Università degli Studi di Sassari

SCUOLA DI DOTTORATO DI RICERCA
Scienze dei Sistemi Agrari e Forestali e delle Produzioni Alimentari

Indirizzo: Produttività delle piante coltivate

Ciclo XXIII

Analisi bioinformatica della struttura genomica di *Arabidopsis thaliana* L.

dr. Salvatore Camiolo

Direttore della Scuola:	Prof. Giuseppe Pulina
Referente di Indirizzo	Prof. Antonino Spanu
Docente Guida	Prof. Andrea Porceddu
Correlatore	Dr. Domenico Rau

Anno accademico 2009/2010

Alla mia famiglia....

Fatti non foste a viver come bruti
ma per seguir virtute e canoscenza
Dante Alighieri

AKNOWLEDGMENTS

First of all I would like to thank Professor Andrea Porceddu, who encouraged me to begin this wonderful experience. A special thank goes to the whole Dipartimento di Scienze Agronomiche e Genetica Vegetale Agraria in the person of Professor Antonino Spanu (to cite just the Director) who cared about my professional growth more than my presence in the department.

I would like to thank both my tutor and co-tutor for the time spent on improving (actually……creating!) my knowledge on genetics, statistics and many other topics that were essential to complete this work.

It is my pleasure to remember also the lab of genetics group members who entertained me during these three years with interesting scientific discussions or just enjoyable chats!

Finally a special thank goes to my wife Egizia who supported and encouraged me, and my little dear Elisa who kept me awake many nights and allowed to focus on this work outside the “normal” work times.

Index

Optimizing transgene expression in plants: the lesson from host genomes.

The impact on plants biotechnology.

Genetic engineering protocols are becoming available for a large number of crop species. In the last decades this technique has been suited to produce novel varieties with disease and insect pests resistance(*1*, *2*), herbicide tolerance(*3*), improved overall yield(*4*), and adaptation to specific soil or environmental conditions(*5*, *6*) (Figure 1). Moreover an increasing number of scientific papers and patents deal with the production of therapeutic proteins in transgenic organisms(*7*).

Figure 1: (a) improved yield of transformed tomatoes obtained by Lipman and co-workers at the Hebrew University(*8*). (b) Transgenic *Arabidosps thaliana* (left) ad control (right) upon rewatering after 8 days drought(*6*). (c) Tobacco plants after treatment with larvae of *H. armigera* (wild type in the left and three transgenic derivatives on the right)(*2*).

Although the use of plants as bio-reactors proved to be extremely advantageous in terms of costs, scalability and appropriate post-translational modifications, currently the majority of the therapeutic proteins are produced mainly in mammalian cells or bacteria where adequate transgene expression could be achieved. A careful evaluation of the entire production path, e.g. from gene integration to protein accumulation, is of primary importance for a successful plant transformation as recently reviewed by Desai et al.(*9*). When the low or even the non-expression of the transgene was analyzed at a molecular level a correlation was

highlighted with its inactivation rather than with its lost(*10*). The highly specific silencing of the transgene may be due to mechanisms such as those evolved to defend the plant against the deleterious effects of expressing a foreign, structurally different and possibly pathogenic DNA. Likewise insertion of heterologous genes can potentially perturb the normal structure and function of the plant genomes, as demonstrated by Elooma et al. in Petunia, where transgenes featuring an anomalous GC (guanine + cytosine) content were recognized and silenced by methylation(*11*).

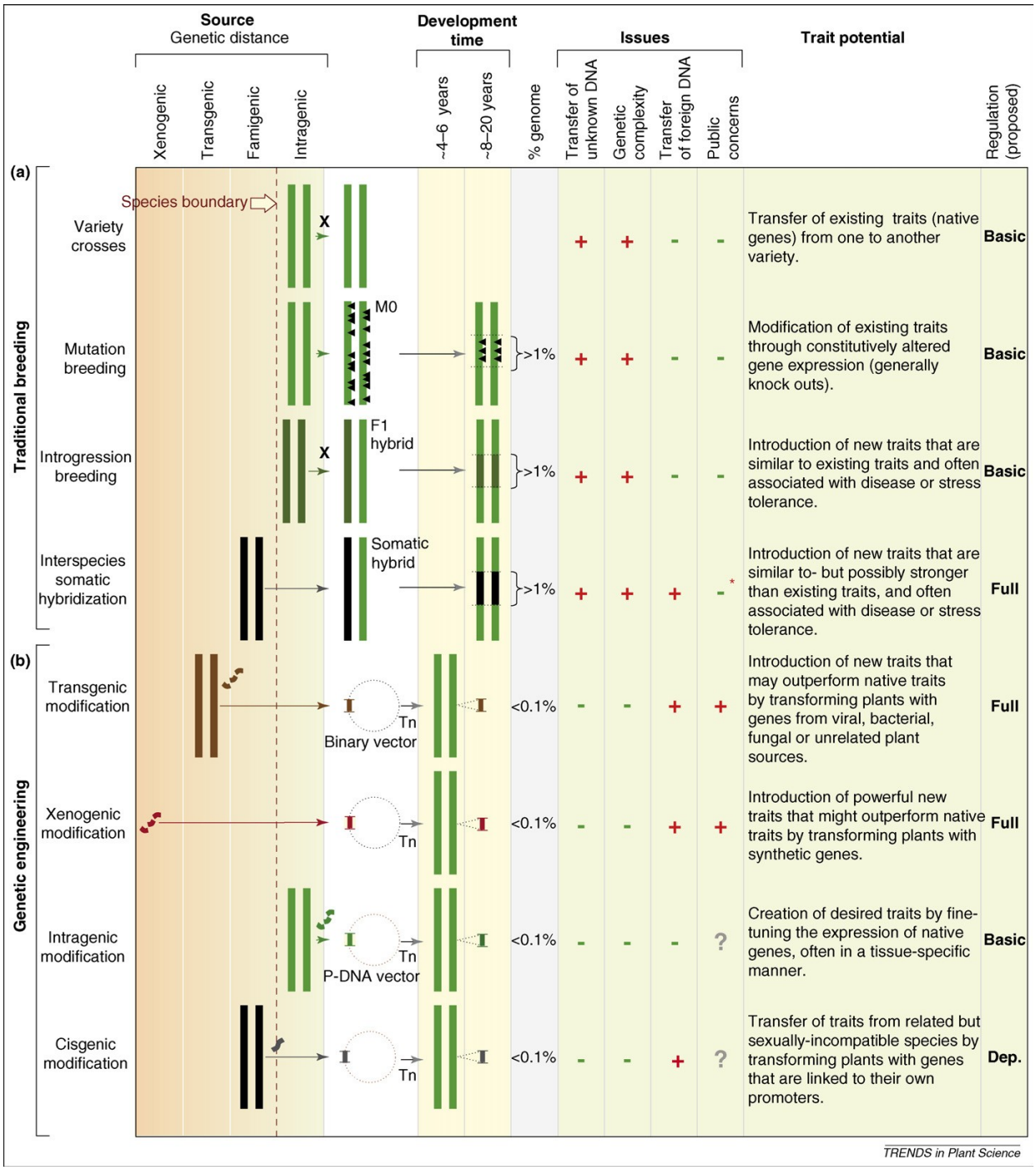


Figure 2: Summary of various plant transformation techniques. The intergenic approach represents a good alternative to traditional methods by minimizing the potential issues and requiring a basic regulation(*12*).

An adequate knowledge of the structure of plants genomes is also at the base of the last extension of traditional plant breeding also known as the intragenic approach (Figure 2) *(12)*. This strategy consists in isolating specific genetic elements from a plant, modulating their structure in according to that of genes with the desired expression profile and introducing them back into plants with the same sexual compatibility group.

Regardless of the aim behind the genetic transformation of a plant, it is therefore clear that the achievement of a stable transgene expression must go hand in hand with a detailed investigation of the genomic structure of the host organism. In facts because of the DNA redundancy a heuristic approach is absolutely not feasible (i.e. a 300 amino acid protein of average amino acid composition would be encoded by more that 10^{100} distinct gene sequences)(*13*).

In this review we report a survey on the structural properties of plants genomes, together with the selective forces that may play a role in shaping composition and structure of genes.

Compositional biases

Knowing the nucleotide composition and arrangement is certainly the first step in understanding the genome structure of plants. It is well known that in eukaryotic organisms the nucleotide composition is not homogeneous across the whole genome. Indeed most eukaryotes feature large DNA regions (up to hundred of kilobases) with fairly constant GC content, known as isochores, that are abruptly interrupted by other areas of diverse average GC composition(*14*). The first study which speculated the presence of isochores in plants dates back to 1989 when Bernardi and co-workers revealed them by the use of cesium chloride ultracentrifugation experiments(*15-17*). Recent advances in DNA sequencing techniques and the institution of many genome projects(*18-20*) led the isochores to be also investigated at a sequence level. A moving window graphic plot of GC content can reveal these structures although long-range patterns could be identified only by eye. On the other hand more sophisticated approach such as the enthropic segmentation algorithm(*21*) or the use of windowless methods(*22*) seemed to provide a more detailed overview (Figure 3). Genes proved to be unequally distributed within isochores and two distinct classes were highlighted(*23*). This differentiation was particularly evident in Gramineae maize, rice and barley, in Solenaceae tobacco, tomato and potato, and, at a much lesser extent in the Brassicacea *Arabidopsis thaliana*. The GC-poorer genes featured few or no intron at all,

longer coding sequences and higher number of exons. Zhang revealed in A. thaliana the presence of three distinct types of isochores by using a windowless sequence investigation: AT-, GC- and centromere-isochores, with the latter being both GC-rich but at different level*(23)*. Investigation of these structures revealed differences in gene density, transposable element (TE) distribution and t-DNA insertion density.

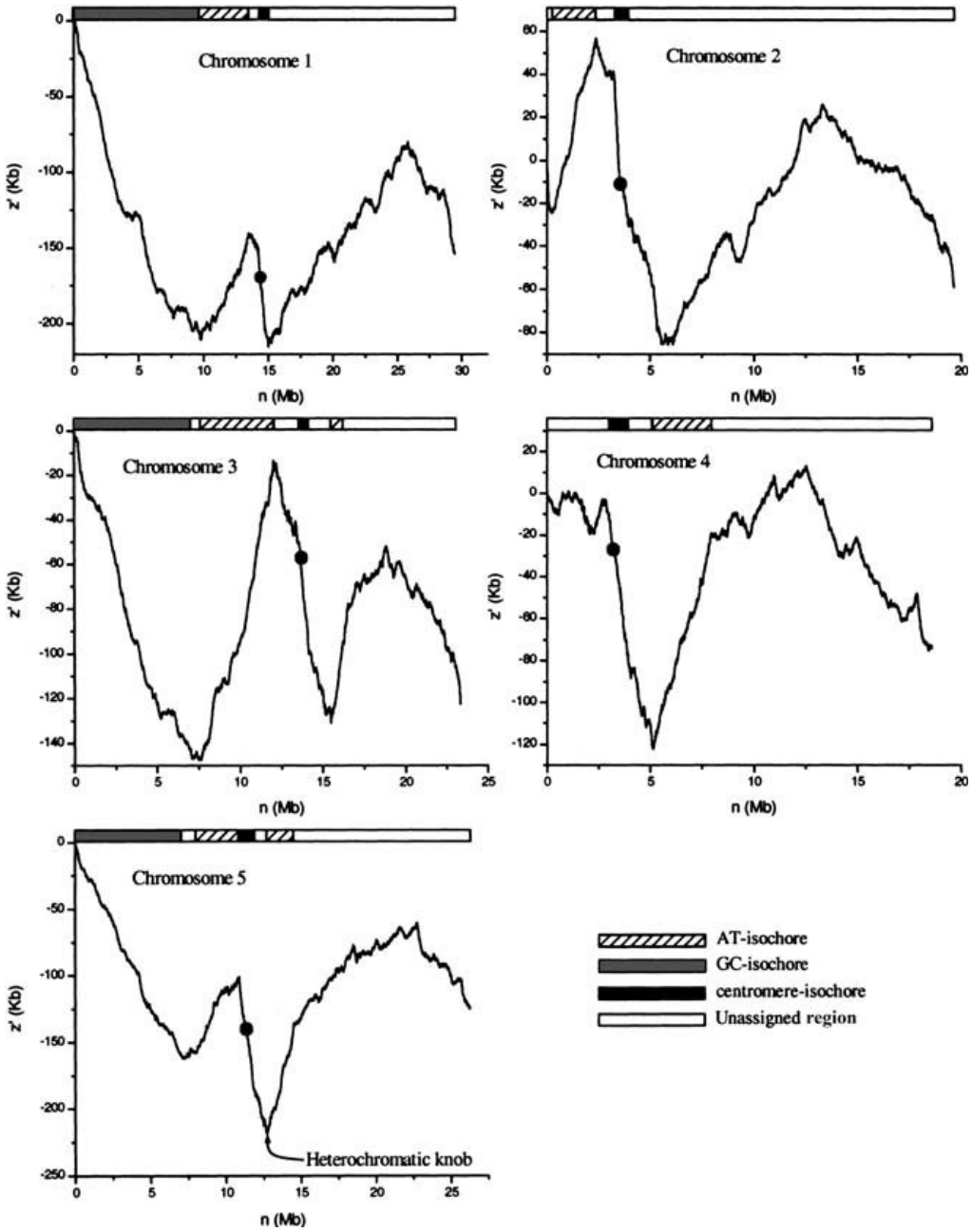


Figure 3: Isochore structures of the 5 chromosomes of *Arabidopsis thaliana* revealed by a window-less approach*(23)*.

Together with the isochore structure, under or over-utilization of particular dinucleotides has been for long time considered the possible results of a genomic signature(*24*). Dinucleotides composition is the response of the whole genome to many different selective pressures and influence the overall nucleotide concentration of both coding and not coding sequences. Moreover dinucleotide biases proved to be independent from GC isochores or from general compositional fractions(*25*). Gentles and co-workers elaborated a dinucleotide relative abundance index which allowed to quantify the signature of DNA sequences for several eukaryotes and used it in order to compare different organisms(*26*). The results of this study revealed that the distance between the dinucleotide biases is highly correlated with the phylogenetic divergence of the analyzed species, with Arabidopsis being more similar to yeast and C. elegans than to human or mouse. Interestingly the genome signature of bacterial phages was found to be highly associated with the nature of the host, proportionally to the extent to which the phage uses the host-cell machinery(*27*). This discovery support the hypothesis that knowing the genome dinucleotide composition of the host organism may be necessary in order to design the tDNA used for plant transformation. Such a task is made difficult by the evidence that significant differences may exist between phylogenetically close organisms. Indeed while an under-representation of the dinucleotide CG was observed in dicot (together with vertebrates and many prokaryotes) but not in monocot, and under-representation of TA seems to be a more general feature of plant genomes(*28*).

Among the possible causes of the dinucleotide genomic signature the transcription associated strand asymmetry seems to play a primary role. The observation of this phenomenon emerged by the realization that many organisms' genomes do not obey the Chargaff's second parity rule(*29*), that is the equimolarity, among the two DNA strands, of adenine + thymine and guanine + cytosine. Interaction of the polymerase enzyme with a specific strand during transcription leaves the second strand more accessible to mutational events such as the de-amination of cytosine(*30, 31*), and more exposed to transcription-coupled repair processes(*32*), This phenomenon was studied in Arabidopsis thaliana by calculating the GC and TA bias in the internal portions of introns which are less subjected to selective pressures comparing to other genic regions. A positive value for the TA skew and, at a lesser extent, for the GC skew, together with a significant difference with the values found in the intergenic regions underlines the presence of a transcription-associated strand asymmetry for this plant organism.

Also replication may be involved in the emergence of a strand asymmetry. Indeed during replication one (leading) strand interact with the DNA polymerase while the other (lagging) results more exposed to mutational and repair processes. This phenomenon is so emphasized in certain organisms that the GC and TA skews produce a clear genomic signal which can be used to reveal the replications origins(*33*, *34*). The co-existence of replication and transcriprion induced strand asymmetries has been proposed in plants with the former producing a weaker signal due to the shortness of replicons(*35*).

The presence of isochores, dinucleotide biases and replication/transcription strand asymmetries seems to highlight a DNA heterogeneity which is primarily ascribable to compositional differences between genomic regions. However Wong and co-worker showed that a large part of the nucleotide (e.g. GC) content variation occurs within the genes, and revealed a negative gradient along the direction of transcription(*36*). Moreover evidence of both codon and amino acid gradients led the author to ascribe this behavior to the co-existence of a transcription-related mutational bias together to a translation related selection. Differences emerged again between monocots and dicots with rice, but not Arabidopsis, showing a GC gradient also in the intronic regions.

Unequal bases distribution within the genes in plant organisms were also observed by Niimura and co-workers who revealed variation of the nucleotide biases in the codon third position across the transcripts of A. thaliana and O. sativa (Figure 4)(*37*).

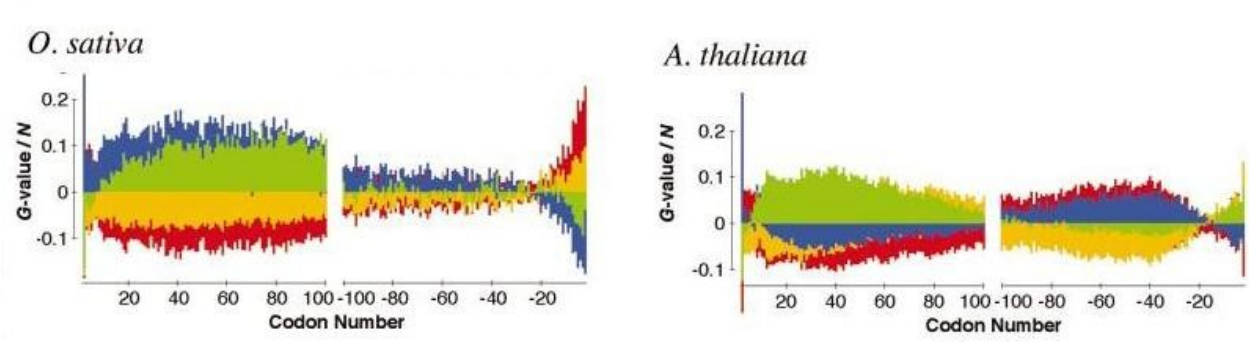


Figure 4: Biases at the third codon position in the use of A (red), T (yellow), G (blue) and C(green) for *Oryza sativa* and *Arabidopsis thaliana*. G-values over the effective number of codons were used for measuring the skews *(37)*.

The codon bias.

All amino acids, with the only exception of methionine and thryptophan, can be coded by 2 to 6 different codons. It is well known that within each amino acid class the use of these

(synonymous) codons results unequal in many species(*38*). The species-specific codon choice represents the so called "codon dialect"*(40)*. When designing the tDNA sequence of a heterelogous gene the choice of the host organisms preferred codons, that is the most used in the most abundant transcripts, proved to be an efficient way to improve the transgene expression(*39*).

One of the first complete surveys on the codon choice in higher plants was reported by Murrey and co-workers back in 1988(*40*). In this work the authors analyzed the sequences of 207 plant genes (mainly from the highly expressed ribulose 1,5 biphosphate small subunit and chlorophyll a and b binding proteins) and managed to highlight a sharp differences in terms of codon choice between monocots and dicots. On the other hand avoidance of CG and TA dinucleotide at the second and third codon position proved to be a common feature of these two species (also reviewed by Campbell et al(*41*)).

Although the pattern of codon usage seems to be associated to the phylogenetic distance(*42*), indicating a slow evolution of this phenomenon, a difference in the choice of synonymous codons was observed also among different genes within the same organism(*43*). As a matter of facts, two classes of genes were isolated by Chiapello according to their codon GC content in Arabidopsis(*44*). Genes with GC-rich coding sequences showed to encode the most abundant proteins of the plant cell, some of them were housekeeping and the corresponding orthologs in other species likewise featured a high codon bias. Contrarily AT-rich coding sequences belong to genes which are generally low expressed, often tissue specific and sometimes involved in the response to various stress conditions.

As previously observed monocots and dicots differ for the overall gene GC content with the former being GC rich and the latter being AT rich(*45*). Analysis of the nucleotide at the third (silent) codon position reflected this difference as demonstrated by a study carried out on 4 monocots and three dicots(*46*). In monocots C and G were always used in the third position of "preferred" (most abundant) codons for two-fold degenerate codon groups, while for four-fold degenerated aminoacid the C was detected more frequently. On the other hand the AT-rich monocots tends to have T at the third position of preferred codons (with the proline coding triplette CCA being the only exception). For both monocots and dicots a general excess of pyrimidines (C and T) over purines (G and A) was also observed*(46)*. Choice of the base at the third codon position seems to depend not only on the overall nucleotide composition of the analyzed organism but also on the 3' flanking position, that is the first nucleotide of the following codon. This phenomenon has been demonstrated by analyzing the two-fold degenerated aminoacids in Arabidopsis. For this organism codon NNC

is preferred (in highly and lowly expressed genes) when the following base is an A or a T. Contrarily NNT is preferred when the following base is a G(*47*).

Recently the codon usage of the world widespread crop *Zea mays* was investigated by analyzing the base content of 7402 cDNA(*48*). As previously observed for *Oryza sativa* an overall high GC content was showed for this organism with the most abundantly used codons ending preferably with either guanine or cytosine. Comparison of the synonymous relative codon usage (RSCU) between high and low expressed genes revealed the presence of 28 "optimal codons" which may provide useful information for maize gene-transformation and gene prediction.

Finally an analysis carried out on 6 plants (two monocots, two dicots, a moss, and a liverwort) revealed the presence of an unequal synonymous codon usage in mitochondrial genomes(*49*). Interestingly mitochondrial genes proved to be more conserved in terms of GC content. A weaker divergence among dicots and monocots and the absence of correlation between the GC3 and GC12 contents make mitochondrial genes peculiar comparing to nuclear ones in plants.

The transcribed non-coding regions.

The 5' and 3' untranslated regions (UTR) are located in the mature mRNA respectively downstream and upstream the coding sequences and play an important role in the expression, stability and accumulation of the transcript. The 5'-UTR is believed to enhance the translational efficiency by expediting the 40S ribosomal subunit migration and the recognition of the translational starting point(*50*). An analysis supported by mutagenesis experiment highlighted the existence of a preferred nucleotide context surrounding the initiation codon in *Nicothiana tabacum*(*51*). Taylor and co-wokers establish that gene Me1 of *Flaveria bidentis*, which code for an important member of the C4 pathway, owes its sheath cell specific accumulation to the 5' UTR sequence whereas the 3' UTR (and possibly an interaction between these two untranslated regions) determine an expression enhancement of almost 100 fold(*52*). The same author confirmed this hypothesis by expressing specifically and efficiently a reporter Gus genes conjugated with the UTR of AhRbcS1from the C4 dicot amaranth in a transgenic *Flaveria bidentis* system(*53*).

The hypothesis that the untranslated regions may play a role in maintaining an efficient gene expression under stress conditions such as heat shock and hypoxia has been recently demonstrated in *Nicothiana bethamiana*(*54*).

For many years introns, together with intergenic sequences, were regarded as junk DNA. However the original proteo-centric view is now being replaced by a model that is centered on the transcript(*55*) in which introns prove to play an important role in many aspects of the protein synthetic pathway(*56*). Indeed they revealed to be a repository of binding sites for transcription factors, therefore acting as classical transcriptional enhancer. They may release trans-acting messages such as microRNA and small molecular RNA. Finally codons are involved in the enhancement of the mRNA accumulation due to the formation of protein complexes in the proximity of the exon-intron junction (EJC) which are believed to facilitate the mRNA migration to the cytosol.

Nowadays the inclusion of heterologous introns in plants expression vectors is widely used to increase foreign gene expression in transgenic systems(*57*). The expression enhancement due to the presence of an intron in the primary transcript is called Intron Mediated Enhancement (IME). The mechanism that underlies the IME effect is complex and far to be totally understood. The expression enhancement, when observed, may vary from two to hundreds fold and can be associated to many factors such as the base composition of the intron and the flanking exon(*58*) the used reporter gene(*59*) together with its promoter(*60*) and the position of the intron(*61*) (although position independent expression enhancement was also observed). Proofs of evidence support the hypothesis that the IME may exert a post-trascriptional effect. This was confirmed by the observation that Arabidopsis plants transformed with a PAT1 intron bearing plasmid showed no increase in the transcription rate(61). Unequal IME effects were then observed among different species (i.e. monocot and dicot(*62*)) and between different tissues of the same organism (Intron Dependent Spatial Expression(*63*)).

Forces shaping the plants genomes.

As observed in many other organisms, genetic code of plants features a non random base distribution together with an unequal synonymous codon usage. But which are the cuases behind these biases?

The presence of signals involved in the DNA life cycle processes such as transcription and replication are known to shape the nucleotide content of many genomic portions including intergenic(*64*), untranslated regions (UTR)(*65*) and introns(*66*). The choice of certain nucleotides near the translation-initiation point proved to be crucial for many eukariotes(*67*). For example minimization of the mRNA stabilization at the transcript early position in order to avoid packed structures that may interfere with ribosome binding and with

the start codon recognition.(*68*) In eukaryotes the 5' termini of mature mRNA must also accommodate the Kozac sequence GCCACCaugG which is involved in the recognition of the first codon "aug"(*69*). Niimura and co-workers speculated that also the second and third bases of the second codon are under some kind of selection(*70*). Indeed in according with the authors the second codon is the most biased of the entire transcript with GCG being preferred in the plant species Arabidopsis and *Oryza*. This may be possibly due to a specific extension of the Kozac sequence in eukaryotes genomes. On the other hand a direct interaction of these regions with the chemical environment has also been reported(*71*).

The splicing machinery requires the use of particular sequences at the exon-intron junction which will necessarily influence the codon choice in these genic regions. In fact Willie and co-workers observed a rapid variation of the nucleotide content near the splice sites in Arabidopsis(*72*). Interestingly while in eukaryotes such as human and *Drosophyla* the interior exons showed a constant nucleotide content, in Arabidopsis monotonically increasing (decreasing) frequency for all the nucleotides were observed (Figure 5).

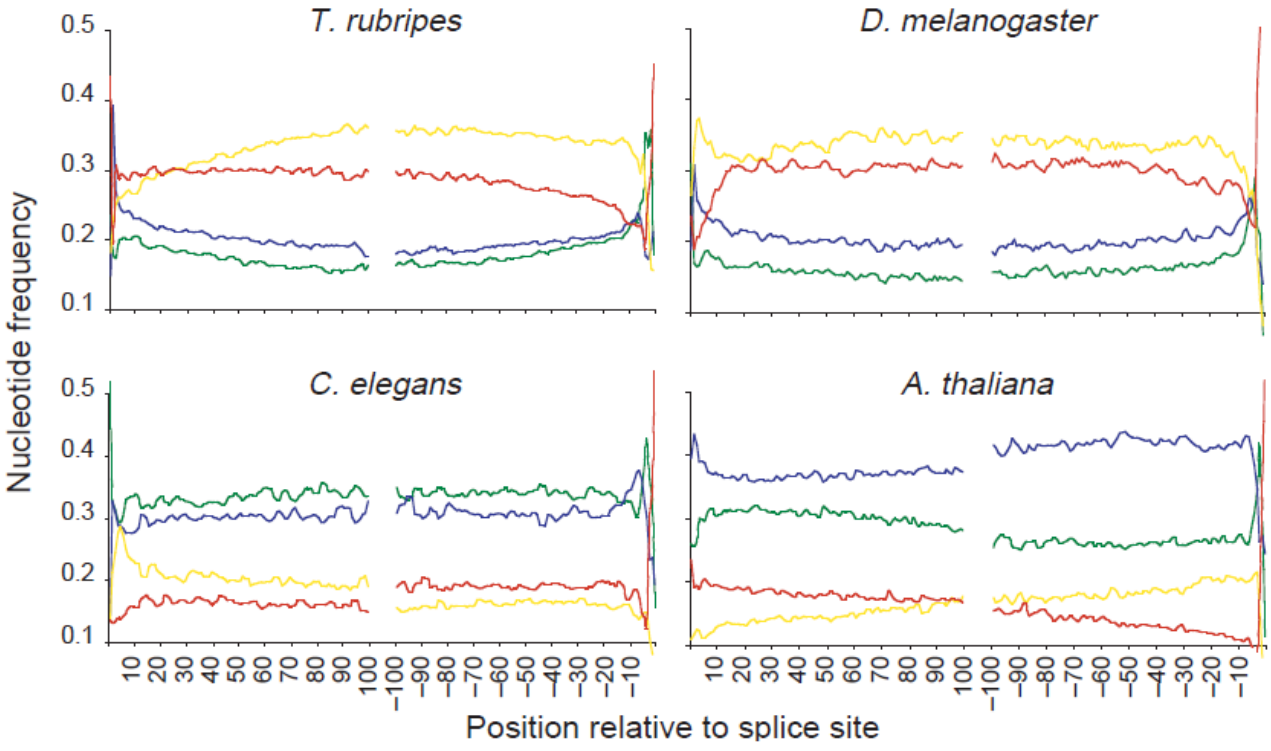


Figure 5: Third position nucleotide frequencies near the splice site for 4 eukaryotes (A, green; C, yellow; G, red; T, blue)*(72)*.

When considering the totality of the transcriptome a relatively small portion of codons are involved in satisfying the structural and functional requirements just described. Nevertheless the unequal usage of the synonymous codons represents a general trend in plant genomes. The reasons behind the codon bias represent a hot topic for scientific community involved in the study of plant evolution. In many eukaryotes synonymous codon usage is thought to be due to a combination of three factors: mutation, genetic drift and translational selection(*73*). Differences in codon bias among species may be due to strength and direction of these three forces(*74*). The sign of translational selection was extensively observed in Arabidopsis genes. As a matter of fact the non- random distribution of the synonymous codons proved to be correlated with gene expression level, revealing a stronger selection for those codons that can improve efficiency and/or accuracy in highly expressed genes(*75, 76*). Moreover the most used codons in highly expressed genes match those expected by the most abundant tRNA genes(*77*). On the other hand the evolution rate of translational selection seems to be weak comparing to other factors. When investigating the reasons behind the differences in codon usage of the recently diverged Brassicaceae *A. thaliana* and *B. oleraceae* (estimated divergence time approx. 25 MYA(*78*)), Wright and co-workers observed that the differentiation was mainly due to a different mutational pressure on these two organisms(*79*). The predominant role of mutation over selection was suggested also for *Oryza sativa* by Wang and Hickey(*80*). The authors argue that after the divergence between monocots and dicots (200 MYA) the formers experienced a more intense mutational pressure which increased their overall GC content. The consequent nucleotide skew can alone explain the bias in the synonymous codon usage masking the possible effect of selection (which anyway exists as recently reported for rice(*81*) and maize(*82*)). This does not happen in Arabidopsis where the absence of a strong mutational bias allows the detection of the still weak translational selection sign. Divergence in the intensity of the mutational pressure can actually explain the differences of the most abundant codons in monocots and dicots, with the former using preferably GC ending codons while the latter preferring AT-ending codons. Interestingly, in spite of these differences, when comparing correlation of codon usage and level of expression in *Zea mays* and Arabidopsis thaliana the trend for these two species is highly similar (Figure 6)(*83*). This shows that not only mutation and selection act with different strength but also in different directions.

A sign of selection was observed also in a long lived perennial plant like *Populus*(*84*). 25 codons were identified as preferred in high expressed genes of 5 lineage of this species, underlining the presence of an ancient common selection. On the other hand the species P.

tremula showed a marginal divergence revealing a higher rate of unpreferred to preferred mutations. Although this behavior seems evocative of a different selection rate, such an event seems unlikely and can be better explained by an in increase of the *P tremula* population size(*85*).

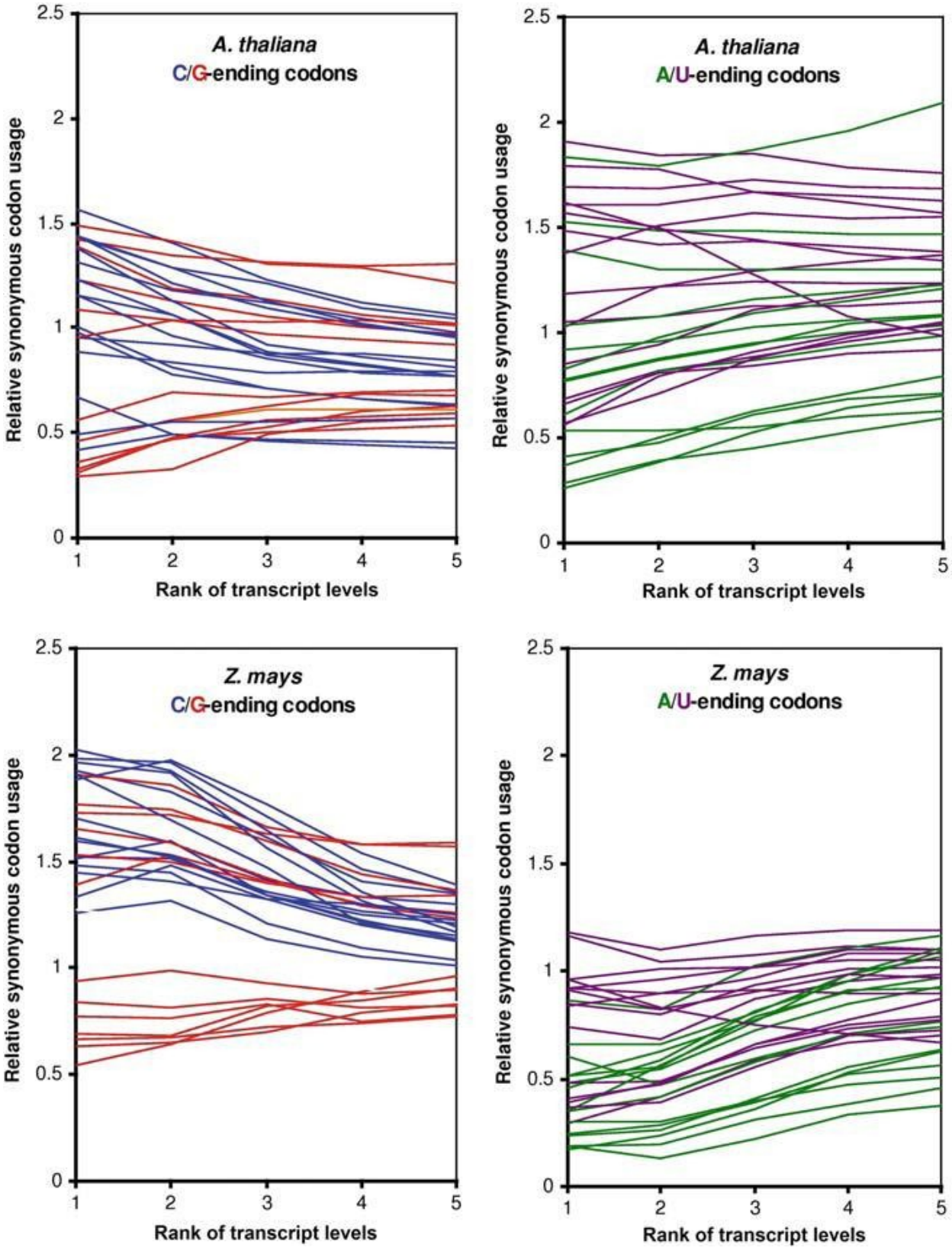


Figure 6: Codon usage variation with the expression level in Arabidopsis thaliana and Zea mays.(*83*)

Evidence of translational selection emerged from the analysis of the primitive moss plant *Physcomitrella patens*(*86*). This species has, on average, a higher GC content (50%) compared to *Arabisopsis* and, interestingly, a relatively elevate number of not biased transcripts (~15%), revealing a more ancient nucleotide composition (Figure 7)(*87*).

