UNIVERSITÀ DEGLI STUDI DI SASSARI

SCUOLA DI DOTTORATO DI RICERCA
Scienze e Biotecnologie
dei Sistemi Agrari e Forestali
e delle Produzioni Alimentari
Indirizzo: Produttività delle Piante Coltivate

Ciclo XXV

Comparative analysis of genic structure in plants

Dr.ssa Giampiera Milia

Direttore della Scuola	Prof.ssa Alba Pusino
Referente di Indirizzo	Prof.ssa Rosella Motzo
Docente Guida	Prof. Andrea Porceddu

Anno accademico 2011- 2012

Index

Preface

The availability of fully sequenced genomes and comprehensive expression studies offer new opportunities for studying the relationships between structures and functions of genes. These topics are relevant for the fields of plant biotechnology since the debate on transgenic plants has raised concerns about the impact of heterologous gene expression and structure on host genome integrity.

In this thesis we adopted comparative approaches to study the degree of variation of genic structures among plant species.

In **chapter 1** we revised the relationship between average expression level of a gene and its architecture in 8 plant species. The results indicate that gene expression is non monotonically related to gene size.

In **chapter 2** we present the results of a thorough characterization of intron size in *A. thaliana* and *V. vinifera*. Our data suggest that *Vitis vinifera* has on average longer introns than *A. thaliana.* Data indicate that Vitis intron tend to become longer due to the insertion of repetitive element but not for an increased rate of small insertion over deletions.

Chapter 3 reports on the evolution of intron-exon organization in eleven species spanning an evolutionary range from green algae to extant dicots. The picture confirmed the tenet that intron losses outnumbered the gains. However we identified several examples suggesting that the general view may be punctated by relevant exceptions

Chapter 1

Title: Non monotonic relation between gene size and gene expression level in plant genes

Abstract

The advent of synthetic biology has refuelled the interest on the relations between gene structure and expression. While compelling evidences obtained in species as diverse as *D. melanogaster*, *C. elegans* and *H. sapiens* indicate a common tendency for highly expressed genes to be compact, debates over the evolutionary mechanisms underlying this phenomenon seem to be far from converging to a universal view. In this work we adopted a rank based approach to reconcile expression data from various platforms and plant species into a unique framework. Our results demonstrate that in plant genes the relations between gene expression level and gene length is universally "non monotonic". However the trends showed topological differences that suggest how the factors at work may weigh differently depending on the genomic context. Some of the presented considerations may have a significant impact on transgene designing.

Introduction

Several analyses carried out on organisms as diverse as bacteria, mammals and plants have led to the conclusion that high level of gene expression are associated to reduced genic sizes (Yang 2009, Urrutia and Hurst. 2003; Camiolo *et al*.2009). Three theories have been proposed in order to explain such a phenomenon: (i) requirement of genome organization (Vinogradov 2004), (ii) mutational bias (Urrutia and Hurst 2003; Comeron 2004) and (iii) selection for economy (Urrutia and Hurst 2003; Seoighe *et al.* 2005).

The first hypothesis, also known as "genome by design", was firstly proposed by Vinogradov *et al.* (2004) and relies on the idea that broadly expressed genes are more compact because they need a lower number of regulative elements. Sironi and co-workers (2006) supported this idea while underlining the role of MSC (multi-species conserved sequences) on the evolution of intron structure. However such a hypothesis was recently challenged by the work of Carmel *et al.* (2009) that reported a much weaker dependence between gene compactness and expression pattern. Accordingly, Li and coworkers (2007) found that high functional/regulatory complexity in genes does not mirror an increase in intron size.

The mutational bias model postulates that genes sharing the same expression pattern tend to cluster into genomic regions (Urrutia and Hurst 2003; Comeron 2004) and thus are subjected to similar rates of insertions and deletions.

Finally the "selection for economy" model relies on the idea that highly expressed genes are shorter due to the need to reduce the energy cost associated to expression (Urrutia and Hurst 2003; Seoighe *et al.* 2005; Carmel and Koonin 2009). Both transcription and translation may be involved in such a process and for this reason exons as well introns evolution must be considered (e.g. reduction in exon size but not in intron size would indicate that selection for economy is driven by translation rather than transcription). Highly expressed genes could require shorter transcripts also in order to minimize the time of transcription when a large amount of mRNA is required in a short period (Rao *et al.* 2010).

Studies aimed at investigating the association between the expression level and the structural features of the genic sequences in plants, have led to contradictory results.

Indeed while Camiolo *et al*. (2009) and Yang (2009), have found that highly expressed genes are also the most compact in *Arabidopsis thaliana*, the opposite conclusion was reached by Ren and co-workers (2006). Reasons for this apparent contradiction could be in part methodological. The expression profile is composed by two distinct, although correlated (Park *et al.* 2012) parameters: the expression breadth (EB) that is a measure of the number of tissues in which a gene is expressed, and the expression level (EL) that is inferred by measures of the transcript abundance. The way the measures of transcript abundances are averaged can lead to results that are, to various degree, influenced by the expression breadth and consequently attribute a different weight to either components of the expression profile (Camiolo *et al*., 2009). In addition transcript abundances can be measured with different techniques (e.g. microarray or sequencing methods) or even different platforms for the same technique and thus their comparisons can be cumbersome.

A considerable amount of both genomic and expression data are now available for plant species thanks to several ongoing genome projects and the design of new expression platforms. In this work we analyze the association between the genic structure of eight plant organisms (e.g. *Arabidopsis thaliana*, *Oryza sativa*, *Medicago truncatula*, *Populus trichocarpa, Glycine max, Zea mays, Vitis vinifera and Solanum lycopersicum*) and the expression profile of their genes with the aim of highlighting both common and peculiar traits and shed light on the evolutionary constraints that can be involved in the observed trends.

Materials and Methods

Expression data.

Arabidopsis thaliana, *Oryza sativa*, *Populus trichocarpa*, *Glycine max,* and *Medicago truncatula* expression atlas were downloaded from the PlExDb website (Dash *et al.* 2012) website. Arabidopsis dataset was retrieved by the series AT40 and consisted of 63 microarray experiments. (http://www.plexdb.org/modules/PD_browse/experiment_browser.php?experiment=AT40. *Oryza sativa* expression profiles were estimated by the experiments series OS5 (http://www.plexdb.org/modules/PD_browse/experiment_browser.php?experiment=OS5) based on the analysis of 15 tissues/experimental stages. *Populus trichocarpa* expression atlas was based on the estimation of the expression values calculated in 9 poplar tissues, each retrieved from the series PT2 (http://www.plexdb.org/modules/PD_browse/experiment_browser.php?experiment=PT2). *Glycine max* expression profiles were estimated by considering the experiment series GM10 which consist of 29 experiments mainly exploring the seed developmental stages (http://www.plexdb.org/modules/PD_browse/experiment_browser.php?experiment=GM10). *Medicago truncatula* expression atlas was retrieved by the series ME1 (http://www.plexdb.org/modules/PD_browse/experiment_browser.php?experiment=ME1) which is based on the analysis of 18 tissues/developmental stages. All the data retrieved from the PlExDb were obtained by using Affymetrix platforms and, with the exception of few *Glycine max* tissues, all the experiments were repeated three times. MAS5.0 was chosen for the normalization of the microarray data. For each probeset, data from different replicates were averaged and an arbitrary cutoff of 100 was used in order to define a gene as expressed (e.g. below this value the gene expression value was set to 0).

Zea mays expression data were downloaded from NCBI (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE27004) and rely on the analysis carried out on 60 tissues/developmental stages, each replicated three times. The Nimblegen Maize Whole-Genome Microarray 385K (VersionV1_4a.53) was used and expression data were normalized by the use of the RMA algorithm. Following the author suggestion, a cutoff of 200 was used to define a gene as “expressed”.

Vitis vinifera expression atlas was obtained by applying a Nimblegen custom array in order to analyze the expression profile of 54 grape tissues (Fasoli *et al.* 2012). In this case the expression values of the "random" probesets were sorted and the value found at the 95th percentile was used as a cutoff in order to define a gene as expressed using a normal kernel smoothing method.

Finally *Solanum lycopersicum* expression data were retrieved by the NCBI website (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE19326) and rely on the analysis of 24 tomato tissues (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE19326). Normalization and analysis of microarray data were performed using GeneSpring GX 7.3 software (Agilent Technologies, URL: http://www.home.agilent.com/). Again for each set of replicated experiments a cut off of 100 was applied to the average expression value in order to define a gene as "expressed".

All data were log transformed in order to reduce the leverage effect of possible outlier expression values.

Information about the association between probe sets and their corresponding locus for poplar were kindly provided by Prof. Chung-Jui Tsai of the University of Georgia. The same information for *Medicago truncatula* was downloaded from the Noble Foundation website (http://bioinfo3.noble.org/medicago/MT3.5/Mt3.5v2_RELEASE_20100723/affy/Mt3.5v2_RELEASE_20100723_affymap.gene2probes.map), while data for soybean was retrieved from http://seedgenenetwork.net/annotate. Mapping information for the remaining species were obtained directly from the specific microarray manufacturers. When more than a single alternative splicing variant was associated to the same probeset one of them was randomly chosen.

Rank-based expression level categorization

Since the available data were obtained from different microarray platforms and from different tissues a ranking approach was used in order to put the expression level measurements on the same scale and make the comparisons more reliable (Carmel and Koonin 2009). Briefly for each organism expression data were combined in a matrix consisting of n_g rows (number of genes) and n_t columns (number of tissues). Each column was then ranked in 30 categories such as the lowest expression value had the value 1 while the highest had the value 30. Such a classification was achieved by calculating the

expression level range (e.g. the difference between the highest and the lowest expression values divided by 30) and then assigning each genes according to a given class based on its expression value. This process produced a new matrix in which the original expression values were replaced by ranks. In order to obtain a single expression values for each gene, an average was performed only considering those tissues in which that gene was actually expressed and the resulting value was rounded to the nearest integer. The expression breadth was measured as the number of tissues in which a gene is expressed divided for the total number of tissues.

Genic parameters

Vitis vinifera (Jaillon *et al.* 2007), *Populus trichocarpa* (Tuskan *et al.* 2006), *Medicago truncatula* (2012), *Glycine max* (Schmutz *et al.* 2010), *Arabidopsis thaliana* (Swarbreck *et al.* 2008), *Zea mays* (Schnable *et al.* 2009) and *Oryza sativa* (Ouyang *et al.* 2007) annotation gff3 file and chromosomes sequences were downloaded from the Phyzome database (Goodstein *et al.* 2012). For *Solanum lycopersicum* genomic data were retrieved from the Sol Genomics network database (Bombarely *et al.* 2011) at http://solgenomics.net/itag/release/2.3/list_files (ITAG2.3_gene_models.gff3 and ITAG2.3_genomic.fasta). Annotations information were used in order to extract introns and exons from the chromosomes raw sequences. Such a task together with the calculation of the sequences length and the number of exons/introns was achieved by using custom C/C++ scripts. For each transcript gene lengths were calculated by summing total length of introns and exons.

Results

The relationships between structural parameters of nuclear genes and the average level of transcript accumulation were investigated in eight plant species (six dicothyledons: *Arabidopsis thaliana*, *Vitis vinifera*, *Populus trichocarpa*, *Medicago truncatula,* and *Solanum lycopersicum*, *Glycine max*; and 2 monocotyledons: *Oryza sativa*, and *Zea mays* species). These species were chosen since they represent a wide range of genic and genomic structures as highlighted in Table 1.

Species	Clade	Chr.	Genome size (MB)	Annotated genes	Av. Num. of introns	Av. Intron length	Av. Exon length
Arabidopsis thaliana	dicot.	5	116	27416	5.9	198.7	441.3
Oryza sativa	monocot.	12	362	56171	4.6	485.5	555.3
Populus trichocarpa	dicot.	19	304	40668	5.0	382.0	427.1
Vitis vinifera	dicot.	19	414	26346	5.6	1005.3	317.3
Medicago truncatula	dicot.	8	375	50962	3.6	490.5	411.3
Glycine max	dicot.	20	915	46367	5.9	504.9	396.9
Solanum lycopersicum	dicot.	12	950	69523	4.7	590.8	350.4
Zea mays	monocot.	10	1969	39656	4.9	716.0	349.3

Table 1: Genomic features of the studied species

Although scatterplots of gene size as function of expression level always showed a non ubiquitous monotonic trend several trend's topologies could be distinguished (Figure 1). Up to a certain level of expression, gene size and expression level were positively associated. Passed that level of expression, the positive association was broken and gene size and expression level became negatively correlated (see Figure 1). Exceptions to such a behavior were noticed for *Medicago truncatula, Glycine max* and *Solanum lycopersicum* that showed an initial flat segment (featuring the absence of an association) followed by a negative trend. The position of the breakpoint (i.e the level of expression at which the slope change occurred) was a distinctive character for the relationships. In fact *A. thaliana* and *P. trichocarpa* showed the breakpoint in correspondence of classes 5 and 10 respectively while for *O. sativa* and *V. vinifera* the breakpoint was observed at class 15. Breakpoint in *Z. mays* scatterplot was observed in correspondence of expression level categories between 15 and 20 while *G. max*, together with *S. lycopersicum*, showed a flat slope until class 15, which was followed by a negative trend. Finally the analysis carried

out on *M. truncatula* did not reveal any association between gene length and expression level.

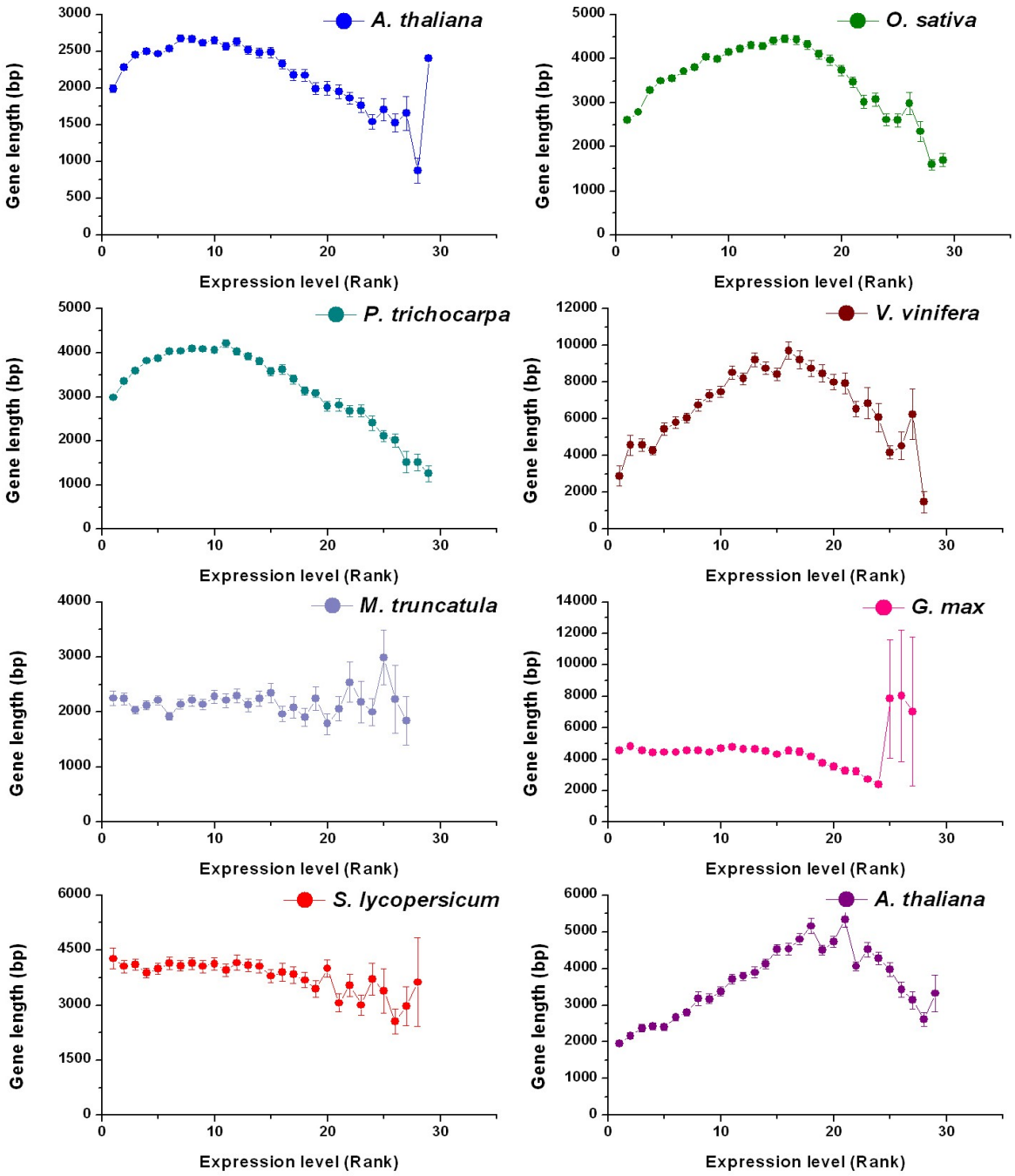


Figure 1: Gene length as function of the expression level category

Because several components can contribute to the total gene length such as the number and average length of introns and exons the analysis was repeated for each component separately. Estimation of the number of exons as a function of the expression

level (Figure 2) led to results that were quite similar to those observed for the total gene length. Indeed, trends and breakpoints positions were almost over imposable to the ones observed for the relationships between gene size and expression level. Interestingly the two Fabaceae showed an absolute absence of any kind of association between number of exons and expression level.

