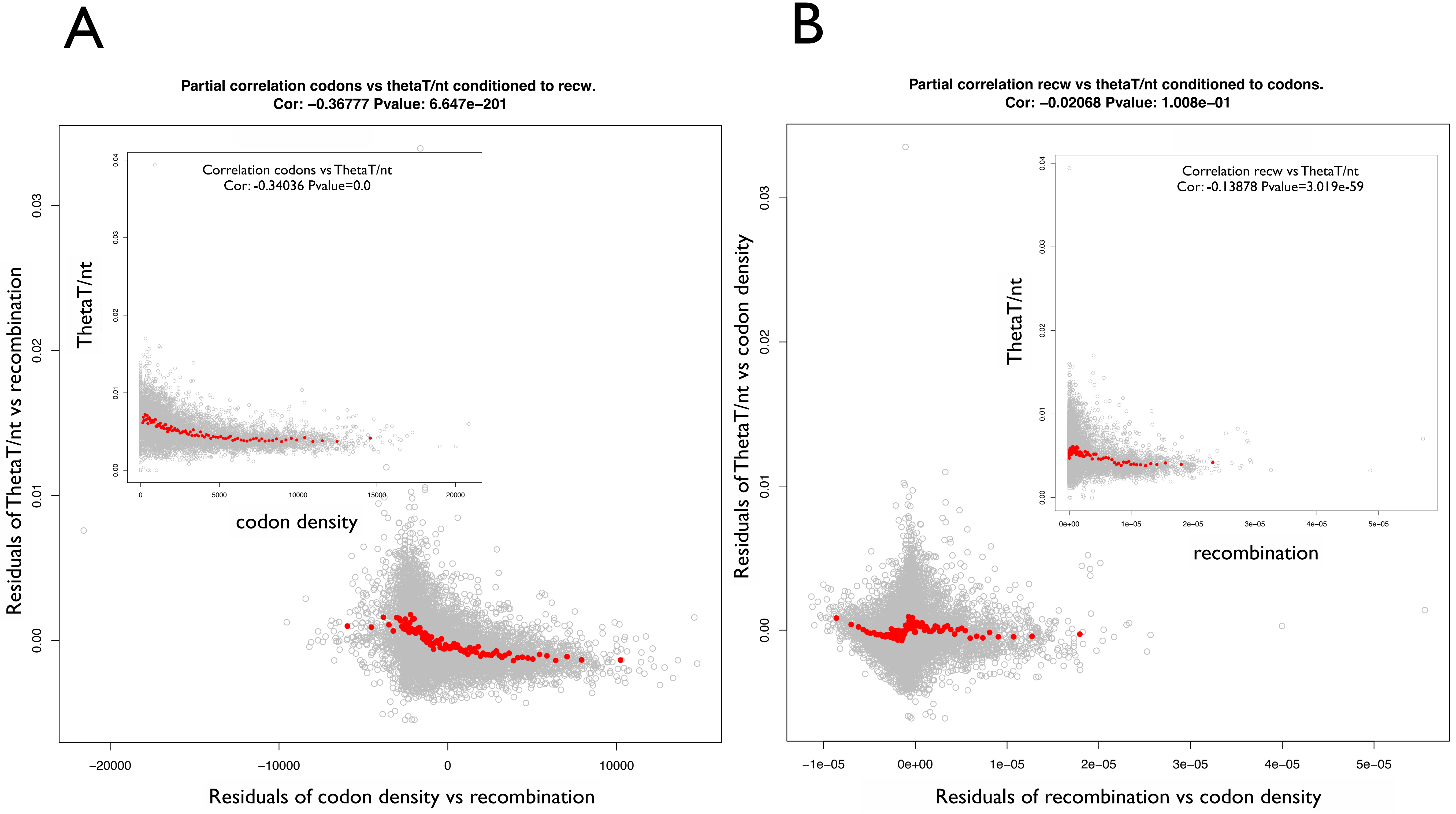
Figure 2. SNP-based (A) and transposon-based (B) phylogenetic relationships among the seven re-sequenced lines. Bootstrap support values are shown in black numbers while in red are the node support numbers obtained from constructing gene trees in subsets of 10Kb across the genome. In parenthesis is shown the rank order in frequency this node is observed across the whole set of trees. TRN is calculated as the sum of the ranks observed in the global tree minus the sum of ranks from an ideal tree (see Supplementary Materials). Bar plots shown in red the rank frequency of the nodes that are represented in the tree.

Figure 3. The left panel (A) shows in the inner plot the correlation of silent polymorphisms vs codon density and in the external plot the same correlation but considering recombination in the partial correlation (R=-0.37, P-value= 6.4E-201). The right panel (B) shows in the inner plot the correlation of silent polymorphisms vs recombination and in the external plot the same correlation but considering codon density in the partial correlation (R=-0.02, P-value= 0.1).

Figure 4. The left panel (A) shows in the inner plot the correlation of silent polymorphisms vs non-synonynmous divergence and in the external plot the same correlation but considering codon density in the partial correlation (R=0.13, P-value<1E-200). The right panel (B) shows in the inner plot the correlation of synonymous polymorphisms vs non-synonymous divergence and in the external plot the same correlation but considering codon density in the partial correlation (R=0.01, P-value= 0.28).