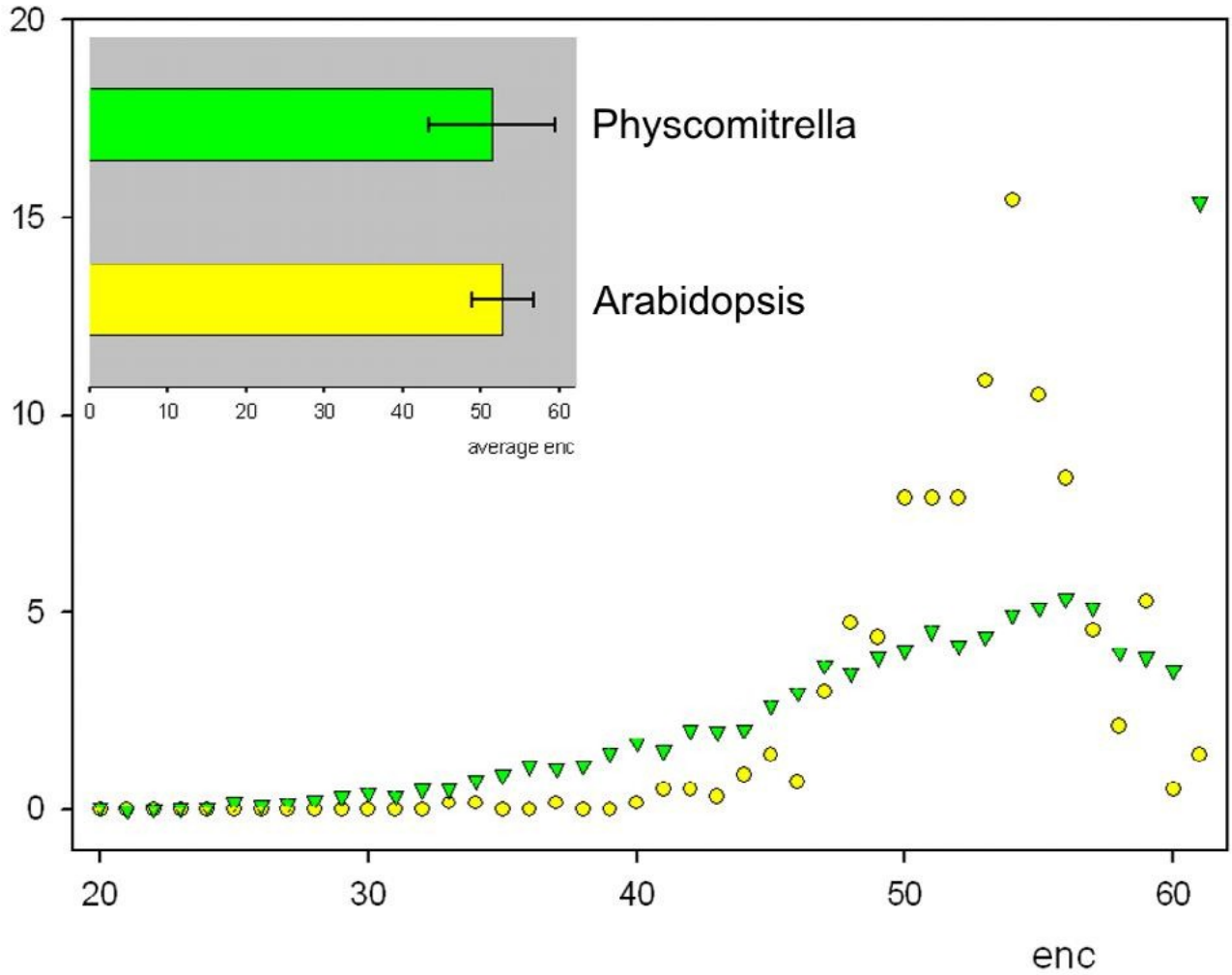


Figure 7: Average and distribution of the codon bias, measured by the effective number of codons (enc) in *Arabidopsis* and *Physcomitrella*.(*87*)

Regardless the mechanisms behind the molecular evolution in plants, it is interesting to note that different gene classes can experience unequal pressures. Analysis of synonymous substitution patterns between rice and *Arabidopsis* homologous pairs revealed that housekeeping genes are generally more conserved than tissue-specific ones(*88*). Moreover

highly expressed genes in housekeeping feature a lower substitution rate when compared to lowly expressed whereas an opposite trend is detected for tissue specific genes. Together these observations led to the conclusion that the translational selection is associated with both expression level and breadth, but some other selective forces may determine the synonymous codon choice in highly expressed tissue-specific genes. An additional constraint is represented by the stabilization of the mRNA helix which was found to be positively correlated with the codon usage but only in tissue-specific genes*(88)*.

If the usage of a certain codon set can be considered as a species specific feature, its role in differentiating several tissues within the same organism is worthy to be mentioned. Recently Whittle and co-workers observed a divergence in the usage of synonymous codons between male and female tissues of *Z. mays*, *T. aestivum* and *B. napus*(*89*), confirming a previously observed gender-specific sign of selection in plants(*90*). Indeed a higher level of codon bias, together with an increase of the GC content, was observed in eggs as compared to sperm of Z. mays, and for genes expressed in ovary as compared to anther in *T. aestivum*. Absence of a significant difference in gene length and function between male and female tissues provided evidence that codon usage is altered by a gender-specific selection in plants.

Conclusions.

Many factors are involved in shaping the genome composition of plants and the equilibrium of several evolutionary forces seems to be at the basis of the species specific structure. Revealing the preferred codon usage has always been considered of primary importance since it allows to improve the efficiency of plants transformation by adapting the tDNA codon to those of the host organism. However this approach is not always successful and it is difficult to understand the reasons behind the non-expression of a trangene since a survey of the “failures” is not reported in literature. Certainly the use of particular codons can not be considered independent from the context in which they are integrated. As a matter of fact it has been demonstrated that 90% of the codons are dependent on the first base of the following triplette and more than 50% is not in accordance with the tetranucleotides presence at a genome level in Arabidopsis(*91*). Moreover particular emphasis should be placed also on the reasons behind the use of non-preferred (rare) codons. Are they the consequence of a random mutational pressure which has not been compensated yet by the slower/weaker translational selection? Although at the best of our knowledge the rare codons usage has not been studied yet in plants, it was found to be associated to pausing events during the

translational process that proved to be essential for the correct folding of the nascent protein in other eukaryotes(*92*).

Reference List

1. A. M. Showalter, S. Heuberger, B. E. Tabashnik, Y. Carriere, B. Coates, *J. Insect Sci.* **9**, 22 (2009).

2. M. Luo *et al.*, *Int. J. Mol. Sci.* **10**, 1896 (2009).

3. M. Pedotti *et al.*, *J. Biol. Chem.* **284**, 36415 (2009).

4. Y. F. Hong *et al.*, *Plant Mol. Biol.* **67**, 347 (2008).

5. C. R. Wang *et al.*, *Planta* **227**, 1127 (2008).

6. D. E. Nelson *et al.*, *Proc. Natl. Acad. Sci. U. S. A* **104**, 16450 (2007).

7. J. K. Ma *et al.*, *EMBO Rep.* **6**, 593 (2005).

8. Z. Lipman. Genetic key discovered to dramatically increase yields and improve taste of hybrid tomato plants. e! Science News - Biology and Nature . 2010.
Ref Type: Magazine Article

9. P. N. Desai, N. Shrivastava, H. Padh, *Biotechnol. Adv.* **28**, 427 (2010).

10. J. Finnegan, D. McElroy, *Nature Biotechnology* **12**, 883 (1994).

11. P. Elomaa *et al.*, *Mol. Gen. Genet.* **248**, 649 (1995).

12. C. M. Rommens, M. A. Haring, K. Swords, H. V. Davies, W. R. Belknap, *Trends Plant Sci.* **12**, 397 (2007).

13. M. Welch, A. Villalobos, C. Gustafsson, J. Minshull, *J. R. Soc. Interface* **6 Suppl 4**, S467 (2009).

14. G. Bernardi, *Annu. Rev. Genet.* **29**, 445 (1995).

15. G. Matassi, L. M. Montero, J. Salinas, G. Bernardi, *Nucleic Acids Res.* **17**, 5273 (1989).

16. L. M. Montero, J. Salinas, G. Matassi, G. Bernardi, *Nucleic Acids Res.* **18**, 1859 (1990).

17. J. Salinas, G. Matassi, L. M. Montero, G. Bernardi, *Nucleic Acids Res.* **16**, 4269 (1988).

18. V. L. Chandler, V. Brendel, *Plant Physiol* **130**, 1594 (2002).

19. Arabidopsis Genome Initiative, *Nature* **408**, 796 (2000).

20. S. Ouyang *et al.*, *Nucleic Acids Res.* **35**, D883 (2007).

21. J. L. Oliver, P. Bernaola-Galvan, P. Carpena, R. Roman-Roldan, *Gene* **276**, 47 (2001).

22. R. Zhang, C. T. Zhang, *J. Mol. Evol.* **59**, 227 (2004).

23. N. Carels, G. Bernardi, *Genetics* **154**, 1819 (2000).

24. S. Karlin, *Curr. Opin. Microbiol.* **1**, 598 (1998).

25. G. J. Russell, P. M. Walker, R. A. Elton, J. H. Subak-Sharpe, *J. Mol. Biol.* **108**, 1 (1976).

26. A. J. Gentles, S. Karlin, *Genome Res.* **11**, 540 (2001).

27. B. E. Blaisdell, A. M. Campbell, S. Karlin, *Proc. Natl. Acad. Sci. U. S. A* **93**, 5854 (1996).

28. S. Karlin, C. Burge, *Trends Genet.* **11**, 283 (1995).

29. E. CHARGAFF, *Fed. Proc.* **10**, 654 (1951).

30. A. Beletskii, A. S. Bhagwat, *Biol. Chem.* **379**, 549 (1998).

31. M. P. Francino, H. Ochman, *Mol. Biol. Evol.* **18**, 1147 (2001).

32. J. Q. Svejstrup, *Nat Rev. Mol. Cell Biol.* **3**, 21 (2002).

33. A. Grigoriev, *Nucleic Acids Res.* **26**, 2286 (1998).

34. A. Zawilak *et al.*, *Nucleic Acids Res.* **29**, 2251 (2001).

35. D. K. Niu, K. Lin, D. Y. Zhang, *J. Mol. Evol.* **57**, 325 (2003).

36. G. K. Wong *et al.*, *Genome Res.* **12**, 851 (2002).

37. Y. Niimura, M. Terabe, T. Gojobori, K. Miura, *Nucleic Acids Res.* **31**, 5195 (2003).

38. S. Aota, T. Gojobori, F. Ishibashi, T. Maruyama, T. Ikemura, *Nucleic Acids Res.* **16 Suppl**, r315 (1988).

39. A. Fuglsang, *Protein Expr. Purif.* **31**, 247 (2003).

40. E. E. Murray, J. Lotzer, M. Eberle, *Nucleic Acids Res.* **17**, 477 (1989).

41. W. H. Campbell, G. Gowri, *Plant Physiol* **92**, 1 (1990).

42. T. Ikemura, *Mol. Biol. Evol.* **2**, 13 (1985).

43. C. Medigue, T. Rouxel, P. Vigier, A. Henaut, A. Danchin, *J. Mol. Biol.* **222**, 851 (1991).

44. H. Chiapello, F. Lisacek, M. Caboche, A. Henaut, *Gene* **209**, GC1 (1998).

45. N. Carels, G. Bernardi, *Genetics* **154**, 1819 (2000).

46. A. Kawabe, N. T. Miyashita, *Genes Genet. Syst.* **78**, 343 (2003).

47. B. R. Morton, S. I. Wright, *Mol. Biol. Evol.* **24**, 122 (2007).

48. H. Liu *et al.*, *Mol. Biol. Rep* **37**, 677 (2010).

49. M. Zhou, X. Li, *Mol. Biol. Rep* **36**, 2039 (2009).

50. M. G. Koziel, N. B. Carozzi, N. Desai, *Plant Mol. Biol.* **32**, 393 (1996).

51. D. R. Gallie, D. E. Sleat, J. W. Watts, P. C. Turner, T. M. Wilson, *Nucleic Acids Res.* **15**, 8693 (1987).

52. S. Ali, W. C. Taylor, *Plant Mol. Biol.* **46**, 251 (2001).

53. M. Patel *et al.*, *Plant Physiol* **136**, 3550 (2004).

54. E. S. Mardanova, L. A. Zamchuk, N. V. Ravin, *Mol. Biol. (Mosk)* **41**, 1002 (2007).

55. T. R. Gingeras, *Genome Res.* **17**, 682 (2007).

56. L. Morello, D. Breviario, *Curr. Genomics* **9**, 227 (2008).

57. A. H. Christensen, P. H. Quail, *Transgenic Res.* **5**, 213 (1996).

58. K. R. Luehrsen, V. Walbot, *Mol. Gen. Genet.* **225**, 81 (1991).

59. N. Rethmeier, J. Seurinck, M. M. Van, M. Cornelissen, *Plant J.* **12**, 895 (1997).

60. V. Vasil, M. Clancy, R. J. Ferl, I. K. Vasil, L. C. Hannah, *Plant Physiol* **91**, 1575 (1989).

61. A. B. Rose, R. L. Last, *Plant J.* **11**, 455 (1997).

62. A. Tanaka *et al.*, *Nucleic Acids Res.* **18**, 6767 (1990).

63. S. Giani, A. Altana, P. Campanoni, L. Morello, D. Breviario, *Transgenic Res.* **18**, 151 (2009).

64. K. Singh, R. C. Foley, L. Onate-Sanchez, *Curr. Opin. Plant Biol.* **5**, 430 (2002).

65. S. Ali, W. C. Taylor, *Plant Mol. Biol.* **46**, 251 (2001).

66. H. H. Le, B. Seraphin, *Cell* **133**, 213 (2008).

67. W. Gu, T. Zhou, C. O. Wilke, *PLoS Comput. Biol.* **6**, e1000664 (2010).

68. S. Bhat *et al.*, *Plant Mol. Biol.* **56**, 761 (2004).

69. M. Kozak, *Nucleic Acids Res.* **12**, 857 (1984).

70. Y. Niimura, M. Terabe, T. Gojobori, K. Miura, *Nucleic Acids Res.* **31**, 5195 (2003).

71. J. L. Ortega *et al.*, *Plant J.* **45**, 832 (2006).

72. E. Willie, J. Majewski, *Trends Genet.* **20**, 534 (2004).

73. H. Akashi, *Gene* **205**, 269 (1997).

74. J. R. Powell, E. Sezzi, E. N. Moriyama, J. M. Gleason, A. Caccone, *J. Mol. Evol.* **57 Suppl 1**, S214 (2003).

75. L. Duret, D. Mouchiroud, *Proc. Natl. Acad. Sci. U. S. A* **96**, 4482 (1999).

76. S. I. Wright, B. Lauga, D. Charlesworth, *Mol. Biol. Evol.* **19**, 1407 (2002).

77. S. I. Wright, C. B. Yau, M. Looseley, B. C. Meyers, *Mol. Biol. Evol.* **21**, 1719 (2004).

78. M. A. Koch, B. Haubold, T. Mitchell-Olds, *Mol. Biol. Evol.* **17**, 1483 (2000).

79. S. I. Wright, G. Iorgovan, S. Misra, M. Mokhtari, *J. Mol. Evol.* **64**, 136 (2007).

80. H. C. Wang, D. A. Hickey, *BMC Evol. Biol.* **7 Suppl 1**, S6 (2007).

81. X. Guo, J. Bao, L. Fan, *FEBS Lett.* **581**, 1015 (2007).

82. H. Liu *et al.*, *Mol. Biol. Rep* **37**, 677 (2010).

83. L. Wang, M. J. Roossinck, *Plant Mol. Biol.* **61**, 699 (2006).

84. P. K. Ingvarsson, *BMC Evol. Biol.* **8**, 307 (2008).

85. R. M. dos, L. Wernisch, *Mol. Biol. Evol.* **26**, 451 (2009).

86. H. K. Stenoien, *Heredity* **94**, 87 (2005).

87. S. A. Rensing, D. Fritzowsky, D. Lang, R. Reski, *BMC Genomics* **6**, 43 (2005).

88. P. Mukhopadhyay, S. Basak, T. C. Ghosh, *DNA Res.* **15**, 347 (2008).

89. C. A. Whittle, M. R. Malik, J. E. Krochko, *BMC Genomics* **8**, 169 (2007).

90. C. Seoighe, C. Gehring, L. D. Hurst, *PLoS Genet.* **1**, e13 (2005).

91. A. Fedorov, S. Saxonov, W. Gilbert, *Nucleic Acids Res.* **30**, 1192 (2002).

92. C. H. Makhoul, E. N. Trifonov, *J. Biomol. Struct. Dyn.* **20**, 413 (2002).

Mutational biases and selective forces shaping the structure of Arabidopsis genes[†]

ABSTRACT

Recently features of gene expression profiles have been associated with structural parameters of gene sequences in organisms representing a diverse set of *taxa*. The emerging picture indicates that natural selection, mediated by gene expression profiles, has a significant role in determining genic structures. However the current situation is less clear in plants as the available data indicates that the effect of natural selection mediated by gene expression is very weak. Moreover, the direction of the patterns in plants appears to contradict those observed in animal genomes.

In the present work we analized expression data for >18000 *Arabidopsis* genes retrieved from public datasets obtained with different technologies (MPSS and high density chip arrays) and compared them with gene parameters.

Our results show that the impact of natural selection mediated by expression on genes sequences is significant and distinguishable from the effects of regional mutational biases. In addition, we provide evidence that the level and the breadth of gene expression are related in opposite ways to many structural parameters of gene sequences. Higher levels of expression abundance are associated with smaller transcripts, consistent with the need to reduce costs of both transcription and translation. Expression breadth, however, shows a contrasting pattern, i.e. longer genes have higher breadth of expression, possibly to ensure those structural features associated with gene plasticity. Based on these results, we propose that the specific balance between these two selective forces play a significant role in shaping the structure of *Arabidopsis* genes.

Please see Supplemental Material for Tables and Figures that are not reported within the text

[†]Published in PLoS One. 2009 Jul 27;4(7):e6356

INTRODUCTION

Several studies conducted in organisms as diverse as humans, *Drosophila melanogaster* and *Caenorhabidtis elegans*, have demonstrated that there is an inverse relation between levels of gene expression and a variety of sequence parameters such as the length of coding sequence, intron number and length*(1-3)*. These patterns have inspired the energetic cost hypothesis under which natural selection would favour shorter transcriptional units to minimize time and cost of gene expression *(1)*. However, alternative interpretations have not been ruled out. For example, the *genomic by design* model postulates that the activity of a gene is a key determinant of its structure. According to this hypothesis, genes that work in only a few tissues and thus require a high level of epigenetic regulation, have a structure particularly rich for non coding sequences to host the necessary regulatory elements *(3,4)*. The observation that tissue specific human genes have more introns and longer intergenic sequences than broadly expressed ones lends credit to this proposition *(4,5)*. A third hypothesis points toward a local mutational bias as the main force controlling for gene structure. For example, Urrutia and Hurst *(2)* have suggested that broadly expressed human genes are positioned in genomic regions more prone to deletions and for this reason are shorter than tissue specific ones. The mutational bias could be even more focused on gene sequences. Highly abundant transcripts would be more prone to reverse transcription and retroposition and this would explain the lower density of intronic sequences in highly transcribed regions *(6)*.

This debate has recently experienced a puzzling turn due to the contradictory results presented for plant genomes. Seoighe et al. have shown that genes expressed in the *Arabidopsis* male gametophyte have shorter introns than genes that are expressed exclusively in the sporophyte. This observation provided the first evidence of a molecular signature of strong gametophytic selection and was considered in agreement with the energetic cost hypothesis *(7)*. Yet, subsequent studies in *Arabidopsis* and rice genomes have shown that the relationships between gene expression level and length in plants depict just the opposite trend: highly expressed genes are the least compact having more and longer introns than lowly expressed ones *(8,9)*. Given these contradictory results, it is not possible to establish to what extent selective pressures in plants differ from previously studied metazoan genomes. To address this issue we compared estimates of gene expression against several gene characteristics. For this purpose we assembled expression data from publicly available oligo-array *(10)* and MPSS (Massively Parallel Signature Sequencing) experiments *(11)*. Our results provide evidence that expression level and breadth are related to gene sequence characteristics

in opposite ways. Higher levels of gene expression are associated with smaller transcripts whereas greater breadth of expression correlates with longer transcripts. The balance between these two selective forces represents a significant determinant of the structure of *Arabidopsis* genes.

METHODS

Sequence information

Sequence information for protein coding genes was obtained from TAIR 8 annotations (ftp://ftp.arabidopsis.org/home/tair/Sequences/). Transposons, pseudogenes, plastid and mitochondrial genes were filtered out from the data set. Although 23528 fully annotated transcript entries were available for the analysis, data for each structural parameter were not obtained for all the genes and therefore the actual number of genes used in comparisons varied as indicated in the text. In all cases in which more than one alternative transcript was predicted, the longest was analysed.

Expression data

We used gene expression data from two web sources. Oligo array data were retrieved from the http://www.ncbi.nlm.nih.gov/sites/entrez (see supplemental table S1 for a complete list of experiments considered) *(10)*. They represent the results of oligonucleotide array experiments performed uniformly with a total of 77 developmental stages belonging to different organs*(10)*. The signals from probes on the chip corresponding to the same gene were normalized using the RMAexpress software*(12)* (also the replicates representing the same developmental stages were averaged. A gene was regarded as expressed if its signal level exceeded a conservative threshold of 75 average difference value.

MPSS data were retrieved from http://mpss.udel.edu/at/ (See Table S2 for a complete list of the experiments considered) and only tags matching a single gene were considered as described by Meyers et al. *(11)*. The five libraries analysed represented five organs: leaves, roots, germinating seedlings, flowers and siliques.

Measures of gene expression

Gene expression profiles were measured in two ways: i) breadth and ii) level of expression. The *breadth of expression* of a gene (EB) takes into account the number of experimental units (organs or developmental stages) in which a gene is expressed. The level

of expression of a gene provides a quantitative estimation of mRNA accumulation in a experimental unit and was obtained in several ways: the peak of expression (pE) which represents the highest value of expression of a gene across all experimental units, and the mean level of expression calculated taking into account all experimental units (hereafter referred to as A) or, only those in which the expression is not zero (hereafter designed as informative experimental units I). In addition, two possible definitions of experimental units were used. In the (O) organ based definition the data from developmental series were pooled together, while in the (DS) definition each developmental stage of an organ represented a unique experimental unit.

By combining expression variables and experimental unit definitions six different methods for measuring gene expression were established. For a complete list of measures of gene expression see table S3 in supplemental material.

DS methods: Each developmental stage represented an experimental unit and expression measures for each gene were averaged either considering all experimental units (DS-A) or only informative ones (DS-I):. DS-pE represents the peak of expression identified taking into account all DS experimental units.

O methods: Each organ represented an experimental unit. Expression measures for a given gene were calculated by averaging the values of either all organs (O-A) or only those with expression different from zero (O-I). O-pE is the peak of expression: i.e the highest value among O-defined experimental units

Expression breadth was calculated considering DS experimental units (DS-EB) or organ defined experimental unit (O-EB).

Statistical analyses

Indexes of expression levels and genic parameters were log transformed prior to analysis. Pearson' s (parametric) correlations (r) were calculated on single genes data and graphs were drawn binning by expression (bin = 5% of the dataset). Multiple regression analyses were performed on standardized variables.

RESULTS

If selection is acting on gene sequences to maximize expression efficiency, we might expect to find a relationship between gene expression profiles and some descriptions of

sequence parameters. Expression profiles were reduced to discrete measures considering all (A-methods) or only informative (I-methods) experimental units which were based on single observations (DS) or organ series (O). Three descriptive variables of gene expression profiles were considered: the average level of expression (EL), the peak of expression (pE) and the breadth of expression (EB).

To study information redundancy we applied cluster analysis on multivariate correlations between estimates obtained with different methods. It is worth to underline here that, in principle, the two averaging methods (A versus I) are expected to have a different propensity for representing quantitative differences between expression profiles. In facts, because the A averaging methods minimizes the weight of single data, A-based measures are expected to be more dependent on expression breadth*(2)*.

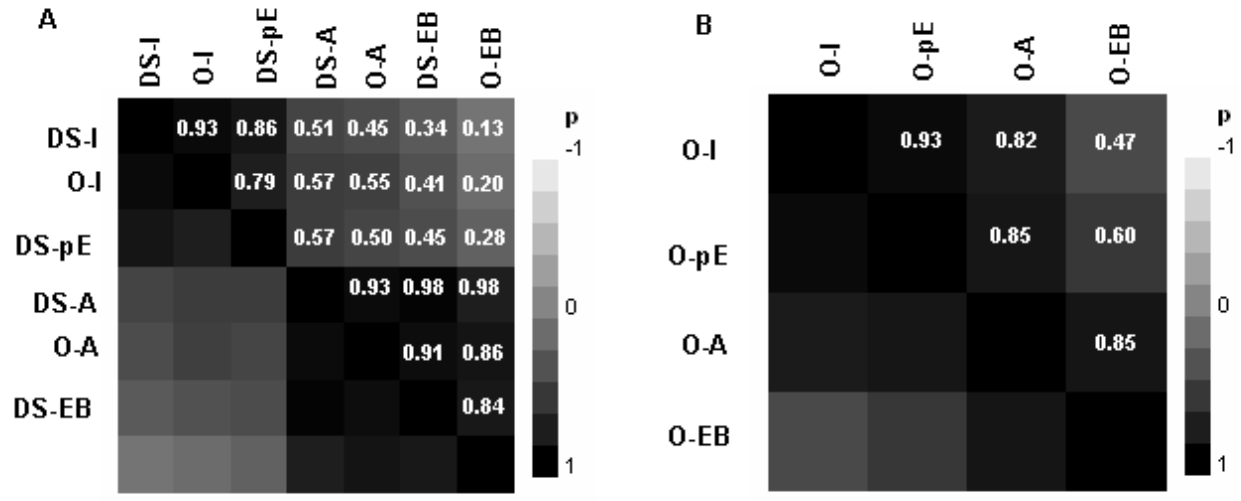


Figure 1: Cluster on the correlations (Pearson's *r*) among the different measures of expression considered both for (A) oligo-array and (B) MPSS. The experimental unit is represented by the developemental stage in the DS method and by the Organ in the O method. Expression profiles were reduced to discrete measures considering all (A-methods) or only informative (I-methods) experimental units.

As a matter of fact, for oligo-array (Figure 1-a), the way of collecting data (A vs I) discriminated between EL methods more than the definition of experimental units (DS vs O). As expected, we found that expression breadth measures were closer to EL measures with A definition, whereas the pE measure was positioned between I and A-EL (Figure 1-a). The results from MPSS (Figure 1-b) were similar, though of a lower resolution.

To represent the whole range of variation we considered here the most independent measures: DS-I and DS-EB for oligo-array and O-I and O-EB methods for MPSS. For

complete information on all six methods the reader is addressed to supplemental tables (table S3).

The length of genic and intergenic regions

Time and costs of transcription are proportional to the length and amount of the transcript that is produced. Thus, according to the selection for energy cost hypothesis, highly expressed genes are likely to experience greater selective pressure for a reduction in transcript length*(1)*.

	Expression level		***Expression breadth***	
	Microarray	*MPSS*	*Microarray*	*MPSS*
	DS-I	O-I	DS-EB_w0	O-EB_w0
Gene characteristics	*r*	*r*	*r*	*r*
Gene length	-0.142*	-0.064*	0.136*	0.108*
5' UTR length	-0.023†	0.030‡	0.154*	0.125*
CDS length	-0.164*	-0.125*	0.014	0.014ns
Intron length	-0.022†	0.018§	0.174*	0.124*
Intron length_w0	-0.102*	0.014ns	0.152*	0.141*
3' UTR length	0.074*	0.141*	0.202*	0.178*
Number of exons	-0.079*	-0.015	0.132*	0.099*
Average exon length	-0.028*	-0.061*	-0.159	-0.084*
Number of introns	-0.079*	-0.015*	0.132*	0.099*
Number introns_w0	-0.099*	-0.016ns	0.090*	0.099*
Average intron length	0.008ns	0.045*	0.146*	0.042*
Average intron length_w0	0.003ns	0.050*	0.060*	0.044*
Intron Density	0.007ns	0.078*	0.188*	0.124*
Intron Density_w0	0.004ns	0.078*	0.151*	0.124*

Table 1: Correlation between structural genomic parameters and both gene expression and expression. Data are presented both for oligo-array and MPSS assays. Statistical significances: ns = not significant; §P<0.05 †P<0.01; ‡P<0.001;*P<0.0001. Sample size of the correlations were comprised between n=12051 and n=14236. **Note1.** For this calculations, gene for which expression was zero were not considered to avoid potential problem arising from defective probes (Eisenberg and Levanon, 2003) . **Note2.** For introns parameters, data are presented both considering and excluding (w0) genes without introns.

We found that indeed expression levels and primary transcript length were negatively correlated (e.g. oligo-array DS-I: r= -0.142 n=14236 P<0.0001) and this was common to almost every transcript component indicating the presence of a generalized effect. In fact, total intron, 5' UTR and, noticeably, also CDS lengths were negatively correlated to expression level (Table 1 and Figure 2). A significant tendency towards reduction was found also for the total number of introns per gene (e.g. oligo-array DS-I r= -0.079 n=14236 P<0.0001; for details see Table 2). The only exception to such a scenario was the positive

correlation between 3' UTR and expression level (e.g. oligo-array DS-I r=0.074 n= 14236 P< 0.0001).

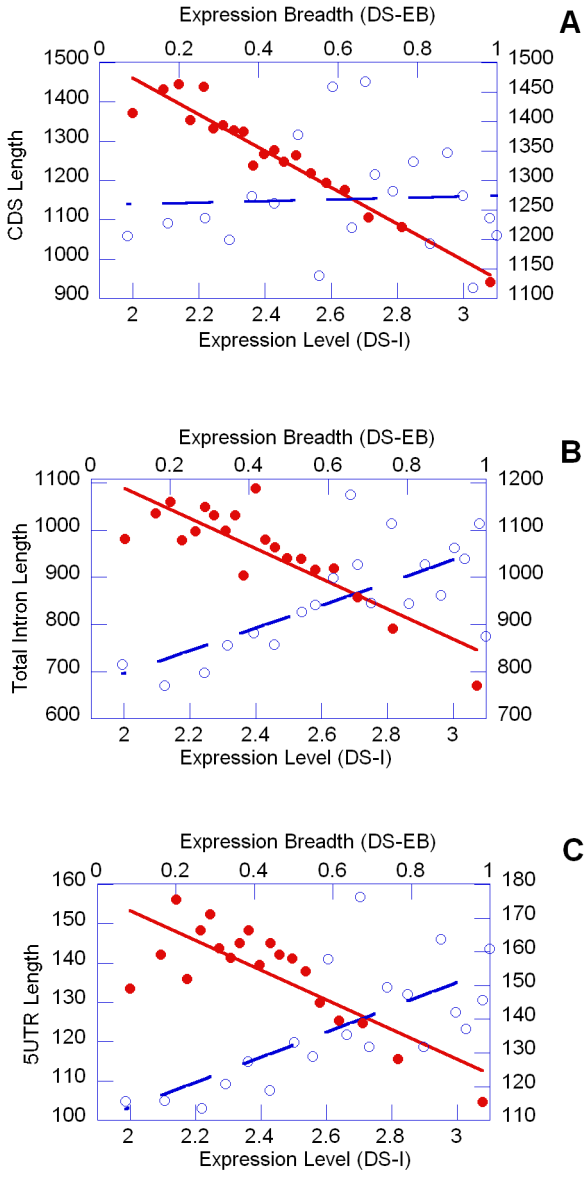


Figure 2: Relationship between the expression profile and (a) the CDS length, (b) the Total Intron Length and (c) 5utr Length. The points indicated by the symbol ● and interpolated by the continuous line refer to the expression level, whereas the points indicated by the symbol ○ and interpolated by a dashed line refer to the expression breadth. In both cases 14236 genes were considered which have been grouped in bins, each representing the 5% of the whole dataset.

The picture emerging from MPSS data was less clear as expected given the lower resolution of the data. Both CDS and total intron length followed the pattern toward reduction

(see Table 1). Intron number was not significantly correlated with expression level and 5'UTR, joined 3'UTR in marking the opposite trend toward expansion (Table 1).

In addition we studied a measure of intron density defined as the number of introns per kb of CDS*(13)*. Expression level was positively correlated to intron density for MPSS experiments and not correlated for oligo-array experiments (Table 1). This result is not caused by low expression of intron less genes because a similar trend was observed also when only intron containing genes were considered in the analysis.

EB measures for both oligo-array and MPSS were positively correlated to transcript length and this tendency was conserved also for 5' UTR, 3' UTR and total intron length and number. Conversely the CDS was not correlated to EB either in oligo-array or in MPSS data. As expected intron density was positively correlated to expression breadth for both oligo-array and MPSS experiments (Table 1).

Finally, we studied the relationships between expression profiles and the length of intergenic sequences. Noteworthy expression breadth, which is associated with gene sequence expansion, was negatively correlated with the length of intergenic sequence (e.g for oligoarray DS-EB r=-0.109 n= 14526 P< 0.0001; see Table 2).Expression level, on the contrary was weakly, positively, correlated to the lengths of intergenic spacers (e.g. for DS-I r=0.027 n=14236 P<0.01).

	Expression level		***Expression breadth***	
	Microarray	*MPSS*	*Microarray*	*MPSS*
Genomic variables	DS-I	O-I	DS-EB_w0	O-EB_w0
GC% Gene	0.033*	0.043*	0.046*	0.084*
GC% 5' UTR	-0.077*	0.021§	0.156*	0.137*
GC% CDS	0.222*	0.254*	0.109*	0.084*
GC% intron_w0	-0.079*	0.025†	0.286*	0.234*
GC% 3' UTR	-0.045*	0.000ns	0.124*	0.124*
Length of intergenic spacers	0.027†	0.017†	-0.109*	-0.063*
Length of intergenic spacers_w0	0.024†	0.016ns	-0.123*	-0.069*
GC% intergenic spacers	0.008ns	0.000ns	0.076*	0.045*
GC% of RNA	0.074*	0.115*	0.121*	0.093*

Table 2: **Correlation between structural genomic parameters and both gene expression level and expression breadth.** Data are presented both for oligo-array and MPSS assays. Statistical significances: ns = not significant; §P<0.05 †P<0.01; ‡P<0.001;*P<0.0001. Sample size of the correlations were comprised between n=12051 and n=14236. **Note.** For this calculations, gene for which expression was zero were not considered to avoid potential problem arising from defective probe (Eisenberg and Levanon, 2003).

GC content

GC content of genes can account for several DNA physical features potentially associated to gene expression. For example, the bendability of DNA increases faster with the

elevation of GC content and curvature drops faster than in random sequence*(14,15)*. The former property is considered to be associated with open chromatin usually linked to active transcription while the latter with condensed chromatin associated to repressed transcription states*(16,18)*. Also, the thermostability of RNA/DNA and RNA/RNA complexes increases faster with the elevation of GC content which suggests implications in transcription regulation or sense/antisense transcript interactions*(14)*.

In Arabidopsis genome the GC content of coding sequence was positively correlated to both expression level and breadth (Table 2). Non coding gene sequences, showed different patterns between expression breadth and level: introns and UTRs were positively correlated to expression breadth and negatively or not even associated to expression level. Also the GC content of intergenic spacers followed a similar pattern, being positively correlated to expression breadth and not associated to expression level.

Direct effects and regional mutation biases.

Because the measures of expression level and expression breadth are highly correlated with each other (see Figure 2) we reconsidered the effects of either of the two measures on gene sequences after correcting for the other. In general, the results of the multiple regressions shown in table 3 confirmed and strengthen the patterns described by the pairwise correlations: i.e the level of expression being negatively correlated to gene parameters and breadth of expression depicting the opposite trend in favour of expansion. The only noticeable exception was the absence of a significant relationship of expression level, calculated by oligo array data, with 3'UTR after correcting for expression breadth.

Recently, Urrutia and Hurst have shown that regional mutational biases may influence the local level of insertions and deletion in the human genome*(2)*. Also compositional issues have been related to structural features. For example, based on GC contents of gene sequence, Carels and Bernardi have identified two classes of genes which exhibit distinctive structural features*(19)*.

In the previous section we have described the relationships between measures of gene expression and the length and the GC content of intergenic sequences. To verify the possibility that these correlations, could explain the relationship between expression profiles and gene characteristics we studied the consequences of including the length and GC content of intergenic sequences as additional independent variables in our models. As shown in table 3, the correction for regional variables showed only a limited impact on the effects of expression profiles on gene characteristics (Table 3).