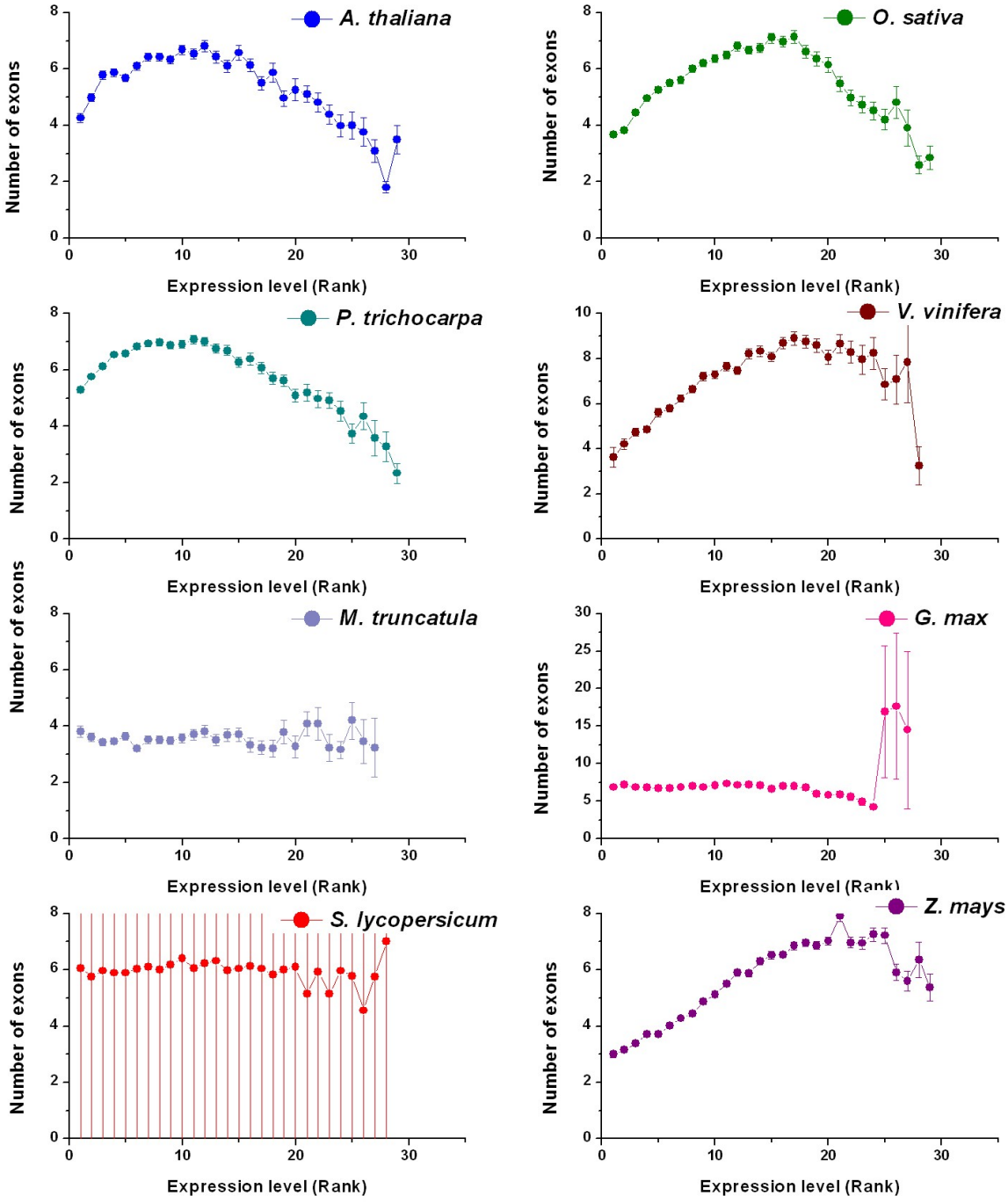


Figure 2: Number of exons as function of the expression level category

Analyses of the association between total intron or total exon lengths with the expression levels were in line with what observed for the number of exons (Figure 3-4).

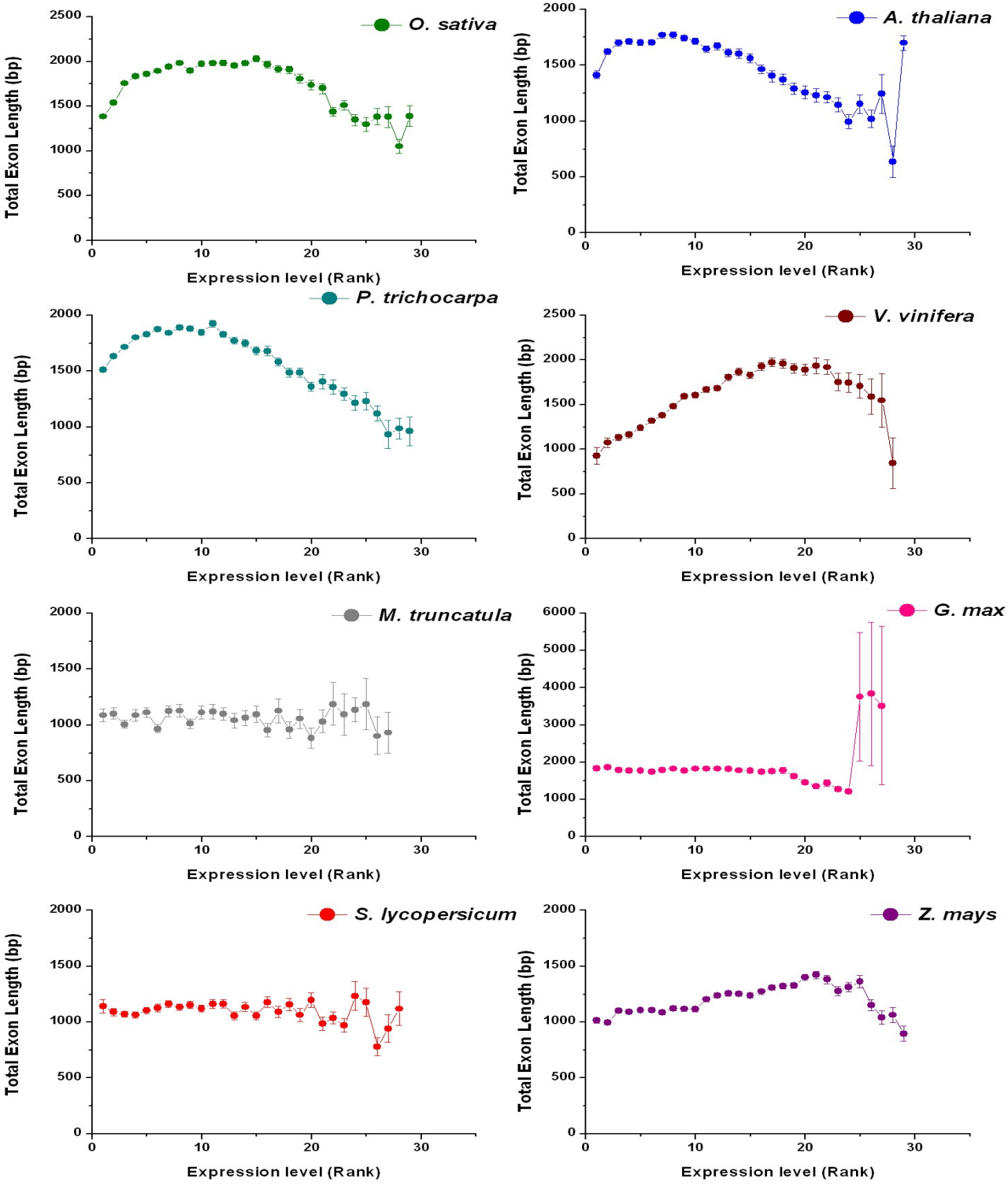


Figure 3: Total exon length as a function of the expression level category

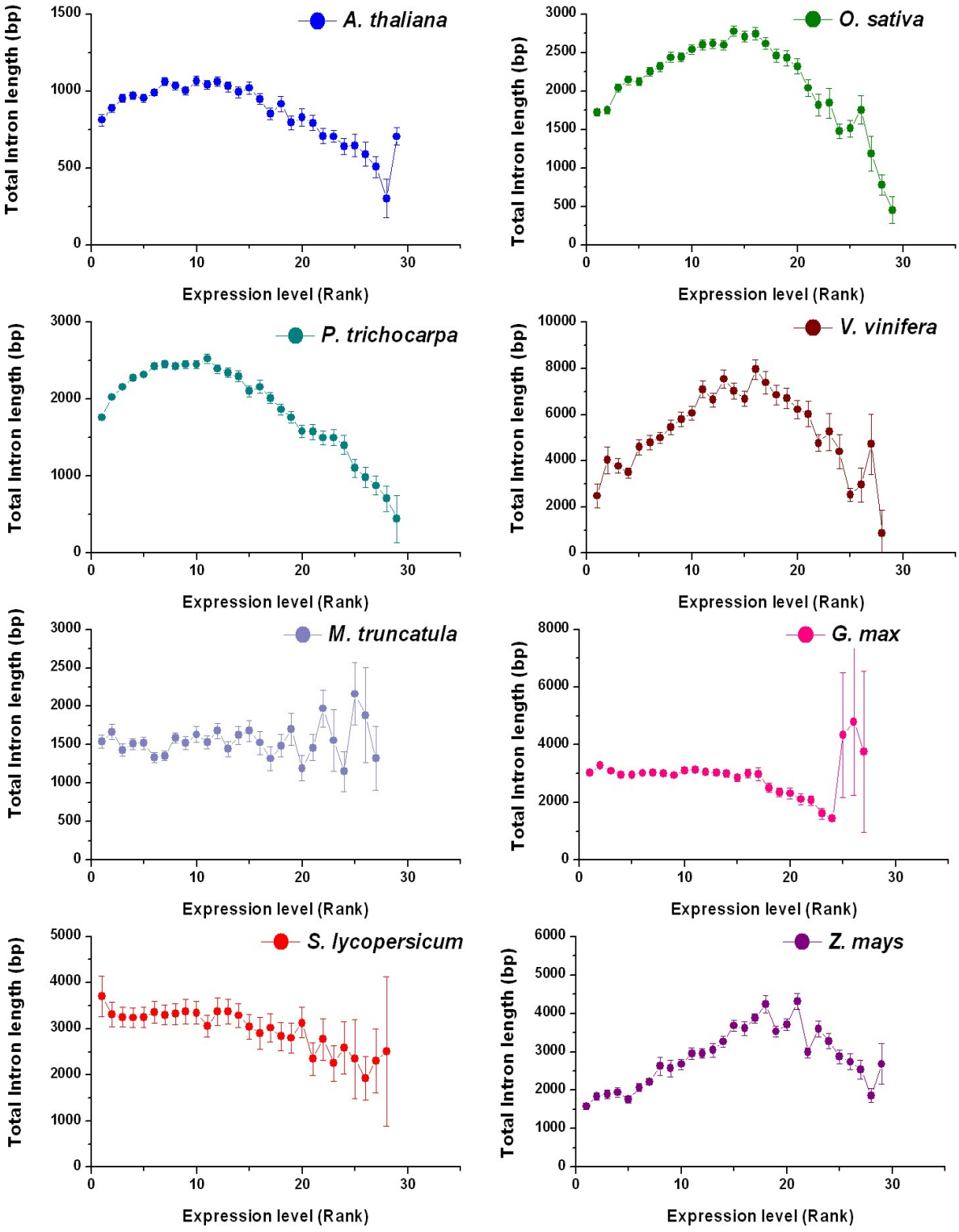


Figure 4: Total intron length as function of the expression level category

A tendency towards compactness was observed at high expression level when considering the average exon length for many of the analyzed organisms. Again *M. truncatula, G. max* and *S. lycopersicum* showed a deviation from this common behavior

(Figure 5). A general size reduction was also observed in introns at high expression level, although *Z. mays, A. thaliana* and *M. truncatula* did not followed such a trend (Figure 6).