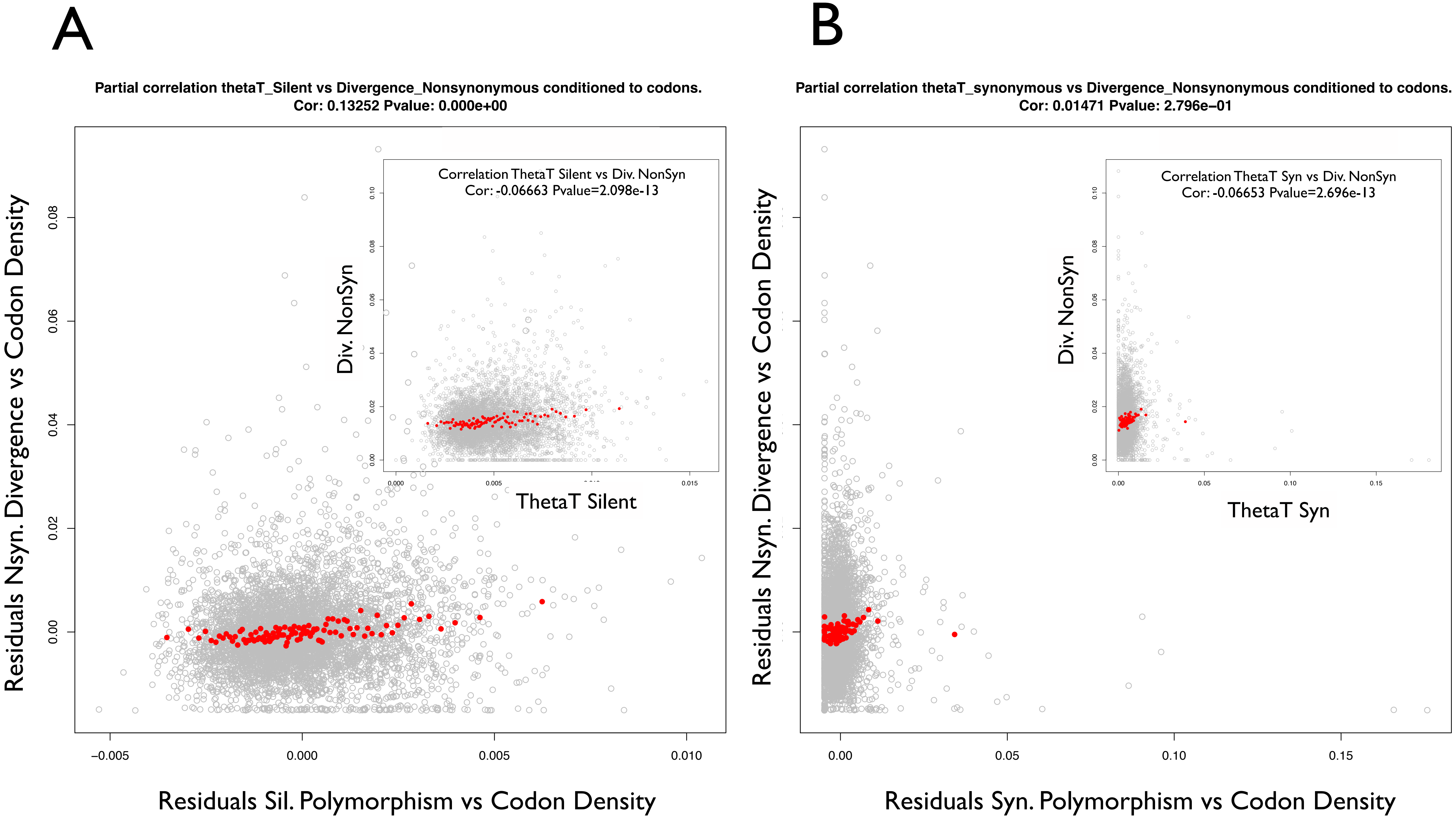

A
chr1
chr 2
chr 3
chr 4
chr 5
chr 6
chr 7
chr 8
chr 9
chr 10
chr 11
chr 12
Silent Variability (pi)
0.002
0.006
0.010
B
Recombination
0.0e+00
1.0e-05
2.0e-05
C
Codon density
0e+00
4e+04
8e+04
D
Fixed TEs
0
50
100
200
300
0.0e+00
5.0e+07
1.0e+08
1.5e+08
2.0e+08
2.5e+08
3.0e+08
Positions


A
Nt
VED
PS
IRK
CAL
SC
TRI
CV
Cucumber (outgroup)
ssp. melo
ssp. agrestis
B
TE
100
100
99
99
47
VED
PS
IRK
CAL
SC
CV
TRI
ssp. melo
ssp. agrestis
0.05



Partial correlation codons vs thetaT/nt conditioned to recw.
Cor: −0.36777 Pvalue: 6.647e−201

Partial correlation recw vs thetaT/nt conditioned to codons.
Cor: −0.02068 Pvalue: 1.008e−01


A
Partial correlation thetaT_Silent vs Divergence_Nonsynonymous conditioned to codons.
Cor: 0.13252 Pvalue: 0.000e+00
Residuals Nsyn. Divergence vs Codon Density
0.00
0.02
0.04
0.06
0.08
-0.005
0.000
0.005
0.010
Residuals Sil. Polymorphism vs Codon Density
Correlation ThetaT Silent vs Div. NonSyn
Cor: -0.06663 Pvalue=2.098e-13
Div. NonSyn
0.00
0.02
0.04
0.06
0.08
0.10
0.000
0.005
0.015
ThetaT Silent
B
Partial correlation thetaT_synonymous vs Divergence_Nonsynonymous conditioned to codons.
Cor: 0.01471 Pvalue: 2.796e-01
Residuals Nsyn. Divergence vs Codon Density
0.00
0.05
0.10
0.15
Residuals Syn. Polymorphism vs Codon Density
Correlation ThetaT Syn vs Div. NonSyn
Cor: -0.06653 Pvalue=2.696e-13
Div. NonSyn
0.00
0.02
0.04
0.06
0.08
0.10
0.00
0.05
0.10
0.15
ThetaT Syn