Microarray

Dependent variable	Independent variables in the models: EL,EB		EL,EB & regional[1]		EL,EB, regional & genic[2]	
	β_{EL}	β_{EB}	β_{EL}	β_{EB}	β_{EL}	β_{EB}
5' length	**-0.084***	**0.181***	**-0.089***	**0.195***	**-0.046***	**0.108***
CDS length	**-0.176***	**0.073***	**-0.179***	**0.078***	**-0.095***	-0.025†
Intron length[3]	**-0.175***	**0.214***	**-0.177***	**0.219***	**-0.065***	**0.095***
3' length	0.005[ns]	**0.188***	0.001[ns]	**0.197***	0.035‡	**0.128***
PT length	**-0.205***	**0.200***	**-0.206***	**0.204***	n.a.	n.a.
Intron number[4]	**-0.150***	**0.144***	**-0.148***	**0.140***	**0.027***	**-0.029***

MPSS

Dependent variable	Independent variables in the models: EL,EB		EL,EB & regional[1]		EL,EB, regional& genic	
	β_{EL}	β_{EB}	β_{EL}	β_{EB}	β_{EL}	β_{EB}
5' length	**-0.035‡**	**0.135***	**-0.041***	**0.143***	-0.021§	**0.085***
CDS length	**-0.155***	**0.086***	**-0.160***	**0.094***	**-0.118***	**0.026†**
Intron length[3]	**-0.066***	**0.169***	**-0.066***	**0.169***	-0.010[ns]	**0.047***
3' length	**0.065***	**0.127***	**0.063***	**0.131***	**0.076***	**0.080***
PT length	**-0.125***	**0.150***	**-0.129***	**0.157***	n.a.	n.a.
Intron number[4]	**-0.080***	**0.138***	**-0.080***	**0.136***	0.020†	-0.010[ns]

Table 3: Multiple regression analysis of EL, EB and several gene parameters. Results from multiple-regression analyses of level of gene expression (EL) and breadth (EB) and length (of 5', CDS, intron, 3' and primary transcript, PT) when controlling for regional and genic effects. The results for the intron number are also shown. All lengths were $\log_{10}$ transformed.
Note. PT = Primary Transcript; n.a. = not applicable;
Significance levels: ns = not significant; §P<0.05 †P<0.01; ‡P<0.001;*P<0.0001. **In bold:** values that are significant after Bonferroni correction for multiple regression. Alpha level was adjusted separately for each type of model: EL,EB: 0.05/2 = 0.025; EL,EB & regional: 0.05/4 = 0.0125; EL,EB, regional + genic = 0.05/9 = 0.0056 (see Mundfrom et al 2006).
[1] regional variables: intergenic spacers length and intergenic spacers GC content. Genes without intergenic spacers were excluded.
[2] genic variables; 5',CDS,Intron,3', primary transcript lengths and intron number. Genes without introns were not included in the analyses.
[3],[4] genes without introns were excluded from the analyses.

Intra-genic effects

Since the lengths of genic regions are known to be highly correlated each other, we investigated the nature of the dependencies between expression profiles and length of genic regions. Each relationship between expression and length of genic regions was, thus, reconsidered after correcting for the effects of other regions of the gene and regional variables. The corrections had a strong impact on some of the relationships and in some cases the beta coefficients were no longer significant after the Bonferroni correction (see Table 3). All together, these results confirmed the scenarios previously described. Expression level was negatively correlated to gene sequences, with the only exception of 3' UTR. The results for expression breadth confirmed its positive relation with both UTRs and introns whereas the correlations for CDS again differed between MPSS (positively correlated) and oligo-array (negatively correlated).

Notably, the largest corrections were noticed for CDS and total intron lengths, while the least varied coefficients were found for the 3' UTR.

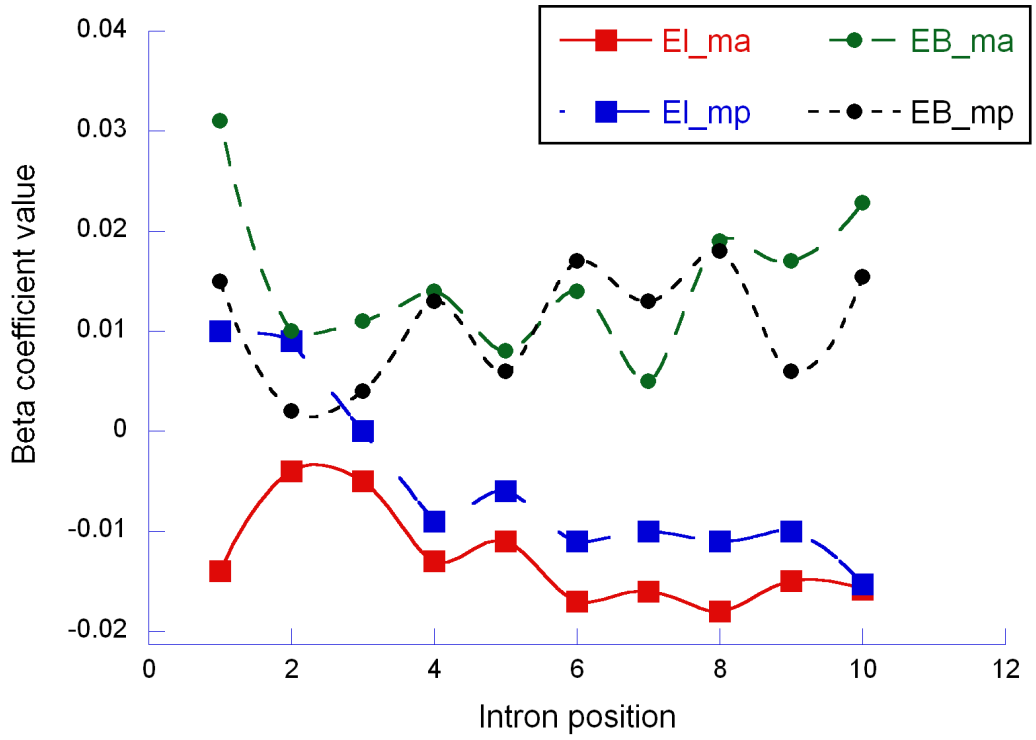


Figure 3: Results of multiple regression analyses between intron length and espression level (EL) and breadth (EB) by intron position**.** The analysis was conducted considering both microarray (ma) and MPSS (mp) data and all genes. Full circles: significant at P<0.05; empty circles: not significant at P>0.05.

To gather a deeper insight on the dynamics of intron length and number variations we analysed the effect of expression profile with respect to intron position. Several studies, that have combined data from multiple species, have found that the first intron is,on average, longer that other introns*(20)*. A proposed explanation for this trend is that introns from the 5' proximal region of a gene have important functional features related to gene expression while introns from the middle or 3' end of genes have progressively lower impacts. Recently, Rose et al. have demonstrated that signals responsible for boosting the expression are most abundant in introns near the transcription start site and that the compositional difference between promoter proximal introns and distal introns can be used to predict the ability of an intron to stimulate expression*(21)*. In a subsequent paper, Bradman and Korf*(20)* have demonstrated that density of enhancing signals decreases with intron order. In the effort to minimize the cost of gene transcription, the action of selection should preferentially be directed toward distal introns. According to this hypothesis the strength of the miniaturization effect on intron length should steadily increase with intron order.

Multiple regression analyses were performed considering intron length as dependent variable and EL and EB as independent variables. For each order of intron position, β coefficients were collected both for EL (β_{EL}) and EB (β_{EB}).

The β_{EL} coefficients depended by intron order (figure 3): proceeding from 5' to 3' ends, the β_{EL} values became more negative remaining significant in spite of the reduction of sample size. For oligo-array, 42.1% (Spearman $\rho = -0.699$, n=10, P=0.0245) of the variance of β_{EL} coefficients was explained by intron position and for MPSS this proportion reached the 78.6% ($\rho = -0.924$, n=10, P=0.0001). No significant trends (Spearman P>0.16) were observed for β_{EB} (only 13.1% and 10.0 % of the variance explained by intron position for microarray and MPSS, respectively). Interestingly, after removing the first intron, negative trends were confirmed for β_{EL} and positive trends were indicated for β_{EB}. Indeed, the proportion of explained variance become 63.3% ($\rho = -0.753$, P=0.0190; for microarray) and 70.6% ($\rho = -0.895$, P=0.0011; for MPSS) for β_{EL} and 34.0% ($\rho = 0.630$, P = 0.0688; for microarray) and 40.1% ($\rho = 0.679$, P = 0.0442, for MPSS) for β_{EB}.

In the previous section we have demonstrated that natural selection mediated by expression level exerts its action preferentially on distal introns owing to their lower importance for gene expression. With an analogous construction we may hypothesize that natural selection acts on CDS reducing, preferentially, the sequences coding for amino acids not strictly essential for protein function. Indeed we found that within CDSs the slope of the linear regression between the length of the regions outside functional domains and expression

level was about 1,7 fold higher than the slope of the linear regression between the length of functional protein domain and expression level (Figure 4). As expected the EB behaved differently with the length of CDS coding for functional domains increasing while the remaining part decreased.

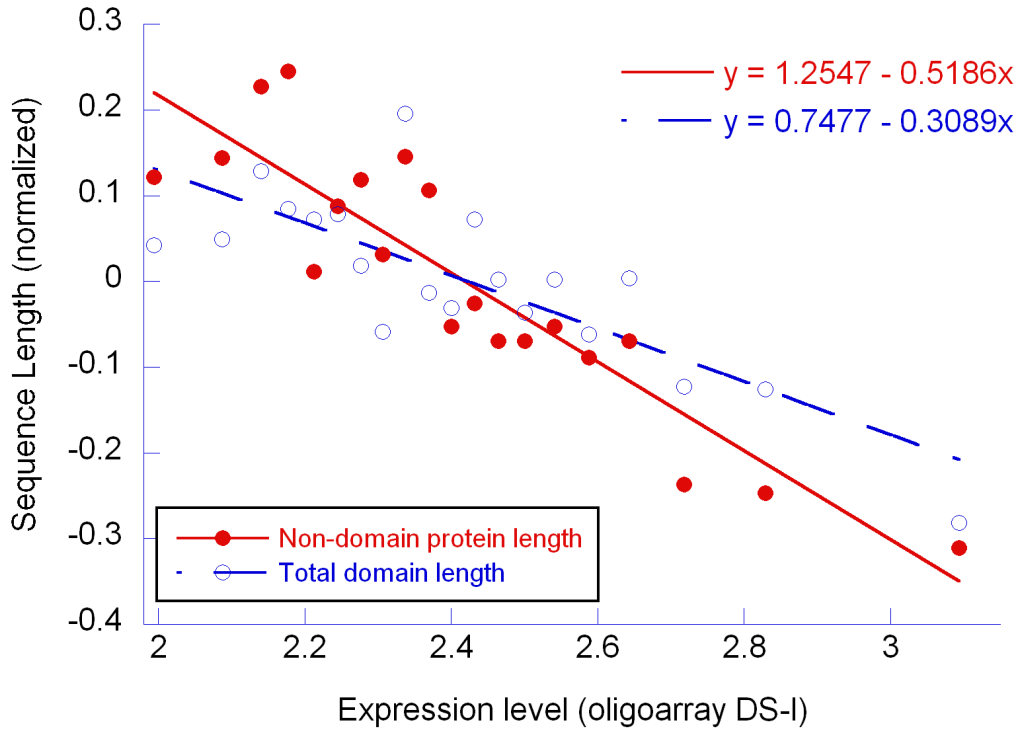


Figure 4: Relation between the expression level (oligoarray DS-I) and the Total Domain Length and Non-domain protein length. Variables on the Y axis have been normalized

DISCUSSION

The results presented here suggest that in *Arabidopsis* genome, structural parameters of genes are related to expression level and expression breadth in opposite directions. The level of expression is assocated with a generalized tendency toward transcript miniaturization while the breadth, i.e the number of experimental units in which a gene is expressed, is associated with sequence expansion.

Are these patterns consistent with the action of natural selection, or can they be explained by neutral processes in the Arabidopsis genome?

To answer this question we resorted to a multiple regression analysis of regional genomic characters that could potentially explain the relationships between expression profiles and gene structure. The considered regional variables only partially accounted for the relationships between the level and breadth of gene expression and gene sequences. These findings were not conclusive in favor of selection, as it can be argued that other regional variables, not explored in our analysis, might account for the remaining parts of the effects.

Instead of embarking in a cumbersome search of these additional regional variables, we adopted a more qualitative approach to identify signatures specific of selection. We began by proposing that selective process acting on genic sequences to maximize expression efficiency should discriminate between sequence dispensability. In contrast, regional mutational biases acting alone should have more generalized effects. Following such an approach, we found two evidenceS of signatures of selection. First, the miniaturization effect of selection in highly expressed genes is stronger in the more distal introns. This evidence is in line with a recent finding in plants that indicates that the signals responsible for boosting expression are more abundant in proximal than in distal introns*(21)*. Second, the action of selection in highly expressed genes is preferentially directed toward sequences outside the regions coding for functional domains. Taken together these observations suggest that natural selection is directly acting on gene sequences to maximize the efficiency of expression and that regional variables may, eventually, concur to this action.

Are these effects similar to those observed in other non plant genomes, or do they represent unique features of plant genomes? Before trying to answer this question, it is important to remember, that the level of resolution of our analysis was achieved considering the least correlated definitions of expression level and breadth. Other definitions of expression level, such as those weighing the experimental units with no expression, produced less defined picture*(8,9)*.

While the miniaturization effect is in line with the selection for economy model, the effect observed on genes with high EB, in *Arabidopsis*, is in apparent striking contrast with observations in other systems. At least three scenarios may have generated this dichotomy: i) the forces shaping plant genes are the same as those affecting other genomes but subject to a different balance ii) the nature of selective forces are unique to plant genomes iii) a mix of the first two scenarios.

Let’s analyse the different options moving from the homologies between different genomes. Average intron density is correlated to expression breadth in both humans*(13)* and *Arabidopsis*. Comeron, has speculated that this association could be related to the influence of

introns on mRNA metabolism and splicing efficiency*(13)*. During splicing, a complex of several proteins called exon-exon junction complexes (EJC) are deposited on processed mRNA in proximity of splicing sites*(22)*. Recent evidence has indicated that this complex would exert a post transcriptional enhancing effect, influencing export efficiency of mRNA to the cytosol and promoting transcriptional elongation and translation. Thus, according to this view, the number and total length of introns per gene results from two opposing selective forces: one favoring their proliferation because of the beneficial effects of EJC on some aspect of general RNA metabolism, and the other working for their reduction because of the cost associated to their transcriptions. An additional force favoring intron number and length increase could be connected to gene plasticity. Kreitman and Wayne have proposed a selective model for intron presence associated with the deleterious effects of linkages between sites under selection, a phenomenon termed Hill-Robertson effect. Indeed, linkage between selected loci can reduce the overall effectiveness of selection and the rate of adaptation*(23)*. This phenomenon also generates indirect selection to increase recombination rates. Introns generally have a reduced frequency of sites under selection compared to exons. Thus, an increase in intron number and length will increase the distance between mutations under the influence of selection in adjacent exons. This will favor recombination events between selected sites on different exons and consequently would improve the responsiveness of gene to selection.

Number and length of introns could also be related to the expression breadth in term of balancing selection. Wegmann et al. have recently demonstrated that human genes producing politype transcripts are expressed in a larger number of tissues and have more introns than genes producing monotype transcript*(24)*. The authors suggested that genes with high expression breadth and high intron number will be more prone to produce new transcript isoforms which could be maintained because ensuring a higher adaptability to various tissues conditions. Following this reasoning we analyzed alternatively spliced Arabidopsis genes and found that they show the tendency to have a wide expression breadth and longer primary transcript with more and longer introns (data not shown). Therefore, in this respect Arabidopsis genes known to produce polytype transcripts are similar to human alternatively spliced genes.

Under the first scenario, different balances between selective forces would account for the differences in selective signatures. For example, it is possible that selection for miniaturization in human genes would overcome the effects of the other forces. A very important genic parameter to consider in this context is the overall gene structure. For

example in the human genome the average intron length is 5.5 kb, which is considerably larger than 152 bp of the average intron length of *Arabidopsis* genes*(8)*. Human genes have on the average also more introns than *Arabidopsis*. Broadly expressed human genes would be more compact than narrowly expressed ones in consequence of the high correlation between expression level and breadth. The observation that the difference in compactness between broadly expressed and narrowly expressed human genes loses significance when the comparison between these two categories is carried out taking into account the level of expression, lends support to this hypothesis*(25)*. In *Arabidopsis* genome, on the contrary, the marginal gain in fitness due to intron reduction and miniaturization in broadly expressed genes could be lower, due to the shorter average length and lower average number of introns per gene. Thus in spite of its high level of correlation with expression level, the breadth of expression shows, in the *Arabidopsis* genome, a prevalent effect toward sequence expansion.

Alternative interpretations may be hypothesized calling into the field plant specific selective forces (scenario 2). For example, compositional differences between plant and human introns may have have different consequences for gene expression regulation. The nucleosome binding sites may have a different distribution between introns and exons in plant genes compared to human genes and this may be a specificity of plant introns favouring the establishment of open chromatin configurations which are typical of genes with high EB.

Further investigations in other plant species having different genome size, life cycle, and reproductive system (for example vegetative vs sexual reproduction) will add elements to either of the two scenarios analyzed or configure new pictures by combining new and old elements (our scenario three).

AKNOWLEDGMENTS

We thank A. Urrutia, G. Bernardi, G. Pesole and T. Brown for critical reading of the manuscript and for their useful suggestions.

Reference List

1. C. I. Castillo-Davis, S. L. Mekhedov, D. L. Hartl, E. V. Koonin, F. A. Kondrashov, *Nat. Genet.* **31**, 415 (2002).
2. A. O. Urrutia, L. D. Hurst, *Genome Res.* **13**, 2260 (2003).
3. A. E. Vinogradov, *Trends Genet.* **20**, 248 (2004).
4. A. E. Vinogradov, *Genome Res.* **16**, 347 (2006).
5. E. Eisenberg, E. Y. Levanon, *Trends Genet.* **19**, 362 (2003).
6. T. Mourier, D. C. Jeffares, *Science* **300**, 1393 (2003).
7. C. Seoighe, C. Gehring, L. D. Hurst, *PLoS. Genet.* **1**, e13 (2005).
8. X. Y. Ren, O. Vorst, M. W. Fiers, W. J. Stiekema, J. P. Nap, *Trends Genet.* **22**, 528 (2006).
9. J. Colinas, S. C. Schmidler, G. Bohrer, B. Iordanov, P. N. Benfey, *PLoS. One.* **3**, e3670 (2008).
10. M. Schmid *et al.*, *Nat. Genet.* **37**, 501 (2005).
11. B. C. Meyers *et al.*, *Genome Res.* **14**, 1641 (2004).
12. B. M. Bolstad, R. A. Irizarry, M. Astrand, T. P. Speed, *Bioinformatics.* **19**, 185 (2003).
13. J. M. Comeron, *Genetics* **167**, 1293 (2004).
14. A. E. Vinogradov, *Nucleic Acids Res.* **31**, 1838 (2003).
15. A. E. Vinogradov, *Mol. Biol. Evol.* **18**, 2195 (2001).
16. M. Z. Radic, K. Lundgren, B. A. Hamkalo, *Cell* **50**, 1101 (1987).
17. C. Anselmi, G. Bocchinfuso, S. P. De, M. Savino, A. Scipioni, *Biophys. J.* **79**, 601 (2000).
18. M. J. Lercher, A. O. Urrutia, L. D. Hurst, *Nat. Genet.* **31**, 180 (2002).
19. N. Carels, G. Bernardi, *Genetics* **154**, 1819 (2000).
20. K. R. Bradnam, I. Korf, *PLoS. One.* **3**, e3093 (2008).
21. A. B. Rose, T. Elfersi, G. Parra, I. Korf, *Plant Cell* **20**, 543 (2008).
22. H. H. Le, B. Seraphin, *Cell* **133**, 213 (2008).
23. M. Kreitman, M. L. Wayne, *EXS* **69**, 157 (1994).
24. D. Wegmann, I. Dupanloup, L. Excoffier, *PLoS. One.* **3**, e3587 (2008).

25. S. W. Li, L. Feng, D. K. Niu, *Biochem. Biophys. Res. Commun.* **360**, 586 (2007).

Supplemental Material

Table S1: List of all oligoarray experiments considered

Series: Seedlings and whole plants

Available at:
http://affy.arabidopsis.info/narrays/experimentpage.pl?experimentid=149

GEO Access Identifier	Experiment Name	Tissue
GSM131471	ATGE_7_A2	seedling, green parts
GSM131472	ATGE_7_B2	
GSM131473	ATGE_7_C2	
GSM131474	ATGE_22_A	whole plant after transition
GSM131475	ATGE_22_B	
GSM131476	ATGE_22_C	
GSM131477	ATGE_23_A	whole plant after transition
GSM131478	ATGE_23_B	
GSM131479	ATGE_23_C	
GSM131480	ATGE_24_A	whole plant after transition
GSM131481	ATGE_24_B	
GSM131482	ATGE_24_C	
GSM131483	ATGE_96_A	seedling, green parts
GSM131484	ATGE_96_B	
GSM131485	ATGE_96_C	
GSM131486	ATGE_97_A	seedling, green parts
GSM131487	ATGE_97_B	
GSM131488	ATGE_97_C	
GSM131489	ATGE_100_A	seedling, green parts
GSM131490	ATGE_100_B	
GSM131491	ATGE_100_C	

Series: Roots

Available at:
http://affy.arabidopsis.info/narrays/experimentpage.pl?experimentid=151

GEO Access Identifier	Experiment Name	Tissue

GSM131555	ATGE_3_A	root
GSM131556	ATGE_3_B	
GSM131557	ATGE_3_C	
GSM131558	ATGE_9_A	root
GSM131559	ATGE_9_B	
GSM131560	ATGE_9_C	
GSM131561	ATGE_93_A	root
GSM131562	ATGE_93_B	
GSM131563	ATGE_93_C	
GSM131564	ATGE_94_A	root
GSM131565	ATGE_94_B	
GSM131566	ATGE_94_C	
GSM131567	ATGE_95_A	root
GSM131568	ATGE_95_B	
GSM131569	ATGE_95_C	
GSM131570	ATGE_98_A	root
GSM131571	ATGE_98_B	
GSM131572	ATGE_98_C	
GSM131573	ATGE_99_A	root
GSM131574	ATGE_99_B	
GSM131575	ATGE_99_C	

Series: Flowers and pollen
Available at:
http://affy.arabidopsis.info/narrays/experimentpage.pl?experimentid=152

GEO Access Identifier	Experiment Name	Tissue	GEO Access Identifier	Experiment Name	Tissue
GSM131576	ATGE_31_A2	flowers stage	GSM131609	ATGE_43_A	flowers stage
GSM131577	ATGE_31_B2	9	GSM131610	ATGE_43_B	15 stamen
GSM131578	ATGE_31_C2		GSM131611	ATGE_43_C	
GSM131579	ATGE_32_A2	flowers stage	GSM131612	ATGE_45_A	flowers stage
GSM131580	ATGE_32_B2	10/11	GSM131613	ATGE_45_B	15 carpels
GSM131581	ATGE_32_C2		GSM131614	ATGE_45_C	
GSM131582	ATGE_33_A	flowers stage	GSM131615	ATGE_53_A	flower stage
GSM131583	ATGE_33_B	12	GSM131616	ATGE_53_B	12 equivalent
GSM131584	ATGE_33_C		GSM131617	ATGE_53_C	
GSM131585	ATGE_34_A	flowers stage	GSM131618	ATGE_54_A	flower stage
GSM131586	ATGE_34_B	12 sepals	GSM131619	ATGE_54_B	12 equivalent
GSM131587	ATGE_34_C		GSM131620	ATGE_54_C	

GSM131588	ATGE_35_A	flowers stage	GSM131621	ATGE_55_A	flower stage
GSM131589	ATGE_35_B	12 petals	GSM131622	ATGE_55_B	12 equivalent
GSM131590	ATGE_35_C		GSM131623	ATGE_55_C	
GSM131591	ATGE_36_A	flowers stage	GSM131624	ATGE_56_A	flower stage
GSM131592	ATGE_36_B	12 stamens	GSM131625	ATGE_56_B	12 equivalent
GSM131593	ATGE_36_C		GSM131626	ATGE_56_C	
GSM131594	ATGE_37_A	flowers stage	GSM131627	ATGE_57_A	flower stage
GSM131595	ATGE_37_B	12 carpels	GSM131628	ATGE_57_B	12 equivalent
GSM131596	ATGE_37_C		GSM131629	ATGE_57_C	
GSM131597	ATGE_39_A	flowers stage	GSM131630	ATGE_58_A	flower stage
GSM131598	ATGE_39_B	15	GSM131631	ATGE_58_B	12 equivalent
GSM131599	ATGE_39_C		GSM131632	ATGE_58_C	
GSM131600	ATGE_40_A	flowers stage	GSM131633	ATGE_59_A	flower stage
GSM131601	ATGE_40_B	15 pedicels	GSM131634	ATGE_59_B	12 equivalent
GSM131602	ATGE_40_C		GSM131635	ATGE_59_C	
GSM131603	ATGE_41_A	flowers stage	GSM131636	ATGE_73_A	mature pollen
GSM131604	ATGE_41_B	15 sepals	GSM131637	ATGE_73_B	
GSM131605	ATGE_41_C		GSM131638	ATGE_73_C	
GSM131606	ATGE_42_A	flowers stage	GSM131639	ATGE_92_A	flower
GSM131607	ATGE_42_B	15 petals	GSM131640	ATGE_92_B	
GSM131608	ATGE_42_C		GSM131641	ATGE_92_C	

Series: Shoots and Stems

Available at:
http://affy.arabidopsis.info/narrays/experimentpage.pl?experimentid=153

GEO Access Identifier	Experiment Name	Tissue	GEO Access Identifier	Experiment Name	Tissue
GSM131643.	ATGE_2_A	hypocotyl	GSM131673	ATGE_49_A	shoot apex,
GSM131644	ATGE_2_B		GSM131674	ATGE_49_B	inflorescence
GSM131645	ATGE_2_C		GSM131675	ATGE_49_C	
GSM131646.	ATGE_4_A	shoot apex,	GSM131676	ATGE_50_A	shoot apex,
GSM131647	ATGE_4_B	vegetative +	GSM131677	ATGE_50_B	inflorescence
GSM131648	ATGE_4_C	young leaves	GSM131678	ATGE_50_C	

GSM131649. GSM131650 GSM131651	ATGE_6_A ATGE_6_B ATGE_6_C	shoot apex, vegetative	GSM131679 GSM131680 GSM131681	ATGE_51_A ATGE_51_B ATGE_51_C	shoot apex, inflorescence
GSM131652. GSM131653 GSM131654	ATGE_8_A ATGE_8_B ATGE_8_C	shoot apex, transition	GSM131682 GSM131683 GSM131684	ATGE_52_A ATGE_52_B ATGE_52_C	shoot apex, inflorescence
GSM131655. GSM131656 GSM131657	ATGE_27_A2 ATGE_27_B2 ATGE_27_C2	stem, 2nd internode	GSM131480 GSM131481 GSM131482	ATGE_24_A ATGE_24_B ATGE_24_C	shoot apex, inflorescence
GSM131658. GSM131659 GSM131660	ATGE_28_A2 ATGE_28_B2 ATGE_28_C2	1st node			
GSM131661. GSM131662 GSM131663	ATGE_29_A2 ATGE_29_B2 ATGE_29_C2	shoot apex, inflorescence			
GSM131664. GSM131665 GSM131666	ATGE_46_A ATGE_46_B ATGE_46_C	shoot apex, inflorescence			
GSM131667 GSM131668 GSM131669	ATGE_47_A ATGE_47_B ATGE_47_C	shoot apex, inflorescence			
GSM131670 GSM131671 GSM131672	ATGE_48_A ATGE_48_B ATGE_48_C	shoot apex, inflorescence			

Series: Siliques and Seeds

Available at:
http://affy.arabidopsis.info/narrays/experimentpage.pl?experimentid=154

GEO Access Identifier	**Experiment Name**	**Tissue**

GSM131685	ATGE_76_A	siliques, w/
GSM131686	ATGE_76_B	seeds stage 3
GSM131687	ATGE_76_C	
GSM131688	ATGE_77_A	siliques, w/
GSM131689	ATGE_77_B	seeds stage 4
GSM131690	ATGE_77_C	
GSM131691	ATGE_78_A	siliques, w/
GSM131692	ATGE_78_B	seeds stage 5
GSM131693	ATGE_78_C	
GSM131694	ATGE_79_A	seeds, stage 6,
GSM131695	ATGE_79_B	w/o siliques
GSM131696	ATGE_79_C	
GSM131697	ATGE_81_A	seeds, stage 7,
GSM131698	ATGE_81_B	w/o siliques
GSM131699	ATGE_81_C	
GSM131700	ATGE_82_A	seeds, stage 8,
GSM131701	ATGE_82_B	w/o siliques
GSM131702	ATGE_82_C	
GSM131703	ATGE_83_A	seeds, stage 9,
GSM131704	ATGE_83_B	w/o siliques
GSM131705	ATGE_83_C	
GSM131706	ATGE_84_A	seeds, stage
GSM131707	ATGE_84_B	10, w/o
GSM131708	ATGE_84_C	siliques

Table S2: List of the MPSS experiments considered

MPSS Access Identifier	Tissue	Experiment Description	http
INF	Influorescence	Infloresence - mixed stage, immature buds, classic MPSS	http://mpss.udel.edu/at/Library.php?lib=2&tag_length=17
LEF	Leaves	Leaves - 21 day, untreated, classic MPSS	http://mpss.udel.edu/at/Library.php?lib=3&tag_length=17
ROF	Root	Root - 21 day, untreated, classic MPSS	http://mpss.udel.edu/at/Library.php?lib=4&tag_length=17
SIF	Silique	24 to 48 hr post-fertilization, classic MPSS	http://mpss.udel.edu/at/Library.php?lib=5&tag_length=17
GSE	Seedlings	Germinating seedlings	http://mpss.udel.edu/at/Library.php?lib=15&tag_length=17

Table S3: Methods used in order to estimate the gene expression levels.

Method	Descriprion
DS-A	The Exeprimental unit is rapresented by the developmental stage. The expression level for each gene is obtained by averaging the expression values of all the experimental units.
DS-I	The Experimental unit is rapresented by the developmental stage. The expression level for each gene is obtained by averaging the expression values of only the experimental units in which the gene is actually expressed
DS-pE	The Experimental unit is rapresented by the developmental stage.The expression level for each gene is equal to the peak of expression taking in to account all experimental units.
O-A	The Experimental unit is rapresented by the organ. The expression level for each gene is obtained by averaging the expression values of all the experimental units.
O-I	The Experimental unit is rapresented by the organ. The expression level for each gene is obtained by averaging the expression values of only the experimental units in which the gene is actually expressed
O-pE	The Experimental unit is rapresented by the organ. The expression level for each gene is equal the peak of expression taking in to account all experimental units.

The effect of local selective pressures in shaping the codon bias of *Arabidopsis thaliana*

ABSTRACT

The unequal use of synonymous codons has been studied in many unicellular species. The emerging picture points to a co-adaptation of codon usage and t-RNA gene copy numbers. Such a point of view, initially transported as a whole to multicellular organisms, has been recently challenged by the observation that t-RNA abundance may vary among tissues.

In the present paper we analyzed the codon usage of Arabidopsis genes expressed in 15 different tissues. MANOVA analysis indicated statistically significant differences in the codon usage among tissues. Cluster analysis of tissues specific genes revealed the highest differences in codon usage frequencies between reproductive and vegetative tissues.

Finally, based on a measure of similarity in codon usage between tissue specific and more broadly expressed genes we identified those tissues that contribute the most in shaping the codon composition of widely expressed genes.

INTRODUCTION

Although synonymous codons correspond to the same aminoacid they are used at different frequencies in genes. Two models of evolution have been proposed to explain such a bias. The neutralist model postulates that the observed pattern of codon bias is determined by local differences of some mutational processes(*1*). A co-adaptation of synonymous codon usage and abundance of tRNA to optimize translation efficiency and accuracy is at the basis of the selective model(*2-5*). Both theoretical considerations and simulation studies have suggested that these two models are not mutually exclusive and thus that codon usage may reflect a balance between selective and mutational pressures as well as stochastic processes acting at population level such as random genetic drift(*6*).

In unicellular organisms such as *Escherichia coli* the co-adaptation based model was proved to explain the preferential use of certain synonymous codons in highly expressed genes(*7*). However analysis carried out on multicellular organisms revealed a more complex picture.

The unequal DNA composition along many eukaryotic genomes (i.e. isochores(*8*)) proved to be associated to the observed codon bias(*9*). However such association failed to explain the bias as even after controlling for local nucleotide composition, Urrutia and co-workers still observed a codon bias in *Homo sapiens*.(*10*) The preference of certain synonymous codon has been associated to the mRNA stability(*11*) as well as to the protein structure possibly mediated by induction of translational pausing.(*12, 13*)

The presence of diverse cell types within the same eukaryote organism also raises many intriguing questions: is the codon bias consistently distributed among different tissues and/or developmental stages? Do tissue specific genes privilege the use of particular synonymous codons? Plotkin and coworkers found a non equal distribution of codon bias in tissue specific genes in human,(*14*) probably due to a co-adaptation to the pool of tRNA available in each tissue.(*15*) This theory was confirmed by the work of Kotlar and coworkers who found a strong positive correlation between the number of tissues in which a gene is expressed and the codon bias, particularly for those codons that are involved in the incorporation of smaller, simpler and more ancient amino acids.(*16*) However Sèmon et al reached a different conclusion by using a diverse statistical analysis and a richer dataset of tissue specific genes.(*17*)

The mutational processes that underlie the neutralist model may be dependent on the expression pattern of genes. Indeed tissue specific genes feature an higher GC content than housekeeping genes in mammal and in plant genomes(*18*). Comeron reported that the

association between gene expression and GC3 (the GC content at the codon third position) is tissue dependent in human(*19*); and explained this finding in terms of local composition and to the transcription-associated mutational biases (TAMB). The tissue-dependent codon bias was also confirmed by the discovery of a gametophytic-specific codon usage in genomes of plants such as *Zea mays* and *Triticum aestivum*(*20*) as well as in Drosophila melanogaster. (*21*)

Recently, Waldman and coworkers reported that tissue specificity in the usage of synonymous codons may be the result of a global, rather than local, tRNA-codon bias co-adaptation.(*22*) This is confirmed by the weaker association between gene expression and codon bias observed in human and is due to the need for the same tRNA pool, to cater for more tissues in multicellular organisms. The authors also highlight that a different adaptation level can be observed not only among several tissues but also at different developmental stages with the human adult tissues being more adapted than the fetal ones*(22)*.

Although an always higher number of papers aim to speculate the reasons behind the unequal distribution of synonymous codon both between and within genomes, many questions still remain unanswered. If tissue specific genes prefer particular synonymous codons, which features will a gene that is expressed in two or more tissues exhibit? Can some tissue exert a leading effect? In the present paper the high annotation quality of the *Arabidopsis thaliana* genome,(*23*) together with the wide availability of expression data for this organism, is exploited to gather a deeper insight in the mechanisms that govern the unequal codon bias distribution within different tissues of plants.

Glossary

T = tissue in which a gene is expressed

G = the entire set of tissues in which a gene is expressed, excluding T

HK = housekeeping genes

TS = tissue specific genes

TE = genes expressed in T and G

PE = genes expressed in T and G, with the peak of expression in T

METHODS

Sequences and expression data

Arabidopsis thaliana coding sequences (CDS) were downloaded from The Arabidopsis Information Resource (TAIR) website (ftp://ftp.arabidopsis.org/home/tair/Sequences/blast_datasets/TAIR8_blastsets/TAIR8_cds_20080412). Expression data relative to 15 A. thaliana tissues were retrieved from the Genome Exression Omnibus repository at NCBI database (http://www.ncbi.nlm.nih.gov/gds/). Each experiment was replicated three times and genes were considered expressed when the corresponding probe set was significantly detected in all replicates and the average expression measurements proved to be >75 (for the complete set of used experiments see Table S1 in supplemental material). Genes were classified in four groups: (a) Housekeeping genes which are ubiquitously expressed in all tissues (HK); (b) tissue specific genes (TS); (c) genes that are expressed in a particular tissue and are simultanisouly expressed in one or more further tissues (TE) (d) genes that show a peak of expression in a particular tissue and are, at the same time, expressed in one or more further tissue (PE). PE and TE genes were firstly classified by the tissue in which they were expressed and then grouped by the number of tissue in which they were actually expressed (expression breadth). As a way of example the class Carpels TE3 comprises all the genes being expressed in Carpel and in two further tissues, whereas Carpels PE3 feature the additional constraint of having the peak of expression in Carpels.