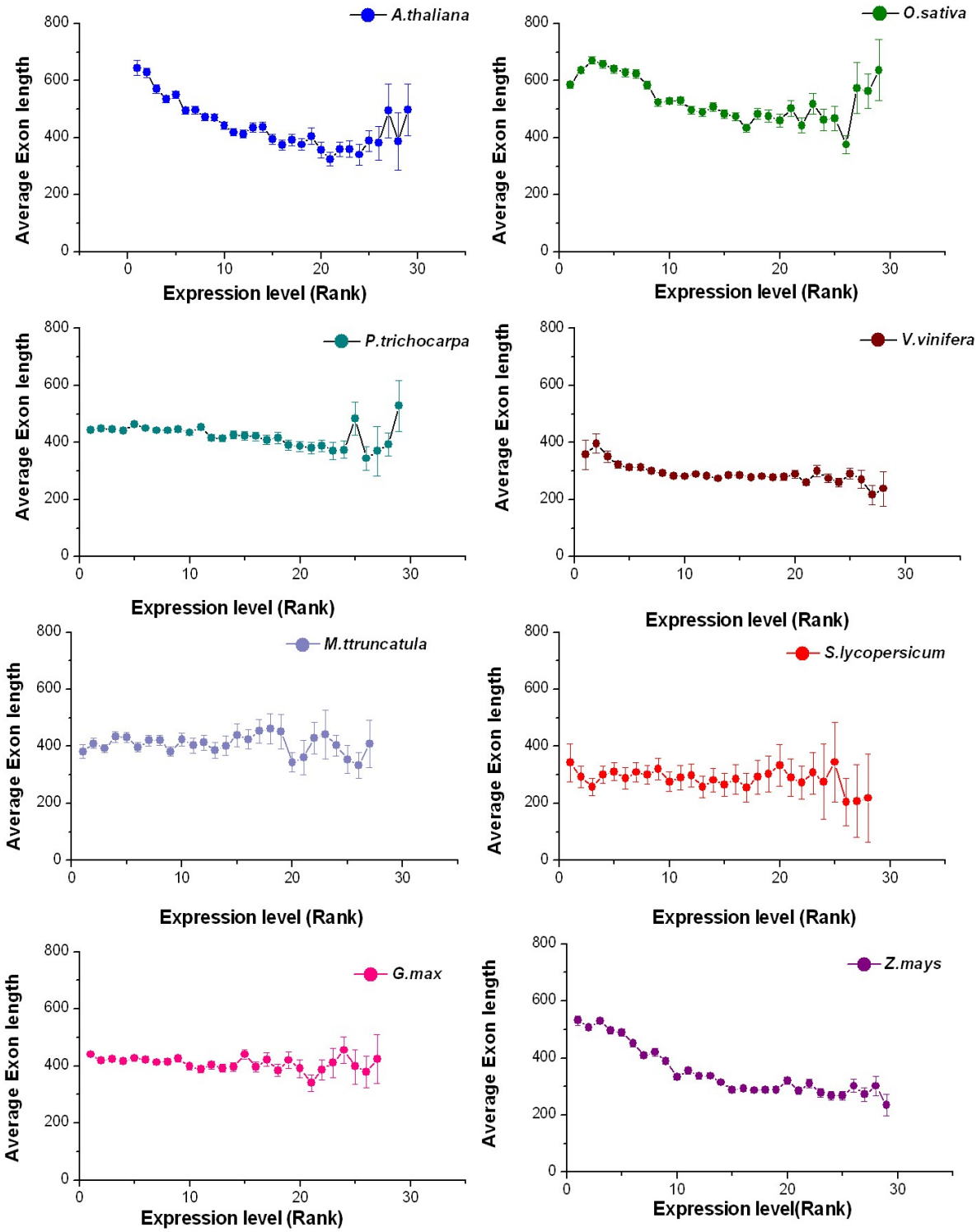


Figure 5: Average exon length as function of the expression level category

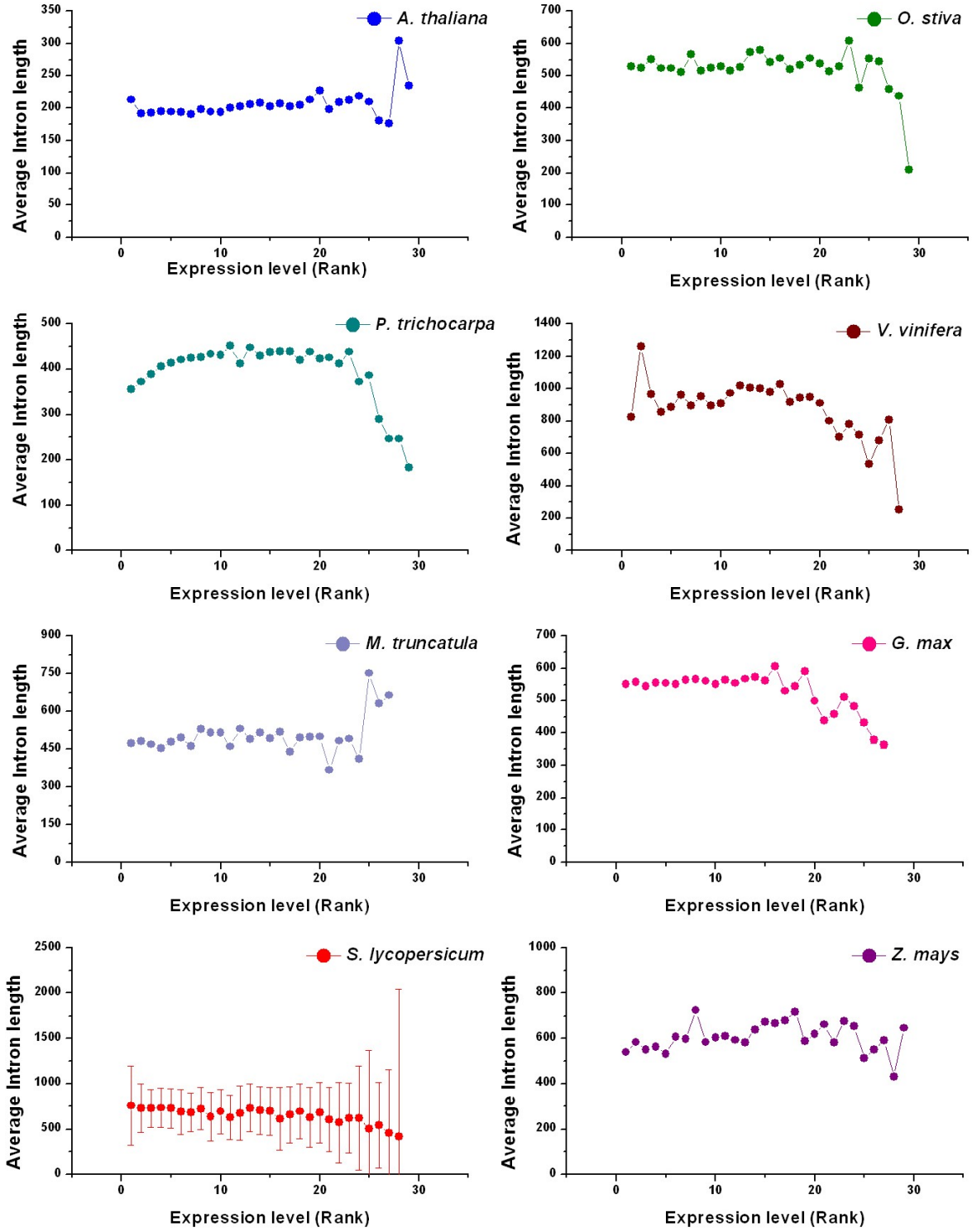


Figure 6: Average intron length as function of the expression level category

Carmel *et al*. (2009) have proposed that the non monotonicity of the relationships of the size of genes as function of their level of expression could be explained by the concomitance of different selective constraints acting on genes. While the negative trends

observed for highly expressed genes may be due to a selection for economy, the genomic design may underlie the positive slopes observed for lowly expressed genes. Indeed, keeping in mind the strong association between expression level and expression breath, accumulation of introns at the first part of the scatter-plots depicted in Figure 2, may reflect the need of increasing the plasticity of genes that necessitate to adapt their expression to the requirements of different tissues (Camiolo *et al.* 2009) Controlling for the expression breadth may help in addressing the causes that underlie the positive association between the number of introns and the expression level. Such an approach could also highlight differences in the structure of lowly and highly expressed genes because of the strong positive correlation between the two components of the expression profile. To test such a hypothesis we re-analyzed the relationships between genic parameters and expression level of those genes that were expressed either in all tissues or in a tissue specific manner (only tissues with more than 9 expressed genes were included). *V. vinifera* and *M. truncatula* were excluded from the analysis because the corresponding dataset performed poorly the classification process.

For all the analyzed species the association between the transcript abundance and the number of exons disappeared in tissue specific genes. On the other hand, in general, monotonic negative trends were observed for housekeeping genes (Figure 7). It is interesting to notice that the dicotyledons *O. sativa* and *Z. mays* showed for the housekeeping genes non monotonic trends that were very similar to the ones observed when the whole dataset was used in the analysis (Figure 7).

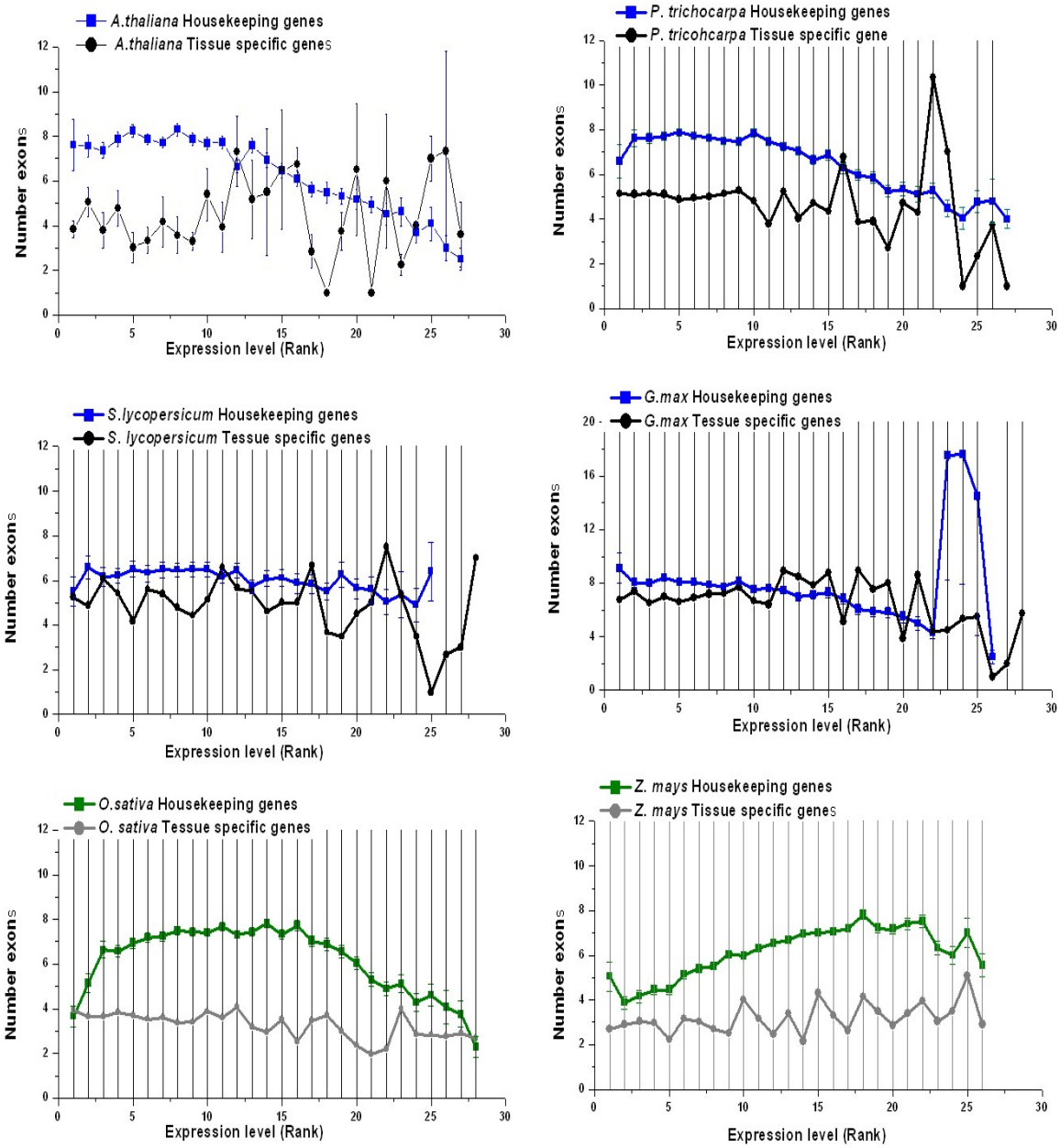


Figure 7: Average number of exons in housekeeping and tissue specific genes as a function of the expression level (Blue / Black eudicots, Green / Gray monocots).

Discussion

In this work we have demonstrated that the relationship between expression level and gene length is "non monotonic" in five of the eight analyzed plant species. Up to a certain level, genes become longer with the increase of the expression level, but at higher expression level both coding and non coding sequences become more compact. These findings are in agreement with the results reported by Yang *et al.* (2009) in Arabidopsis and Rice and by Carmel and Koonin (2009) in species as diverse as *D. melanogaster*, *C. elegans* and *H. sapiens* and as a whole suggest that the "non monotonicity of the trends" should be regarded as a feature common to different genome contexts.

In our study we found three species, *G. max, M. truncatula* and *S. lycopersicum* that are characterized by an atypical non monotonic relationship with an initial flat segment (featuring the absence association) followed by a decreasing segment. However we warn that before these apparently "anomalous" trends are claimed as of a new typology more information on the more common category should be gathered. The slope change (breakpoint) occurred at points corresponding to different ranks of gene expression in the analyzed plant species.

Although it should be always kept in mind that the position of the breakpoint could be influenced by differences in the dynamic range of the techniques used to measure gene expression, we consider that such an observation has an important biological meaning. Indeed similar observations were reported by Carmel and Koonin (2009) in non plant species and more recently by Park and collaborators (2012) in human and mouse.

For sake of simplicity the trends of Arabidopsis and Populus, that showed the breakpoint at expression level category of 10-15, are herefater referred as of the "early breakpoint" trend categories, Rice and Vitis trends (breakpoint at expression level of 15) as of the "intermediate breakpoint" category and maize trend as of the "late breakpoint" category.

It has been hypothesized that highly expressed genes may experience strong selective forces toward gene compactness due to the energetic costs associated with the transcription process. In contrast genes that exhibit intermediate levels of expression breadths may require the most complex signals for regulation and become the longest. The non linear trend could be explained by a combination of both selective forces toward

efficient cellular expression and for more regulatory sequences necessary for complex gene regulation. This hypothesized scenario redirects all the attentions toward the factors determining the balance between these forces. Because there is no room to think that a basic process such as transcription would have different (absolute) energetic costs in different cellular environment we are allowed to think that the genomic context plays a strong effect. One hypothesis could be that species experiencing strong selection for genome size reduction would also make stronger efforts to minimizing the cost associated to gene expression. However observations on differential intensity or direction of evolutive forces acting on genic or non genic regions within a same species caution about such a conclusion. As a matter of fact Wendel *et al.* (2002) have found that the the forces determining genome size in *Gossypium* have only a limited impact on introns. Moreover studies that combined intron data from multiple species have indicated that beside the total intron length also the distributions of the average size of introns within genes could be subjected to variations among species. The evolutionary forces working toward genome expansion may have a different impact in genic and non genic sequence and their relative importance may vary between species. Different class of retrotransposon elements have been reported to species-specific abundance or genomic distributions. For example, a comparison of gene structure between Arabidopsis and Vitis has demonstrated that in spite of a highly level of exon-intron structure conservation, Vitis introns are particulary enriched in LINE elements (see chapter 2). It is possible that this transposable elements have coevolved with some of the mechanisms tuning gene expression in Vitis and therefore have become essential elements for gene expression regulation.

In conclusion, we propose that the position of the breakpoint in the trends representing the relation between gene size and expression level is a feature influenced by several genomic factors whose relative importance may vary with the genomic context.

Literature Cited

2012 Medicago: Department of Energy Joint Genome Institute (JGI) before scientific publication. *Medicago: Department of Energy Joint Genome Institute (JGI) before scientific publication* .

Bombarely A., Menda N., Tecle I. Y., Buels R. M., Strickler S, Fischer-York,T., Pujar A., Leto J., Gosselin J.and Mueller L. A. 2011 The Sol Genomics Network (solgenomics.net): growing tomatoes using Perl. *Nucleic Acids Res.* **39**: D1149-D1155.

Camiolo S., Rau D. and Porceddu A. 2009 Mutational biases and selective forces shaping the structure of Arabidopsis genes. *PLoS One.* **4**: e6356.

Carmel L., and Koonin E.V., 2009 A universal nonmonotonic relationship between gene compactness and expression levels in multicellular eukaryotes. *Genome Biol.Evol.* **1**: 382-390.

Comeron, J. M. 2004 Selective and mutational patterns associated with gene expression in humans: influences on synonymous composition and intron presence. *Genetics* **167**: 1293-1304.

Dash S., Van H. J., Hong L., Wise R. P., and Dickerson J. A., 2012 PLEXdb: gene expression resources for plants and plant pathogens. *Nucleic Acids Res.* **40**: D1194-D1201.

Fasoli M., Santo, S. D., Zenoni, S., Tornielli, G. B., Farina, L., Zamboni, A., Porceddu, A., Venturini, L., Bicego, M., Murino, V., Ferrarini, A., Delledonne, M. and Pezzotti, M. 2012 The Grapevine Expression Atlas Reveals a Deep Transcriptome Shift Driving the Entire Plant into a Maturation Program*Plant Cell.*

Goodstein D. M., Shu S., Howson R., Neupane R., Hayes R. D. *et al.* 2012 Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* **40**: D1178-D1186.

Jaillon O., Aury J.-M., Noel B., Policriti A., Clepet C., Casagrande A., Choisne N., Aubourg S., Vitulo N., Jubin C., Vezzi A., Legeai F., Hugueney P., Dasilva C., Horner D., Mica E., Jublot D., Poulain J., Bruyère C., Billault A., Segurens B., Gouyvenoux M., Ugarte E., Cattonaro F., Anthouard V., Vico V., Fabbro C. D., Alaux M., Gaspero G. D., Dumas V., Felice N., Paillard S., Juman I., Moroldo M., Scalabrin S., Canaguier A., Clainche I. L., Malacrida G., Durand, E., Pesole G., Laucou V., Chatelet P., Merdinoglu D., Delledonne M., Pezzotti M., Lecharny A., Scarpelli C., Artiguenave F., Pè M. E., Valle G., Morgante M., Caboche M., Adam-Blondon A.-F., Weissenbach J., Quétier F., Wincker P. and French-Italian Public Consortium for Grapevine Genome Characterization for Grapevine Genome Characterization 2007 The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**: 463-467

Li S.W., Feng L. and Niu D.K., 2007 Selection for the miniaturization of highly expressed genes. *Biochem Biophys Res Commun.* **360**: 586-592

Ouyang S., Zhu W., Hamilton J., Lin, Campbell M., Childs K., Thibaud-Nissen F., Malek R. L., Lee Y., Zheng L. Orvis J., Haas B., Wortman J. and Buell C. R. 2007 The TIGR Rice Genome Annotation Resource: improvements and new features. *Nucleic Acids Res.* **35**: D883-D887.

Park J., Xu K., Park T., and Yi S. V. 2012 What are the determinants of gene expression levels and breadths in the human genome? *Hum.Mol.Genet.* **21**: 46-56.

Pozzoli, U., Menozzi G.,. Comi G. P, Cagliani R., Bresolin N. and Sironi M. 2007 Intron size in mammals: complexity comes to terms with economy. *Trends Genet.* **23**: 20-24.

Rao Y. S., Wang Z. F., Chai X. W., Wu G. Z., Zhou M. Nie Q. H. and Zhang X. Q. 2010 Selection for the compactness of highly expressed genes in Gallus gallus. *Biol.Direct.* **5**: 35.

Ren, X. Y., Vorst O., Fiers M. W., Stiekema W. J. and J. P. Nap 2006 In plants, highly expressed genes are the least compact. *Trends Genet.* **22**: 528-532.

Schmutz J., Cannon S. B., Schlueter J., Ma J., Mitros T. Nelson W., Hyten D.L., Song Q., Thelen J.J., Cheng J., Xu D., Hellsten U., May G.D., Yu Y., Sakurai T., Umezawa T., Bhattacharyya M.K., Sandhu D., Valliyodan B., Lindquist E., Peto M., Grant D., Shu S., Goodstein D., Barry K., Futrell-Griggs M., Abernathy B., Du J., Tian Z., Zhu L., Gill N., Joshi T., Libault M., Sethuraman A., Zhang X.C., Shinozaki K., Nguyen H.T., Wing R.A., Cregan P., Specht J., Grimwood J., Rokhsar D., Stacey G., Shoemaker R.C, Jackson S.A. 2010 Genome sequence of the palaeopolyploid soybean. *Nature* **463**: 178-183.

Schnable P. S., Ware D., Fulton R. S., Stein J. C., Wei F., Pasternak S., Liang C., Zhang J., Fulton L., Graves T.A., Minx P., Reily A.D., Courtney L., Kruchowski S.S., Tomlinson C., Strong C., Delehaunty K., Fronick C., Courtney B., Rock S.M., Belter E., Du F., Kim K., Abbott R.M., Cotton M., Levy A., Marchetto P., Ochoa K., Jackson S.M., Gillam B., Chen W., Yan L., Higginbotham J., Cardenas M., Waligorski J., Applebaum E., Phelps L., Falcone J., Kanchi K., Thane T., Scimone A., Thane N., Henke J., Wang T., Ruppert J., Shah N., Rotter K., Hodges J., Ingenthron E., Cordes M., Kohlberg S., Sgro J., Delgado B., Mead K., Chinwalla A., Leonard S., Crouse K., Collura K., Kudrna D., Currie J., He R., Angelova A., Rajasekar S., Mueller T., Lomeli R., Scara G., Ko A., Delaney K., Wissotski M., Lopez G., Campos D., Braidotti M., Ashley E., Golser W., Kim H., Lee S., Lin J., Dujmic Z., Kim W., Talag J., Zuccolo A., Fan C., Sebastian A., Kramer M., Spiegel L., Nascimento L., Zutavern T., Miller B., Ambroise C., Muller S., Spooner W., Narechania A., Ren L., Wei S., Kumari S., Faga B., Levy M.J., McMahan L., Van Buren P., Vaughn M.W., Ying K., Yeh C.T., Emrich S.J., Jia Y., Kalyanaraman A, Hsia AP, Barbazuk WB, Baucom R.S., Brutnell T.P., Carpita N.C., Chaparro C., Chia J.M., Deragon J.M., Estill J.C., Fu Y, Jeddeloh J.A., Han Y, Lee H, Li P, Lisch D.R., Liu S, Liu Z, Nagel DH, McCann MC, SanMiguel P, Myers A.M., Nettleton D., Nguyen J., Penning B.W., Ponnala L., Schneider K.L., Schwartz DC, Sharma A, Soderlund C., Springer N.M., Sun Q., Wang H., Waterman M., Westerman R., Wolfgruber T.K., Yang L., Yu Y., Zhang L., Zhou S., Zhu Q., Bennetzen J.L., Dawe R.K., Jiang J., Jiang N., Presting G.G., Wessler S.R., Aluru S., Martienssen R.A., Clifton S.W., McCombie W.R., Wing R.A.,

Wilson R.K. . 2009 The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**: 1112-1115.

Seoighe C., Gehring C., and Hurst L. D., 2005 Gametophytic selection in Arabidopsis thaliana supports the selective model of intron length reduction. *PLoS Genet.* **1**: e13.

Sironi M., Menozzi G., Comi G. P., Cereda M., Cagliani R., Bresolin N. and Pozzoli U., 2006 Gene function and expression level influence the insertion/fixation dynamics of distinct transposon families in mammalian introns. *Genome Biol.***7:** R120

Swarbreck D., Wilks C., Lamesch P., Berardini T. Z., Garcia-Hernandez M. *et al.* 2008 The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.* **36**: D1009-D1014.

Tuskan, G. A., Difazio S., Jansson S., Bohlmann J., Grigoriev I., Hellsten U., Putnam N., Ralph S., Rombauts S., Salamov A., Schein J., Sterck L., Aerts A., Bhalerao R. R., Bhalerao R. P., Blaudez, D., Boerjan W., Brun A., Brunner A., Busov V., Campbell M., Carlson J., Chalot M., Chapman J., Chen G.-L., Cooper, D., Coutinho P. M., Couturier J., Covert S., Cronk Q., Cunningham R., Davis J., Degroeve S., Déjardin A., Depamphilis C., Detter J., Dirks B., Dubchak I., ,Duplessis S., Ehlting J., Ellis B., Gendler K., Goodstein D., Gribskov M., Grimwood J., Groover A., Gunter L., Hamberger B., Heinze B., Helariutta Y., Henrissat B., Holligan D., Holt R., Huang W., Islam-Faridi N., Jones S., Jones-Rhoades M., Jorgensen R., Joshi C., Kangasjärvi J., Karlsson J., Kelleher C., Kirkpatrick R., Kirst M., Kohler A., Kalluri U., Larimer F., Leebens-Mack J., Leplé J.-C., Locascio P., Lou Y., Lucas S., Martin F., Montanini B., Napoli C., Nelson D. R., Nelson C., Nieminen K., Nilsson O., Pereda V., Peter G., Philippe R., Pilate G., Poliakov A., Razumovskaya J., Richardson P., Rinaldi C., Ritland K., Rouzé P., Ryaboy D., Schmutz J. Schrader, J., Segerman B., Shin H. Siddiqui A., Sterky F., Terry A., Tsai C.-J., Uberbacher E., Unneberg P., Vahala J., Wall K., Wessler S., Yang G., Yin T., Douglas C., Marra M., Sandberg G., de Peer Y. V. and Rokhsar D. 2006 The genome of black cottonwood, Populus trichocarpa (Torr. and Gray). *Science* **313**: 1596-1604.

Urrutia A. O., and Hurst L. D., 2003 The signature of selection mediated by expression on human genes. *Genome Res.* **13**: 2260-2264.

Vinogradov A. E., 2004 Compactness of human housekeeping genes: selection for economy or genomic design? *Trends Genet.* **20**: 248-253.

Wendel J. F., Cronn R. C., Alvarez I., Liu B., Small R. L. and Senchina D. S. 2002 Intron size and genome size in plants. *Mol Biol Evol*. **19**: 2346-2352

Yang H., 2009 In plants, expression breadth and expression level distinctly and non-linearly correlate with gene structure. *Biol.Direct.* **4**: 45.

Chapter 2

Title: Comparative analysis of intron size variation in *Arabidopsis thaliana* and *Vitis vinifera*.

Abstract

Several comparative analyses have converged on indicating that the exon-intron architectures is highly conserved in plant genes. However large variations in the size of single genic regions have been documented in several taxa and mark a sharp contrast with such a view. Even more puzzling is the observation that such a variation is not always congruent with findings reported for other genic or non genic regions within the same genome. In the present paper we present a thorough investigation of intron size variation in two eudicots: *Vitis vinifera* and *Arabidopsis thaliana.* Data obtained from comparisons of orthologous genes with fully conserved intron positions clearly indicate that grape introns are ubiquitously longer that the thale cress orthologous. The size difference, seems not to be explained by a higher frequency of transposon element insertions or microsatellite presence. Estimation of the balances between short insertions deletions rates in neutrally evolving genomic regions indicated *Arabidopsis thaliana* but not *Vitis vinifera* genome is subjected to a high pressure for size miniaturization.

Introduction

Comparative analyses of fully annotated genomic sequences have shed light on the degree of evolutionary conservation of genic structural features. Most of the analyses of exon-intron organization have converged to a high level of conservation between closely related genes. Roy and Penny (2007) have demonstrated that as much as 94.7% of intron positions in regions of conserved coding alignments are conserved between *Arabidopsis thaliana* and *Oryza sativa* orthologous genes. The non conserved introns represented more frequently cases of intron loss than of intron gains. A similar picture was reported by Fawcett *et al*. (2012) in a comparison of exon intron structure between *Arabidopsis lyrata* and *Arabidopsis thaliana* and by Lin *et al*. (2006) in paralogous Rice genes. However, in spite of the high level of conservation of genic architectures, large size variations of single structural regions have been reported. Wendel and coworkers (2002) reported that total intron size could vary six fold in a wide evolutionary range of plant species.

Variation in genome size among organisms is thought as being associated to congruent changes across different classes of non coding DNA i.e. introns and intergenic DNA. However observation on differential intensity or direction of evolutive forces acting on non coding DNA have been also documented. For example the forces determining genome size in Gossypium spp would have only a limited impact on introns (Wendel *et al*. 2002). On the other hand studies that have combined intron data from multiple species indicated that beside the total intron length also the distributions of the average size of introns within genes could be subjected to variations among species. For example Bradnam and Korf (2008) demonstrated that in *O. sativa* genes the first intron is on average longer then other introns while the opposite behaviour was observed in maize. This observation suggests that not only total intron size but also the size of single introns within a gene may be subjected to different selective constrains and therefore that the forces acting on introns may have a different outcome depending on the order of the intron within the gene.

The two main factors that can influence the length of introns are transposable elements (TE) and microsatellites sequences. TE elements have been studied in many species and are thought to be the leading factors influencing the length of non coding DNA and genome size in some species (Li *et al.,* 2004). Their distribution between different classes of non coding DNA exhibit striking deviations from expectations based on

neutrality with patterns that can be substantially different even between species. Wrigth *et al.* (2003) showed a strong under representation of TE elements in Arabidopsis introns suggesting that the ratio between coding and non coding DNA of a transcript is under a strong purifying selection. Observation carried out in other species seem to figure out just the opposite scenario. Jaillon *et al*. (2003) have reported that in grape genes TE are accumulated to higher extent in introns than in other non coding genomic sequences.

In maize, Baucom *et al.* (2009) reported that different classes of TE define specific niches within the genome with several non LTR occupying preferentially intronic sequences. Also SSR have been credited for a non random distribution across different genomic regions. Morgante *et al*. (2007) have demonstrated that in 5 plant species representing a 50 fold range of genome size, SSR accumulate preferentially in transcribed regions. The distribution within genic non coding sequence seems to be biased in favor of an higher presence within introns. Studies conducted in several species suggest that in general the forces acting on intron and intergenic length are not independent of those influencing abundance and length of microsatellite.

Although positive associations between TE distribution and or SSR dynamic to intron length have been documented, it is unlikely that these factors alone may explain the variation of the average intron size between species or between genes at different genomic regions. Theoretical and simulations studies have indicated that the recombination between strongly selected loci is expected to enhance the efficiency of natural selection. This findings lead to the suggestion the accumulation of neutrally evolving DNA in gene rich regions should be positively selected. A similar reasoning could be adapted to mutational forces which should be biased in favor of insertion over deletion in low recombination regions of the genome. Confirmations of the theoretical expectations have been not always straightforward depending on several additional factors including the mating system, the meiotic behavior of species, the breeding history etc. Wrigth *et al* have reported that recombination does not correlate with TE abundance in *A. thaliana* (Wrigth *et al.* 2003). Moreover the substantial depletion of TE element within introns indicate that selection for expression efficiency may have a more important role than interfering selection in Arabidopsis genes. Other studies have highlighted the difficulties of measuring mutational bias in different category of non coding DNA. The analysis of indel polymorphism in pseudogenes have been used as estimate of insertion versus deletion bias based on the

assumption that these sequence are free of selective constrains and therefore the observed indel patterns are a faithful representation of the indel mutation process (Ophir and Graur 1997). However the estimate carried out on pseudogenes although accurate do not take into account the effect of transcription on mutation rate and or mutation repair. On the other hand direct measurement of mutational bias within introns are hampered by the presence of strongly selected elements such as those involved in splicing or in gene expression regulation. In the present work we carried out a detailed analysis of possible causes for the average length difference of introns form *A. thaliana* and *V. vinifera*.

Materials and Methods

Sequence data and annotations.

Arabidopsis thaliana genic annotations was downloaded from TAIR (Swarbreck 2008), Vitis annotation and sequences were obtained from Phytozome (Goodstein 2012). The list of orthologous proteins were identified using the InParanoid software (http://inparanoid.sbc.su.se/cgi/-bin/index.cgi). Only genes showing one to one orthologous were considered.

Identification of SC orthologous dataset.

We defined SC-orthologous (StructurallyConserved-orthologous) as those Arabidopsis-Vitis gene pairs which were i) univocally identified as orthologous according to the criteria explained above and ii) presented the same intron phases and intron exon junctions at corresponding positions of the pairwise protein alignments. All protein alignments were performed using the Muscle software (Edgar R.C. 2004) with standard settings. Mapping of introns in protein sequences and phase determination was obtained with an in-house developed perl-based pipeline (Figure 1). In brief, the position of exon-intron-exon junction was mapped in the three aligned protein sequences. An intron position was identified as conserved in orthologous genes if the exon-exon junction showed the same phase of a corresponding codon in the three orthologous coding sequence. (Figure S1 Supplemental Materials)

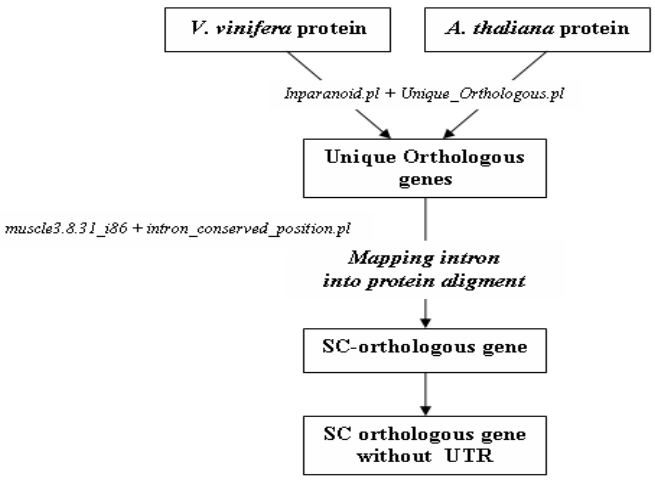


Figure 1: Identification of SC-Orthologous pipeline

Repetitive element identification

Repetitive elements in *Vitis vinifera* introns were identified using the Repeat masker software (Smit and Hubley 2008-2010) with standard settings. A manually curated dataset of rtepetitive elements, kindly provided by Prof Morgante (Institute of Applied Genomics) was used as query library. Classification of unannotated repetitive sequence was carried out using RECON software (Bao and Eddy 2002). The sequences of elements belonging to the two families with the highest number of annotated LINE (Long INterspersed Elements) sequences were aligned using the clustal software to produce a consensus sequence.

Pseudogene identification

We screened the annotation file of *V. vinifera*, which reports the gene model predictions of all grape genes, to identify all intron containing genes. The cDNA sequences of these genes were then used as query against the whole grape genome using the FASTA 3.4 with a ktup of 6 (Benovoy and.Drouin 2006) The produced alignment were filtered using a C script that identified sequences with a length higher than 70% of the query and having a stretch of at least 6 nucleotides (A/T) at not more than 80 bases from the 3' end of the putative processed pseudogenes.

Insertion/deletion analysis in pseudogenes

The assessment of indel polarity was carried out according to the procedure described by Ophir and Graur (1997). In brief, we first analyzed the alignments of pseudogenes with functional paralogous genes and functional orthologous. Whenever a gap in the alignment appeared in both the paralogous and orthologous sequences an insertion was inferred in the pseudgene. Similarly if at a given position, a gap was inferred in the pseudogene but not in the paralogous or orthologous sequences, an insertion was inferred in the pseudogene. Gaps at the end of pseudogene (truncations) were considered as due to abortive transcript as reported by Ophir and Graur (1997). (see Figure S2 Supplemental Materials). The indel bias was defined as the ratio between the number of insertions to deletions per site accumulated during evolutions between sequences.

The degree of divergence (or evolutionary distance) between a processed pseudogene and its functional paralogous was used as an estimate of the age of the

pseudogene. The evolutionary distance was calculated by using Kimura's two parameters model (Kimura 1980) as well as the numbers of transition and transversions per site between two sequences.

Results

Introns are the main contributors to the length difference between grape and Arabidopsis genes

On average, grape has four times longer genes than *Arabidopsis thaliana* (Figure 2). Such a difference is prevalently due to a higher frequency of long genes in the grape genome. While essentially no Arabidopsis gene is longer than 6 kb, as much a 32.54% of grape genes exceeds this size.

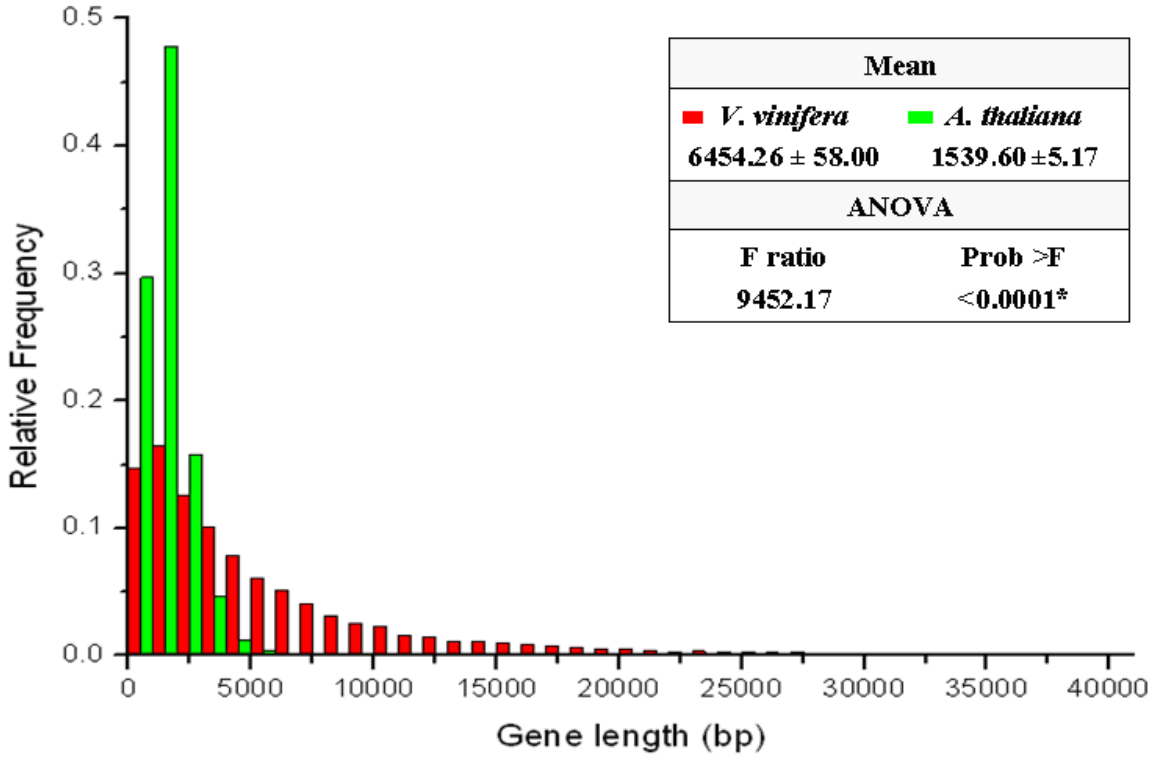


Figure 2: Gene length distributions of *V.vinifera* (red) and *A.thaliana* (green) gene length

To analyze how the gene length difference is distributed across regions, separate comparisons were carried out for untranslated sequences (both 5'and 3'UTRs), exons and introns. Most of the gene length difference was due to introns (see Table 1 a). In facts, only 2.25 % of the average gene length difference was attributable to total exons length opposed to a noticeable 97.75% due to introns (see Table 1a).