Measures of codon bias

Two punctual measures of codon bias were used. RSCU estimates how each synonymous codon deviates from the random use within each amino acid class.(*3*) The choice of a punctual measure in general, and of the RSCU index in particular, allowed to consider the codon usage free of amino acid usage bias. Although this index is widely used some of its weaknesses must be considered. Indeed if a particular amino acid is under represented within a group of sequences, few relative synonymous codons are available for a correct bias estimation. For this reason two different indexes were used to compute the punctual codon bias within a gene class: (a) the gRSCU where RSCU is calculated for each gene (as reported in the original paper) and then averaged among all genes composing the class; (b) the pRSCU which is calculated by concatenating all the genes within a class, and then computing the RSCU of the resulting super-sequence. Some of the analyses reported in the present paper has been repeated by using both indexes in order to confirm the most outstanding results

Codon adaptation index (CAI)(*24*) was also used as a measure of codon bias. This parameter estimates the portion of used synonymous codon which correspond to the most abundant tRNA in *Arabidopsis thaliana*.

Tissue specific Euclidean distance.

A gene featuring an expression breadth >1 is expressed in a particular tissue T and in a further set of tissues G, comprising from 1 to 14. Within each class of expression breadth and each tissue, Euclidean distances of the TE genes were calculated from both T and G tissue specific genes, computing for the latter the average distance from the whole set of tissues. The results obtained for all the TE genes of each tissue (Figure S1) were averaged and available in Supplemental Material. Distances obtained for the single tissues were further mediated among all the tissues in order to obtain the plot reported in Figure 3. The same analysis was then carried out by applying the same approach at the PE genes (single tissue plots available in Figure S2).

The statistical analysis

Cluster analysis was performed using Jmp v.7.0. Data were standardized and then submitted to a hierarchical clustering with a Ward method.

Genes were re-classified by using a discriminant analysis. A quadratic model was chosen and the study was carried out by using the software S-plus. The original 58 variable dataset was cut of the redundant information derived from the two-fold degenerated amino acids. The quadratic model approach was chosen in order to take into account the relatively small number of variables (RSCU) together with their mutual correlation, and the high number of entries (genes) in the dataset.

The over-representation of particular gene ontology categories was tested by the use of the Cytoscape(*25*) plug-in BINGO(*26*). An hypergeometric test was used in combination with the multiple testing correction of Benjamini and Hochberg (FDR) with a significance level of 0.05

RESULTS

Tissue specific codon usage

The extent of codon bias in different tissues was studied using a multivariate approach. In practice, gRSCU indexes were calculated for genes expressed exclusively in one tissue (hereafter called tissue specific genes). TS-gRSCU indexes were then obtained for each codon and for each tissue by averaging the corresponding gRSCU values of genes expressed in that given tissue. Similarly gRSCU were calculated for genes ubiquitously expressed (HK genes) and the HK-gRSCU indexes were obtained as mean of the corresponding gRSCU (Supplemental material Table S2).

A Manova analysis was carried out on TS-RSCU in order to highlight possible differences in the codon usage among the set of considered tissues. Four different multivariate tests were performed which confirmed a significant difference in the gRSCU values within the analysed groups (Manova statistics reported in Table S4).

A cluster analysis was carried out on a dataset comprising TS-gRSCU and HK-gRSCU in order to reveal possible similarities in codon usage between different tissues. The results (Figure 1) showed a discernible differentiation between HK and TS genes. Moreover tissue specific genes split in two distinct groups hereafter referred to as Cluster A and Cluster B with Rosette leaf departing from this classification. An almost equivalent picture emerged when the same analysis was carried out by using a pRSCU analysis (see Methods; Figure S3).

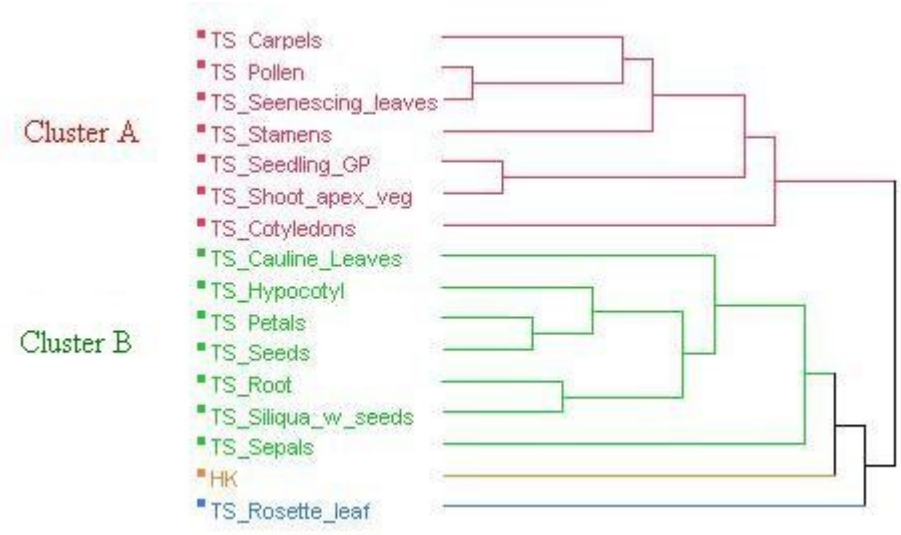


Figure 1: Cluster analysis of genes specifically expressed in 15 tissues (TS-gRSCU) and housekeeping (HK-gRSCU). The analysis was carried out by using the RSCU as variables.

While marked differences between housekeeping and tissue specific genes were expected(*16*, *27*), we did not find a plausible explanation for separation of Cluster A and B. The presence within the two groups of mainly reproductive (A) and vegetative (B) tissues seems evocative of what observed in human where differences in codon bias led to translation efficiency variations between adult and fetal stages.(*22*) It is well known that certain codons are preferred near splice sites as part of the exonic splicing enhancer (ESEs) and therefore this may have an impact on codon bias.(*28*) Such phenomenon should be particularly strong for genes featuring short exons. For this reason the average exon lengths were compared by ANOVA for the genes belonging to Cluster A and B (Table 1). Indeed exons in Cluster A proved to be, on average, significantly shorter than those in Cluster B ($p<0.05$).

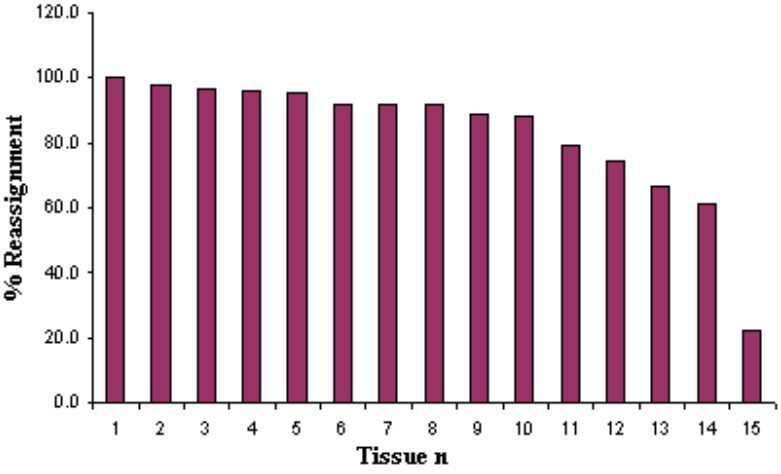
Cluster	Num. Exons	Av. Exon Length	CDS Length	Exp. Level	CAI
A					
Carpels	6.33	227.01	1437.76	2.1	0.48
Pollen	5.3	248.79	1423.93	2.29	0.44
Senescing leaves	5.99	263.9	1684.39	2.05	0.44
Stamen	5.82	219.62	1298.2	2.32	0.46
Seedling	7.37	227.16	1740.06	2	0.44
Shoot Apex	8.25	211.13	1741.85	1.99	0.46
Cotyledons	4.44	228.79	1016.83	2	0.49
Average	6.21	232.34	1477.57	2.11	0.46
B					
Cauline leaves	4.81	218.06	1090.32	1.96	0.46
Hypocotyl	4.46	272.85	1230.98	2.02	0.47
Petals	4.26	277.79	1221.38	2.08	0.45
Seeds	4.15	267.97	1111.55	2.45	0.46
Root	4.06	279.91	1145.43	2.15	0.49
Siliqua w/ seeds	4.62	256.32	1197.62	2.25	0.48
Sepals	4.21	246.59	1074.43	2.13	0.47
Average	4.37	259.93	1153.10	2.15	0.47
ANOVA (Prob>F)	**0.0027***	**0.0246***	**0.0089****	**0.6192**	**0.2995**

Table 1: ANOVA analysis of cluster A and B tissue specific genes.

Since the association between codon bias and gene size was also reported,(*4*) the CDS length was used as variable for an ANOVA analysis. Again a significant difference between the two clusters emerged, with genes of cluster A having, on average, longer coding sequences than those of Cluster B.

One may argue that differences in sequences length between the two clusters can be ascribed to a variation of the expression level as these two parameters are strongly correlated. (*29*) However no statistically significant differences were observed between the two groups of tissues when considering the corresponding expression level.

Moreover the CAI (codon adaptation index), also known to be involved in shaping the synonymous codon usage, does not seem to contribute in differentiating the two clusters.



n	Tissue
1	Seedling_GP
2	Shoot_Apex_Veg
3	Petals
4	Sepals
5	Cauline_Leaves
6	Hypocotyl
7	Senescing_Leaves
8	Stamens
9	Carpel
10	Cotyledons
11	Siliqua
12	Pollen
13	Root
14	Seeds
15	Rosette_Leaf

Figure 2: Quadratic Discriminant analysis of the tissue specific genes.

Because codon choice is believed to be associated to gene function in plants,(*30*) the gene ontology was considered as possible source of differentiation between Cluster A and B. The Cytoscape plug-in BINGO (see method) was used to determine which gene ontology categories were possibly overrepresented within Clusters A or B. The results (Table 2) reveal that while gene belonging to A seems to be significantly more involved in hydrolytic activities, Cluster B feature an over presence of genes implicated in redox processes. Moreover the biological processes “response to chemical status” and “response to oxidative stress” are overrepresented in cluster B while both groups feature a significantly outstanding number of genes that code for endomembrane proteins.

Cluster	Gene ontology	Corr-pVal	Cluster Freq	Total_freq
	Molecular Function			
A				
	Catalytic Activity	0.00	0.373	0.291
	DNA Binding	0.00	0.119	0.077
	Hydrolase Activity	0.00	0.147	0.101
	Hydrol. Activity act.glycosyl bond	0.00	0.037	0.016
	transferase activity	0.00	0.141	0.099
	Transcription factor activity	0.00	0.088	0.054
	Hydrol. Activity act. On O-glycosil	0.01	0.032	0.014
	lipase activity	0.03	0.016	0.005
	carboxylesterase activity	0.03	0.024	0.011
	trnsition metal ion binding	0.03	0.073	0.047
	hydrol. Activity acting on ester bonds	0.03	0.055	0.033
	transcription regulator activity	0.03	0.094	0.065
	cation binding	0.03	0.080	0.054
	matal ion binding	0.03	0.085	0.059
	ion binding	0.03	0.085	0.059
	carotenoid deoxygenase activity	0.04	0.003	0.000
	9-cis-epoxycarothenoid dioxigenase	0.04	0.003	0.000
	lipid binding	0.04	0.016	0.006
B				
	Oxidoreductase activity	0.00	0.090	0.042
	peroxidase activity	0.00	0.021	0.004
	Oxidoreduct. Activ. On peroxide	0.00	0.021	0.004
	antioxidant activity	0.00	0.021	0.004
	manganese ion binding	0.01	0.006	0.001
	magnesium ion binding	0.01	0.008	0.001
	catalityc activity	0.02	0.344	0.291
	nutrient reservoir activity	0.02	0.009	0.002
	Biological Process			
A				
	None			
B				
	response to chemical status	0.001	0.076	0.041
	response to oxidative stress	0.004	0.023	0.007
	Cellular Component			
A				
	Endomembrane system	0.019	0.257	0.199
	Endoplasmic reticulum part	0.019	0.010	0.001
	Endoplasmic reticulum lumen	0.024	0.004	0.000
B				
	Endomembrane system	0.024	0.253	0.199

Table 2: Gene Ontology classes that resulted to be over-represented in either Cluster A or B

In order to investigate the variables that most accounted for the groups emerged by the cluster analysis shown in Figure 1, an ANOVA was performed on each codon gRSCU. Approximately 90% of the considered codons showed to be used differently between the two clusters (Table 3).

Finally to speculate whether the synonymous codons usage may be considered as a distinctive mark of each tissue, a quadratic discriminant analysis was performed by using the gRSCUs as variables. The results (Figure 2) showed a high number of correct genes reassignment among the tissues specific genes, with Rosette leaf representing again an exception probably due to the low number of genes specifically expressed in this tissue.

Codon	R^2	p	Codon	R^2	p	Codon	R^2	p
ttt(Phe)	**0.016449**	**<0.0001**	agc(Ser)	0.004903	0.2349	**aac(Asn)**	**0.01946**	**<0.0001**
ttc(Phe)	**0.01656**	**<0.0001**	**cct(Pro)**	0.026424	**<0.0001**	**aaa(Lys)**	**0.071196**	**<0.0001**
tta(Leu)	**0.096163**	**<0.0001**	ccc(Pro)	0.004927	0.2329	**aag(Lys)**	**0.071661**	**<0.0001**
ttg(Leu)	**0.00749**	**<0.05**	cca(Pro)	0.004319	0.3629	**gat(Asp)**	**0.021522**	**<0.0001**
ctt(Leu)	**0.023313**	**<0.0001**	**ccg(Pro)**	**0.032057**	**<0.0001**	**gac(Asp)**	**0.021388**	**<0.0001**
ctc(Leu)	**0.019036**	**<0.0001**	**act(Thr)**	**0.042748**	**<0.0001**	**gaa(glu)**	**0.027998**	**<0.0001**
cta(Leu)	**0.033194**	**<0.0001**	**acc(Thr)**	**0.011079**	**<0.01**	**gag(glu)**	**0.028274**	**<0.0001**
ctg(Leu)	**0.034373**	**<0.0001**	**aca(Thr)**	**0.012613**	**<0.0001**	**tgt(Cys)**	**0.007596**	**<0.05**
att(Ile)	**0.027477**	**<0.0001**	**acg(Thr)**	**0.045759**	**<0.0001**	**tgc(Cys)**	**0.007619**	**<0.05**
atc(Ile)	**0.024351**	**<0.0001**	**gct(Ala)**	**0.063394**	**<0.0001**	**aga(Arg)**	**0.045633**	**<0.0001**
ata(Ile)	**0.070791**	**<0.0001**	**gcc(Ala)**	**0.017254**	**<0.0001**	**agg(Arg)**	**0.01705**	**<0.0001**
gtt(Val)	**0.036805**	**<0.0001**	**gca(Ala)**	**0.015341**	**<0.0001**	**cga(Arg)**	**0.009239**	**<0.01**
gtc(Val)	**0.009834**	**<0.01**	**gcg(Ala)**	**0.034893**	**<0.0001**	**cgt(Arg)**	**0.043514**	**<0.0001**
gta(Val)	**0.03335**	**<0.0001**	**tat(Tyr)**	**0.013605**	**<0.0001**	cgc(Arg)	0.006593	0.0524
gtg(Val)	**0.00662**	**<0.05**	**tac(Tyr)**	**0.013644**	**<0.0001**	**cgg(Arg)**	**0.007939**	**<0.05**
tct(Ser)	**0.021868**	**<0.0001**	**cat(His)**	**0.009421**	**<0.01**	**ggt(gly)**	**0.031535**	**<0.0001**
tcc(Ser)	0.006174	0.0765	**cac(His)**	**0.009444**	**<0.01**	**ggc(gly)**	**0.009336**	**<0.01**
tca(Ser)	**0.012747**	**<0.0001**	**caa(gln)**	**0.098618**	**<0.0001**	**gga(gly)**	**0.010547**	**<0.01**
tcg(Ser)	**0.01524**	**<0.0001**	**cag(gln)**	**0.098618**	**<0.0001**	**ggg(gly)**	**0.014105**	**<0.0001**
agt(Ser)	**0.007146**	**<0.05**	**aat(Asn)**	**0.019435**	**<0.0001**			

Table 3: Codons which contributed in separating the groups during the cluster analysis (ANOVA test performed on the dataset comprising the tissue specific and housekeeping genes).

The Tissue effect

If synonymous codon usage is tissue dependent(*14*), what is the expected codon composition of a gene expressed in several tissues? In order to reveal the importance of a possible tissue-specific selection on widely expressed genes, a comparative analysis of the

Euclidean distances TE-gRSCU from each TS-gRSCU was performed as described in the Methods section. The results indicate that genes that have in common the expression in a given tissue are closer to genes exclusively expressed in that tissue than any other tissue specific genes (Figure 3 – white bars). This trend tends to wane in higher expression breadth classes and no differences are observed at values > 7. Such a behavior seems evocative of some kind of advantage in using certain codons in order to be expressed in a specific tissue. If this is true then the trends observed in Figure 3 should be emphasize if considering genes that in tissue T have the peak of expression. After repeating the analysis using PE genes the emerged picture revealed marked differences in the Euclidean distances from T and G tissues for the entire range of expression breadth classes (Figure 3 – red bars). Curves presented in Figure 3 represent an average trend of all the considered tissues. Nevertheless differences may emerge when considering the single tissues (see Supplmental Material) and the divergence between the curves representing the T and G distances may have different intensity or even direction. As a way of example distances from T genes are always higher than those from G genes in Cotyledons whereas the opposite trend is observed in Pollen (Figure S2).

The ratio between the T and G distances for the expression breadth class 2 may be considered as an index of the influence a tissue can exert to promote a certain codon composition to a gene that in that tissue is expressed, or in other words, what which is hereby referred to as the “tissue effect”. When plotting this ratio for all the considered tissue a heterogeneous scenario emerged.

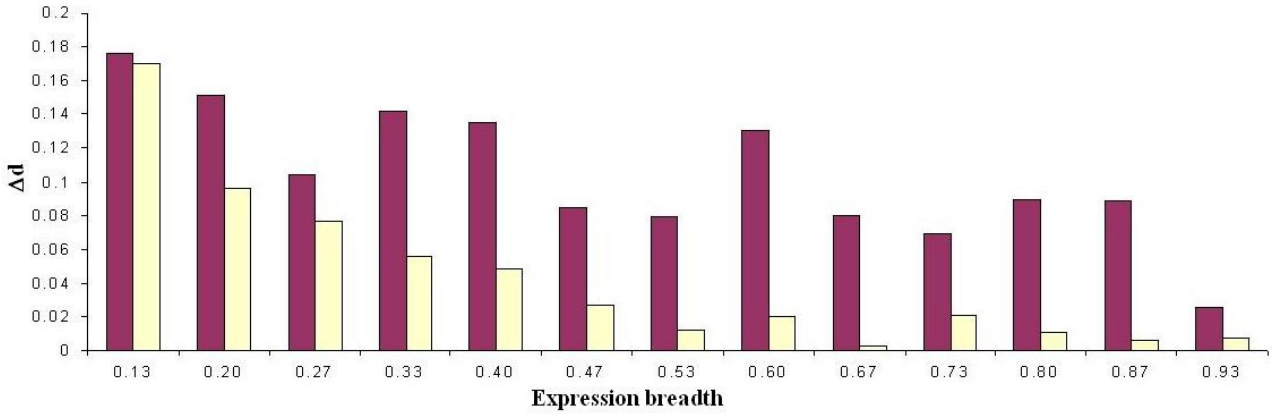
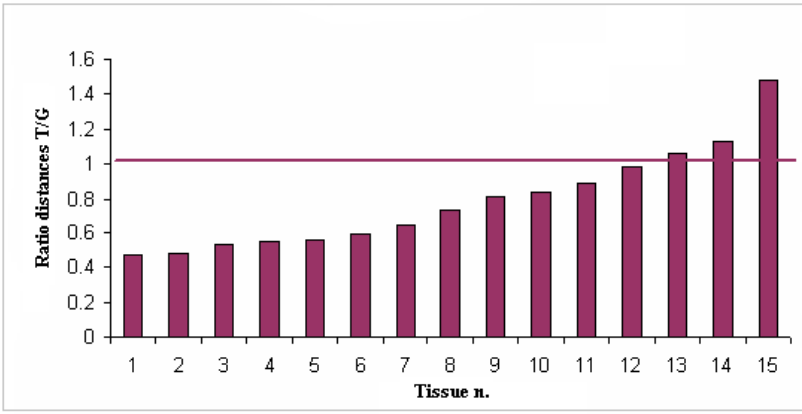


Figure 3: Difference between the Euclidean distances of the widely expressed genes TE (white) and PE (red) from T and G tissue specific genes.

80% of the analyzed tissues exhibited a ratio < 1 with Root and Seedling being < 0.5. On the other hand Sepals, Rosette leaf and, at a higher extent, Cotyledons showed the opposite trend revealing no “tissue effect”.



n.	Tissue
1	Root
2	Seedling_GP
3	Pollen
4	Shoot_Apex_Veg
5	Senescing_Leaves
6	Seeds_wo_siliqua
7	Stamens
8	Carpels
9	Petals
10	Siliqua_w_seeds4
11	Hypocotyl
12	Cauline_leaves
13	Sepals
14	Rosette_leaf4
15	Cotyledons

Figure 4: Ratio distances between the codon RSCU of TE genes from T and G tissue specific genes.

DISCUSSION

The present study aim to reveal possible diversities in the synonymous codon usage of 15 *Arabidopsis thaliana* tissues. RSCU for 58 codons were used as a mean of estimating the codon bias within the considered tissues, with each gene being characterized by a specific set of values in a multivariate space. Group

First a Manova analysis was performed on genes specifically expressed in 15 tissues which revealed a significant difference in the synonymous codon usage among the considered groups (Wilk’s Lambda test $p < 0.001$).

Afterwards a cluster analysis was carried out by considering the average synonymous codon usage of tissue specific and housekeeping genes, in order to reveal possible differences between these groups. The result (Figure 1) clearly shows a marked divergence between HK and TS which was in accordance with the recently reported observation that different selection constraints may act on these two classes(*27*). Moreover two distinct goups emerged, namely Cluster A and Cluster B, each composed of seven tissues. Interestingly Cluster A features a high portion of reproductive tissues while B shows the presence of mainly vegetative tissues. Several ANOVA were carried out in order to get an insight on the reasons

that may be at the base of the differentiation of the two clusters. Variables that are known to be associated with the codon bias such as the exon junction sequences constraints(*31, 32*), the protein length(*4*), the expression level(*33*) and the use of codons corresponding to the most abundant isoacceptor tRNA(*34*) were considered. The results revealed a significant difference between the two groups in terms of genic sequences length. Cluster A tissues proved to have, on average, a higher number of exons with a shorter average length compared to cluster B. These features can bias considerably the codon usage since they both go in the direction of increasing the coding portion involved in the splicing mechanism(*32*). The higher number of exons, in spite of the shorter length, produces on average longer proteins in tissues belonging to Cluster A. The coding sequence length was recently reported to be associated with the expression level in *Arabidopsis*(*29*) and, in turn, highly expressed genes are known to feature an elevated codon bias in many organisms(*33*). Taken together, these observations may lead to assume a translational indirect effect on the synonymous codon usage. However this hypothesis seems to be unlikely as confirmed by the non significant difference in the average expression revealed by the ANOVA analysis. Such observation is in agreement with the theory proposed by Duret and co-workers who suggested a protein length dependent codon bias which was free from effects due to the expression level(*35*).

The tRNAs gene copy number is considered a reliable index of the isoacceptors molecules accumulation of in the cytosol of many organisms, with the most abundant tRNA being complementary to the so called “favored” codons(*34*). Furthermore levels of tRNAs’ concentration was found to be tissue dependent in human(*15*) and the use of particular codons was proposed to reflect the gene co-adaptation to a local isoacceptor abundance.(*14*) Such a phenomenon, if present in plants, does not seem to contribute in differentiating Cluster A and B. Indeed tissue specific genes tend to use similarly the predicted *Arabidopsis* favored codons as confirmed by the ANOVA analysis which revealed a not significant difference in the CAI values between the two groups. Moreover, codons responsible of differentiating the two groups (Table 3) do not reflect a biased use of tRNA with high gene copy number.

These observations seem to point towards a direct involvement of the sequences length in determining a codon bias divergence. However the variables that significantly differentiated Cluster A and B can not explain, when accounted singularly, the observed codon usage. Indeed, tissues such as Pollen and Sepals should exhibit, in theory, a similar codon bias in according to the comparable average exon length (aprrox. 250np). Similarly this should be valid for Seedling and Shoot Apex which share an average coding sequence length of approximately 1700 bp. If so, a correct reassignment of the genes to the corresponding

tissue specific classes based exclusively on the gRSCU values would be unlikely. Contrarily a gene reclassification based on a quadratic discriminant analysis revealed a close to 100% correct reassignment for Seedling and Shoot Apex and Sepals, while an approx. 80% reassignment was achieved for Cauline and Sepals.

Although the results of the gene ontology analysis (Table 2) may be evocative of an association between codon bias and gene function, as firstly proposed by Chiapello,(*30*) an analysis of the specific synonymous codon usage within particular gene families will be necessary. These findings probably underlie either a combination of shaping forces or the involvement of selective pressures which are different from the ones analyzed in the present paper.

Cluster and quadratic discriminant analyses seem to suggest that the codon bias can be used as a mean of differentiating genes specifically expressed in different tissues and this leads to the hypothesis that the codon structure may be somehow shaped by local forces. At the best of our knowledge a tissue specific synonymous codon usage in plants was suggested only for reproductive tissues in *Zea mais* and *Triticum aestivum* by Whittle and co-workers(*36*). The authors reported a greater codon bias in female rather than in male organs and gametes and suggested this difference to be due to a gender specific translational selection since no differences were observed in terms of mutational bias, protein length or in biological function of the considered genes.

If a particular codon structure is advantageous for a gene to be expressed in a certain tissue, one may wonder which set of codon will be used by genes with wider expression. One possible hypothesis would be that several selective forces may act concurrently, although it is reasonable to think that certain tissues can exert a dominant influence. In order to speculate whether a generic tissue T influences the codon structure of a gene that is expressed in T and in a further set of tissues G its Euclidean distance in terms of 58 gRSCU values was calculated from both the T-specific and the G specific genes (see Methods). Such analysis was performed for genes featuring an expression breadth comprised between 2 and 14, and the trends obtained for each tissue (Figure S3) were averaged in order to obtain the curves reported in Figure 4. Results clearly show that, in general, a gene expressed in a tissue T is more similar to the T-specific genes rather than to the G-specific and this is especially true at low expression breadth values. Convergence of the two trends represented in Figure 3 (white bars) and consequent overlap for expression breadth > 6 are consistent with a codon structure that is diverse from both T- and G-specific genes and is more similar to the HK.

When the analysis was repeated using only genes that in T had the peak of expression, the convergence rate was minor and lower Euclidean distances from T- rather than G-specific genes were observed for the whole range of examined expression breadth classes (Figure 3 – red bars). This observation represents a further proof of evidence that the use of a specific codons' set may expedite gene expression in a certain tissue, since this theory would certainly be more valid for genes that in that tissue are more expressed than anywhere else in the plant.

Given a tissue-specific involvement in shaping the codon bias of a gene another intriguing question need to be answered: are the tissue contributions equally responsible of the final synonymous codon usage of a widely expressed gene? Do tissues with leading effects exist? The more a tissue influences the usage of a particular codons' set the higher will be the differences between the two distances reported in Figure 3 at expression breadth 2 in the single tissue plots (Figure S2). The ratio between these two values is reported in Figure 4 from which a heterogeneous picture emerges. Indeed tissue such as Pollen, Root and Seedling exert a leading effect on shaping the codon structure of the genes that are expressed in these tissues whereas no significant effects seems to derive from the expression in Sepals (ratio ~ 1). On the other hand the opposite trend is observed for genes expressed in Cotyledons, that is genes featuring a peak of expression in this tissue and contemporaneously expressed in a second tissue G, have a codon structure that resemble more the one observed in the G-specific genes. These observations lead to the conclusion that a tissue effect exists which may contribute in shaping the codon usage of Arabidopsis genes, however such influence is highly variable in terms of both strength and direction.

CONCLUSION

The aim of this study was to investigate the possible effects of a tissue specific influence on the synonymous codon usage in *Arabidopsis thaliana*. Proofs of evidence exist that seems to indicate a direct involvement of local pressures behind the gene structure. Indeed tissue specific genes use, on average, a synonymous codon set that proved to be so peculiar to allow a correct reassignment (up to 100%) for most of the analyzed tissues via a quadratic discriminant analysis. Widely expressed genes seem to undergo the influence of the different tissues in which it is expressed. On the other hand these tissue specific shaping forces are not equally strong and some tissue more than other can have a dominant effect on the final gene structure.

Reference List

1. J. L. King, T. H. Jukes, *Science* **164**, 788 (1969).
2. M. Gouy, C. Gautier, *Nucleic Acids Res.* **10**, 7055 (1982).
3. P. M. Sharp, T. M. Tuohy, K. R. Mosurski, *Nucleic Acids Res.* **14**, 5125 (1986).
4. L. Duret, D. Mouchiroud, *Proc. Natl. Acad. Sci. U. S. A* **96**, 4482 (1999).
5. M. Stenico, A. T. Lloyd, P. M. Sharp, *Nucleic Acids Res.* **22**, 2437 (1994).
6. P. M. Sharp, M. Stenico, J. F. Peden, A. T. Lloyd, *Biochem. Soc. Trans.* **21**, 835 (1993).
7. T. Ikemura, *J. Mol. Biol.* **146**, 1 (1981).
8. G. Bernardi, *Annu. Rev. Genet.* **29**, 445 (1995).
9. L. Duret, L. D. Hurst, *Mol. Biol. Evol.* **18**, 757 (2001).
10. A. O. Urrutia, L. D. Hurst, *Genetics* **159**, 1191 (2001).
11. A. G. Nackley *et al.*, *Science* **314**, 1930 (2006).
12. C. Kimchi-Sarfaty *et al.*, *Science* **315**, 525 (2007).
13. J. L. Parmley, L. D. Hurst, *Bioessays* **29**, 515 (2007).
14. J. B. Plotkin, H. Robins, A. J. Levine, *Proc. Natl. Acad. Sci. U. S. A* **101**, 12588 (2004).
15. K. A. Dittmar, J. M. Goodenbour, T. Pan, *PLoS Genet.* **2**, e221 (2006).
16. D. Kotlar, Y. Lavner, *BMC Genomics* **7**, 67 (2006).
17. M. Semon, J. R. Lobry, L. Duret, *Mol. Biol. Evol.* **23**, 523 (2006).
18. A. E. Vinogradov, *Nucleic Acids Res.* **31**, 5212 (2003).
19. J. M. Comeron, *Genetics* **167**, 1293 (2004).
20. C. A. Whittle, M. R. Malik, J. E. Krochko, *BMC Genomics* **8**, 169 (2007).
21. T. M. Hambuch, J. Parsch, *Genetics* **170**, 1691 (2005).
22. Y. Y. Waldman, T. Tuller, T. Shlomi, R. Sharan, E. Ruppin, *Nucleic Acids Res.* (2010).
23. D. Swarbreck *et al.*, *Nucleic Acids Res.* **36**, D1009 (2008).
24. P. M. Sharp, W. H. Li, *Nucleic Acids Res.* **15**, 1281 (1987).
25. P. Shannon *et al.*, *Genome Res.* **13**, 2498 (2003).
26. S. Maere, K. Heymans, M. Kuiper, *Bioinformatics.* **21**, 3448 (2005).

27. P. Mukhopadhyay, S. Basak, T. C. Ghosh, *DNA Res.* **15**, 347 (2008).

28. J. L. Parmley, L. D. Hurst, *Mol. Biol. Evol.* **24**, 1600 (2007).

29. S. Camiolo, D. Rau, A. Porceddu, *PLoS One.* **4**, e6356 (2009).

30. H. Chiapello, F. Lisacek, M. Caboche, A. Henaut, *Gene* **209**, GC1 (1998).

31. H. H. Le, B. Seraphin, *Cell* **133**, 213 (2008).

32. E. Willie, J. Majewski, *Trends Genet.* **20**, 534 (2004).

33. M. dos Reis, L. Wernisch, *Mol. Biol. Evol.* **26**, 451 (2009).

34. S. I. Wright, C. B. Yau, M. Looseley, B. C. Meyers, *Mol. Biol. Evol.* **21**, 1719 (2004).

35. L. Duret, D. Mouchiroud, *Proc. Natl. Acad. Sci. U. S. A* **96**, 4482 (1999).

36. C. A. Whittle, M. R. Malik, J. E. Krochko, *BMC Genomics* **8**, 169 (2007).

Supplemental material

Tissue	Replicates files	Tissue	Replicates files	Tissue	Replicates files
Carpels	GSM131594 GSM131595 GSM131596	**Pollen**	GSM131636 GSM131637 GSM131638	**Senescing Leaves**	GSM131537 GSM131538 GSM131539
Cauline Leaves	GSM131540 GSM131541 GSM131542	**Root**	GSM131655 GSM131656 GSM131657	**Sepals**	GSM131585 GSM131586 GSM131587
Cotyledons	GSM131495 GSM131496 GSM131497	**Rosette Leaf**	GSM131510 GSM131511 GSM131512	**Shoot Apex**	GSM131649 GSM131650 GSM131651
Hypocotyl	GSM131643 GSM131644 GSM131645	**Seedling Green Part**	GSM131471 GSM131472 GSM131473	**Siliqua w/ seeds**	GSM131688 GSM131689 GSM131690
Petals	GSM131588 GSM131589 GSM131590	**Seed**	GSM131694 GSM131695 GSM131696	**Stamen**	GSM131591 GSM131592 GSM131593

Table S1: Microarray experiment files used for the determination of gene expression (3 replicates for each tissue). Files can be freely downloaded from the website http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=FILE_NAME (where FILE_NAME has to be substituted with the file name reported in table).