	Mean		ANOVA			Correlation	
	V.vinifera	*A.thaliana*	**F Ratio**	**Prob > F**	**R²**	**P value**	**Slope**
All genes							**(a)**
CDS length	1138.13 ±7.83	1233.76 ± 4.88	116.65	<0.0001*			
5'UTR length	265.11 ± 2.64	152.14 ± 0.85	2517.01	<0.0001*			
3'UTR length	326.50 ± 2.33	236.60 ± 0.95	1722.19	<0.0001*			
Number of exons	6.15 ± 0.03	5.24 ± 0.03	479.93	<0.0001*			
Total exons lengh	1482.37 ± 8.31	1563.09 ± 5.42	72.28	<0.0001*			
Average exon length	317.30 ± 2.82	608.40 ± 4.31	2474.31	<0.0000*			
Number of introns	5.55 ± 0.04	5.82 ± 0.03	34.62	<0.0001*			
Total introns lengh	5361.26 ± 58.22	964.09 ± 5.50	6938.75	<0.0001*			
Average intron length	1005.52 ± 9.57	201.90 ± 2.22	8038.99	<0.0001*			
Unique_orthologous							**(b)**
Gene length	8766.47 ± 111.07	1779.20 ±9.94	3926.08	<0.0001*	0.49	<0.0001	0.10
CDS length	1404.65 ± 9.36	1445.97 ± 9.69	9.40	<0.002*	0.97	<0.0001	1.01
5'UTR length	219.35 ± 2.82	140.68 ± 1.22	767.21	<0.0001*	0.40	<0.0001	0.37
3'UTR length	326.77 ± 2.76	228.27 ± 1.28	1163.55	<0.0001*	0.55	<0.0001	0.47
Number of exons	7.97 ± 0.58	7.12 ± 0.06	111.40	<0.0001*	0.95	0.0000*	0.90
Total exons length	1803.34 ± 9.98	1779.20 ± 9.94	2.093	0.08	0.95	<0.000*	0.96
Average exon length	306.06 ± 2.40	428.06 ± 4.52	568.24	<0.0001*	0.65	0.0000*	1.30
Number of introns	7.20 ± 0.06	6.77 ± 0.06	26.17	<0.0001*	0.95	<0.0001	0.90
Total introns lenght	7191.03 ± 109.31	1093.98. ± 9.50	2882.03	<0.0001*	0.56	<0.0001	0.08
SC_orthologous							**(c)**
Gene length	7279.18 ± 248.62	1413.50 ± 19.21	553.30	<0.0001*	0.47	<0.0001*	0.10
CDS length	1057.82 ± 18.10	1094.10± 18.30	1.98	0.16	0.98	0.0000*	1.02
5'UTR length	169.73 ± 6.04	117.19 ± 2.83	64.98	<0.0001*	0.18	<0.0001*	0.43
3'UTR length	308.65 ± 8.04	219.63 ± 2.84	116.76	<0.0001*	0.14	<0.0001*	0.35
Number of exons	6.67 ± 0.13	6.67 ± 0.13	0.00	1.00	1.00	<0.0000*	1
Total exons length	1478.76 ± 20.57	1413.55 ± 19.21	5.37	0.02	0.89	<0.0000*	0.93
Average exon length	278.61 ± 4.76	266.70 ± 4.53	3.27	0.70	0.88	<0.0000*	0.92
Number of introns	5.67 ± 0.13	5.67 ± 0.13	0.00	1.00	1.00	0.0000*	1
Total introns lenght	6229.90 ± 151.96	929.53 ± 19.62	410.07	<0.0001*	0.54	0.0000*	0.08

Table 1: Statistical analysis on the lengths of the different genic regions (coding and non-coding) between *V. vinifera* and *A. thaliana*, in (**a**) all genes, (**b**) 10913 unique orthologous, and (**c**) 1184 SC ortholougos genes

Unexpectedly, the total number of introns was significantly higher in *Arabidopsis thaliana* than in grape but this difference alone was too low to explain the average intron length difference between these two species. Also the length of Arabidopsis coding sequences exceeded that of grape but the difference was compensated by the two fold longer grape untranslated regions. (Table 1b, Figure 3)

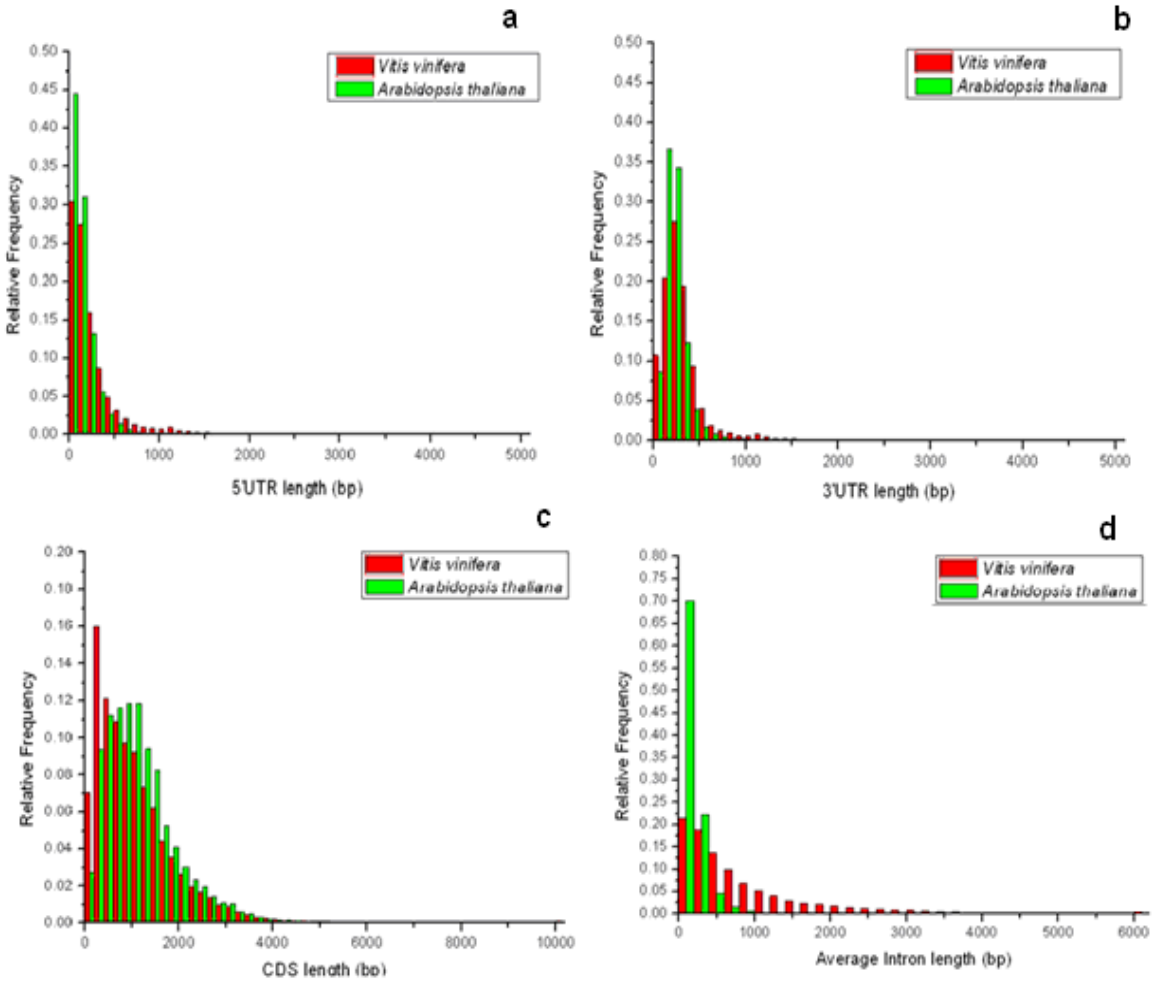


Figure 3: length distributions of 5'UTR (**a**), 3'UTR (**b**), CDS (**c**), average intron length (d), of *V. vinifera* (red) and *A. thaliana* (green) genes

Velasco and coworkes (2007) have reported that for about 16859 grape transripts was not possible to find an Arabidopsis orthologous. To investigate whether the reported intron length differences were due to genes specific for either species, univocally defined pairs of Arabidopsis/Vitis orthologous genes were compared. Interestingly, also these pairwise comparisons confirmed what observed for the whole dataset (see Table 3). The decomposition of length differences into single orthologous genic regions confirmed the pattern observed for the whole dataset for 5' and 3' UTRs, total intron length and coding sequences but not for the total number of introns which showed the inverse relation being, now, more numerous in grape than in Arabidopsis (Table 3). Moreover, it was very interesting to note that the total length of the coding sequences and the total number of introns were highly correlated in the two species suggesting a high level conservation of gene models architectures (see Table 1b).

Prompted by these observations we decided to focus our attention on genic sequences from the translation starting codon to the termination codon. Introns were mapped in protein sequences and orthologous gene pairs with conserved intron positions and phase were selected (Figure S1 Supplemental Material).

After filtering the gene pairs with introns in the UTR the final dataset contained 1184 orthologous gene pairs with a conserved exon-intron organization (SC-orthologous). Length comparisons of SC orthologous confirmed the picture observed for the whole orthologous set with an obvious exception for the intron number difference which was null by construction.

Interestingly a linear model could explain as much as 98% of the relation between Vitis and Arabidopsis coding sequence length of SC-orthologous (see Table 1c). By contrast the linear model describing the relation between total intron length of SC-orthologous gene pairs explained not more than half of the total variation. (Table 1c)

Korf *et al* (2008) have shown that the first intron of Arabidopsis genes is, on average, the longest. A functional explanation of a such intron size distribution has been offered by Rose (2002) who identified sequence repeats within proximal introns conferring an extraordinary transcriptional and translational enhancer activity. No evidence of similar patterns of intron sizes were identified within grape genes. Nevertheless we analyzed the comparisons of introns of SC-orthologous genes to identify a relation between intron order in genes and intron length difference. To this end SC orthologous with 10 introns were analyzed (250 SC-orthologous genes). Only introns of orders 4, 6 and 7 fitted a linear model but in no cases the relation could explain more than half of the total variation (Table 2). A similar finding was confirmed for SC orthologous with 4 and 6 introns (355 and 297 SC-orthologous genes respectively) (Table 2). This finding suggested that the main factor responsible the orthologous intron length divergenge between Arabidopsis and Vitis has a little systematic nature

	Intron length								
	Class with 10 introns			Class with 6 introns			Class with 4 introns		
	Correlation			Correlation			Correlation		
	Slope	R^2	Prob	Slope	R^2	Prob	Slope	R^2	Prob
N1	0.05	0.22	0.3090	0.06	0.24	<0.0001*	0.06	0.22	<.0001*
N2	0.018	0.14	0.2050	0.04	0.21	<0.0001*	0.07	0.20	<.0001*
N3	0.02	0.15	0.9955	0.03	0.17	<0.0001*	0.05	0.16	<.0001*
N4	0.03	0.35	0.0019*	0.03	0.13	<0.0001*	0.04	0.10	0.2042
N5	0.13	0.48	0.0021*	0.04	0.20	<0.0001*			
N6	0.02	0.13	0.7136	0.02	0.13	<0.0001*			
N7	0.05	0.26	0.3179						
N8	0.12	0.34	0.1441						
N9	0.01	0.24	<0.0001*						
N10	0.01	0.08	0.7491						
Total intron length	0.10	0.74	0.1301	0.07	0.40	0.1038	0.09	0.35	0.0620

Table 2: Intron length of SC-orthologous genes genes between *V. vinifera* and *A. thaliana* within 5'UTR and 3'UTR, in class with 10, 6 and 4 introns.

Jaillon *et al.* (2007) have reported that on average 12.7% of intron sequences are represented by repetitive elements. As a matter of fact Jiang and Goertzen (2011) demonstrated that transposable elements insertion in introns may explain the exceptional size of 39 grape genes. Based on these considerations we analyzed the contribution of repetitive elements to total intron length of grape genes. Introns of the SC orthologous gene set were homology searched against a manually curated library of grape repetitive elements. Globally the repetitive elements covered 42.44% of the total intron sequences analyzed. The most abundant elements were type I retrotransposon followed by LINE and type II transposons. Simple repeats occupied 0.53% of intron sequences (Table S1 Supplemental Materials).

The linear relation between total intron length of pairs SC orthologous improved its fit from 55% to 75% (data not shown) after the elimination of repetitive sequence from grape introns and the average intron size difference decreased from about six-fold to three fold (see Figure 4 for the SC-orthologous genes with ten introns).

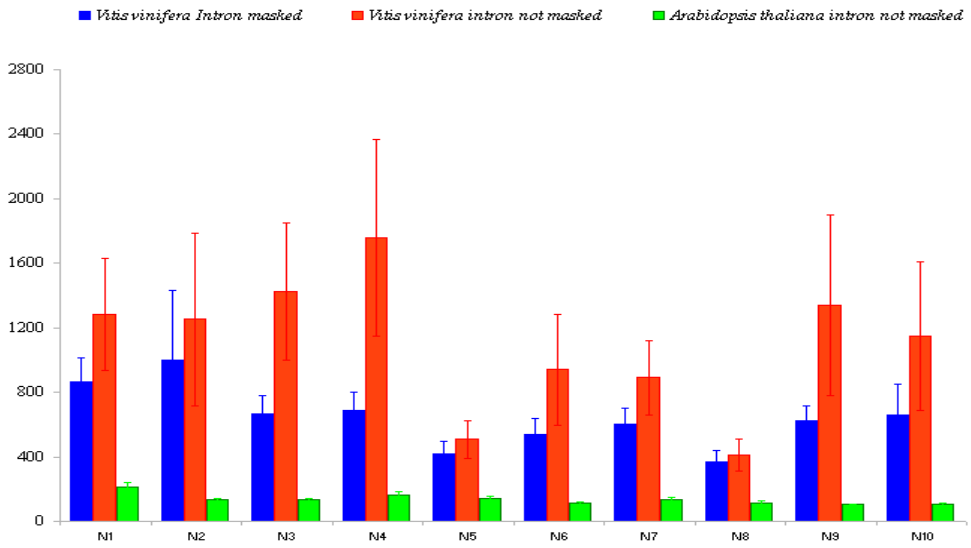


Figure 4: Intron length variation between genes of *A. thaliana* and *V. vinifera* (masked and not masked introns). SC- orthologous with 10 introns

Mutational bias

The data presented above leave open several possible explanations all tuned on the direction and intensity of mutational patterns. In the more contrasted picture an opposite direction of mutational forces could be hypothesized: i.e Vitis genome would be under positive selection for higher genome size while Arabidopsis genome could be under a strong selection for genome size miniaturizations. Of course also difference in intensities of mutational patterns pointing in the same direction could account for the observed picture (Vitis is under a less strong selection than Arabidopsis for size reduction of viceversa for size increase). Fawcett *et al.* (2012) have inferred an higher rate of intron loss in *A. thaliana* then in the close relative *A. lyrata*. These data were proposed as evidences of selection for gene miniaturization. Accordingly Hu *et al.* (2011) have analyzed the pattern of deletions and insertions still segregating in *A. thaliana* suggesting a pervasive selection for a smaller genome in this species. Based on this data we speculated that Vitis genome could be under a less severe selection for size reduction and this could account for the higher size of introns.

The average mutational bias in the *V. vinifera* genome was estimated by studying insertion and deletion (indel) rates in processed pseudogenes. Indel polarity was inferred by studying sequence alignments between Vitis functional genes and Vitis processed pseudogenes together with orthologous sequence from other species (Ophir and Graur 1997). An insertion/deletion was inferred as occurred in the pseudogene after the

retroposition event if the analyzed marker state was not found neither in the paralogous nor in the orthologous sequence. *A. thaliana, O. sativa* and *P. trichocarpa* were used as source of orthologous sequences.

Interestingly the average indel bias calculated considering the mutations in the pseudogenes was more in form of deletion then insertion (Table 4 and Figure 5).

	Functional Orthologous genes from		
	A.thaliana	***O.sativa***	***P.trichocarpa***
Average retropseudogene_length	918.85 ± 34.50	861.57 ± 31.93	832.19 ± 28.08
Average number of insertion	1.74 ± 0.11	1.14 ± 0.10	1.52 ± 0.11
Average number of insertion length	10.08 ± 0.86	8.55 ± 1.26	9.48 ± 1.03
Average number of deletion	3.8 ± 0.21	2.86 ± 0.19	3.54 ± 0.20
Average deletion length	58.02 ± 5.7	42.72 ± 4.89	54.19 ± 6.26
Average number of pirimidin transition	66.77 ± 2.83	61.13 ± 2.67	58.84 ± 3.12
Average number of purine trensition	89.68 ± 4.31	80.58 ± 4.03	75.71 ± 4.74
Average number transvertions	104.26 ± 5.37	93.59 ± 5.09	84.88 ± 5.64
Average estimated age	0.26 ± 0.01	0.25 ± 0.01	0.24 ± 0.01

Table 4: Retropseudogenes genes of *V.vinifera*. This data was calculated with their functional paralogue and functional orthologous in *A. thaliana, O. sativa* and *P. trichocarpa*

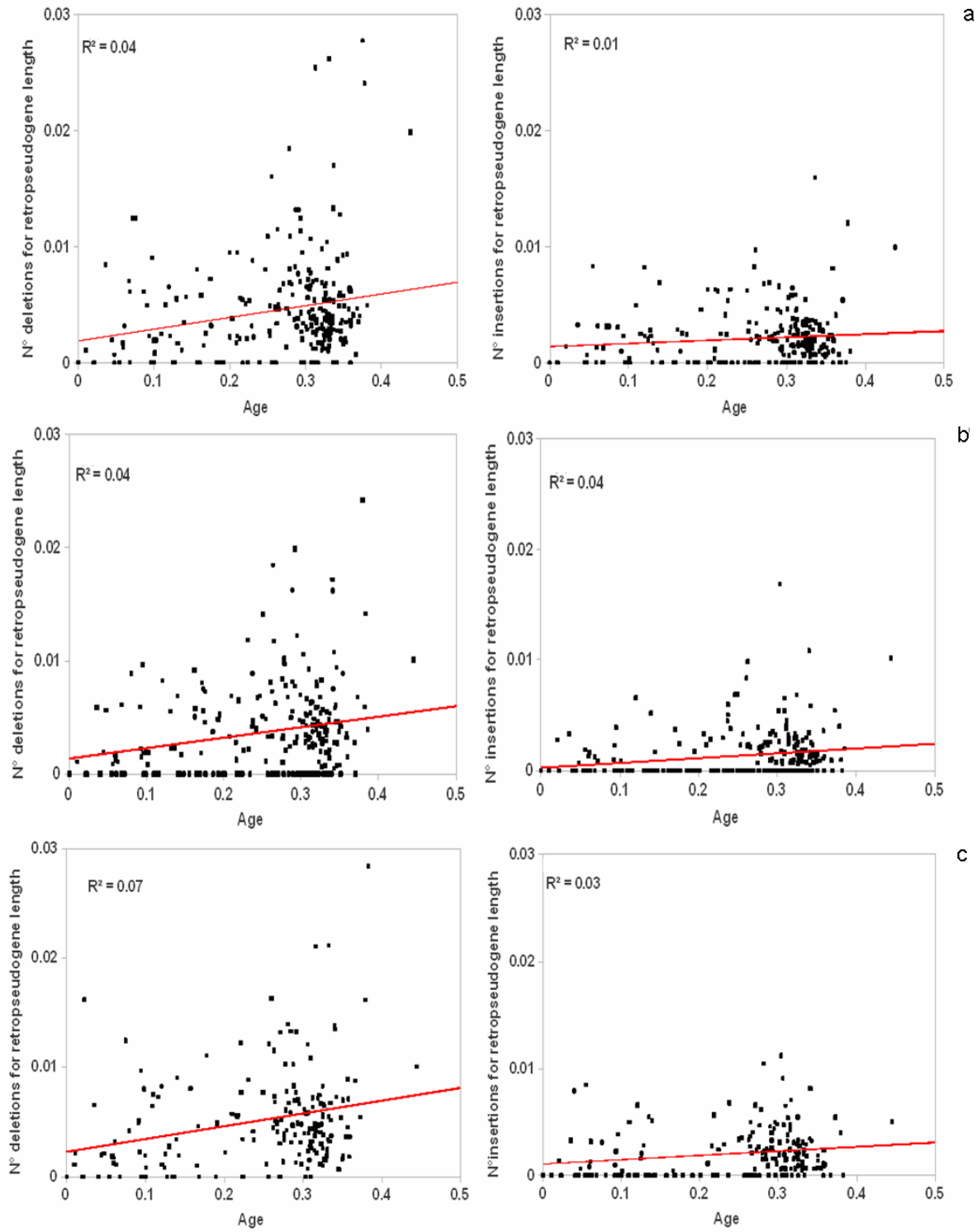


Figure 5: Number of deletions and insertions per site, calculated with orthologous genes *A.thaliana* (a), *O.sativa* (b) and *P.trichocarpa* (c).

We calculated also the age of pseudogenes by considering the proportion of nucleotides substitutions between the pseudogene and the functional paralogous (Table 4). If the bias were a characteristic of grape genome evolution then we should expect that

older pseudogenes should have a lower indel bias (i.e more deletions than insertions). The relation between the inferred age of pseudogenes and indel bias was approximated by a linear model (Figure 6). The identified trend showed a mild negative slope but its linear fit to data was not significant ($p>0.05$).

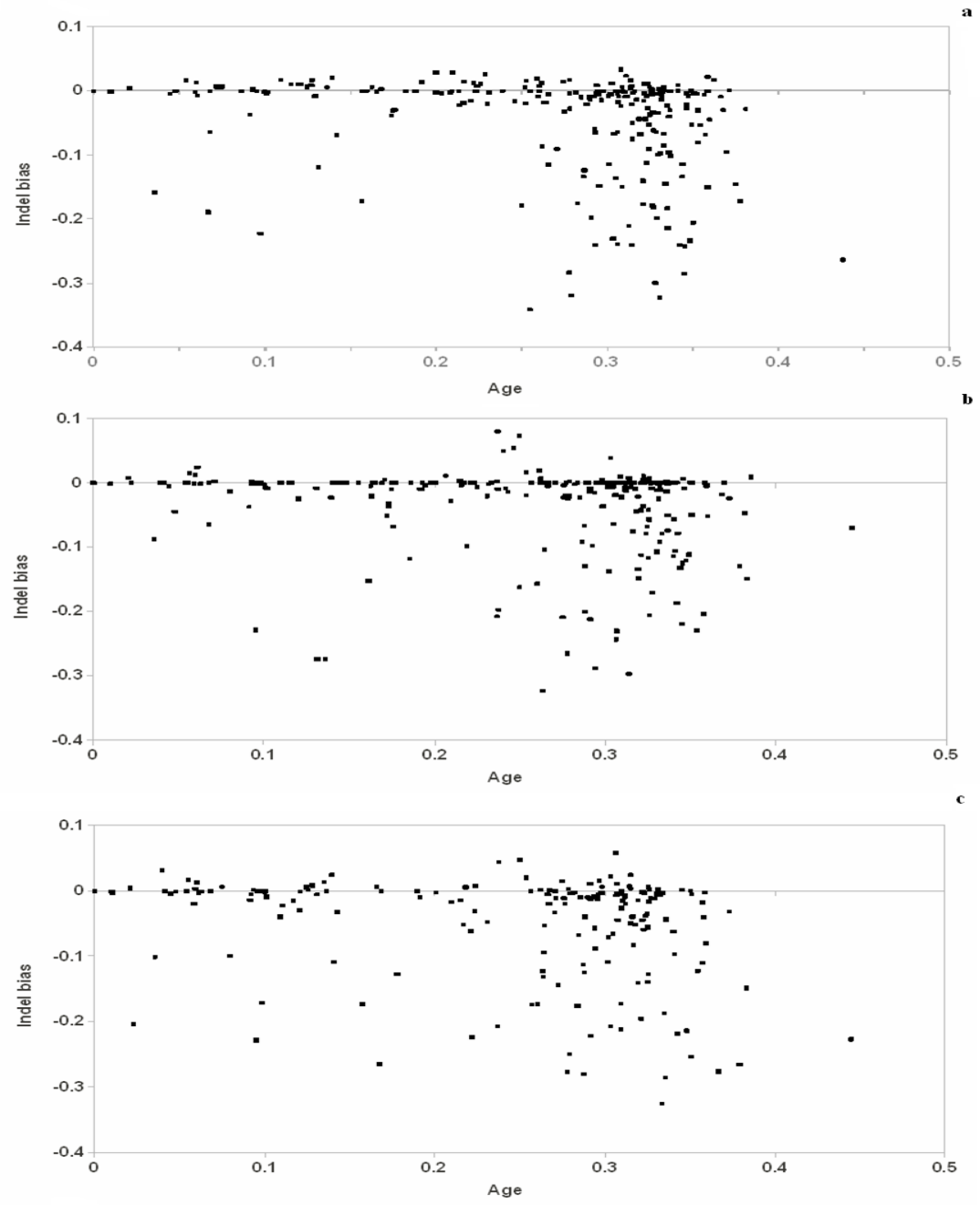


Figure 6: Indel bias as function of retropseudogenes age, calculated with orthologous genes of *A. thaliana*(a), *O.sativa* (b) and *P.trichocarpa* (c).

Next we analyzed the indel bias of retroelement. For these analyses we focused on two families of LINE because most of these elements behave as pseudogenes (Vinogradov 2004). However, we could not find a significant relation between the divergence of LINE sequence which is an estimate of the time passed from the insertion in the intron and the indel bias (Figure S3 and S4). Altogether these data suggest that the mutational bias in Vitis genome is not under selection for size increase.

Discussion

Introns of *Vitis vinifera* genes are exceptionally long in comparison to corresponding genic regions of other eudicots species. The total intron length of grape genes is on average six fold longer than the average total intron length of *Arabidopsis thaliana* genes in spite a genome size difference between these two species of about three times (Jaillon *et al.,* 2007). Because the dynamic of intron number evolution may have proceeded with a different pace in these two species since their divergence from the common ancestor, we analyzed the genetic structure of orthologous genes with completely conserved intron-exon structures. Even in these comparisons grape introns were confirmed longer than Arabidopsis orthologous. Two are the main factors that can influence intron size: the insertion of transposon elements and the presence of microstallite. Jaillon *et al*. have demonstrated that grape introns show higher content of repetitive sequence than Rice (Jaillon et al., 2007). Accordingly, Jiang and Goertzen (2011) showed that retro-transposon insertion are the cause for the exceptional length of introns of 39 grape genes.

To investigate whether the insertion of repetitive elements could account for the exceptional size of grape introns we repeated our comparisons after filtering out repetitive sequence from grape introns. The length difference was reduced of one third. It was very interesting to note that the exclusion of repetitive sequence increased the percentage of length size variance explained by a linear model. This finding pointed to a systematic factor as possible cause for the length variation.

One possibility is that the balance between insertion and deletions in introns may differ between species. To verify such a hypothesis we first sought for pseudogene in grape genome. Pseudogenes are defined as non functional genomic sequence with significant sequence similarity to functional RNA or protein coding genes. Sequence comparison between the functional paralogous and the pseudogene allows to identify regions with small insertion and/or deletions. However because the functional gene and the pseudogene co-evolved, it is often impossible to determine whether the difference in the sequence alignment reflects the substitution in pseudogenes or in the functional genes. The directionality (or polarity) of an indel event is determined by using functional orthologous as a reference (Ophir and Graur 1997). Using this approach we could draw a relation between indel bias and time from the formation of the pseudogene. Analysis carried out on

Vitis pseudogene did not find a clear trend of indel bias in either direction. On contrast the indel bias in Arabidopsis, was reported as significantly biased toward deletion (Fowecett *at al.;* Hu *et al.* 2011)

This finding negate the first of our hypotheses on putative difference in indel bias: i.e Vitis is under selection pressure for genome size increase while the mutational bias in Arabidopsis genome is working in the opposite direction. Unfortunately, in absence of quantitative estimation of indel bias in Arabidopsis pseudogene we cannot define a detailed picture for alternative hypothesis dealing with difference in mutational biases intensities (but not direction) in the two genomes. However we are tempted to speculate that small deletions are more favored (over insertion) in Arabidopsis than in Vitis genome. However caution must be paid in this context as the indel bias estimated by analyses of pseudogenes could be a poor predictor of the mutational processes acting on introns. An immediate difference is that pseudogenes, unlike introns, are seldom transcribed and therefore the mutagenic effect of transcription is disregarded. A possibility to circumvent this problem is represented by the analysis of retro-elements present in introns (Vinogradov 2002). Jaillon *et al.* (2007) have reported that intron sequences are particularly rich of LINE elements.

We therefore decided to analyze the mutational bias of LINE sequences residing in grape genes. In this case, the polarity of indels was determined through a comparison of insertion and deletions with a consensus LINE sequence. Even in this case we could not find a clear indication on the trend of indel bias with the age of the LINE. Again caution must be paid as we cannot ignore the possible influences of selection on the evolution of LINE inserted in grape introns. However, it is worth mentioning that the analysis of indel bias in LINE inserted in intergenic sequence produced similar results. In conclusion, we have demonstrated that two different factors are responsible for the size difference between Vitis and Arabidopsis introns. In one hand the insertion of repetitive elements, may account for one third of the size difference, while most of the remaining difference could be explained by an higher rate of size reduction in Arabidopsis. It was interesting to note that beside the insertion of repetitive element we did not find clear indication of indel rates biased toward insertion in Vitis. This finding suggest that the high frequency of repetitive elements in Vitis introns should not be interpretetd as part of a general response to selection for a larger genome size. In our opinion combined analysis of data on density of repetitive sequences in introns and pattern of gene expression could reveal insightful clues

in this matter.

Acknowledgments

We are indebited with Prof Michele Morgante and Dr. Simone Scalabrin for kindly providing us with the grape repetitive element library.

We would like to acknowledge R. Ophir and professor Shiu for helpful suggestions on pseudogene analysis

Literature Cited

Bao Z., Eddy S.R. 2002 Automated de novo identification of repeat sequence families i n sequenced genomes. *Genome Res*. **12**: 1269-1276

Baucom R. S., Estill J. C., Chaparro C., Upshaw N. Jogi, A., Deragon J.-M., Westerman R. P., Sanmiguel P. J. and Bennetzen J. L. 2009 Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. *PLoS Genet*. **5**: e1000732

Benovoy D. and Drouin G. 2006 Processed pseudogenes, processed genes, and spontaneous mutations in the Arabidopsis genome. *J Mol Evol*. **62:** 511-522
Bradnam K. R. and Korf I. 2008 Longer first introns are a general property of eukaryotic gene structure. *PLoS One*. **3**: e3093

EdgarR. C. 2004 MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res***. 32:** 1792-1797

Goodstein D. M., Shu S., Howson R., Neupane R., Hayes R. D., Fazo J., Mitros T., Dirks W., Hellsten U., Putnam N. and Rokhsar D. S. 2012 Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res*. **40**: D1178-D1186

Hu T. T., Pattyn P., Bakker E. G., Cao J., Cheng J.-F., Clark R. M., Fahlgren N., Fawcett, J. A., Grimwood J., Gundlach H., Haberer G., Hollister J. D., Ossowski S., Ottilar R. P., Salamov A. A., Schneeberger K., Spannagl M., Wang X., Yang L., Nasrallah M. E., Bergelson J., Carrington J. C., Gaut B. S., Schmutz J., Mayer K. F. X., de Peer Y. V., Grigoriev I. V., Nordborg M., Weigel D. and Guo Y.-L. 2011 The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nat Genet* **43:** 476-481

Huala E., Dickerman A. W., Garcia-Hernandez M., Weems D., Reiser L., LaFond F., Hanley D., Kiphart D., Zhuang M., Huang W., Mueller L. A., Bhattacharyya D., Bhaya D., Sobral B. W., Beavis W., Meinke D. W., Town C. D., Somerville C. and Rhee S. Y. 2001 The Arabidopsis Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic Acids Res*. **29**: 102-105

Jaillon O., Aury J.-M., Noel B., Policriti A., Clepet C., Casagrande A., Choisne N., Aubourg S., Vitulo N., Jubin C., Vezzi A., Legeai F., Hugueney P., Dasilva C., Horner D., Mica E., Jublot D., Poulain J., Bruyère C., Billault A., Segurens B., Gouyvenoux M., Ugarte E., Cattonaro F., Anthouard V., Vico V., Fabbro C. D., Alaux M., Gaspero G. D., Dumas V., Felice N., Paillard S., Juman I., Moroldo M., Scalabrin S., Canaguier A., Clainche I. L., Malacrida G., Durand, E., Pesole G., Laucou V., Chatelet P., Merdinoglu D., Delledonne M., Pezzotti M., Lecharny A., Scarpelli C., Artiguenave F., Pè M. E., Valle G., Morgante M., Caboche M., Adam-Blondon A.-F., Weissenbach J., Quétier F., Wincker P. and French-Italian Public Consortium for Grapevine Genome Characterization for Grapevine Genome Characterization 2007 The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**: 463-467

Jiang K. and Goertzen L. R. 2011 Spliceosomal intron size expansion in domesticated grapevine (Vitis vinifera). *BMC Res Notes* **4**: 52

Kimura M. 1980 A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol***. 16:** 111-120

Li Y.-C., Korol A. B., Fahima T. and Nevo E. 2004 Microsatellites within genes: structure, function, and evolution. *Mol Biol Evol*. **21:** 991-1007

Morgante M., Paoli E. D. and Radovic S. 2007 Transposable elements and the plant pan-genomes. *Curr Opin Plant Biol*. **10**: 149-155

O'Brien K. P., Remm M. and Sonnhammer E. L. L. 2005 Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res* **33**: D476-D480

Ophir R. and Graur D. 1977 Patterns and rates of indel evolution in processed pseudogenes from humans and murids. *Gene*. **205**: 191-202

Rhee S. Y., Beavis W., Berardini T. Z., Chen G., Dixon D., Doyle A., Garcia-Hernandez M.,Huala E., Lander G., Montoya M., Miller N., Mueller L. A., Mundodi S., Reiser L., Tacklind J., Weems, D. C., Wu, Y., Xu I., Yoo D., Yoon J. and Zhang P. 2003 The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Res*. **31**: 224-228

Roy S.W. and Penny D. 2007 Patterns of intron loss and gain in plants: intronn loss-dominated evolution and genome-wide comparison of O.sativa and A.thaliana. *Mol. Biol. Evol.* **24:**171-181

Rose A. B. 2002 Requirements for intron-mediated enhancement of gene expression in Arabidopsis. RNA*, Molecular and Cellular Biology*. **8**: 1444-1453

Smit A.F.A., Hubley,R. RepeatModeler Open-1.0. 2008-2010 <http://www.repeatmasker.org>.