	Carpels	Cauline leaves	Cotyledon	Hypocotyl	Petals	Pollen	Root	Rosette leaf	Seedling	Seed	Senescing leaves	Sepals	Shoot Apex	Siliqua	Stamens	HK
N. Genes	94	23	17	77	54	243	332	9	50	373	188	24	98	165	121	1738
ttt(Phe)	1.01	0.94	1.05	1	1.01	1.05	0.92	1.07	1.1	1	1.06	0.95	1.05	0.97	1.05	0.94
ttc(Phe)	0.99	1.06	0.95	1	0.99	0.95	1.08	0.93	0.9	0.95	0.94	1.05	0.91	1.03	0.95	1.06
tta(Leu)	0.95	0.97	0.87	0.8	0.96	1	0.77	0.71	0.84	0.88	0.85	0.72	0.78	0.86	0.92	0.58
ttg(Leu)	1.28	1.32	1.46	1.29	1.22	1.44	1.28	1.27	1.42	1.33	1.42	1.25	1.43	1.3	1.31	1.39
ctt(Leu)	1.45	1.32	1.51	1.68	1.5	1.4	1.5	1.2	1.61	1.47	1.52	1.35	1.51	1.54	1.47	1.67
ctc(Leu)	1.07	1.11	1.31	1.14	1.1	0.96	1.26	1.23	0.83	1.03	0.98	1.43	0.9	1.18	1.05	1.14
cta(Leu)	0.64	0.61	0.57	0.64	0.67	0.66	0.68	1.17	0.64	0.6	0.63	0.83	0.67	0.64	0.73	0.52
ctg(Leu)	0.61	0.66	0.27	0.46	0.55	0.54	0.51	0.43	0.66	0.54	0.59	0.43	0.6	0.48	0.52	0.7
att(Ile)	1.18	1.04	1.13	1.13	1.18	1.23	1.05	1.21	1.3	1.11	1.16	1.19	1.28	1.15	1.08	1.26
atc(Ile)	1.08	1.22	1.17	1	1.04	1.04	1.23	1.14	0.86	1.05	1.05	1.09	0.91	1.15	1.09	1.19
ata(Ile)	0.74	0.74	0.7	0.86	0.78	0.74	0.72	0.65	0.84	0.76	0.8	0.73	0.75	0.69	0.83	0.54
gtt(Val)	1.48	1.58	1.58	1.47	1.5	1.62	1.49	1.29	1.67	1.48	1.62	1.16	1.65	1.48	1.43	1.69
gtc(Val)	0.79	0.79	0.75	0.88	0.77	0.74	0.88	0.83	0.68	0.77	0.71	0.86	0.67	0.84	0.84	0.81
gta(Val)	0.64	0.65	0.65	0.57	0.66	0.62	0.55	0.56	0.63	0.59	0.58	0.67	0.59	0.59	0.67	0.47
gtg(Val)	1.08	0.98	1.03	1.09	1.08	1.03	1.08	1.32	1.01	1.05	1.08	1.31	1.02	1.09	1.06	1.02
tct(Ser)	1.71	1.49	1.73	1.57	1.57	1.65	1.56	1.44	1.67	1.55	1.57	1.79	1.63	1.59	1.65	1.8
tcc(Ser)	0.86	0.75	0.69	0.77	0.73	0.73	0.83	0.47	0.71	0.76	0.77	0.83	0.68	0.85	0.67	0.79
tca(Ser)	1.18	1.23	1.16	1.24	1.18	1.23	1.22	1.52	1.16	1.13	1.24	1.03	1.27	1.19	1.37	1.12
tcg(Ser)	0.52	0.71	0.65	0.71	0.77	0.68	0.66	0.82	0.67	0.73	0.56	0.63	0.55	0.69	0.66	0.6
agt(Ser)	0.88	0.88	0.97	0.93	0.91	0.95	0.86	0.89	1.04	0.91	1	0.91	1.04	0.92	0.91	0.89
agc(Ser)	0.85	0.93	0.81	0.79	0.84	0.76	0.87	0.86	0.76	0.76	0.86	0.8	0.71	0.77	0.75	0.81
cct(Pro)	1.48	1.55	1.51	1.36	1.36	1.46	1.45	1.1	1.64	1.37	1.52	1.39	1.51	1.53	1.44	1.65
ccc(Pro)	0.42	0.39	0.33	0.37	0.4	0.39	0.44	0.37	0.44	0.44	0.42	0.54	0.4	0.39	0.46	0.46
cca(Pro)	1.37	1.38	1.46	1.34	1.27	1.39	1.3	1.75	1.28	1.26	1.33	1.32	1.33	1.28	1.32	1.28
ccg(Pro)	0.74	0.68	0.7	0.93	0.97	0.76	0.8	0.78	0.65	0.82	0.73	0.75	0.68	0.8	0.79	0.6

Table S2: gRSCU for tissue specific genes of 15 *Arabidopsis* tissues. Values correspond to the average single gene gRSCU within each tissue class. The number of genes composing each class and the statistics relative to the housekeeping genes are also reported.

	Carpels	Cauline leaves	Cotyledon	Hypocotyl	Petals	Pollen	Root	Rosette leaf	Seedling	Seed	Senescing leaves	Sepals	Shoot Apex	Siliqua	Stamens	HK
act(Thr)	1.33	1.34	1.19	1.29	1.2	1.28	1.22	1.29	1.49	1.18	1.34	1.17	1.39	1.19	1.18	1.48
acc(Thr)	0.79	0.8	0.69	0.81	0.92	0.77	0.89	0.79	0.67	0.79	0.79	0.92	0.71	0.9	0.83	0.89
aca(Thr)	1.2	1.05	1.26	1.23	1.18	1.25	1.19	1.06	1.3	1.14	1.27	1.25	1.28	1.21	1.23	1.1
acg(Thr)	0.68	0.81	0.86	0.67	0.7	0.71	0.7	0.86	0.54	0.79	0.61	0.66	0.54	0.69	0.76	0.52
gct(Ala)	1.63	1.66	1.75	1.68	1.6	1.69	1.53	1.64	1.75	1.55	1.68	1.56	1.75	1.66	1.51	1.89
gcc(Ala)	0.6	0.71	0.55	0.65	0.69	0.58	0.78	0.43	0.55	0.67	0.6	0.76	0.53	0.71	0.71	0.64
gca(Ala)	1.18	1.06	1.11	1.07	0.95	1.1	1.03	1	1.13	0.97	1.09	1.02	1.11	1.02	1.19	0.98
gcg(Ala)	0.59	0.57	0.59	0.6	0.76	0.63	0.66	0.94	0.56	0.71	0.63	0.65	0.54	0.61	0.59	0.49
tat(Tyr)	1.01	0.9	1.21	0.95	0.93	1.08	0.91	0.98	1.15	0.99	1.03	0.9	1.08	0.96	0.98	0.96
tac(Tyr)	0.99	1.1	0.79	1.05	1.07	0.92	1.09	1.02	0.85	0.96	0.97	1.1	0.89	1.04	1.02	1.05
cat(His)	1.23	1.12	1.41	1.22	1.14	1.27	1.15	1.08	1.29	1.13	1.29	1.05	1.21	1.17	1.26	1.17
cac(His)	0.77	0.88	0.59	0.78	0.86	0.73	0.85	0.92	0.71	0.81	0.71	0.95	0.75	0.83	0.74	0.83
caa(gln)	1.15	1.3	1.18	1.23	1.2	1.18	1.26	1.13	1.09	1.14	1.21	1.26	1.1	1.3	1.22	0.95
cag(gln)	0.85	0.7	0.82	0.77	0.8	0.82	0.74	0.87	0.91	0.8	0.79	0.74	0.87	0.7	0.78	1.05
aat(Asn)	1.06	0.86	0.99	0.97	1.02	1.08	0.89	1.12	1.15	0.95	1.06	0.85	1.07	0.97	0.99	0.98
aac(Asn)	0.94	1.14	1.01	1.03	0.98	0.92	1.11	0.88	0.85	0.99	0.94	1.15	0.9	1.03	1.01	1.02
aaa(Lys)	1	0.98	0.92	1.01	1.06	1.01	0.99	0.79	1.01	0.97	1.03	0.99	0.95	1.05	1.11	0.85
aag(Lys)	1	1.02	1.08	0.99	0.94	0.99	1.01	1.21	0.99	0.98	0.97	1.01	1.01	0.95	0.89	1.15
gat(Asp)	1.31	1.35	1.28	1.3	1.4	1.4	1.24	1.39	1.41	1.3	1.36	1.12	1.34	1.31	1.31	1.36
gac(Asp)	0.69	0.65	0.72	0.7	0.6	0.6	0.76	0.61	0.59	0.65	0.64	0.88	0.62	0.69	0.69	0.64
gaa(glu)	1.04	1.04	0.98	1.02	1.01	1.07	1.02	1.03	1.08	0.99	1.06	1.04	1.05	1.07	1.06	0.94
gag(glu)	0.96	0.96	1.02	0.98	0.99	0.93	0.98	0.97	0.92	0.96	0.94	0.96	0.92	0.93	0.94	1.06
tgt(Cys)	1.22	1.07	1.06	1.16	1.23	1.21	1.16	1.25	1.13	1.16	1.21	1.34	1.16	1.21	1.2	1.12
tgc(Cys)	0.78	0.93	0.94	0.84	0.77	0.79	0.84	0.75	0.87	0.78	0.79	0.66	0.8	0.79	0.8	0.88
aga(Arg)	2.07	2.22	2.2	2.39	2.33	2.18	2.2	2.59	2.05	2.09	2.3	2.09	2.1	2.13	2.19	1.8
agg(Arg)	1.19	1.17	1.03	1.19	1.14	1.24	1.14	1.4	1.31	1.21	1.2	1.15	1.27	1.16	1.15	1.4
cga(Arg)	0.81	0.65	0.71	0.77	0.7	0.7	0.67	0.65	0.74	0.69	0.68	0.59	0.66	0.64	0.66	0.6
cgt(Arg)	0.91	0.96	0.83	0.77	0.9	0.95	1.07	0.7	0.9	0.9	0.89	1.16	0.86	1	0.96	1.28
cgc(Arg)	0.46	0.41	0.52	0.4	0.4	0.38	0.43	0.28	0.45	0.41	0.39	0.4	0.42	0.57	0.49	0.47
cgg(Arg)	0.56	0.58	0.72	0.47	0.51	0.55	0.5	0.38	0.54	0.54	0.53	0.6	0.57	0.49	0.55	0.46

Table S2: continued.

	Carpels	Cauline leaves	Cotyledon	Hypocotyl	Petals	Pollen	Root	Rosette leaf	Seedling	Seed	Senescing leaves	Sepals	Shoot Apex	Siliqua	Stamens	HK
ggt(gly)	1.31	1.14	1.22	1.3	1.24	1.3	1.22	1.4	1.36	1.28	1.25	1.2	1.31	1.36	1.22	1.46
ggc(gly)	0.61	0.62	0.54	0.6	0.57	0.55	0.62	0.53	0.51	0.6	0.58	0.79	0.57	0.61	0.58	0.54
gga(gly)	1.46	1.68	1.76	1.41	1.49	1.54	1.54	1.4	1.44	1.39	1.5	1.43	1.38	1.44	1.61	1.45
ggg(gly)	0.62	0.56	0.47	0.68	0.7	0.61	0.62	0.68	0.69	0.61	0.67	0.58	0.66	0.59	0.58	0.55

Table S2: continued.

	Carpels	Cauline leaves	Cotyledon	Hypocotyl	Petals	Pollen	Root	Rosette leaf	Seedling	Seed	Senescing leaves	Sepals	Shoot Apex	Siliqua	Stamens	HK
ttt(Phe)	1.02	0.98	1.05	1.03	1.08	1.06	0.94	0.97	1.13	1.04	1.06	0.94	1.11	0.98	1.03	1.02
ttc(Phe)	0.98	1.02	0.95	0.97	0.92	0.94	1.06	1.03	0.87	0.96	0.94	1.06	0.89	1.02	0.97	0.98
tta(Leu)	0.87	0.97	0.87	0.82	0.97	0.94	0.79	0.8	0.87	0.91	0.82	0.8	0.84	0.88	0.93	0.67
ttg(Leu)	1.36	1.28	1.42	1.3	1.3	1.46	1.27	1.29	1.42	1.36	1.39	1.28	1.46	1.3	1.28	1.37
ctt(Leu)	1.46	1.41	1.52	1.68	1.51	1.47	1.49	1.4	1.58	1.5	1.52	1.36	1.5	1.55	1.51	1.66
ctc(Leu)	0.95	1.14	1.24	1.06	0.96	0.9	1.21	1.22	0.78	1.02	0.96	1.24	0.85	1.11	0.99	0.98
cta(Leu)	0.68	0.63	0.64	0.61	0.66	0.65	0.7	0.8	0.64	0.63	0.67	0.79	0.67	0.64	0.72	0.56
ctg(Leu)	0.69	0.57	0.31	0.53	0.6	0.58	0.53	0.47	0.71	0.58	0.65	0.52	0.66	0.52	0.57	0.75
att(Ile)	1.24	1.06	1.22	1.16	1.15	1.23	1.05	1.17	1.3	1.18	1.19	1.11	1.35	1.16	1.11	1.32
atc(Ile)	0.98	1.1	1.06	0.99	1	0.99	1.19	1.25	0.87	1.05	0.99	1.07	0.88	1.13	1.05	1.07
ata(Ile)	0.77	0.84	0.72	0.85	0.85	0.78	0.76	0.58	0.83	0.77	0.82	0.82	0.77	0.71	0.85	0.6
gtt(Val)	1.58	1.51	1.68	1.53	1.6	1.61	1.49	1.41	1.72	1.54	1.64	1.27	1.7	1.5	1.44	1.71
gtc(Val)	0.71	0.84	0.75	0.81	0.7	0.71	0.88	0.87	0.64	0.76	0.7	0.9	0.67	0.81	0.79	0.76
gta(Val)	0.63	0.67	0.63	0.58	0.69	0.66	0.59	0.55	0.63	0.61	0.6	0.58	0.62	0.61	0.69	0.53
gtg(Val)	1.08	0.97	0.94	1.08	1.02	1.01	1.04	1.16	1.01	1.09	1.05	1.26	1.01	1.08	1.07	1
tct(Ser)	1.77	1.48	1.76	1.58	1.73	1.68	1.61	1.52	1.73	1.63	1.63	1.55	1.68	1.63	1.67	1.78
tcc(Ser)	0.71	0.74	0.64	0.79	0.7	0.72	0.82	0.6	0.72	0.73	0.74	0.82	0.65	0.81	0.72	0.74
tca(Ser)	1.25	1.27	1.23	1.27	1.24	1.29	1.24	1.35	1.19	1.21	1.29	1.1	1.36	1.24	1.34	1.21
tcg(Ser)	0.53	0.63	0.61	0.65	0.68	0.6	0.64	0.86	0.62	0.7	0.51	0.71	0.53	0.65	0.6	0.56
agt(Ser)	0.97	0.93	0.97	0.93	0.89	0.97	0.86	0.92	1	0.97	1.01	0.99	1.06	0.92	0.92	0.94
agc(Ser)	0.78	0.96	0.79	0.78	0.76	0.74	0.82	0.75	0.74	0.76	0.81	0.83	0.73	0.75	0.74	0.76
cct(Pro)	1.51	1.54	1.53	1.41	1.45	1.52	1.37	1.19	1.66	1.45	1.53	1.43	1.52	1.51	1.47	1.67
ccc(Pro)	0.43	0.43	0.35	0.4	0.4	0.4	0.39	0.32	0.41	0.42	0.43	0.48	0.42	0.38	0.46	0.45
cca(Pro)	1.41	1.38	1.53	1.35	1.3	1.39	1.54	1.51	1.33	1.32	1.38	1.3	1.39	1.34	1.34	1.31
ccg(Pro)	0.65	0.65	0.58	0.84	0.86	0.69	0.7	0.98	0.6	0.82	0.66	0.79	0.67	0.77	0.73	0.57

Table S3: pRSCU for tissue specific genes of 15 *Arabidopsis* tissues. Values correspond to RSCU of the super-sequence obtained by concatenating the set of sequences within each tissue class. Statistics relative to the housekeeping genes are also reported.

	Carpels	Cauline leaves	Cotyledon	Hypocotyl	Petals	Pollen	Root	Rosette leaf	Seedling	Seed	Senescing leaves	Sepals	Shoot Apex	Siliqua	Stamens	HK
act(Thr)	1.39	1.41	1.24	1.26	1.22	1.3	1.21	1.2	1.44	1.24	1.33	1.18	1.38	1.24	1.23	1.49
acc(Thr)	0.76	0.79	0.63	0.84	0.87	0.74	0.88	0.71	0.67	0.79	0.76	0.91	0.7	0.9	0.81	0.83
aca(Thr)	1.27	1.16	1.4	1.28	1.22	1.32	1.23	1.11	1.33	1.22	1.32	1.17	1.4	1.24	1.3	1.19
acg(Thr)	0.58	0.64	0.73	0.61	0.7	0.64	0.68	0.98	0.56	0.75	0.59	0.74	0.53	0.63	0.66	0.49
gct(Ala)	1.74	1.6	1.71	1.67	1.72	1.68	1.54	1.59	1.78	1.6	1.68	1.48	1.8	1.66	1.58	1.88
gcc(Ala)	0.6	0.63	0.54	0.6	0.62	0.55	0.78	0.43	0.55	0.66	0.6	0.8	0.53	0.69	0.71	0.61
gca(Ala)	1.18	1.17	1.11	1.13	1.05	1.19	1.05	0.98	1.16	1.04	1.14	1.03	1.2	1.04	1.18	1.06
gcg(Ala)	0.49	0.6	0.64	0.61	0.6	0.58	0.63	1	0.52	0.7	0.58	0.68	0.48	0.61	0.54	0.46
tat(Tyr)	1.07	0.95	1.05	0.96	1.08	1.08	0.88	0.93	1.19	1.05	1.04	0.87	1.15	0.99	0.97	1.02
tac(Tyr)	0.93	1.05	0.95	1.04	0.92	0.92	1.12	1.07	0.81	0.95	0.96	1.13	0.85	1.01	1.03	0.98
cat(His)	1.29	1.19	1.51	1.27	1.13	1.29	1.17	1.1	1.34	1.21	1.29	1.07	1.27	1.19	1.27	1.23
cac(His)	0.71	0.81	0.49	0.73	0.87	0.71	0.83	0.9	0.66	0.79	0.71	0.93	0.73	0.81	0.73	0.77
caa(gln)	1.07	1.24	1.27	1.18	1.16	1.17	1.24	1.22	1.1	1.17	1.18	1.24	1.12	1.26	1.22	0.97
cag(gln)	0.93	0.76	0.73	0.82	0.84	0.83	0.76	0.78	0.9	0.83	0.82	0.76	0.88	0.74	0.78	1.03
aat(Asn)	1.07	0.98	1	0.96	1.03	1.09	0.9	1.03	1.18	1.03	1.05	0.87	1.13	0.97	0.99	1.07
aac(Asn)	0.93	1.02	1	1.04	0.97	0.91	1.1	0.97	0.82	0.97	0.95	1.13	0.87	1.03	1.01	0.93
aaa(Lys)	0.99	1.01	0.98	1.02	1.08	1.01	1	0.88	1.02	0.98	1.03	1.03	0.96	1.06	1.09	0.86
aag(Lys)	1.01	0.99	1.02	0.98	0.92	0.99	1	1.13	0.98	1.02	0.97	0.97	1.04	0.94	0.91	1.14
gat(Asp)	1.35	1.39	1.38	1.34	1.38	1.41	1.25	1.37	1.42	1.35	1.38	1.15	1.42	1.31	1.32	1.39
gac(Asp)	0.65	0.61	0.62	0.66	0.62	0.59	0.75	0.63	0.58	0.65	0.62	0.85	0.58	0.69	0.68	0.61
gaa(glu)	1.06	1	1.12	1	0.99	1.07	1.03	1.04	1.08	1.03	1.06	1.02	1.08	1.06	1.1	0.99
gag(glu)	0.94	1	0.88	1	1.01	0.93	0.97	0.96	0.92	0.97	0.94	0.98	0.92	0.94	0.9	1.01
tgt(Cys)	1.19	1.07	1.09	1.13	1.25	1.2	1.17	1.39	1.22	1.19	1.2	1.24	1.17	1.21	1.22	1.14
tgc(Cys)	0.81	0.93	0.91	0.88	0.75	0.8	0.83	0.61	0.78	0.81	0.8	0.76	0.83	0.79	0.78	0.86
aga(Arg)	2.18	2.3	2.37	2.34	2.39	2.26	2.22	2.33	2.04	2.17	2.32	2.05	2.17	2.29	2.21	1.83
agg(Arg)	1.27	1.13	1.06	1.22	1.19	1.21	1.14	1.27	1.34	1.24	1.22	1.11	1.32	1.12	1.13	1.37
cga(Arg)	0.7	0.75	0.82	0.71	0.78	0.72	0.7	0.71	0.76	0.73	0.67	0.68	0.71	0.67	0.71	0.64
cgt(Arg)	0.88	0.92	0.9	0.79	0.77	0.9	0.99	0.86	0.91	0.89	0.86	1.1	0.83	0.98	0.91	1.2
cgc(Arg)	0.44	0.43	0.29	0.39	0.35	0.37	0.42	0.35	0.43	0.38	0.39	0.42	0.41	0.43	0.45	0.46
cgg(Arg)	0.53	0.47	0.57	0.55	0.52	0.54	0.53	0.48	0.53	0.6	0.55	0.63	0.57	0.51	0.59	0.51

Table S3: continued.

	Carpels	Cauline leaves	Cotyledon	Hypocotyl	Petals	Pollen	Root	Rosette leaf	Seedling	Seed	Senescing leaves	Sepals	Shoot Apex	Siliqua	Stamens	HK
ggt(gly)	1.32	1.14	1.28	1.3	1.3	1.31	1.25	1.28	1.37	1.32	1.24	1.34	1.32	1.32	1.25	1.47
ggc(gly)	0.61	0.57	0.59	0.56	0.52	0.55	0.6	0.57	0.52	0.58	0.58	0.64	0.58	0.59	0.57	0.53
gga(gly)	1.44	1.67	1.63	1.47	1.47	1.5	1.53	1.51	1.44	1.45	1.49	1.48	1.44	1.5	1.61	1.42
ggg(gly)	0.63	0.62	0.5	0.67	0.71	0.64	0.61	0.63	0.68	0.65	0.7	0.54	0.65	0.6	0.58	0.59

Table S3: continued.

Manova Analysis Statistics

Test	Value	Approx. F	NumDF	DenDF	Prob>F
Wilks' Lambda	0.5913	1.1672	826	24533	0.0007
Pillai's Trace	0.50885	1.1558	826	25312	0.0015
Hotelling-Lawley	0.54314	1.1791	826	25104	0.0003
Roy's Max Root	0.14767	4.5253	59	1808	<.0001

Table S4: Summary of the Manova multivariate tests carried out on TS-gRSCU.

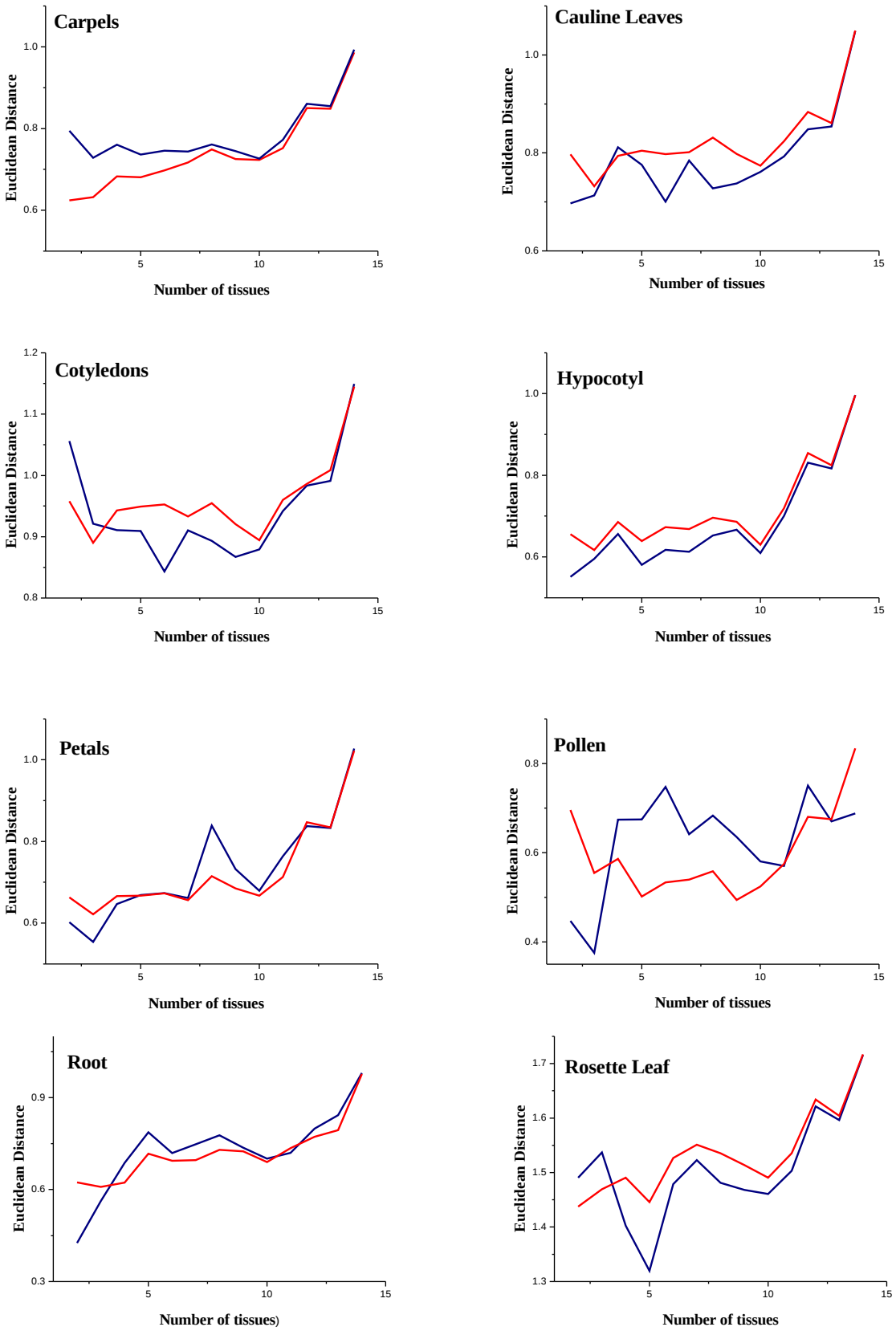


Figure S1: Average euclidean distances of TE-genes from T-specific (blue) and G-specific (red) genes.

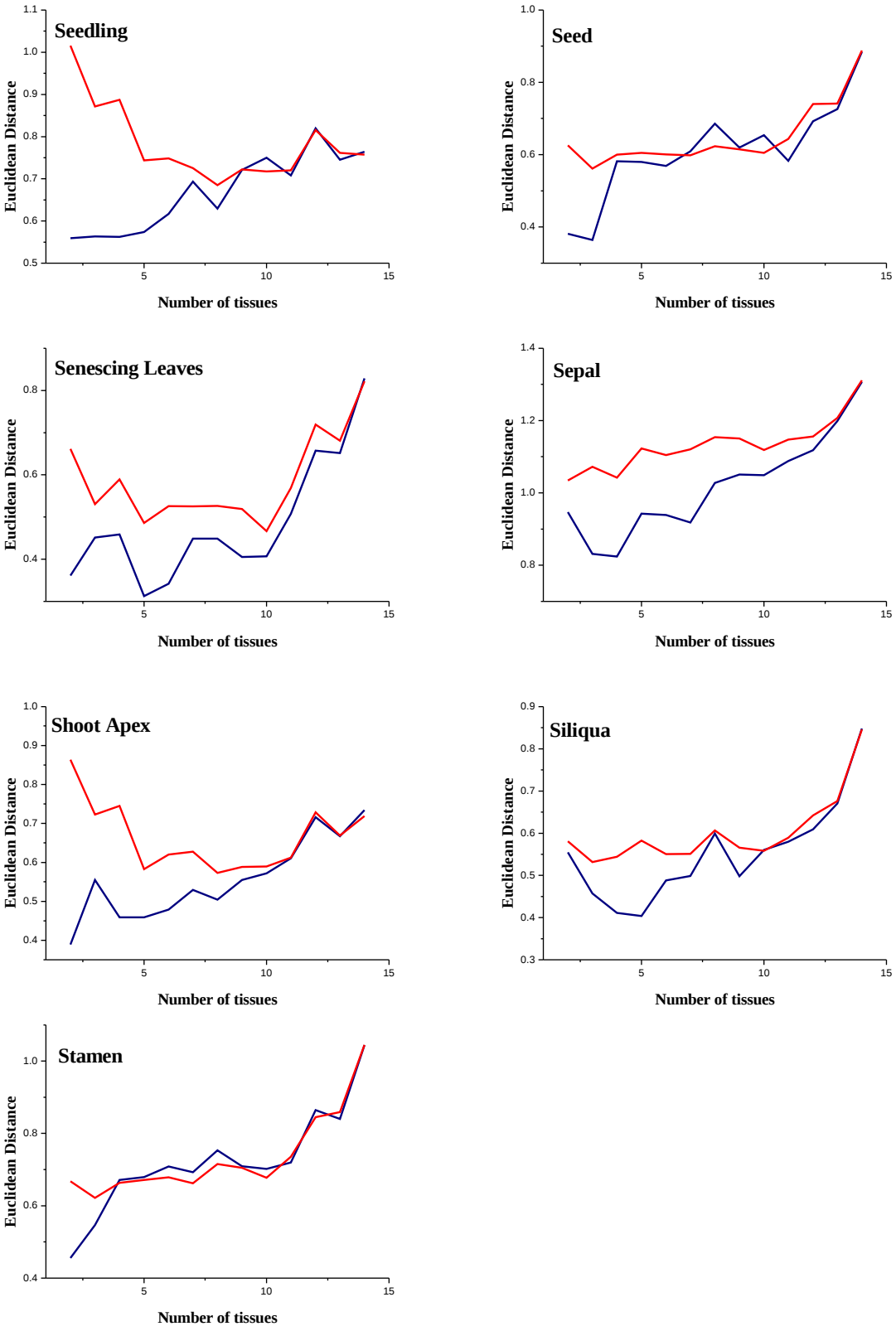


Figure S1: continued

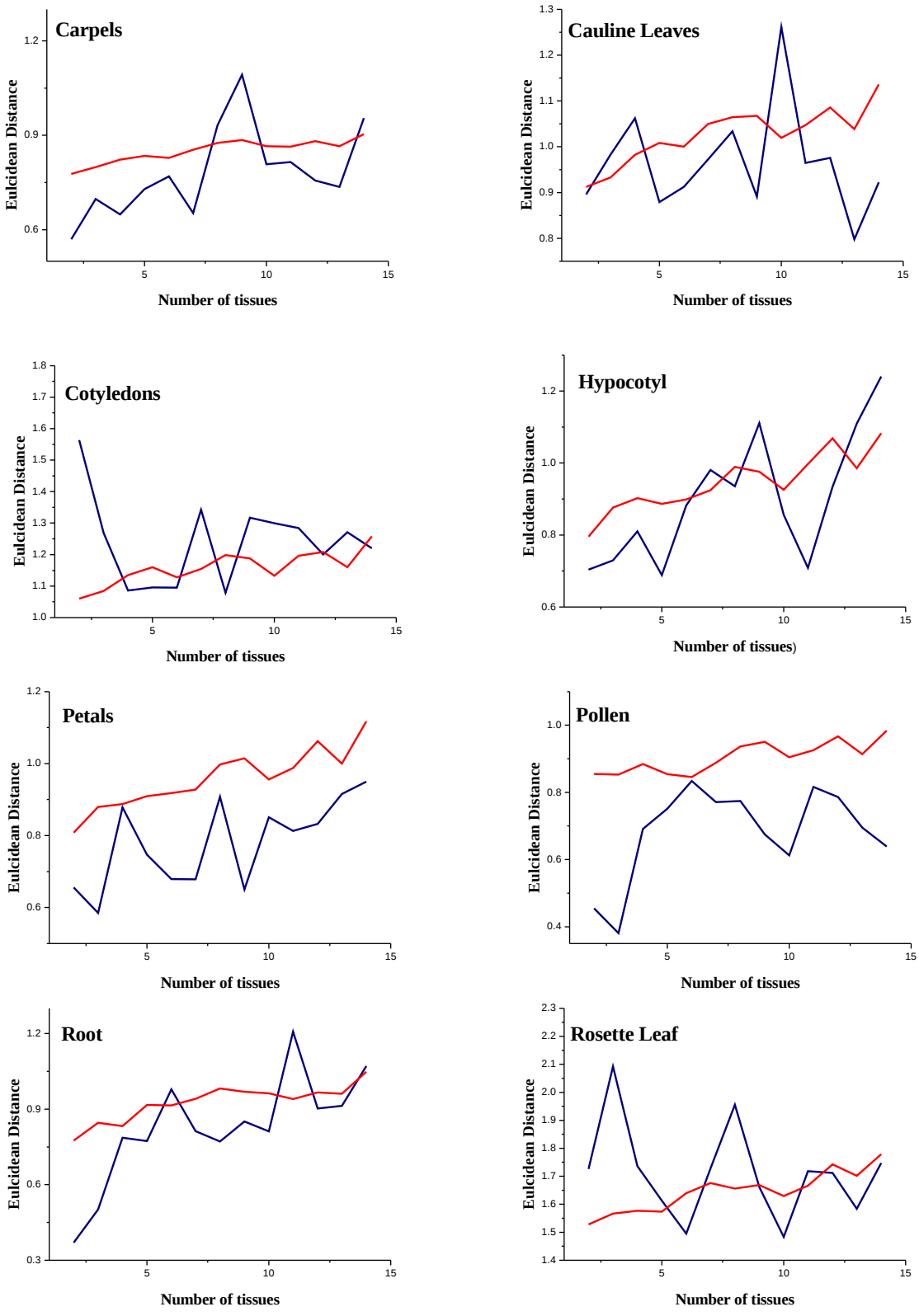
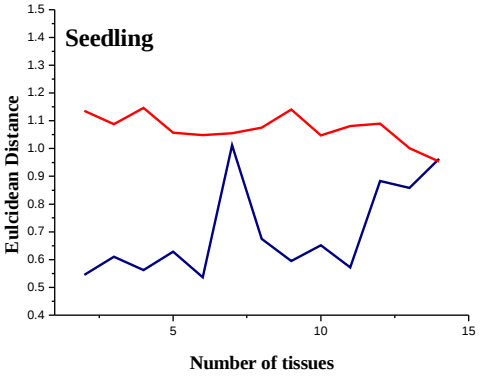
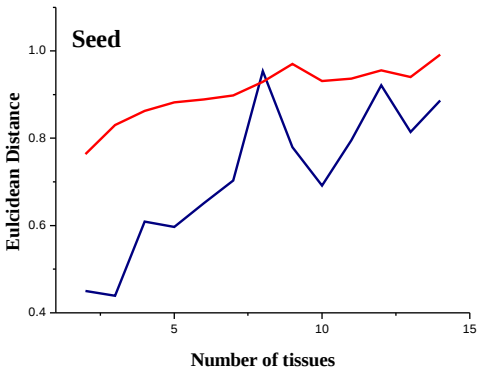
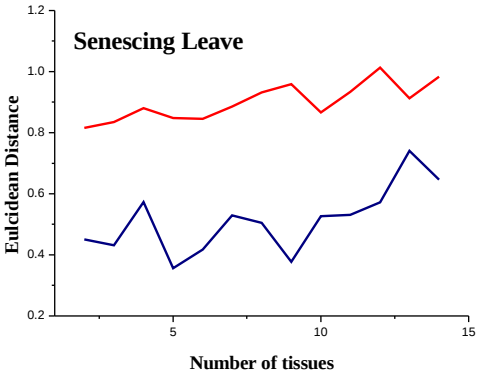
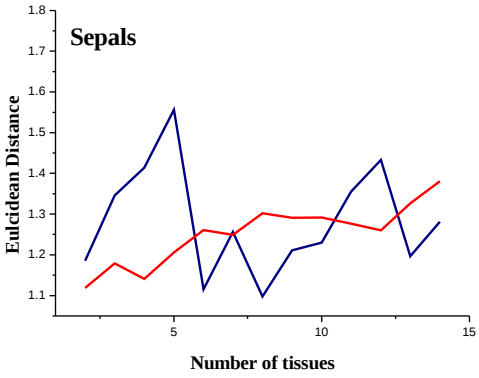


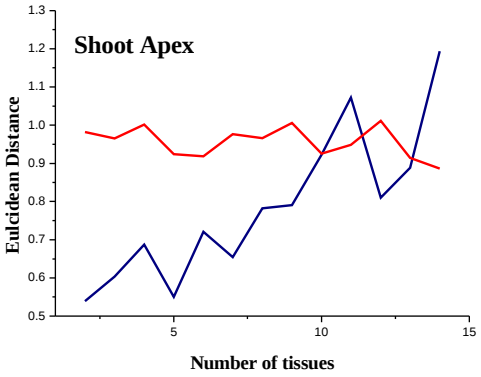
Figure S2: Average euclidean distances of PE-genes from T-specific (blue) and G-specific (red) genes.











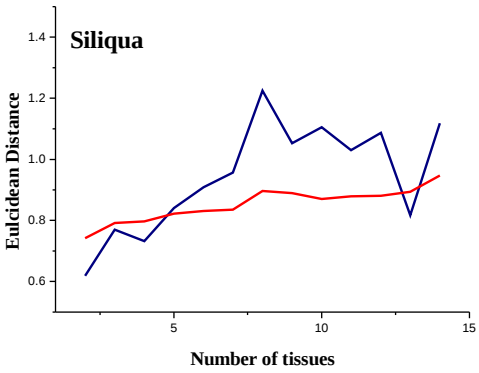


Figure S2: continued

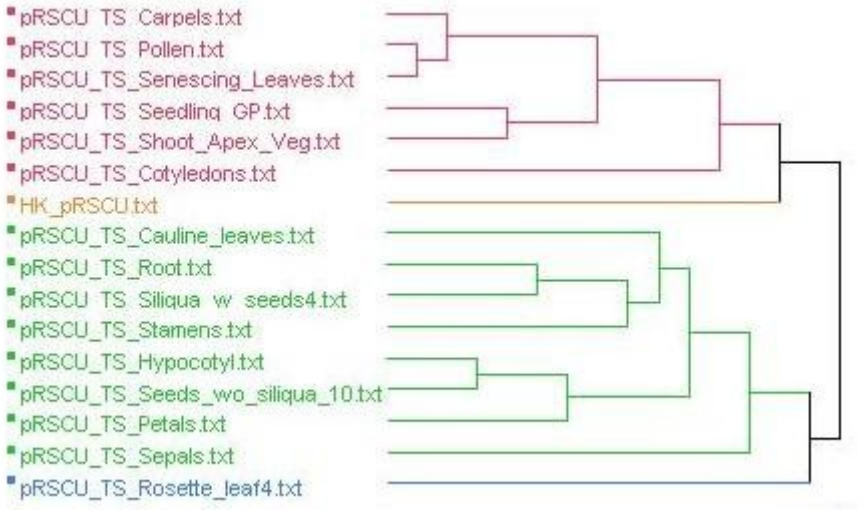


Figure S3: Cluster analysis performed on a dataset comprising TS-RSCU and HK-RSCU obtained by calculating the RSCU on the "pooled" super sequences (pRSCU).