Swarbreck D., Wilks C., Lamesch P., Berardini T. Z., Garcia-Hernandez M., Foerster H., Li D., Meyer T., Muller R., Ploetz L., Radenbaugh A., Singh S., Swing V., Tissier C., Zhang P. and Huala E. 2008 The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.* **36**: D1009-D1014

Velasco R., Zharkikh, A., Troggio, M., Cartwright D. A., Cestaro A., Pruss D., Pindo M., Fitzgerald L. M., Vezzulli S., Reid J., Malacarne G., Iliev D., Coppola G. Wardell B., Micheletti D., Macalma T., Facci M., Mitchell J. T., Perazzolli M., Eldredge G., Gatto P., Oyzerski R., Moretto M., Gutin N., Stefanini M., Chen Y., Segala C., Davenport C., Demattè L. Mraz A., Battilana J., Stormo K., Costa F. Tao Q. Si-Ammour A., Harkins T., Lackey A., Perbost C., Taillon B., Stella A., Solovyev V., Fawcett J. A., Sterck L., Vandepoele K., Grando S. M., Toppo S., Moser C., Lanchbury J., Bogden R., Skolnick M., Sgaramella V., Bhatnagar S. K., Fontana P., Gutin A., de Peer Y. V., Salamini F. and Viola R. 2007 A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS One*. **2**: e1326

Wendel J. F., Cronn R. C., Alvarez I., Liu B., Small R. L. and Senchina D. S.2002 Intron size and genome size in plants. *Mol Biol Evol*. **19**: 2346-2352

Wicker T., Sabot F., Hua-Van A., Bennetzen J. L., Capy P., Chalhoub B., Flavell A., Leroy P.,Morgante M., Panaud O., Paux E., SanMiguel P. and Schulman A. H. 2007 A unified classification system for eukaryotic transposable elements. *Nat Rev Genet*. **8**: 973-982

Vinogradov A.E. 2002 Growth and decline of introns. Trends in Genetics. **18:** 232:236

Wright S. I., Agrawal N. and Bureau T. E. 2003 Effects of recombination rate and gene density on transposable element distributions in Arabidopsis thaliana. *Genome Res*. **13**: 1897-1903

Supplemental Materials

Figure S1: Multiple protein alignment. The introns with common position and phase into protein alignments are evidenced. Introns than don't interrupt the codon (**phase0**) are marked in red, introns are positioned between first and second bases of codon (**phase1**) are marked in blue, and intron collocated between second and third base of codon **(phase2**) are marked in green.

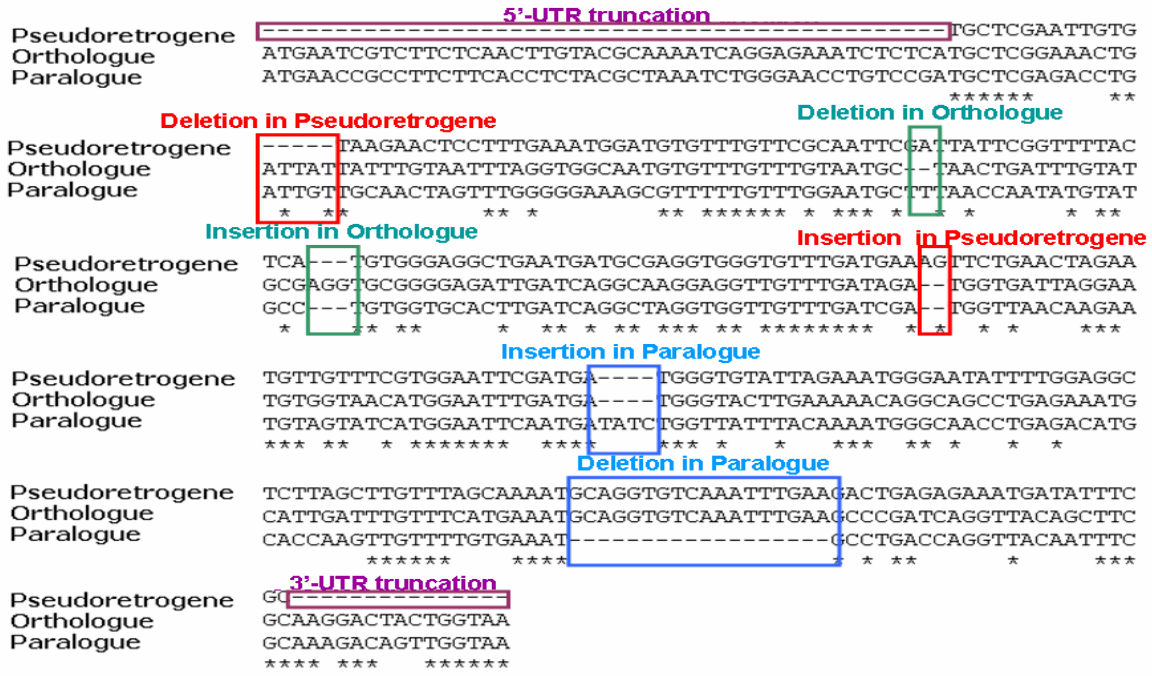
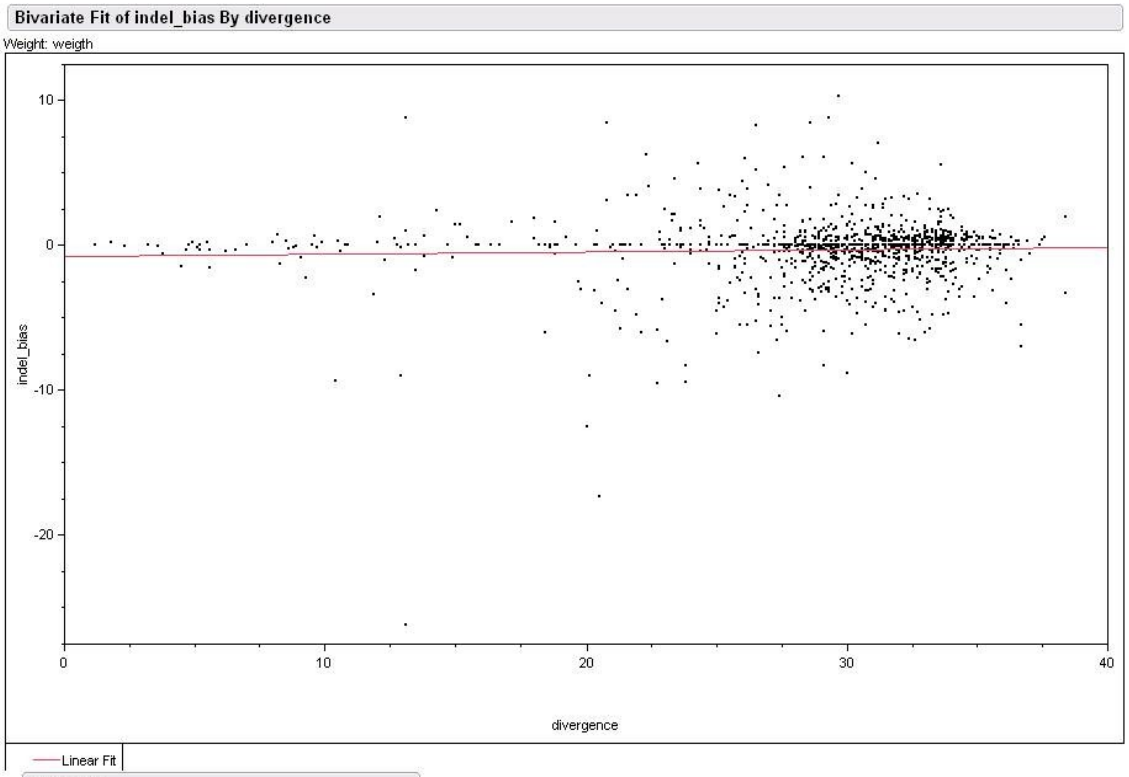


Figure S2: Identification Rate indel in Pseudogene



Linear Fit

indel_bias = -0.718497 + 0.0137131*divergence

Summary of Fit

RSquare	0.003933
RSquare Adj	0.003083
Root Mean Square Error	41.75763
Mean of Response	-0.32592
Observations (or Sum Wgts)	911992

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	227	624106.7	2749.37	1.8303
Pure Error	945	1419509.0	1502.13	Prob > F
Total Error	1172	2043615.7		<.0001*
				Max RSq
				0.3081

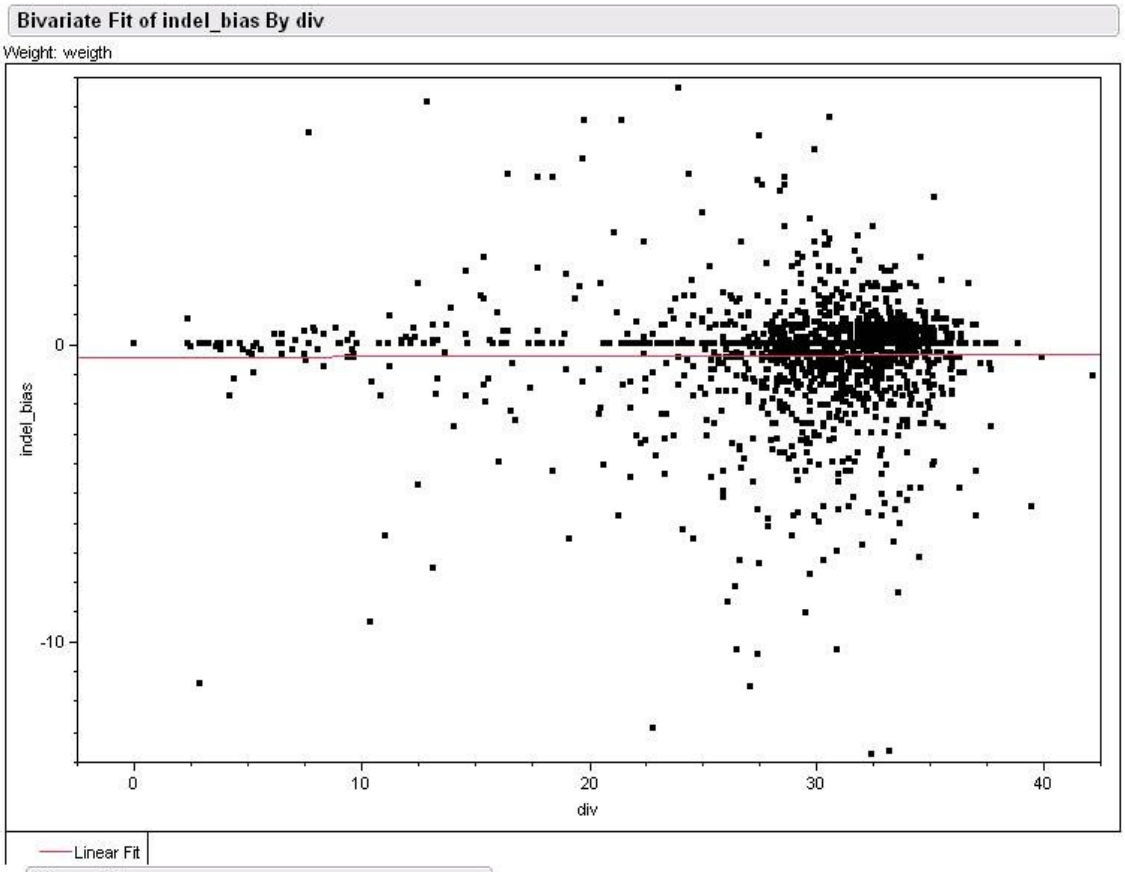
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	8068.6	8068.61	4.6273
Error	1172	2043615.7	1743.70	Prob > F
C. Total	1173	2051684.3		0.0317*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.718497	0.187664	-3.83	0.0001*
divergence	0.0137131	0.006375	2.15	0.0317*

Figure S3: relation between LINE divergence and indel bias. Only LINE mapping in introns were considered for this analysis.



Linear Fit

indel_bias = -0.421045 + 0.0022849*div

Summary of Fit

RSquare	0.000164
RSquare Adj	-0.00046
Root Mean Square Error	39.97555
Mean of Response	-0.35661
Observations (or Sum Wgts)	1370593

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	423.1	423.11	0.2648
Error	1615	2580841.6	1598.04	Prob > F
C. Total	1616	2581264.7		0.6069

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.421045	0.129788	-3.24	0.0012*
div	0.0022849	0.004441	0.51	0.6069

Figure S4: relation between LINE divergence and indel bias. Only LINE mapping in intergenic sequences were considered for this analysis.

	Introns of *V.vinifera*	Introns of SC-orthologous genes of *V.vinifera*
TOT Bases masked (%)	32.26	29.90
LINE (%)	13.18	16.99
LTR (%)	14.54	10.41
DNA elements (%)	2.21	1.23
Simple repeats (%)	0.59	0.53
Low complexity (%)	2.28	1.40
Unclassified (%)	0.06	0.04

Table S1: Percentage of repetitive elements in Vitis introns masked with RepeatMask software

Chapter 3

Title: The dynamic of intron-exon structure during angiosperm evolution

Abstract

The ever increasing availability of fully sequenced genomes offers new opportunity for analyzing fundamental questions in evolutionary biology such as the dynamic of intron loss and gains in eukaryotes. In this work, we present detailed analyses of exon-intron structure in genes of plant species covering the evolutionary range from green algae to modern eudicots. As a whole the picture confirmed the tenet that losses of intron outnumbered the gains. However we identified several examples suggesting that the general view may be punctuated by relevant exceptions. First, the divergence of angiosperm was associated to a significant high rate of intron gains. The dynamic of intron appearance was subsequently resettled to an higher rate of losses as repeatedly confirmed by analyses involving both monocots and eudicots. However cases of increased rates of losses versus gains were identified for *Populus trichocarpa* and *Arabidopsis thaliana* genes. Altogether these evidences suggest that exon-intron losses is an highly dynamic trait which has probably played a major role during angiosperm evolution.

Introduction

The evolutionary history of gene structure and the selective forces that shape intron-exon evolution, are poorly understood questions in evolutionary biology. Most of the attempts made to explain the variability in intron densities across species in terms of rates of intron loss and gain have been synthesized in two alternative theories. (Jeffares *et al*. 2006; Roy 2006; Roy and Gilbert 2006). The intron early (IE) theory predicts that introns appeared at early stage of life evolution and then disappeared from prokaryotes while being retained in eukaryotes. By contrast, according to the introns-late (IL) model the phylogenetic restriction of spliceosomal introns to eukaryotes would reflect their more recent insertion into originally intron less genes after the divergence of eukaryotes and prokaryotes. The sharp contrast between these scenarios is exacerbated by the observations that the massive variation in intron number among eukaryotic species shows no simple phylogenetic patterns with intron-rich and intron-poor species interspersed in the eukaryotic phylogenetic tree. Such a observation implies recurrent episodes of intron loss and/or gain (Roy and Gilbert, 2006).

Whether a specific intron is conserved or lost during evolution may depend on multiple factors whose relevance may vary with the genomic context. Introns may contain regulatory sequences, entire genes or pieces of genes and in some cases can be required for alternative splicing, the processing and export of the mRNA and translational efficiency (Le Hir *et al.*2003; Nott *et al*. 2004; Cenik *et al*. 2011). The elimination of such a type of introns is likely to be selected against, whereas the gain of an intron that happen to improve the function of a gene is likely to be fixed. Recent genome wide analyses have showed that the position of some spliceosomal intron is conserved among species evolutionary as distant as plant and animals (Fedorov *et al.* 2002; Roy and Gilbert 2006). However, examples suggesting that this apparently universal pattern of intron-exon structure conservation may be contrasted by the action of evolutionary forces such as those pointing to a reduction of the length of primary transcript or to an enhancement of recombination rate in gene rich genomic regions have also been documented. An example in case was recently reported for two Arabidopsis species which differ significantly for genome size, probably due to different rates of genome reduction driven by selection. *Arabidopsis thaliana* has lost six times more introns than *Arabidopsis lyrata* since the

divergence of these two species, and gained very few introns (Fawcett *et al.* 2011). The higher rate of intron loss in *A. thaliana* is in line with the smaller genome size of this species as compared to *A. lyrata* (Fawcett *et al.* 2012).