Toward the definition of a compositional signature for plant genes.

ABSTRACT

Genomic DNA sequences display compositional heterogeneity on many scales. In the present paper we used an approach based on ensemble averages to describe tendencies and anomalies in the occurrence of mono di, and tri nucleotides as a function of position in structural regions of plant genes. The analysis of the trends highlighted several compositional features peculiar of either monocots or eudicots that were remarkably uniform within these evolutionary clades. The most evident of these features appeared in the form of gradients of base content along the direction of transcription. Assessment of dinucleotide and trinucleotide biases were referred to expectation based on random union of the components units. The graphs reproducing these biases as function of position highlighted compositional hallmarks that will contribute to characterize the compositional signature of plant genes. Examination of single gene trends confirmed that these features were shared by genes of different length, genomic orientation and overall base composition suggesting that these are essential attributes for gene organization.

INTRODUCTION

Compositional heterogeneity is a common feature of eukaryotic genomes(*1-3*). Many biochemical and molecular studies have focused on descriptions of the structural organization of such a variation at different genomic scales. Large blocks of DNA of homogeneous G+C content were described in warm blooded vertebrates as the main high-scale compositional units(*4*). These blocks, termed isochors, proved to be strongly associated with the genome organization(*5*). Indeed observations from the mammalian genomes indicated that isochores correlated with gene density(*4*), repetitive DNA elements distribution(*6*), chromosomal bands(*7, 8*) and potentially also with replication time in the genome(*7*). Subsequent studies conducted on long genomic sequences have extended the perception of compositional heterogeneity to many multicellular organism belonging to a wide variety of evolutionary taxa, indicating, however, the existence of large genomic regions which do not fit with the classic isochore model(*9*).

The availability of many genomic sequences has provided us with the unprecedented opportunity to perform compositional studies at low-scale level. These analyses are, in time, highlighting a number of genomic features associated to basic cellular mechanisms. Huvet and co-workers(*10*) have recently reported that in more than one quarter of the human genome the nucleotide compositional skew presents characteristic patterns consisting of succession of “N-shaped” structures. Based on these observations the authors have proposed a new model of gene organization which integrates transcription, replication, and chromatin structure(*10*).

Other features that have not yet found a clear causal link are currently interpreted as punctuation marks for low-scale genomic organization. For example spikes in GC content have been associated to the boundaries of transcriptional units in warm blooded and invertebrate species(*11*). An analogous feature seems to be present in plants and fungi genomes where spikes of GC compositional strand bias identify transcriptional start sites(*12*).

Analyses at intragenic scale level have mainly focused on non random usage of synonymous codons. Reports from several species have revealed a relation between synonymous codon usage and position in the coding sequence. For example, enterobacterial genes avoid some codons near the start sites perhaps to limit the formation of secondary structure in the messenger which could interfere with ribosome binding site near the start of translation(*13*).

Hong and co-workers have demonstrated how the pattern of codon usage bias along genes may have different features among species(*14*). In yeast and several prokaryotic species it increases along translational direction which is consistent with purifying selection against nonsense errors.

Drosophila melanogaster codon usage bias is high at the ends and lower in the middle of coding sequence probably as a consequence of the Hill Robertson effect.

Several analyses conducted on coding and genic sequence have indicated the existence of a base compositional bias at the termini of plant genes(*15*, *16*). Various explanations have been proposed for such bias, including bias in codon and aminoacid usage, and mutation related process. However, peculiarities of each species and intrinsic limitations of the experimental setups have, so far, prevented the convergence of this data on a well defined picture. Niimura and co-workers showed that base appearance at the codon third position of the terminal regions of both *Arabidopsis* and *Oryza* genes is extremely biased(*15*). Unfortunately as the analysis considered exclusively the third codon position mainly the involvement of bias in synonymous codon usage could be tested. Wong and co-workers have reported that in the first 1.5 kb of monocots but not eudicots coding sequences there is a negative gradient of G+C content proceeding along the translation direction(*16*). This compositional bias was observed, although with different intensity, for all three codon bases and therefore affected both synonymous codon and aminoacid usages. In this work we present a detailed analysis of compositional bias as function of position in structural regions of plant genes. The investigation considered three different degree of compositional complexity, from mono-nucleotide to trinucleotides and for each of these, the biases were calculated under the hypothesis of random union of the component units. The emerging picture offers novel elements which will be instrumental for defining the compositional signatures of plant genes.

MATHERIALS AND METHODS

Sequence datasets

DNA sequences of two eudicots (*Arabidopsis thaliana*, *Vitis vinifera*) and two monocots (*Oryza sativa* and *Brachypodium distaychon*.) were used in this study. All dataset were filtered out for i) transposons and ii) pseudogenes sequences, iii) mitochondrial and chloroplast genes.. The cDNA sequences of *Vitis vinifera*, were downloaded from the NCBI FTP server (ftp://ftp.ncbi.nih.gov/). The genomic sequence of *Arabidopsis thaliana* (TAIR 8) and rice (v6) were downloaded from the http://www.arabidopsis.org and respectively. The number of transcripts used for the analysis were 23536 for *Arabidopsis thaliana*, 31088 for *Oryza*, 32255 for *Brachypodium* and 56478 for *Vitis*.

Ensemble graphs

Mononucleotide. The compositional profiles along sequences were computed using either sliding (step 1) or adjacent window of different sizes (33, 51 and 99 bases) from which the same picture emerged. The ensemble profile, for each analyzed dataset, was generated by averaging the corresponding base content of each window of all genes at each position along the sequences. The ensemble graphs (referred to genomic coordinates) were calculated on sequences with either intron or exon masked. They were averaged considering the counts of windows without masked sequences. The analyses were carried out both along and opposite to the translational direction using the start and stop codons as reference respectively.

Dinucleotide

Dinucleotide bias was estimated through the odds ratio $\rho(3)$ f(XY)/f(X)*f(Y) where f(X) and f(Y) denotes the frequency of the nucleotide X and Y at respectively, and f(XY) is the frequency of the dinucleotide XY in the sequence window under study.

Dinucleotides frequencies at different frames were computed considering the base codon indicated by the frame subscript. For example for dinucleotides 1_2 it was considered the first and second base of each codon. Accordingly the ρ values were calculated taking into account base frequency at the positions indicated by the frame subscript.

Trinucleotide

Trinucleotide bias $\gamma XYZ(3)$ was estimated through the odds ratio f(XYZ)*(f(X)*f(Y)*f(Z))/f(XY)*f(YZ)*f(XNZ)), where f(XYZ) is the frequency of the trinucleotide XYZ, f(X), f(Y) and f(Z) are the frequency of mononucleotides, and f(XY), f(YZ), f(XNZ) are the frequency of the dinucleotides identifying the given trinucleotide.

For trinucleotide of different frames (1_2_3, 2_3_1 and 3_1_2) we used the same procedures explained for dinucleotides.

Piecewise regression

Piecewise regression is the process of fitting data to possible more than one linear function. To calculate piecewise regression we wrote a C program that selects the most statistically significant linear model that consists of up to two linear equations and calculates the values of the independent variable where the slopes of the linear functions change (breakpoint). The models identified were adopted only when the improvement in explanation of data over simple regression

could not have arisen by chance. To test this hypothesis we used the ANOVA procedure according to Ryan et al(*17*).

RESULTS

To gather a first insight on spatial compositional patterns of plant genes we studied the global profile of mono-nucleotides as a function of position along genic sequences of several monocots (*Oryza sativa* and *Brachypodium distachyon*) and eudicots species (*Arabidopsis thaliana*, *Vitis vinifera*).

Coding sequence trends

The mono-nucleotide graphs revealed interesting compositional features that in some cases allowed to distinguish between monocots and dicots species.

Eudicots. Guanine content of eudicots genes increased along the direction of translation to reach soon a plateau and then slowly decreased for the whole length of the sequence with the exception of the 3' end where a positive spike was evident (figure 1a). Cytosine profiles steeply decreased along the direction of translation to reach a plateau level which was modified only in the second half of the gene leaving place to a steadily increasing trend (figure 1a). Adenine profiles were the most regular of eudicots genes increasing steadily throughout the whole gene length. Finally thymine profiles, first increased and then steadily decreased along the direction of translation (figure 1a).

Monocots. Guanine and cytosine depicted concave-shaped profiles in all analyzed monocots species (figure 1b). The higher content of guanine over cytosine was almost uniformly distributed over the whole length of monocots coding sequences with the exception of the 5' end region where cytosines were more frequent than guanines.

Nearly convexes profiles were observed for both adenine and thymine, with an adenine content being almost constantly higher than thymine. The only exception to such a behaviour was again at the 5' end of the cds where the two profiles were nearly identical.

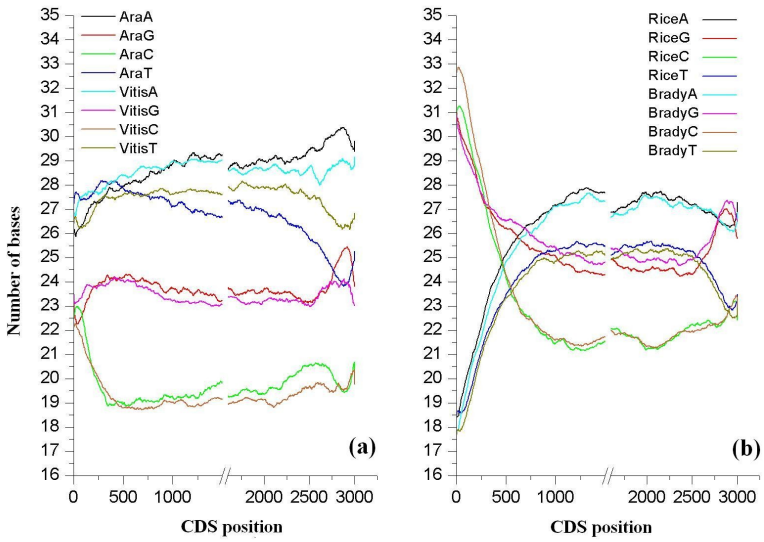


Figure 1: Overall mononucleotides content as function of position in eudocots (*Vitis* and *Arabidopsis*) and monocots (*Oryza* and *Brachypodium*) coding sequences and averaged over all sequences in the dataset with a 99 bp adiacent window.

To analyze the contribution of different codon position to the compositional profiles of figure 1, each point of the plots was resolved in three components attributable to the first second and third base of codons (figure 2a-b).

For each base, the results for the first, second and third bases are similar. In all cases, most of total base variation was attributable to the third codon positions followed by the second and first codon positions. Such an effect was particularly evident for thymine followed by cytosine and adenine and at last guanine.

Ensemble graphs are representative of single gene trends.

Because the above graphs were constructed by considering the ensemble averages of single window positions it is important to verify that they describe the general trends of genes and do not represent artefacts due to compositional differences between groups of genes distinguishable for structural or genomic features. The whole analysis was therefore repeated for the most deeply annotated *Arabidopsis* and *Oryza* on sequences datasets partitioned based on various criteria (gene length, orientation on helix and GC content of coding sequences). The relations between base composition and absolute position in coding sequence (i.e. the distance of each window from the first translation codon) were similar across datasets partitioned in length classes (figure 3a-b). These findings ruled out the hypothesis that the ensemble graphs could have reproduced artefacts due to compositional differences between genes of different length.

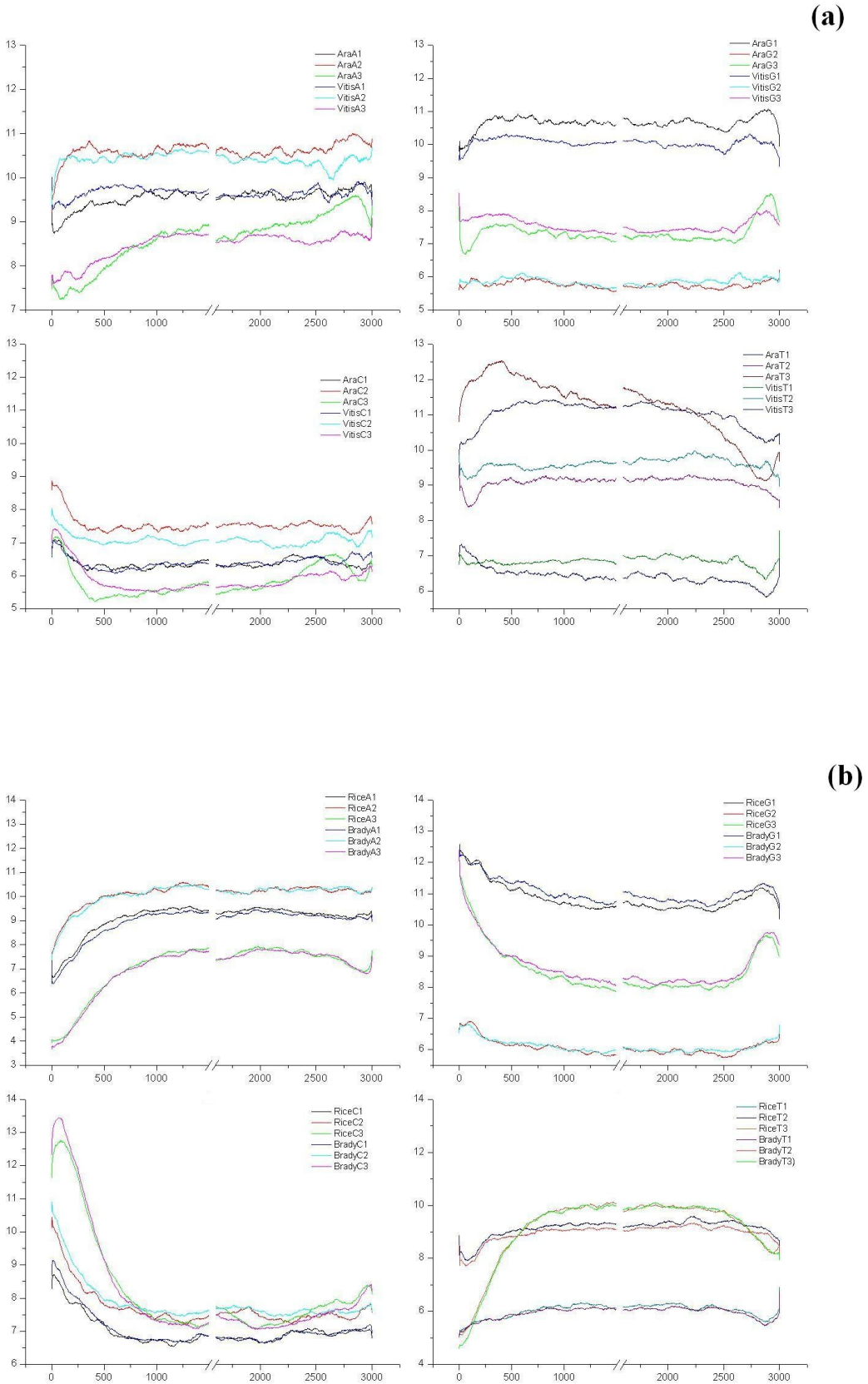


Figure 2: Overall first, second and third codon nucleotide content as function of position in eudicots(a) and monocots(b) coding sequences and averaged over all sequences in the dataset with a 33 bp adiacent window.

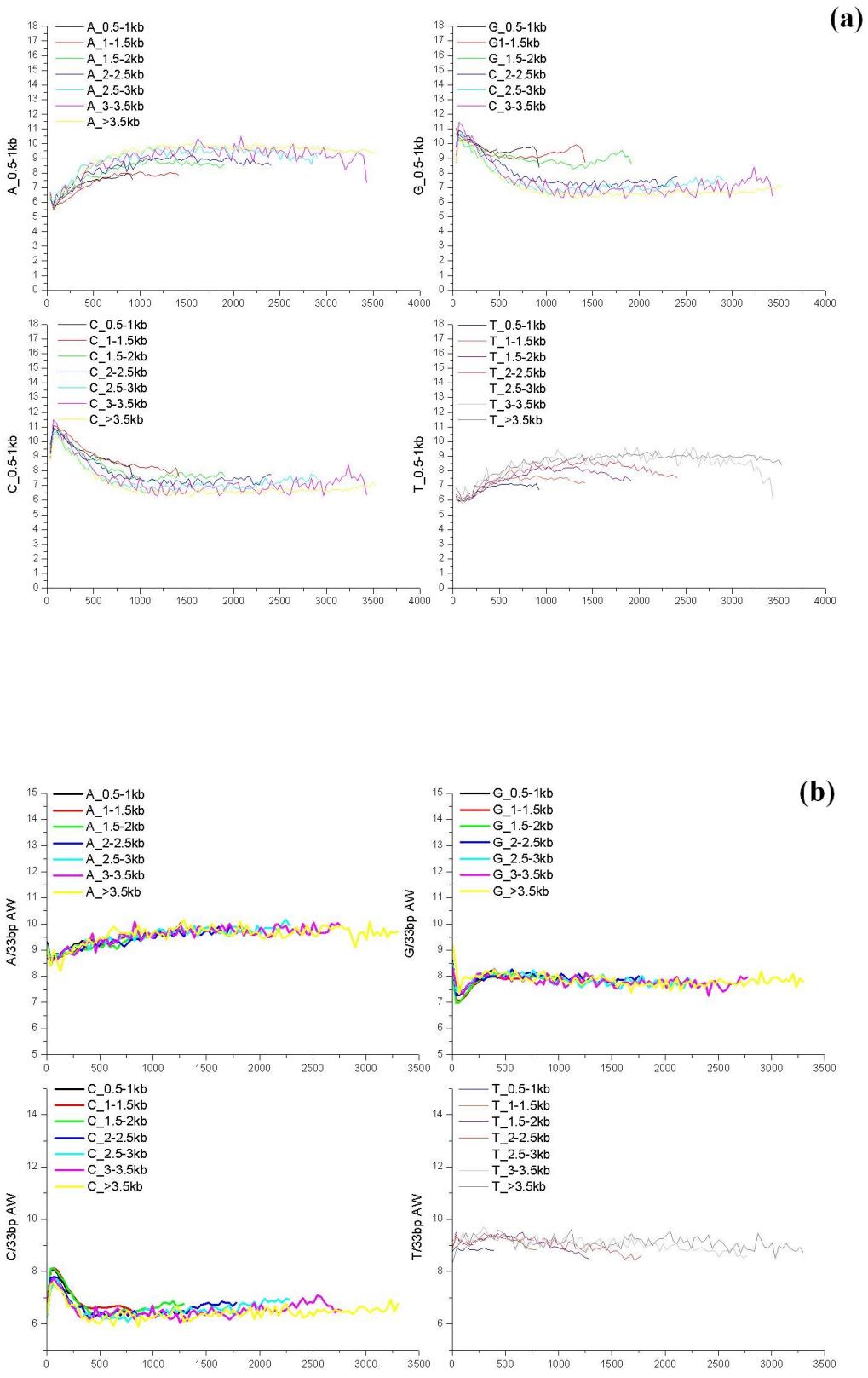


Figure 3: Mononucleotide content of *Arabidopsis* (a) and *Oryza* (b) as function of position starting from the first translation codon and proceeding along translation direction and averaged over all sequences in each dataset with an adicent window of 33bp. The sequences were partitioned in seven dataset based on their length.

Other partition criteria considered the orientation of genes in the chromosomes (i.e forward or reverse) and total G+C content of coding sequences. In all cases the shape of the trends is qualitatively similar to those observed for the original dataset (see Supplemental figures S1a-b and S2-a-b).

The analysis was then scaled at single gene level. Because most of the ensemble trends could be fitted by two-linear models, we analyzed single gene trends by piecewise segmented regression. This analysis returns the segmented linear model which allows the largest improvement in explanation of data over the single linear regression. As a matter of fact, an high proportion (between 61.9 and 70.7%) of single gene trends were better described in the first two kb of length by complex functions with two linear rather than by a simple linear regression (Tables 1 and S3). A combined analysis of the slopes of the fitted segments and of the position of breakpoint indicated that a high proportion of single gene models were compatible with the ensemble graphs. Based on these results we conclude that the ensemble models are *bona fide* representation of genic trends.

	Slope 1								
									
	A	**G**	**C**	**T**	**A**	**G**	**C**	**T**	
gmented Regressi	14.1	1.5	1.1	10.5	7.2	32.6	43.5	9.5	 **Slope 2**
	42.2	14.1	12.5	40.6	0.7	13.7	13.6	1.3	
Linear	13.3	2.3	1.7	2.2	1.6	11	7.8	12	
Not Fitted	20.9	24.8	19.7	23.9					

Table 1: Percentages of the genes that significantly fitted the segmented regression and the linear models. For the segmented regression values where classified by considering the sign of the slope 1 and slope 2 trends.

Trends of genic untranslated sequences

UTR. The 5' UTR graphs were aligned taking as reference the nucleotide preceding the translation starting codon. The gradients were, in general, very mild with the exception of the last windows which showed the most significant variations (see Figure 4). Thus that there are striking differences in base composition between the 5'UTR and the coding sequences.

Ensemble 3'UTR trends calculated for 500 bp sequences after the translation stop codon were barely detectable. The most relevant variations were observed for the thymine content in *Arabidopsis* which decreased proceeding toward the 3'end (Figure 4).

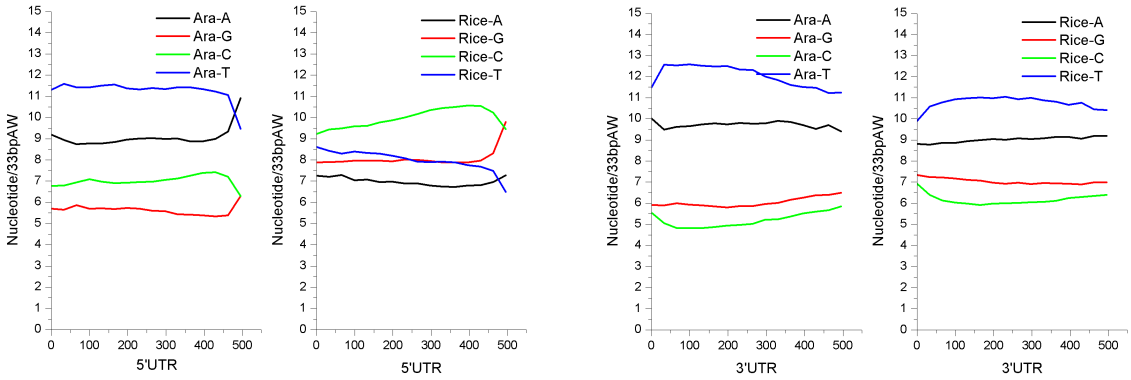


Figure 4. Ensemble graphs of 5' and 3' UTR untranslated sequences. The adiacent window was of 33bp.

Introns. To compare the compositional gradients of introns to those of CDS, ensemble averages were calculated on genic sequence with exon masked. The graph for both *Arabidopsis* and rice showed trends resembling those observed for coding sequences. However, the absolute values of the slopes were lower than those calculated for coding sequences (Figure 5). A similar finding was reported by Wong et al.(*16*) for the G+C content of introns of several plant species. It is worth remembering that this experimental setup evaluates intron's composition as function of genomic position whereas cds ensemble are referred to cDNA coordinates The reported difference between graphs, thus, could be due to the contextual differences between the two analyses. To verify such hypothesis we calculated ensemble graphs of exons referred to genomic coordinates. In practice, we analyzed genic sequence after masking introns. The ensemble graphs referred to genomic coordinates are reported in figure 5a-b. Beside the expected differences in the intercepts, due to compositional diversity between introns and exons, the two trends were remarkably similar.

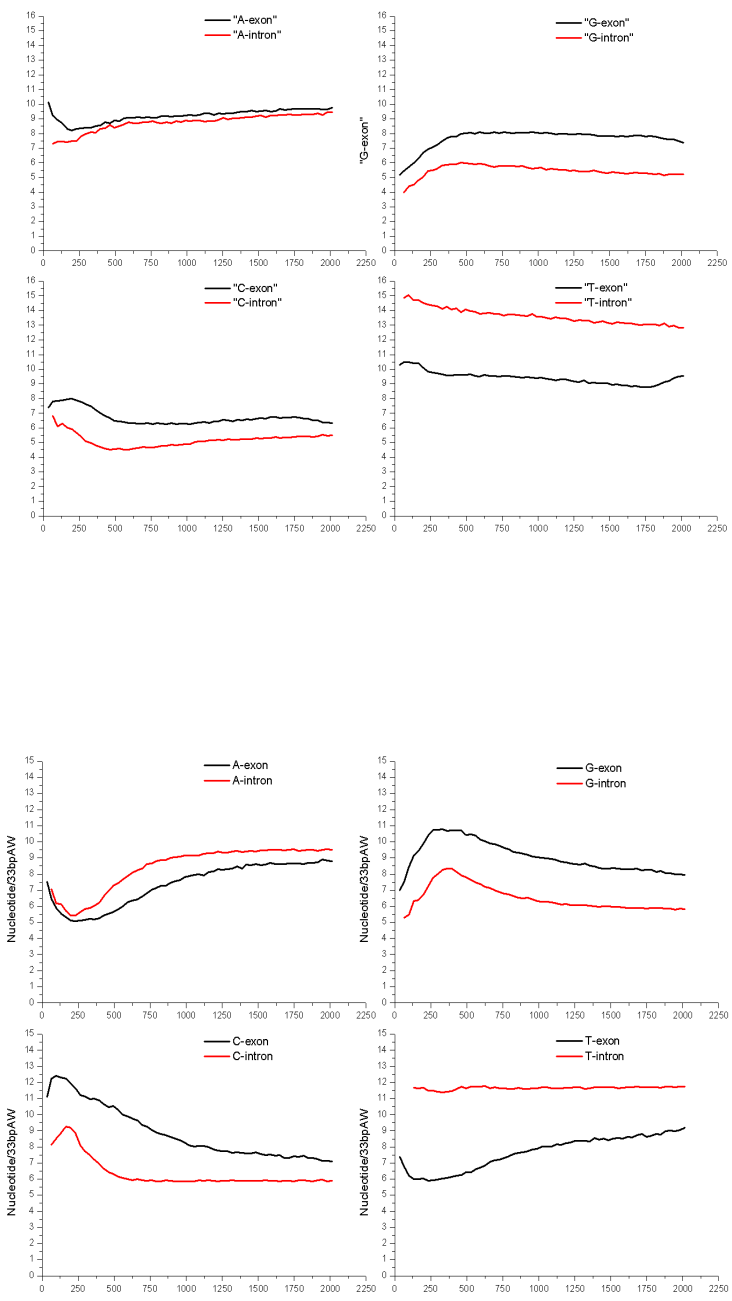


Figure 5. Mononucleotide contents of Arabidopsis (a) and rice (b) genic sequence with either intron (red) or exon (black) masked. The adiacent window was of 33bp.

Dinucleotide trends

As next step in our analyses we studied dinucleotide bias as function of position in genic sequences. The average dinucleotide bias at each position was analyzed using the odds ratio ρ(*18*).

This index, measures the abundance of dinucleotides relative to what would be expected from the random union of mononucleotide(*18*) and the value 1 is expected when no bias is observed.

Both theoretical and empirical studies have indicated that if a given dinucleotide has a ρ index value ≤ 0.78 then this dinucleotide is significantly underrepresented (suppressed) whereas values ≥ 1.23 indicates over-representation(*18*). For coding sequences the average dinucleotide bias of coding sequence was calculated separately for the three reading frame. Interestingly, in both *Arabidopsis* and rice there was a pervasive suppression of CG and TA dinucleotide in all three reading frames (table 2).

	Arabidopsis			***Oryza***		
Dinucleotide	**1_2**	**2_3**	**3_1**	**1_2**	**2_3**	**3_1**
CG	0.71	0.56	0.64	0.75	0.64	0.71
TA	0.42	0.67	0.69	0.47	0.67	0.68
CA	0.97	1.23	1.30	0.96	1.43	1.27
TG	0.82	1.18	1.21	0.89	1.16	1.33
AA	1.16	1.08	1.03	1.65	1.03	1.10
AC	0.75	0.96	0.90	0.74	0.89	0.86
GA	1.24	1.09	1.22	1.19	0.90	1.04
GT	0.73	0.94	0.92	0.72	0.89	0.89
TC	1.47	1.19	1.12	1.42	1.08	0.92
TT	1.39	1.03	1.00	1.29	1.03	0.96
CC	1.09	0.83	0.95	1.00	0.88	1.03
AG	1.16	1.15	1.08	1.12	1.12	0.96
AT	0.93	0.87	0.93	0.98	1.00	1.13
GG	0.88	1.10	0.88	1.13	1.01	0.92
GC	0.88	0.99	0.98	0.99	1.21	1.16
CT	1.14	1.20	1.22	1.23	1.08	1.08

Table 2: Dinucleotide biases observed in the coding sequences of *Arabisopsis* and *Oryza*.

These tendencies were confirmed for introns of both species with the only difference that the average CG under-representation in introns was stronger than in exons while the opposite was found for TA (table 3) Other interesting differences were revealed by the analysis of dinucleotide trends: while TA suppression was almost constant, CG under-representation increased proceeding along the direction of transcription (for a complete picture of the observed trends see figures S3-abc and S4-abc).

	Arabidopsis	Oryza

Dinucleotide	Intron	Exon	Intron	Exon
AA	**1.15**	**1.12**	**1.14**	**1.10**
AG	**1.03**	**1.11**	0.98	1.03
AC	0.93	0.88	0.89	0.85
AT	0.91	0.88	0.98	**1.04**
GA	**1.14**	**1.24***	**1.01**	**1.11**
GG	0.95	0.96	**1.04**	**0.96**
GC	0.94	0.91	**1.13**	**1.11**
GT	0.95	0.86	0.91	0.84
CA	**1.13**	**1.10**	**1.18**	**1.12**
CG	$0.56^{\dagger}$	$0.70^{\dagger}$	$0.59^{\dagger}$	0.86
CC	0.97	0.96	**1.09**	0.96
CT	**1.11**	**1.16**	**1.03**	**1.06**
TA	0.79	$0.63^{\dagger}$	0.80	$0.67^{\dagger}$
TG	**1.18**	**1.13**	**1.20**	**1.11**
TC	**1.09**	**1.21**	0.96	**1.08**
TT	**1.04**	**1.11**	**1.06**	**1.10**

Table 3: Average dinucleotide biases in introns and exons of *Arabidopsis* and *Oryza* (bold = overrepresented, plain = underrepresented, *=significantely over-represented, †. = significantely under-represented)

TG and AC over-representation in introns can be related to CG suppression under the the methylation -deamination- mutation scenario (i.e. CpG islands tend to be easily methylated at the Cytosine residue with a consequent increase in the rate of C -> T mutation). However TG and AC biases in CDS did not mirror accurately the pattern of CG suppression: TG over-representation at 2_3 was lower than at 3_1 in spite of the higher suppression of CG at frames 2_3 than 3_1. A group of five dinucleotides, AA, TC, GT, TT, TC were biased in frame 1_2 only. Such a pattern was probably related to the requirements of specific aminoacids, a hypothesis that was supported also by the virtual absence of bias in intron sequences.

Another group of five dinucleotide (CC, AG, AT, GG, GC,CT) showed no significant bias in all the three reading frames of cds and in introns.

The last two dinucleotides GA and CT showed a different bias in the two species. GA was biased in frames 1_2 and 2_3 of *Arabidopsis* cds but completely unbiased in rice. CT was overrepresented in frame 2_3 and 3_1 in *Arabidopsis* and to lower extent in frame 1_2 while the opposite trend was showed in rice, biased in frame 1_2 but not in frames 2_3 and 3_1. These features are likely to reflect specific difference in the global dinucleotide signature of the two species or differences in codon usage. As a matter of fact, both dinucleotide are slightly overrepresented in *Arabidopsis* introns but are unbiased in those of rice.

Tri-nucleotide trends.

The complete lists of average values of the index of tri nucleotide bias for both *Arabidopsis* and rice coding sequences and introns are reported in table S1a-b. On average, the bias of trinucleotide was low if compared to that of either mono and dinucleotides This was particulary evident for trinucleotide in introns of both species. The most over-represented trinucleotide showed a ρ index of 1.14 in rice and 1.12 in *Arabidopsis* while the most suppressed showed values of 0,878 (CAG) in *Arabidopsis* and 0,930 (AAG) in rice (table 4). In coding sequences the distribution of the bias was highly dependent on the class of the trinucleotide relatively to the reading frame. In fact, the values of the correlations between γ_{XYZ} of the same frame in the two species were always higher than those between γ_{XYZ} with different frames within either species.

	Arabidopsis	Rice	
trinucleotidi			
CCG	**1.14**	TAG	**1.12**
CCA	**1.11**	GAT	**1.12**
GAA	**1.11**	GAG	**1.09**
TAG	**1.1**	GCC	**1.08**
GAG	**1.08**	CCG	**1.06**
CGC	**0.81**	AAG	**0.92**
GCG	**0.87**	GGT	**0.94**
CAG	**0.87**	TTG	**0.94**
CCT	**0.88**	CAG	**0.94**
AGG	**0.92**	GTC	**0.94**

Table 4: Trinucleotides biases in introns of *Arabidopsis* and *Oryza sativa*.

DISCUSSION

Compositional heterogeneity may be described at many genomic scales. In the present paper we used an approach based on ensemble averages to compile descriptions of compositional trends in transcribed genic sequences of several eudicots and monocots species. Two provocative considerations were evoked from the first glance of the graphs. First, the shapes of the trends were

different between monocots and eudicots but surprisingly uniform within the two evolutionary clades. Second, the variations observed in the first part of the sequences petered out toward the 3'end of the genes.

Both deductions are in agreement with previous findings reported by Wong and co-worker in a study on the G+C content of genic regions of several monocot and eudicot species(*16*). However, that study did not detail on differences between single bases. For example, the guanine content decreases firmly in monocots and increases, although weakly in eudicots. Thymine increases in the first part of the cds in both eudicots and monocots and then decreases in the eudicots but not in monocots. Before starting to speculate on the emergence of these differences we cannot elude a fundamental question: what proportion of all single genes trends fit to the ensemble models?

In answering this question it should be kept in mind that most of the factors influencing single sequence's compositions are highly variable and consequently the assignment of single gene trends to stringently defined categories can be a hardly tractable goal. On the light of these considerations, we concluded that the definition of the exact proportion of genes fitting to the ensemble models should first leave place to more qualitative assessment. The most informative part of ensemble graphs were approximated to linear models and then single gene trends were studied for their consistency with the models. With this experimental setup we could demonstrate that the i) shape of compositional trends is independent from sequence length, genomic orientation or overall GC content and ii) that most of the single gene parameters distributions are in agreement with the expectations based on the models described by the ensemble approach. Inherently we should emphasize that the breakpoint distributions covered a wide range of values and that the classes of highest frequency did not obviously correspond to what foreseen by the ensemble model. Whether this is due to the high variability of the evolutionary forces shaping the gradients or is the results of factors of other nature such as for example the genomic position of the genes is still to be determined. Based on these results we conclude that the ensemble graphs can be considered as *master* models for plant genes. With this, we do not rule out that other models may coexist or that selected groups of genes may be better represented by ensemble graphs slightly different from those described in the present study.

With these propositions in mind we can move to the next questions: What are the evolutionary forces responsible for these compositional arrangements?

A caveat of all further considerations should be the polarity of the trends which may help in limiting the range of choices. Indeed mutational or selective forces acting at DNA level are expected to exert an effect of opposite direction in genes with either forward or reverse orientation.

The trends calculated on genic sequences grouped based on their orientation in the genome did not reveal these types of evidences. Because the polarity of the trends was in the same direction of transcription/translation we will restrict our discussion to forces related to these two basic cellular processes.

Observations conducted in other systems may help in constructing insightful analogies. Eyre and Walker has shown that the first portion of *E.coli* genes is compositionally different from the remaining part(*5*). The explanations proposed envisaged selection for i) the mRNA secondary structure near the ribosome binding site and ii) the use of suboptimal codons to regulate gene expression. Evident biases near the translation initiation codon of coding sequences of seven eukaryote genomes, including *Arabidopsis* and *Oryza*, has been reported by Niimura et al (2003) and were explained in terms of selection for efficiency of translation initiation(*18*). Our data featured a strong variations in base composition at the beginning of CDS and therefore do not rule out the involvement of this type of selection in the generation of the gradients. But if so, to account for the difference in compositional features between these two species, we should hypothesize that the target of this type of selection is rather different in monocots and eudicots. Indeed, Gu et al have recently represented a universal trend of reduced mRNA stability near the translation sites in 340 species including eukaryotes and prokaryotes(*19*). Very interesting rice and *Arabidopsis* genes marked a contrasting behavior in this respect. The sequence near the translation starting codon showed a thermodynamic stability slightly higher than that expected by chance in rice genes and moderately reduced in *Arabidopsis*. Furthermore, while in most species, the genes with higher codon bias had lower mRNA stability at their 5' end, highly biased rice genes showed very stable mRNA at the 5' end and the opposite was observed for the low biased genes. These finding supports the hypothesis that the selection near the translation starting codon have singular feature in rice and perhaps may be used as argument to explain some of the difference found in the first portion of *Arabidopsis* and rice graphs. However we question whether a different deal of selection for translation initiation efficiency may generates difference distributed along quite long portion of genes. Moreover we cannot figure out in such context the significance of the gradients observed in introns.

De-Rose-Wilson et al.(*19*) have recently proposed that transcription related mutation (TCR) contributes significantly to rate difference between intergenic and transcribed sequence in *Arabidopsis* genome. Similar conclusion may be reached considering the pattern of strand asymmetry in intergenic and genic sequence in rice (our unpublished results). These findings coupled to recent insights on the transcription coupled repair process may provide ground for explanations of some of the gradients. Experiments conducted in animal systems have indicated that

the speed of the transcription coupled repair may be dependent on the position of the lesion within the transcribed sequence a feature that may be even related to differences in the subunits involved. It is therefore possible that gradient in speed or even fidelity of TCR may contribute significantly to the establishment of compositional gradients within genic regions.

The pattern of dinucleotide bias in coding sequence and in introns discovered other important compositional features of plant genes. Studies carried out on coding or non coding sequence as well as at whole genome level(*20*) have documented, for example, a pervasive under-representation of the CG and TA dinucleotides in plants. This study confirmed these tendencies and underlined new feature of this phenomenon. For example CG suppression was more severe in introns then in exon probably reflecting a higher mutation rate in non coding genic sequences. Moreover the suppression increased proceeding along the direction of transcription in both monocots and eudicots. This effect was only partially mirrored by an over-representation of TG suggesting that other causes than the classical methylation-deamination-mutation scenario could explain the dependence of CG underepresentaion with position in gene sequences. It is known that CG dinucleotides posses the highest thermodynamic stacking energy; the variation of CG suppression along the direction of transcription may therefore aid DNA untwisting during transcription. Other dinucleotides showed different biases in the 1_2 compared to 2_3 and 3_2 reading frames of cds likely reflecting different amino-acid or codon usage. The overall trinucleotide bias was on average quite low in introns confirming the hypothesis that the forces maintaining structural features of DNA act prevalently at level of mono or dinucleotides. Interestingly also the bias at the reading frame 1_2_3 of most trinucleotides was rather low suggesting that a non trivial part of the codon usage bias of plant genes can be explained by compositional features related to DNA structural properties.

In conclusion, we presented a detailed investigation of the compositional features of genic sequence which identified master model of compositional trends. These models can be considered as a “genomic reference” to describe the compositional feature of groups of genes related for some structural or functional features. We are confident that some of these contrasts will find robust associations with some of the compositional features highlighted in this study and therefore may be instrumental in identifying casual links.

Reference List

1. S. Karlin, I. Ladunga, *Proc Natl Acad Sci U S A* **91**, 12832 (1994).
2. S. Karlin, I. Ladunga, B. E. Blaisdell, *Proc Natl Acad Sci U S A* **91**, 12837 (1994).
3. S. Karlin, A. M. Campbell, J. Mrazek, *Annu Rev Genet* **32**, 185 (1998).
4. G. Bernardi *et al.*, *Science* **228**, 953 (1985).
5. A. Eyre-Walker, L. D. Hurst, *Nat Rev Genet* **2**, 549 (2001).
6. O. Clay, S. Caccio, S. Zoubak, D. Mouchiroud, G. Bernardi, *Mol Phylogenet Evol* **5**, 2 (1996).
7. J. Filipski, *FEBS Lett* **217**, 184 (1987).
8. S. Saccone *et al.*, *Proc Natl Acad Sci U S A* **90**, 11929 (1993).
9. A. Nekrutenko, W. H. Li, *Genome Res* **10**, 1986 (2000).
10. M. Huvet *et al.*, *Genome Res* **17**, 1278 (2007).
11. L. Zhang, S. Kasif, C. R. Cantor, N. E. Broude, *Proc Natl Acad Sci U S A* **101**, 16855 (2004).
12. S. Fujimori, T. Washio, M. Tomita, *BMC Genomics* **6**, 26 (2005).
13. M. Bulmer, *Genetics* **129**, 897 (1991).
14. H. Qin, W. B. Wu, J. M. Comeron, M. Kreitman, W. H. Li, *Genetics* **168**, 2245 (2004).
15. Y. Niimura, M. Terabe, T. Gojobori, K. Miura, *Nucleic Acids Res* **31**, 5195 (2003).
16. G. K. Wong *et al.*, *Genome Res* **12**, 851 (2002).
17. S. E. Ryan, L. S. Porth, "A Tutorial on the Piecewise Regression Approach Applied to Bedload Transport Data" (General Technical Report RMRS-GTR-189, 2007).
18. C. Burge, A. M. Campbell, S. Karlin, *Proc Natl Acad Sci U S A* **89**, 1358 (1992).
19. L. J. Rose-Wilson, B. S. Gaut, *BMC Evol Biol* **7**, 66 (2007).
20. A. F. De, S. Marchetti, *Nucleic Acids Res* **28**, 3339 (2000).

Supplemental Material

Ensemble graphs are representative of single gene trends.
The original Arabidopsis and rice coding sequences datasets were partitioned considering gene orientation in the genomes (forward or reverse) and GC content. The shape of the trends were similar to those observed for the original datasets.
Dataset partition considering gene orientation.

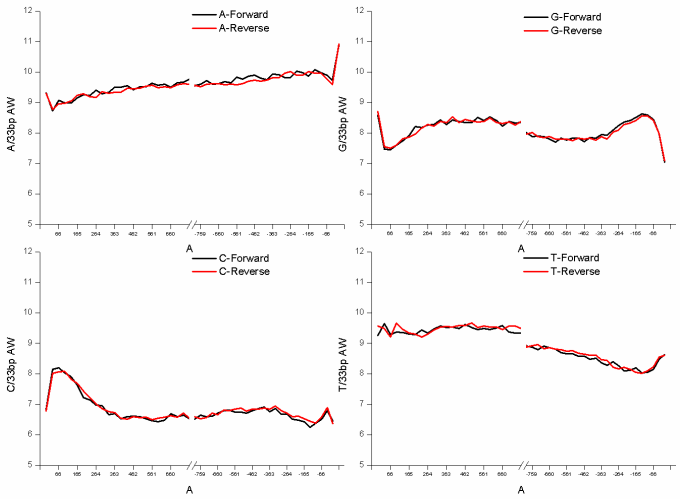


Figure S1a. Overall mononucleotide compositional trends of Arabidopsis coding sequences as a function of position and averaged over all sequences of each dataset with an adjacent window of 33 bp. The trends are calculated from the dataset including cds of genes with forward (-) or reverse (-) orientation.

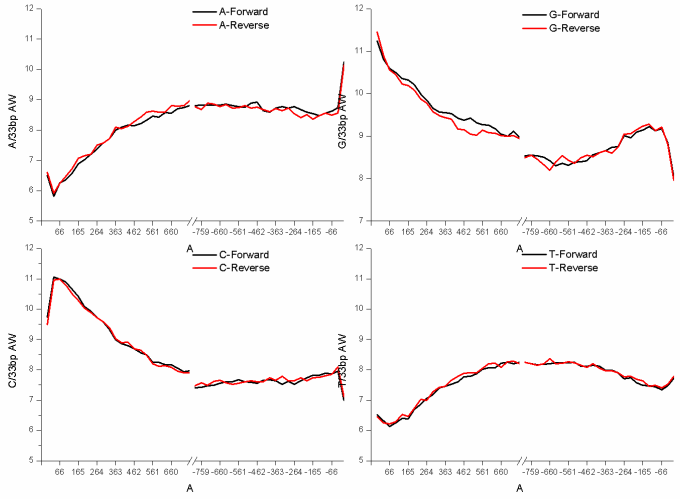


Figure S1b. Overall mononucleotide compositional trends of Oryza coding sequences as a function of position and averaged over all sequences of each dataset with an adjacent window of 33 bp. The trends are calculated from the dataset including cds of genes with forward (-) or reverse (-) orientation.

Dataset partition considering the G+C content of genes

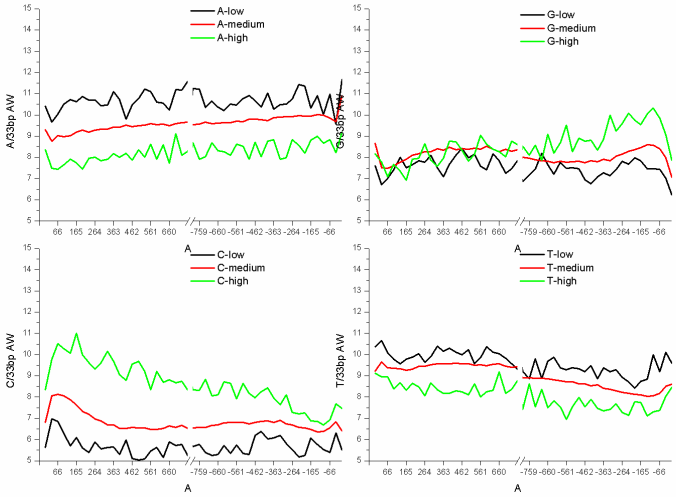


Figure S2a. Overall mononucleotide compositional trends of Arabidopsis coding sequences as a function of position and averaged over all sequences of each dataset with an adjacent window of 33 bp. The dataset "low" included sequence with a G+C content <=0.40. The dataset medium included sequences with a G+C content > 0.40 and <= 0.50. Sequence with a G+C > 0.50 were included in the dataset "high".

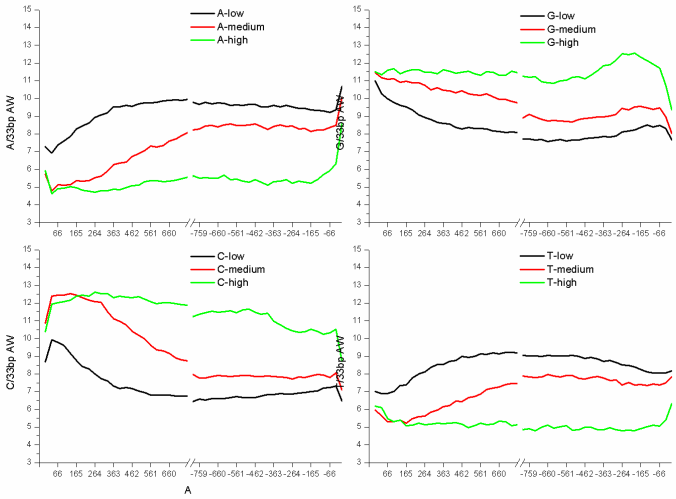
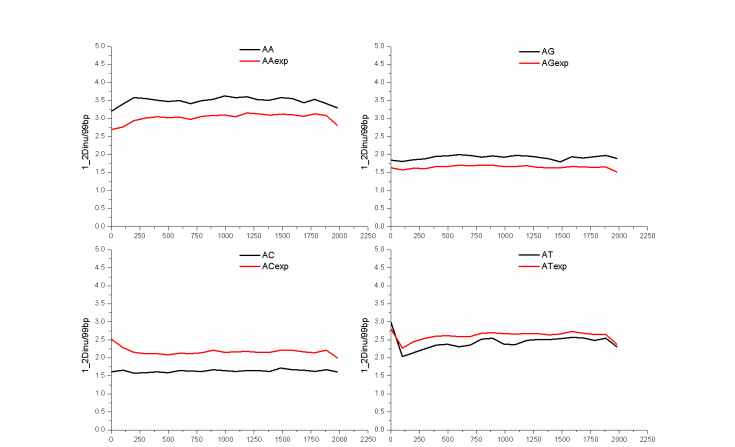


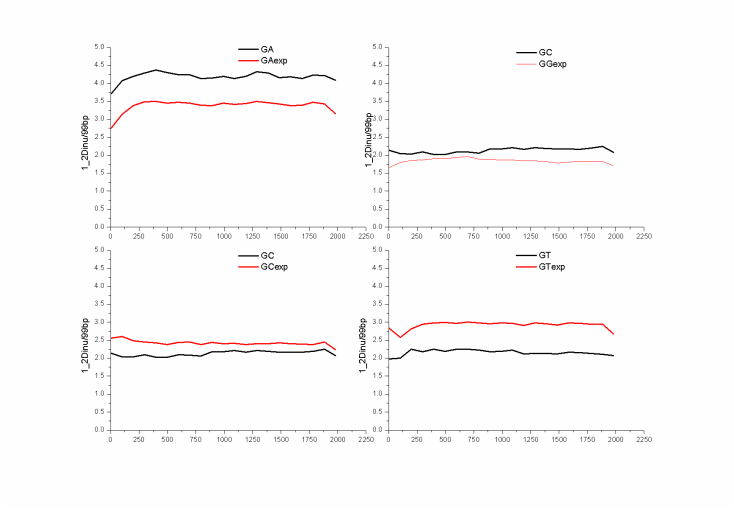
Figure S2b. Overall mononucleotide compositional trends of Oryza coding sequences as a function of position and averaged over all sequences of each dataset with an adjacent window of 33 bp. The dataset "low" included sequence with a G+C content <=0.40. The dataset medium included sequences with a G+C content >0.40 and <= 0.50. Sequence with a G+C > 0.5 were included in the dataset "high".

Dinucleotide trends in coding sequences.

Dinucleotide trends in coding sequence were calculated as a function of position and averaged over all sequence in each dataset with an adjacent window of 99 bp. The expected content for each window position was calculated as pXY/fXY according to Karlin et al ().

Arabidopsis dinucleotides 1_2





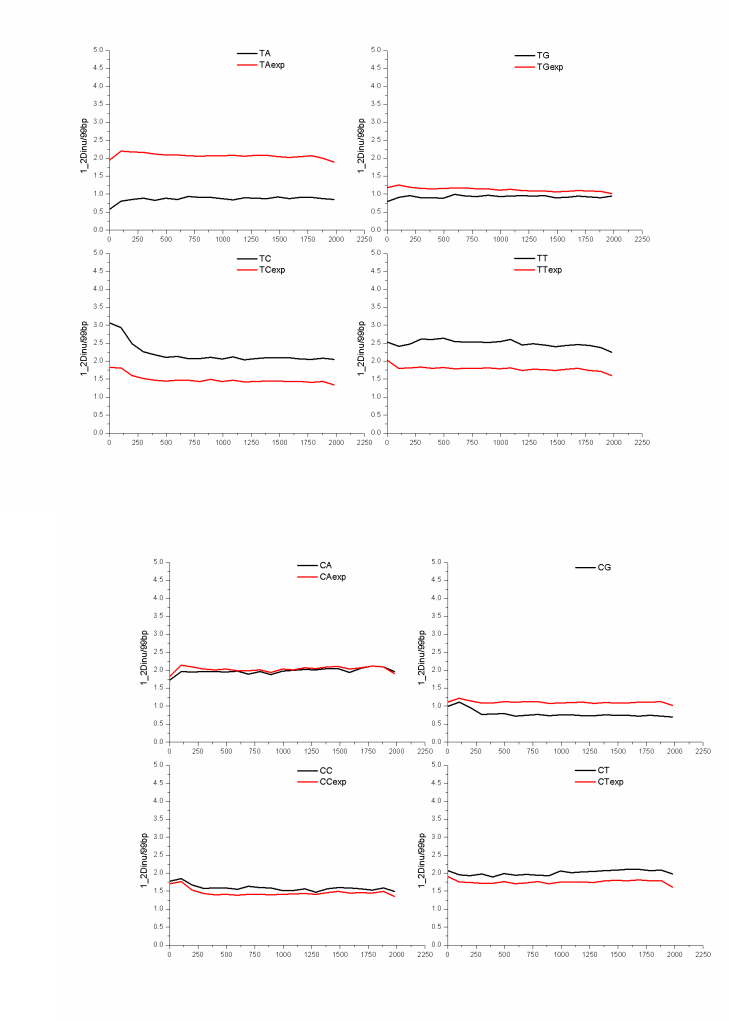
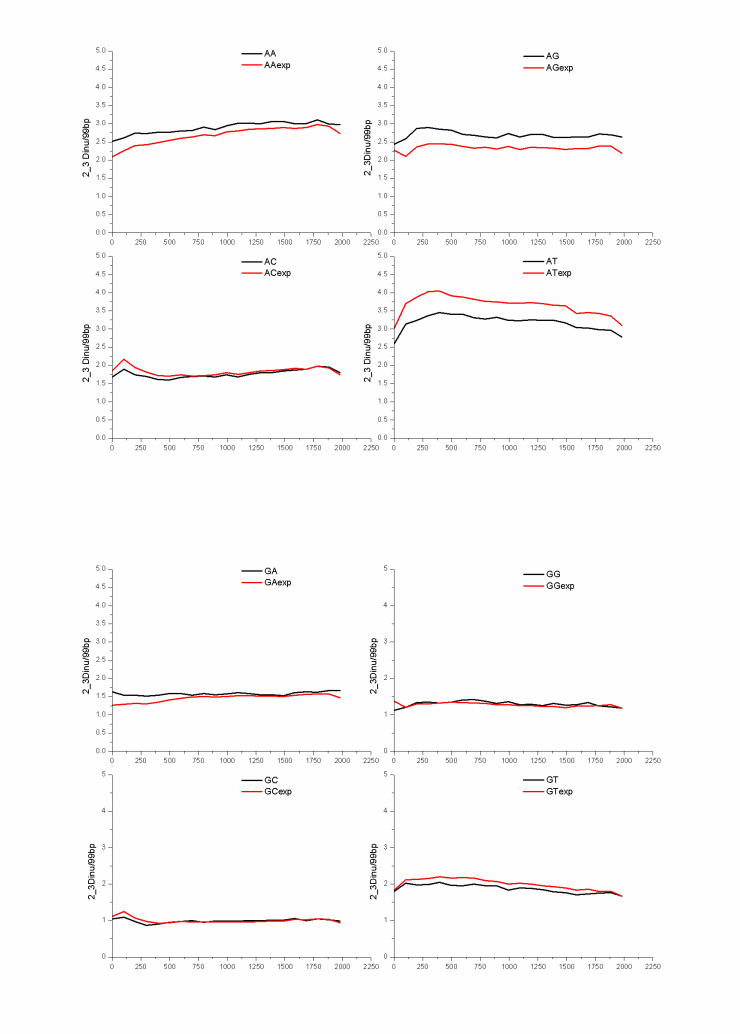


Figure S3a. Dinucleotide content of the first 2 kb of Arabidopsis cds. The dinucleotide were calculated taking into account the first and second position of each codon

Arabidopsis dinucleotides 2_3


AA
AAexp
AG
AGexp
AC
ACexp
AT
ATexp
GA
GAexp
GG
GGexp
GC
GCexp
GT
GTexp
2_3 Dinu/98bp
2_3Dinu/98bp

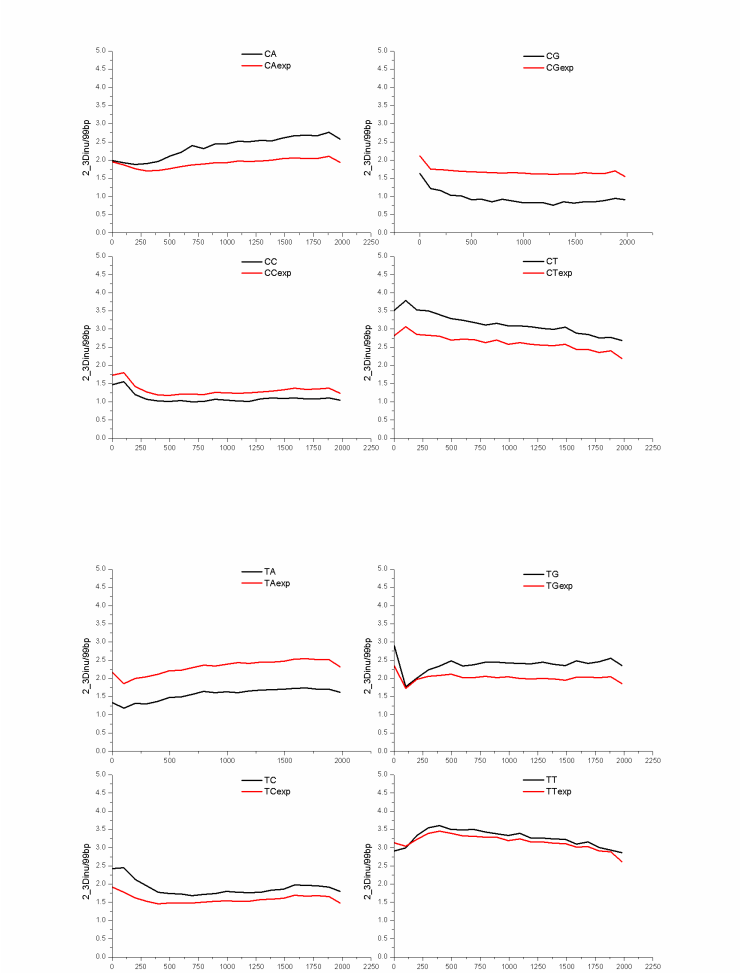
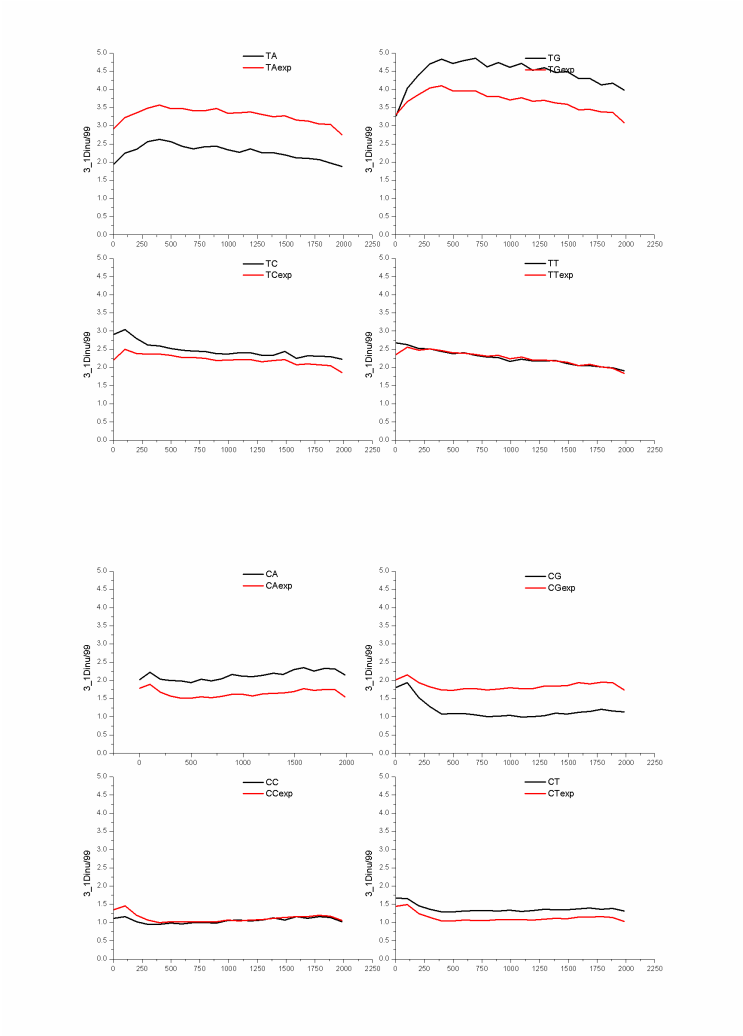


Figure S3b. Dinucleotide (2_3) content of the first 2 kb of Arabidopsis cds. The dinucleotide were calculated taking into account the second and third position of each codon.

Arabidopsis dinucleotides 3_1


TA
TAexp
TG
TGexp
TC
TCexp
TT
TTexp
CA
CAexp
CG
CGexp
CC
CCexp
CT
CTexp
3_1Dinu99

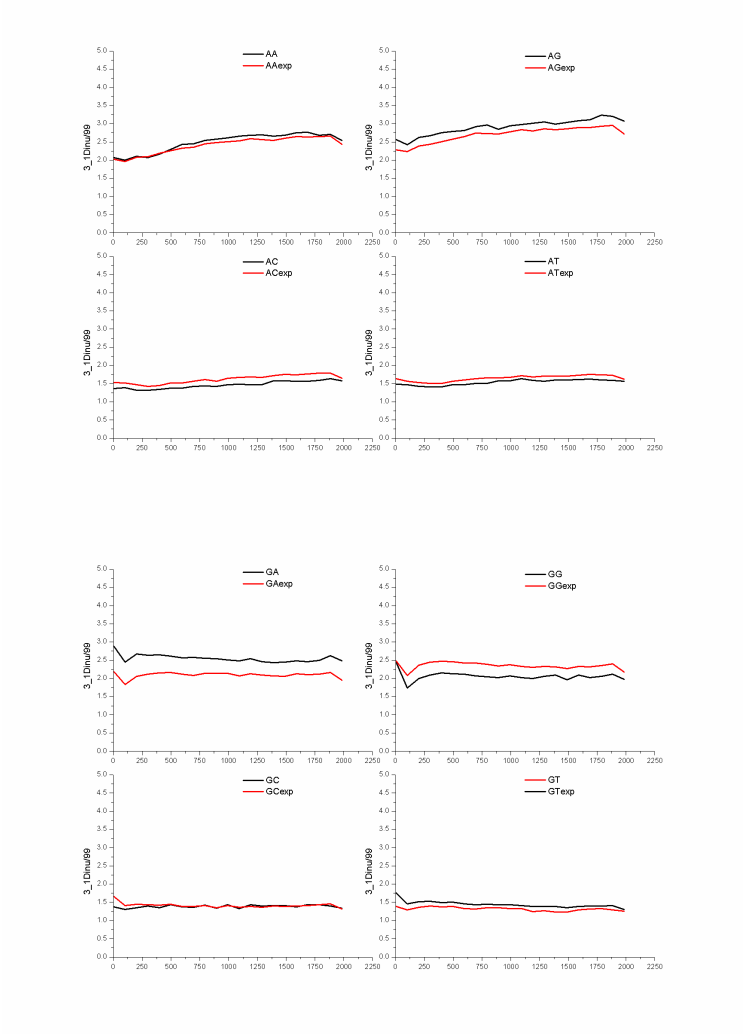
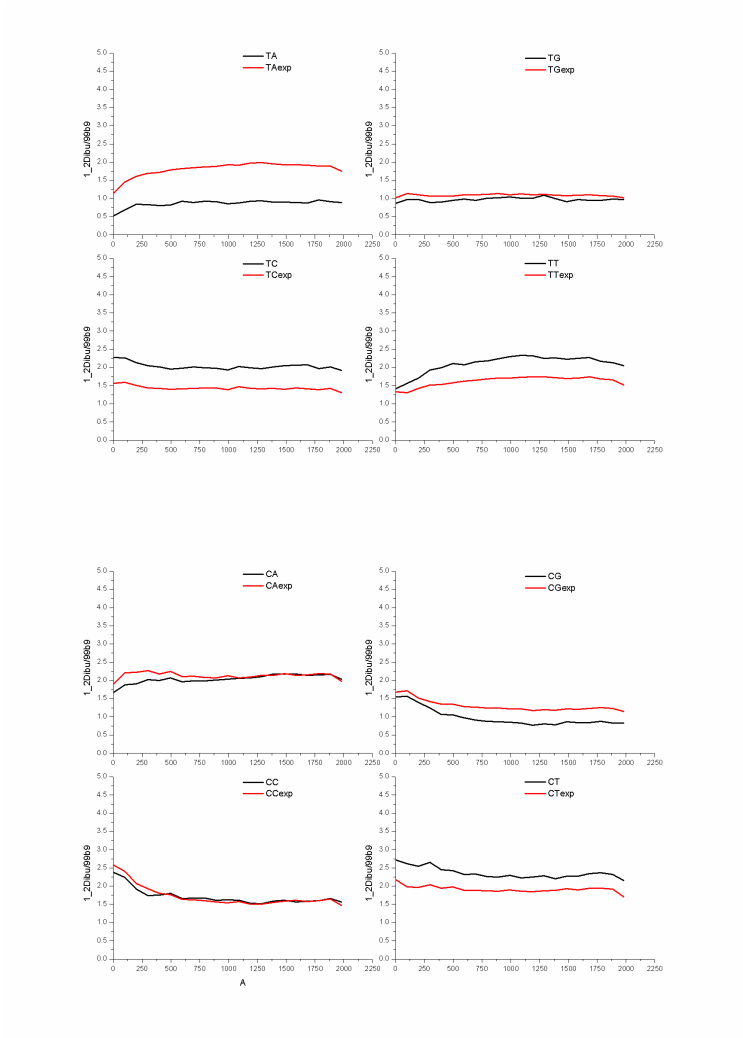


Figure S3c. Dinucleotide content of the first 2 kb of Arabidopsis cds. The dinucleotide were calculated taking into account the third position of a codon and the first of the subsequent codon.

Oryza dinucleotides 1_2


TA
TAexp
TG
TGexp
TC
TCexp
TT
TTexp
CA
CAexp
CG
CGexp
CC
CCexp
CT
CTexp
1_2Dbu/98e9
1_2Dbu/58e9
A

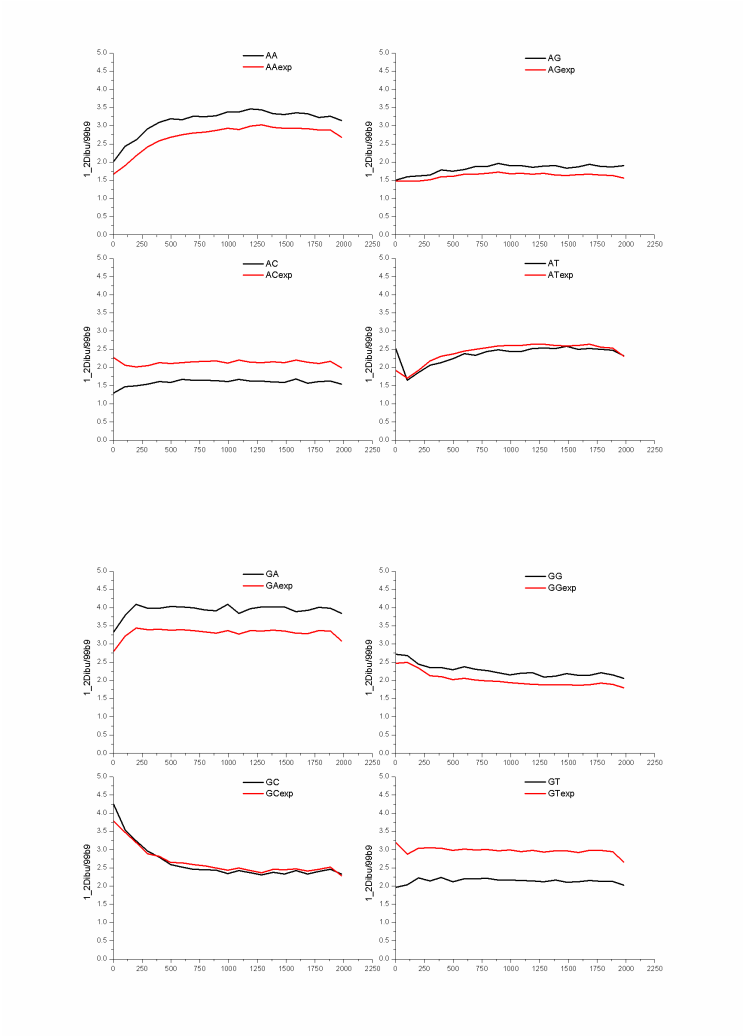
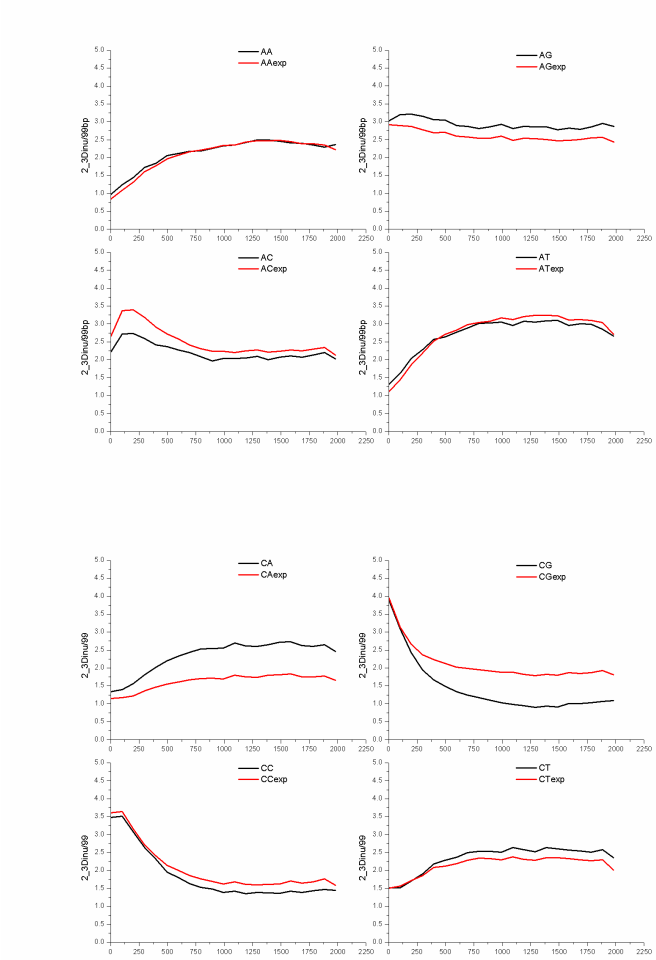


Figure S4a. Dinucleotide content of the first 2 kb of Oryza cds. The dinucleotide were calculated taking into account the first and second position of each codon.

Oryza dinucleotides 2_3


AA
AAexp
AG
AGexp
AC
ACexp
AT
ATexp
2_3Dinu/99bp
CA
CAexp
CG
CGexp
CC
CCexp
CT
CTexp
2_3Dinu/99

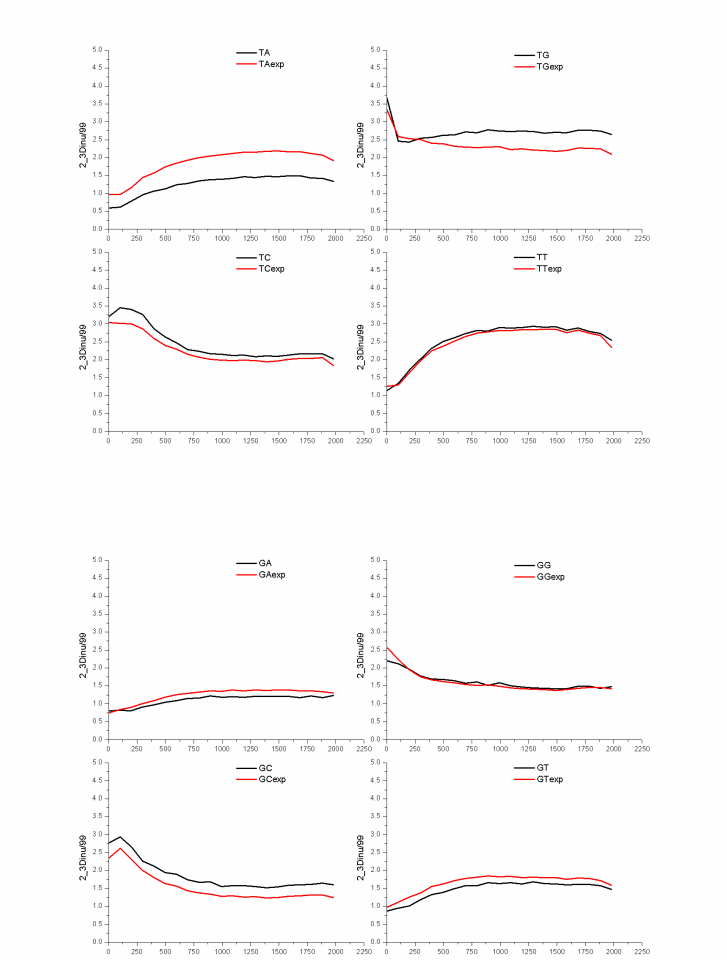
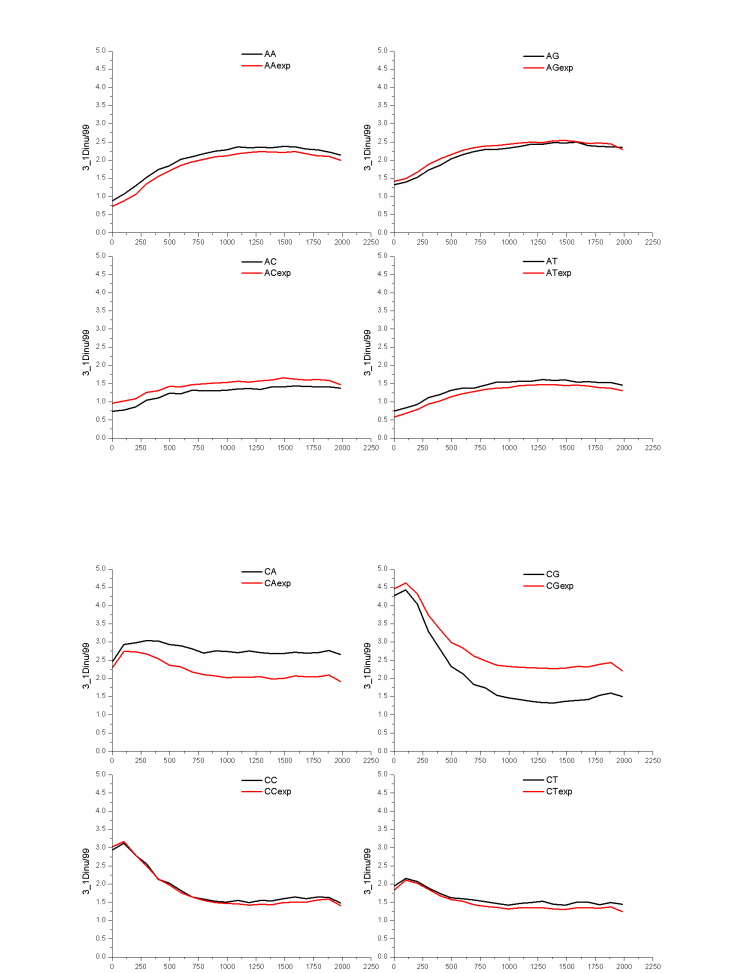


Figure S4b. Dinucleotide content of the first 2 kb of Oryza cds. The dinucleotide were calculated taking into account the second and third position of each codon.

Oryza dinucleotides 3_1


AA
AAexp
AG
AGexp
AC
ACexp
AT
ATexp
CA
CAexp
CG
CGexp
CC
CCexp
CT
CTexp
3_1Dinu99

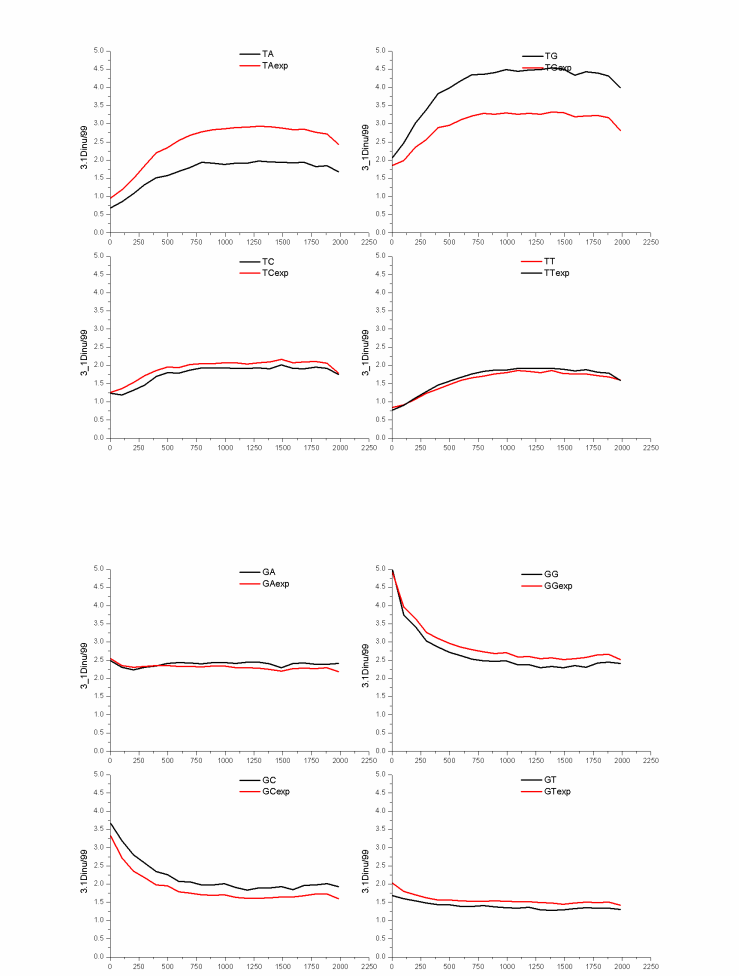
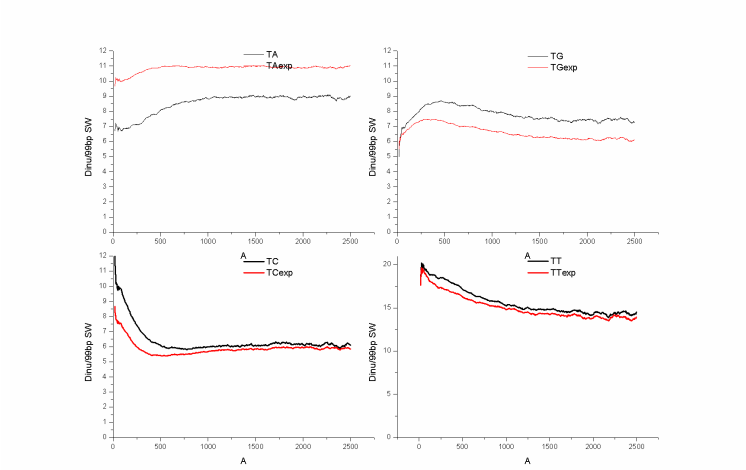


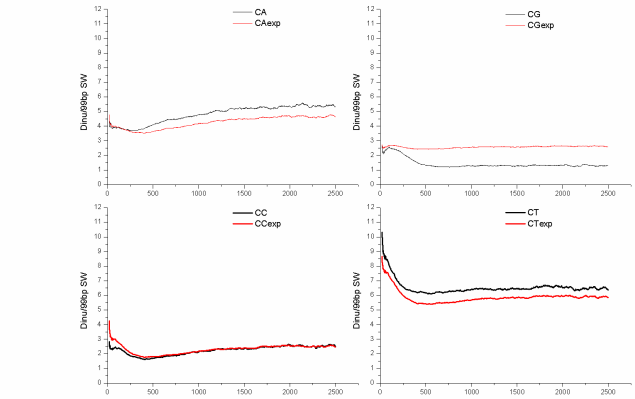
Figure S4c. Dinucleotide content of the first 2 kb of Oryza cds. The dinucleotide were calculated taking into account the third position of a codon and the first of the subsequent codon.

Dinucleotide content in introns

Dinucleotide trends in introns were calculated as a function of position and averaged over all sequence in each dataset with a sliding window of 99 bp and a step of two.

Dinucleotide in Arabidopsis' introns.


TA
TAexp
TG
TGexp
TC
TCexp
TT
TTexp
Dinu/98bp SW
A


CA
CAexp
CG
CGexp
CC
CCexp
CT
CTexp
Dinu/98bp SW

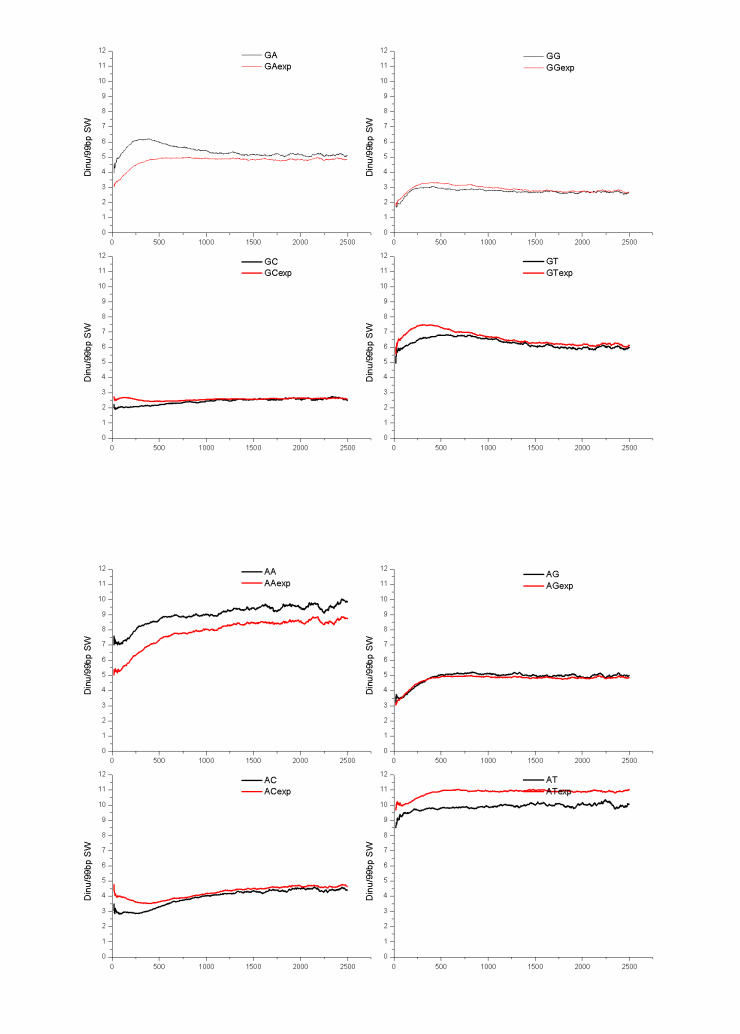
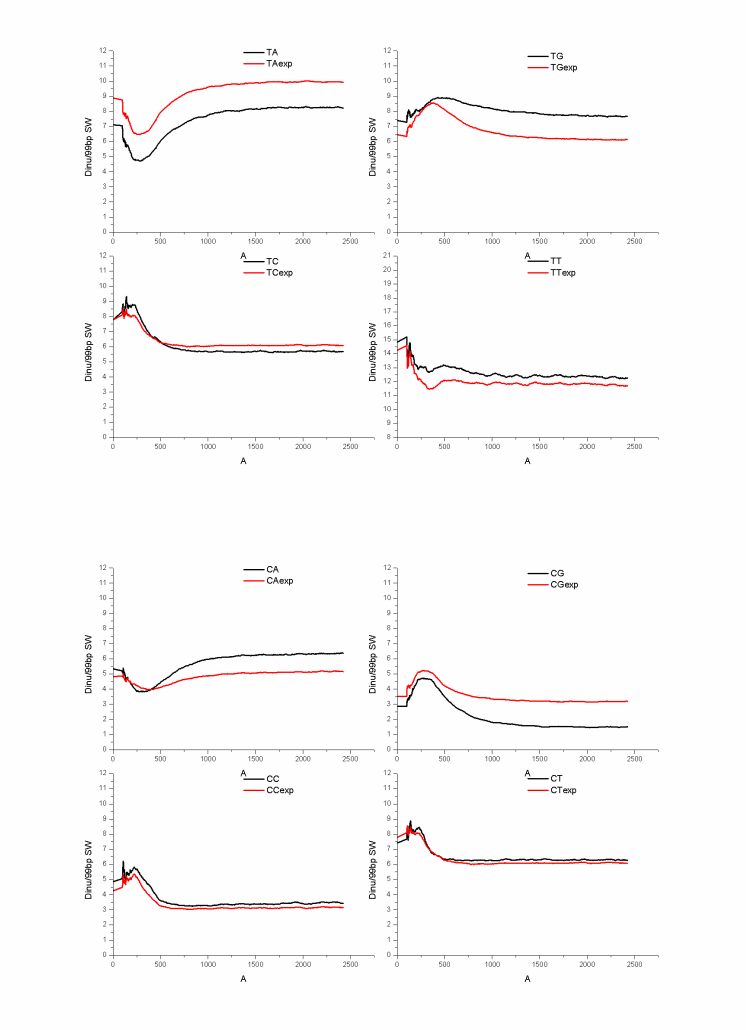


Figure S5a: Dinucleotide content of the first 2.5 kb of *Arabidopsis* introns

Dinucleotide in Arabidopsis' introns.


TA
TAexp
TG
TGexp
TC
TCexp
TT
TTexp
CA
CAexp
CG
CGexp
CC
CCexp
CT
CTexp
Dinu/98bp SW
A

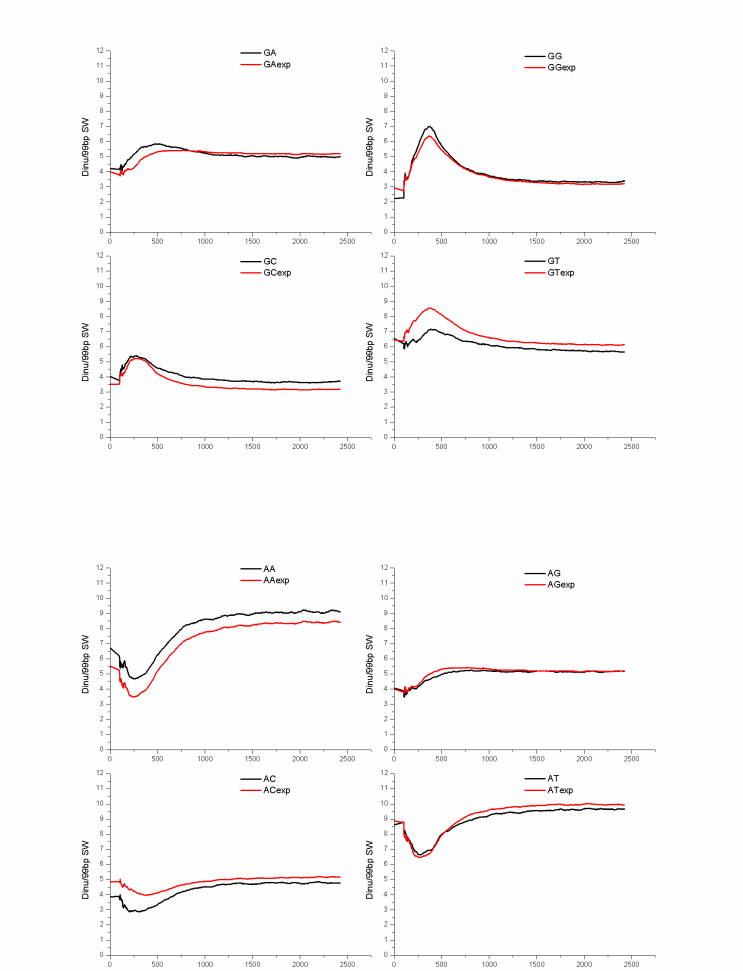


Figure S5b: Dinucleotide content of the first 2.5 kb of *Oryza* introns

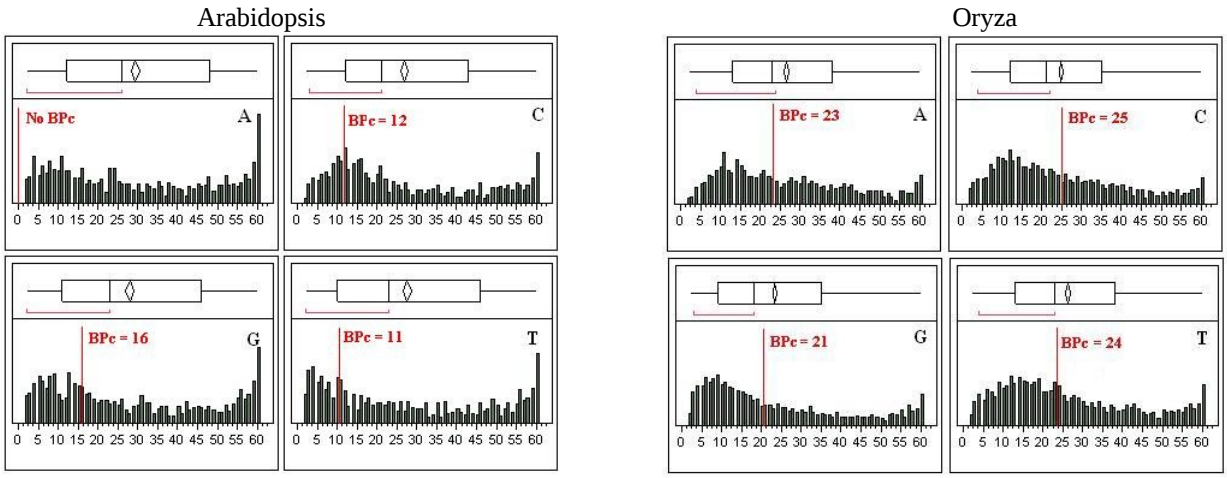


Figure S6: Break point distribution for the 4 bases of Arabidopsis and Oryza.

Tab S1a. Average trinucleotide bias of Arabidopsis coding sequences. The trinucleotide contents were calculated for each windoe position and averaged over all sequences longer than 2kb using a window of 99bp. The γ index were calculated according to Karlin ()

	1_2_3		2_3_1		3_1_2			1_2_3		2_3_1		3_1_2	
	average	σ	average	σ	average	σ		average	σ	average	σ	average	σ
AAA	0.977 ±	0.01 7	0.876 ±	0.02 2	0.928 ±	0.03 3	**CAA**	1.130 ±	0.02 4	1.118 ±	0.02 8	1.054 ±	0.02 3
AAG	1.027 ±	0.04 1	1.129 ±	0.01 0	1.075 ±	0.01 9	**CAG**	1.197 ±	0.02 9	0.880 ±	0.02 8	0.849 ±	0.02 9
AAC	0.973 ±	0.02 6	1.001 ±	0.02 8	1.041 ±	0.02 5	**CAC**	0.858 ±	0.03 2	1.025 ±	0.03 9	1.046 ±	0.02 0
AAT	0.968 ±	0.03 4	0.941 ±	0.04 3	1.019 ±	0.04 5	**CAT**	0.796 ±	0.02 0	1.022 ±	0.02 2	1.009 ±	0.02 7
AGA	1.101 ±	0.03 3	1.034 ±	0.02 0	1.027 ±	0.02 0	**CGA**	0.891 ±	0.04 4	0.938 ±	0.02 5	1.010 ±	0.04 4
AGG	0.787 ±	0.04 7	0.874 ±	0.04 3	0.882 ±	0.02 2	**CGG**	1.010 ±	0.08 0	1.094 ±	0.05 7	1.162 ±	0.05 4
AGC	1.042 ±	0.04 6	1.082 ±	0.02 9	1.007 ±	0.02 3	**CGC**	1.013 ±	0.06 6	0.897 ±	0.04 6	0.914 ±	0.03 8
AGT	1.043 ±	0.02 8	1.038 ±	0.02 6	1.019 ±	0.03 1	**CGT**	1.081 ±	0.03 0	1.091 ±	0.04 9	0.912 ±	0.04 8
ACA	0.959 ±	0.02 4	0.999 ±	0.01 5	0.989 ±	0.03 4	**CCA**	1.014 ±	0.02 8	1.104 ±	0.02 5	1.073 ±	0.03 3
ACG	0.886 ±	0.03 7	0.970 ±	0.03 4	0.976 ±	0.05 3	**CCG**	1.200 ±	0.07 7	1.175 ±	0.06 8	1.330 ±	0.06 1
ACC	1.151 ±	0.05 5	1.041 ±	0.04 4	1.015 ±	0.03 4	**CCC**	0.805 ±	0.07 9	0.831 ±	0.08 3	0.927 ±	0.06 1
ACT	1.078 ±	0.02 4	1.013 ±	0.03 9	1.026 ±	0.04 3	**CCT**	0.964 ±	0.01 4	0.829 ±	0.03 8	0.859 ±	0.03 2
ATA	0.968 ±	0.06 6	1.022 ±	0.02 9	0.984 ±	0.06 6	**CTA**	0.906 ±	0.04 0	1.000 ±	0.02 2	1.003 ±	0.02 9
ATG	1.037 ±	0.06 0	1.011 ±	0.02 1	1.075 ±	0.05 3	**CTG**	0.740 ±	0.06 1	0.942 ±	0.01 6	0.960 ±	0.05 4
ATC	0.926 ±	0.04 8	0.942 ±	0.01 9	0.999 ±	0.02 1	**CTC**	1.207 ±	0.04 2	1.056 ±	0.01 9	1.014 ±	0.02 2
ATT	1.028 ±	0.05 4	1.030 ±	0.02 3	1.002 ±	0.03 0	**CTT**	1.142 ±	0.02 6	1.037 ±	0.02 0	1.000 ±	0.01 9
GAA	0.979 ±	0.02 1	1.042 ±	0.02 3	0.996 ±	0.02 0	**TAA**	0.000 ±	0.00 0	0.990 ±	0.00 5	0.983 ±	0.02 4
GAG	1.080 ±	0.01 8	1.035 ±	0.04 2	1.129 ±	0.03 6	**TAG**	0.000 ±	0.08 0	0.906 ±	0.03 1	0.970 ±	0.03 3
GAC	0.945 ±	0.01 7	0.930 ±	0.02 8	0.950 ±	0.02 1	**TAC**	1.914 ±	0.06 5	1.066 ±	0.04 9	1.010 ±	0.04 0
GAT	0.989 ±	7	0.929 ±	7	0.921 ±	4	**TAT**	1.730 ±	5	1.137 ±	8	1.059 ±	3

			0.02			0.02			0.02			0.00			0.02			0.00
GGA	1.184	±	1	1.009	±	9	1.004	±	5	**TGA**	0.000	±	0	0.974	±	9	0.995	±	9
			0.04			0.06			0.05				0.15			0.04			0.03
GGG	0.792	±	3	0.922	±	8	0.817	±	7	**TGG**	1.938	±	1	1.140	±	2	1.095	±	2
			0.07			0.05			0.05				0.06			0.04			0.02
GGC	0.974	±	3	1.049	±	3	1.163	±	0	**TGC**	1.007	±	0	0.911	±	7	0.927	±	9
			0.02			0.04			0.03				0.04			0.02			0.01
GGT	0.978	±	3	1.015	±	2	1.001	±	4	**TGT**	0.971	±	5	0.945	±	5	1.023	±	8
			0.02			0.03			0.04				0.03			0.01			0.02
GCA	0.853	±	8	0.915	±	6	0.955	±	1	**TCA**	1.451	±	4	0.996	±	7	1.014	±	3
			0.05			0.04			0.05				0.04			0.04			0.02
GCG	0.973	±	8	0.964	±	6	0.966	±	3	**TCG**	1.024	±	8	0.945	±	6	0.890	±	5
			0.02			0.07			0.04				0.03			0.02			0.02
GCC	1.196	±	6	1.136	±	8	0.998	±	4	**TCC**	0.810	±	5	0.988	±	9	1.021	±	5
			0.02			0.04			0.02				0.01			0.03			0.01
GCT	1.013	±	3	1.010	±	2	1.040	±	5	**TCT**	0.908	±	5	1.089	±	4	1.014	±	8
			0.02			0.03			0.05				0.09			0.02			0.03
GTA	0.842	±	5	0.988	±	3	0.965	±	3	**TTA**	1.564	±	3	0.997	±	8	1.039	±	5
			0.04			0.01			0.04				0.07			0.01			0.03
GTG	0.992	±	1	1.033	±	8	1.017	±	5	**TTG**	1.159	±	0	1.030	±	3	0.961	±	0
			0.03			0.02			0.02				0.03			0.02			0.01
GTC	1.025	±	0	0.967	±	9	0.971	±	8	**TTC**	0.835	±	3	1.022	±	7	1.006	±	4
			0.03			0.03			0.02				0.02			0.01			0.02
GTT	1.116	±	6	1.046	±	4	1.074	±	5	**TTT**	0.757	±	1	0.887	±	9	0.950	±	2

Tab S1b. Average trinucleotide bias of Oryza coding sequences. The trinucleotide contents were calculated for each windoe position and averaged over all sequences longer than 2kb using a window of 99 bp. The γ index were calculated according to Karlin ()

	1_2_3			**2_3_1**			**3_1_2**				**1_2_3**			**2_3_1**			**3_1_2**		
	average		σ	average		σ	average		σ		average		σ	average		σ	average		σ
			0.03			0.02			0.01				0.06			0.04			0.02
AAA	0.958	±	7	0.917	±	0	0.971	±	6	**CAA**	1.181	±	0	1.069	±	4	1.067	±	3
			0.03			0.01			0.04				0.02			0.03			0.02
AAG	1.059	±	1	1.128	±	8	1.092	±	9	**CAG**	1.095	±	1	0.888	±	3	0.853	±	3
			0.02			0.03			0.03				0.05			0.03			0.02
AAC	0.979	±	9	0.995	±	1	0.988	±	4	**CAC**	0.851	±	0	1.080	±	5	1.067	±	0
			0.03			0.02			0.02				0.02			0.02			0.02
AAT	0.954	±	1	0.925	±	8	0.971	±	3	**CAT**	0.847	±	2	1.009	±	3	0.993	±	0
			0.03			0.02			0.01				0.07			0.06			0.02
AGA	1.133	±	1	1.044	±	8	1.036	±	7	**CGA**	0.845	±	3	0.796	±	9	0.910	±	3
			0.04			0.01			0.06				0.05			0.02			0.04
AGG	0.907	±	4	1.002	±	8	0.962	±	0	**CGG**	0.981	±	5	1.025	±	4	1.111	±	1
			0.04			0.02			0.03				0.05			0.05			0.03
AGC	0.968	±	2	0.959	±	3	0.946	±	5	**CGC**	1.151	±	0	1.201	±	9	1.092	±	5
			0.03			0.02			0.02				0.07			0.04			0.02
AGT	1.054	±	6	0.979	±	8	1.012	±	6	**CGT**	0.969	±	1	1.005	±	5	0.948	±	1
			0.04			0.01			0.05				0.05			0.01			0.02
ACA	0.987	±	2	1.007	±	7	1.054	±	0	**CCA**	1.048	±	7	1.092	±	8	1.011	±	5

ACG	0.819	±	0.055	0.855	±	0.029	0.828	±	0.065	**CCG**	1.082	±	0.042	1.146	±	0.027	1.222	±	0.042
ACC	1.161	±	0.061	1.088	±	0.035	1.093	±	0.053	**CCC**	0.784	±	0.039	0.780	±	0.038	0.906	±	0.045
ACT	1.048	±	0.041	1.067	±	0.032	0.975	±	0.034	**CCT**	1.080	±	0.065	0.927	±	0.041	0.968	±	0.042
ATA	1.113	±	0.070	1.029	±	0.039	0.947	±	0.044	**CTA**	0.858	±	0.043	1.029	±	0.037	1.118	±	0.031
ATG	0.973	±	0.058	0.989	±	0.023	1.009	±	0.040	**CTG**	0.877	±	0.029	0.946	±	0.028	0.986	±	0.058
ATC	0.945	±	0.051	0.958	±	0.018	1.000	±	0.018	**CTC**	1.175	±	0.033	1.087	±	0.041	0.974	±	0.025
ATT	1.020	±	0.057	1.048	±	0.028	1.038	±	0.043	**CTT**	1.064	±	0.034	0.988	±	0.025	0.965	±	0.017
GAA	1.016	±	0.033	1.027	±	0.056	0.999	±	0.018	**TAA**	0.000	±	0.001	0.983	±	0.023	0.907	±	0.030
GAG	1.037	±	0.029	1.075	±	0.054	1.145	±	0.040	**TAG**	0.000	±	0.000	0.904	±	0.026	0.965	±	0.035
GAC	0.929	±	0.035	0.850	±	0.048	0.927	±	0.033	**TAC**	1.837	±	0.011	1.053	±	0.035	1.047	±	0.022
GAT	0.999	±	0.019	0.943	±	0.033	0.930	±	0.017	**TAT**	1.567	±	0.058	1.138	±	0.035	1.130	±	0.011
GGA	1.240	±	0.039	1.079	±	0.069	1.022	±	0.024	**TGA**	0.000	±	0.000	1.010	±	0.012	1.018	±	0.017
GGG	0.790	±	0.018	0.910	±	0.034	0.809	±	0.022	**TGG**	1.827	±	0.059	1.040	±	0.025	1.083	±	0.035
GGC	1.042	±	0.016	1.020	±	0.058	1.115	±	0.029	**TGC**	0.881	±	0.021	0.916	±	0.021	0.885	±	0.035
GGT	1.013	±	0.033	1.010	±	0.041	1.013	±	0.025	**TGT**	0.933	±	0.034	1.034	±	0.025	1.028	±	0.025
GCA	0.833	±	0.032	0.898	±	0.022	0.931	±	0.019	**TCA**	1.413	±	0.050	1.005	±	0.011	1.042	±	0.029
GCG	1.094	±	0.034	1.054	±	0.027	1.075	±	0.038	**TCG**	1.045	±	0.039	1.000	±	0.025	0.818	±	0.037
GCC	1.162	±	0.034	1.091	±	0.035	0.979	±	0.036	**TCC**	0.814	±	0.035	1.011	±	0.036	1.092	±	0.035
GCT	0.956	±	0.043	0.961	±	0.029	1.028	±	0.020	**TCT**	0.921	±	0.017	1.014	±	0.023	0.987	±	0.028
GTA	0.881	±	0.040	0.929	±	0.027	1.041	±	0.045	**TTA**	1.335	±	0.049	0.996	±	0.028	0.890	±	0.032
GTG	1.077	±	0.046	1.071	±	0.017	1.123	±	0.040	**TTG**	1.124	±	0.084	1.021	±	0.018	0.906	±	0.028
GTC	0.934	±	0.038	0.908	±	0.036	0.976	±	0.020	**TTC**	0.904	±	0.034	1.012	±	0.018	1.041	±	0.022
GTT	1.076	±	0.036	1.086	±	0.017	0.991	±	0.025	**TTT**	0.850	±	0.060	0.911	±	0.030	1.028	±	0.028

Tab S. Average trinucleotide bias of intron sequences. The trinucleotide contents were calculated for each windoe position and averaged over all sequences The γ index were calculated according to Karlin ()

Arabidopsis thaliana	Oryza sativa

	gindex				gindex		
	average		σ		average		σ
AAA	0.986	±	0.014	**CAA**	1.011	±	0.026
AAG	0.992	±	0.015	**CAG**	0.878	±	0.025
AAC	1.012	±	0.029	**CAC**	1.012	±	0.023
AAT	0.991	±	0.019	**CAT**	1.040	±	0.024
AGA	0.958	±	0.023	**CGA**	1.028	±	0.055
AGG	0.923	±	0.058	**CGG**	1.072	±	0.140
AGC	1.052	±	0.067	**CGC**	0.812	±	0.050
AGT	1.042	±	0.043	**CGT**	1.016	±	0.022
ACA	0.958	±	0.045	**CCA**	1.115	±	0.024
ACG	1.054	±	0.038	**CCG**	1.151	±	0.085
ACC	1.045	±	0.048	**CCC**	0.968	±	0.044
ACT	1.007	±	0.038	**CCT**	0.887	±	0.022
ATA	1.019	±	0.036	**CTA**	0.963	±	0.036
ATG	1.032	±	0.026	**CTG**	1.017	±	0.035
ATC	0.971	±	0.025	**CTC**	1.024	±	0.016
ATT	1.000	±	0.026	**CTT**	1.007	±	0.014
GAA	1.111	±	0.026	**TAA**	0.949	±	0.015
GAG	1.087	±	0.056	**TAG**	1.103	±	0.050
GAC	1.044	±	0.022	**TAC**	1.013	±	0.051
GAT	1.070	±	0.030	**TAT**	0.997	±	0.037
GGA	0.984	±	0.054	**TGA**	1.041	±	0.015
GGG	1.049	±	0.144	**TGG**	1.006	±	0.043
GGC	1.081	±	0.046	**TGC**	0.988	±	0.023
GGT	0.971	±	0.027	**TGT**	0.983	±	0.025

	gindex				gindex		
	average		σ		average		σ
AAA	1.028	±	0.029	**CAA**	0.961	±	0.022
AAG	0.930	±	0.026	**CAG**	0.948	±	0.014
AAC	1.025	±	0.027	**CAC**	1.034	±	0.029
AAT	0.981	±	0.015	**CAT**	1.057	±	0.015
AGA	0.955	±	0.018	**CGA**	0.985	±	0.029
AGG	0.986	±	0.020	**CGG**	1.049	±	0.059
AGC	0.993	±	0.017	**CGC**	0.995	±	0.022
AGT	1.053	±	0.021	**CGT**	0.983	±	0.034
ACA	0.987	±	0.074	**CCA**	1.027	±	0.019
ACG	0.962	±	0.044	**CCG**	1.070	±	0.040
ACC	1.004	±	0.021	**CCC**	0.960	±	0.023
ACT	1.018	±	0.031	**CCT**	0.975	±	0.014
ATA	0.990	±	0.033	**CTA**	1.015	±	0.022
ATG	1.062	±	0.012	**CTG**	1.025	±	0.010
ATC	1.005	±	0.031	**CTC**	1.033	±	0.015
ATT	0.983	±	0.010	**CTT**	0.957	±	0.009
GAA	1.051	±	0.038	**TAA**	0.975	±	0.030
GAG	1.097	±	0.072	**TAG**	1.125	±	0.058
GAC	1.016	±	0.028	**TAC**	1.011	±	0.044
GAT	1.115	±	0.055	**TAT**	0.960	±	0.066
GGA	1.031	±	0.025	**TGA**	1.042	±	0.017
GGG	1.021	±	0.028	**TGG**	0.982	±	0.014
GGC	1.039	±	0.057	**TGC**	0.994	±	0.021
GGT	0.943	±	0.017	**TGT**	0.987	±	0.012

GCA	0.966	±	0.043	**TCA**	1.021	±	0.010	**GCA**	0.962	±	0.017	**TCA**	1.038	±	0.034
GCG	0.876	±	0.057	**TCG**	0.955	±	0.023	**GCG**	1.014	±	0.036	**TCG**	0.962	±	0.030
GCC	1.074	±	0.082	**TCC**	0.966	±	0.024	**GCC**	1.086	±	0.060	**TCC**	0.981	±	0.029
GCT	1.011	±	0.018	**TC**		±	0.009	**GCT**		±	0.031	**TCT**	1.009	±	0.032
GTA	0.964	±	0.037	**TT**		±	0.018	**GTA**		±	0.034	**TTA**	1.010	±	0.014
GTG	1.003	±	0.029	**TTG**	0.973	±	0.015	**GTG**	1.009	±	0.013	**TTG**	0.946	±	0.015
GTC	0.993	±	0.015	**TTC**	1.007	±	0.017	**GTC**	0.950	±	0.043	**TTC**	0.988	±	0.018
GTT	1.031	±	0.012	**TTT**	0.982	±	0.010	**GTT**	1.028	±	0.016	**TTT**	1.019	±	0.008

Slope 1

A G C T **A G C T**

Slope 2

Segmented Regression	14.1	1.5	1.1	10.5	7.2	32.6	43.5	9.5
	42.2	14.1	12.5	40.6	0.7	13.7	13.6	1.3
Linear Regression	13.3	2.3	1.7	2.2	1.6	11.0	7.8	12.0
Not Fitted	20.9	24.8	19.7	23.9				

Table S3:
Percentage of significant fitting to the segmented and linear regression for *Ozyza sativa*.