Plants present very interesting features for this type of studies as closely related lineages present a great variability of life history traits that may have an impact on structural features of genes including number and size of introns. To cite an example, the reproductive systems ranging from fully vegetative to fully sexual (and within the latter from completely autogamous to allogamous) can be expected to interfere with the number and size of introns. Kreitman and Comeron (1999) have proposed that intron sequences experiencing a limited selective contrains may enhance recombination between the more selctively contrained exons. Under this view two species having similar genomic characteristics but a different reproductive systems should have a different size or even number of introns. Despite this consideration the pattern of intron loss and gain in plants have remained relatively unexplored. Recent study has confirmed early observations conducted in plant genes and showing a lower rate of intron gain compared to intron loss. However most studies have focused on comparisons of few species and no information about the dynamic of intron evolution across the evolutionary tree of plant species is to date available. In the present study we present a genome wide analysis of intron evolution in 6 dicot species, 3 monocots, one moss and two algae. The choice of the analyzed species was carried out based on a phylogenetic criterion such that the loss versus gain of unique introns could be inferred by the analysis of orthologous introns in evolutionary more ancient species serving as outgroup.

Materials and Methods

Data sets

Whole genome sequences, protein sequences and the annotation of protein coding genes were downloadable from Phytozome database (Goodstein *et al.* 2012). The quality of gene model predictions was tested by analyzing the protein deduced from cds sequences that were reconstructed based on the coordinates reported in the gff3 file. Genes with cds with internal stop codons or encoding for proteins different from those reported in the protein data file downloaded from phytozome were not considered in further analyses.

Orthologous genes identifications.

The orthologous sequences were identified as those having the best reciprocal Blastp hit with a score lower than 10^{-5} and whose alignment covered at least half of both protein sequences.

Intron position mapping and classification

Each pair of putatively orthologous protein sequences was aligned using Muscle (Edgar R.C. 2004) with default parameters and intron positions were mapped in the resulting alignments reporting for each intron the codon phase.

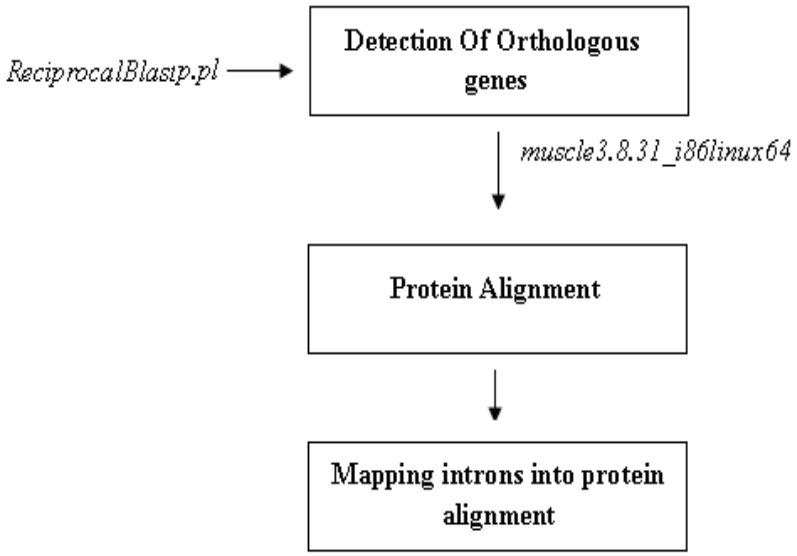


Figure 1: Intron position mapping and classification pipeline

The analyses of mapped introns were carried out taking into account several filters. For each intron we determined whether the 15 aligned aminoacids in both directions (not containing gapped positions) showed at least 50% aminoacid identity (Filter 1).
Species-specific introns that fell opposite to a gap of 6 or more aminoacid in the other species were not considered (Filter 2). Moreover neighboring intron positions within five alignments positions were not considerd (Filter3). Intron positions within five positions from alignments borders were not retained (Filter 4). Introns with non canonical splice sites were not considered (Filter 5) (For a schematic representation of filters see Figure S1, Supplemental Material)

Introns which showed the same position and phase in both aligned proteins and that passed the five filters described above were considered “conserved”.

Accordingly, introns that failed to pass all the five filters were classified as “filtered”. Introns which passed filters but which were present only in one of the two aligned sequence were considered as “unique”. (Figure S2 , Supplemental Material)

Finally unique introns were considered as lost or gained based on the analysis of ortholgous genes in species that served as outgroup for each specific comparisons. (Figure S3, Supplemental Material)

Results and discussion

To study the pattern of intron evolution, pairwise alignments of orthologous proteins from six dicots, three monocots, one moss and two algae were analyzed. The list of analyzed species is reported in Figure 1.

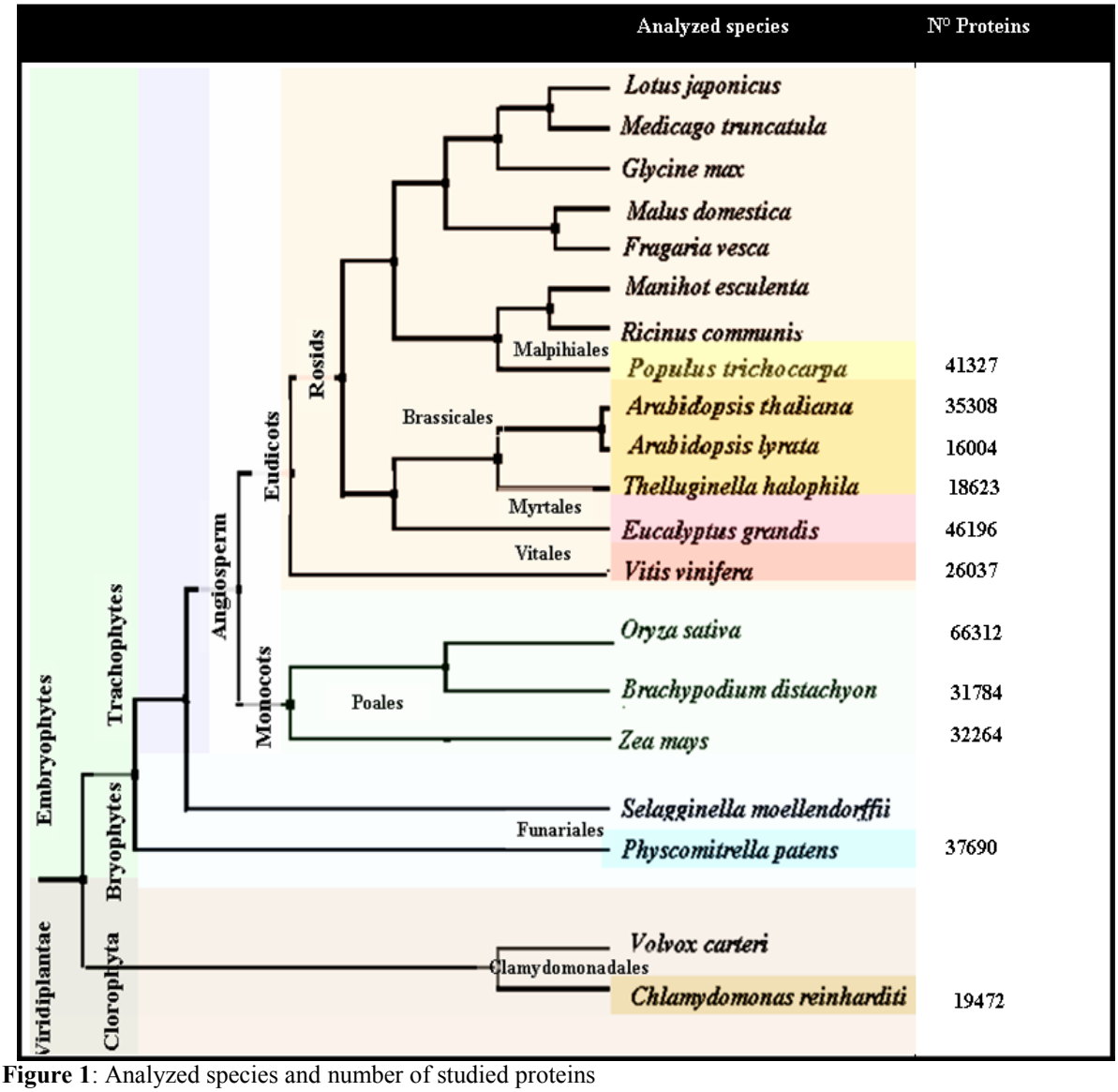


Figure 1: Analyzed species and number of studied proteins

Introns were classified as "conserved" when they mapped in the same position and with the same codon phase in aligned proteins and as "unique" when present only in one of the aligned proteins (see Figure S2 Supplemental Material and Material and Method for

details). Introns which mapped in poorly aligned regions or that showed anomalous splicing sites were classified as filtered (see Figure S1).

On average about 96.88% of the total number of mapped introns were classified as conserved in pairwise comparisons (see Table 1). The highest level of intron position conservation was found for the comparisons between *Arabidopsis thaliana* and *Arabidopsis lyrata* (99.87%) and the lowest for the comparison between *Populus trichocarpa* and *Zea mays* (60.59%).

The average number of unique introns per pairwise comparison was 2.67% and with a maximum of 11.82 % (in comparisons of *Z. mays* to *P. patens*) and a minimum of 0.13 % between *A. thaliana* and *A. lyrata*. In general, there was a strong inverse correlation between the percentage of conserved and unique introns per pairwise comparison.

					Conserved					**(a)**
	PP	**OS**	**BD**	**ZM**	**VV**	**EG**	**PT**	**AT**	**AL**	**TH**
PP		89.07%	89.18%	88.18%	89.03%	89.11%	88.86%	88.85%	88.53%	88.87%
OS	95.27%		99.30%	99.10%	98.37%	98.16%	98.35%	97.97%	98.00%	98.22%
BD	95.20%	98.83%		98.81%	95.73%	98.03%	98.11%	97.81%	97.85%	97.93%
ZM	94.67%	98.55%	98.78%		98.02%	97.88%	60.59%	97.85%	97.54%	98.14%
VV	96.91%	99.64%	99.63%	99.59%		99.73%	99.84%	99.64%	99.60%	99.62%
EG	95.94%	99.53%	99.51%	99.62%	99.74%		99.82%	99.56%	99.53%	97.80%
PT	95.49%	99.27%	99.32%	99.20%	99.46%	99.38%		99.25%	99.08%	99.27%
AT	94.12%	97.53%	97.67%	97.68%	97.76%	97.58%	97.75%		98.85%	99.42%
AL	94.11%	97.82%	97.93%	98.06%	98.23%	97.04%	97.94%	99.87%		99.55%
TH	94.30%	97.98%	97.88%	97.92%	97.90%	99.56%	98.13%	99.77%	99.55%	
					Unique					**(b)**
	PP	**OS**	**BD**	**ZM**	**VV**	**EG**	**PT**	**AT**	**AL**	**TH**
PP		10.93%	10.82%	11.82%	10.96%	10.89%	11.14%	11.15%	11.47%	11.13%
OS	4.73%		0.70%	0.90%	1.63%	1.84%	1.65%	2.03%	2.00%	1.78%
BD	4.80%	1.17%		1.19%	1.78%	2.39%	1.89%	2.19%	2.15%	2.07%
ZM	5.33%	1.45%	1.22%		1.98%	2.12%	1.21%	2.15%	2.46%	1.86%
VV	3.09%	0.36%	0.36%	0.41%		0.27%	0.16%	0.31%	0.40%	0.38%
EG	4.06%	0.47%	0.49%	0.37%	0.26%		0.18%	0.44%	0.47%	2.20%
PT	4.51%	1.75%	0.68%	0.80%	0.54%	0.64%		0.75%	0.92%	0.73%
AT	5.87%	2.47%	2.33%	2.32%	2.24%	2.42%	2.01%		1.15%	0.58%
AL	5.89%	2.18%	2.07%	1.94%	1.77%	2.34%	2.06%	0.13%		0.45%
TH	5.70%	2.02%	2.09%	2.08%	2.10%	0.44%	1.87%	0.23%	0.45%	

Table 1: Number of introns classified as conserved (**a**) and unique (**b**) in parwise comparison: (**PP** = *P. patens*, **OS** = *O. sativa*, **BD** = *B. distachyion*, **ZM** = *Z. mays*, **VV** = *V. vinifera,* **EG** = *E. grandis,* **PT** = *P. trichocarpa*, **AT** = *A. thaliana,* **AL** = *A. lyrata* and **TH** = *T. halophila*)

Intron evolution dynamic at the time of land conquest by green plants

As expected, the species with the lowest average number of conserved introns in pairwise comparison was the moss *Physcomitrella patens* (see Table 1). The observation that the percentage of *P. patens* introns classified as conserved in comparisons with species as distant as monocots or dicots was rather invariable, suggested that intron loss/gain proceeded at a roughly constant rate in the various angiosperm lineages (see Table 1). Comparisons of gene structure of *P. patens* genes with orthologous from the monocots *Z. mays*, *O. sativa* and *B. distachyon* indicates that 331 introns were the cases of introns present in the monocots but absent in *P. patens* (Table 2a). On the contrary, only 135 were the cases of introns present in *P. patens* genes but absent in the three monocots considered (Table 2a).

The data obtained from the analyses of multiple comparisons involving *P. patens* and several trios of dicots species confirmed the picture (Table 2c and Table 2d).

Two possible scenarios may be envisaged from these data: i) the intron dynamic in *P. patens* was dominated by cases of losses or ii) the differentiation of angiosperms was concomitant with a burst of introns gains. To discriminate between these two hypotheses we analyzed alignments of *P. patens* proteins to orthologous proteins from two angiosperms and the algae *Clamydomonas reinhardti* which served as outgroup. Notably the ratio of intron positions gained versus lost by the two angiosperm was 153 to 17 respectively (see Table 2b). A similar picture was observed when *V. carteri* was used as outgroup instead of *Clamydomonas reinhardti* (see Table S1 Supplemental Material). Such a finding suggests that after the conquest of land by green plants but before the divergence of monocots, the common ancestor of angiosperms experienced a dramatic increase in intron gains.

Intron loss has dominated the evolution of the angiosperm genes

The pattern deduced from the analysis of intron conservation in pairwise comparisons between angiosperm orthologous proteins reflected the phylogenetic distance between species. For example, on average, more than 98% of mapped introns were classified as conserved in pairwise comparisons between orthologous proteins of gramineae species.

	(a)					(b)					(c)					(d)				
	PP	ZM	OS	BD	Nº introns	CR	PP	OS	AT	Nº introns	PP	VV	EG	PT	Nº introns	PP	AT	AL	TH	Nº introns
Conserved	+	+	+	+	2517	+	+	+	+	307	+	+	+	+	4959	+	+	+	+	7242
Lost	-	+	+	+	331	-	+	+	+	609	-	+	+	+	705	-	+	+	+	833
	+	-	+	+	4	+	-	+	+	47	+	-	+	+	4	+	-	+	+	19
	+	+	-	+	3	+	+	-	+	3	+	+	-	+	8	+	+	-	+	9
	+	+	+	-	3	+	+	+	-	11	+	+	+	-	33	+	+	+	-	12
Gain	+	-	-	-	135	+	-	-	-	902	+	-	-	-	163	+	-	-	-	416
	-	+	-	-	1	-	+	-	-	35	-	+	-	-	1	-	+	-	-	1
	-	-	+	-	0	-	-	+	-	8	-	-	+	-	4	-	-	+	-	1
	-	-	-	+	0	-	-	-	+	3	-	-	-	+	1	-	-	-	+	1
Others	+	-	-	+	3	+	-	-	+	2	+	-	-	+	2	+	-	-	+	15
	+	+	-	-	2	+	+	-	-	17	+	+	-	-	0	+	+	-	-	7
	+	-	+	-	0	+	-	+	-	3	+	-	+	-	0	+	-	+	-	1
		+	+	-	5		+	+	-	19		+	+	-	13		+	+	-	2
	-	+	-	+	1	-	+	-	+	10	-	+	-	+	5	-	+	-	+	0
	-	-	+	+	0	-	-	+	+	153	-	-	+	+	3	-	-	+	+	2
Analyzed Introns					30005					2122					8192					8554

Table2: Intron gains or losses between reciprocal orthologous. "+" indicates the presence of intron , and "-" absence of introns
(a) 917 reciprocal orthologous between *P. patens* (PP), *O. sativa* (OS), *Z. mays* (ZM) and *B. distachyon* (BD),
(b) 950 reciprocal orthologous between *C. reinharditi* (CR) *P. patens* (PP), *O. sativa* (OS) and *A. thaliana* (AT)
(c) 2261 reciprocal orthologous between *P. patens* (PP), *V. vinifera* (VV) *E. grandis* (EG) and *P. trichocarpac* (PT)
(d) 2173 reciprocal orthologous between *P. patens* (PP), *A. thaliana* (AT), *A. lyrata* (AL) and *T. halophila* (TH)

Gene structure evolution in Graminaceae

Intron dynamic in gramineae genes was dominated by intron losses. Multiple comparisons between *O. sativa*, *B. distachyon, Z. mays* and *P. patens* orthologous genes indicated that 81.2 % of intron positions were conserved (see Table 2a). In all three monocots species analyzed we found an higher tendency to loose introns then to generate new intron positions (see Table 2a and Table 3a). Table S2 reports the locus names and intron positions classified as gained in either of the monocots analyzed. Interestingly the ratio loss/gain was comparable in the three species suggesting that the differentiation of grasses species was not accompanied by dramatic changes in gene structures. Such a observation is unexpected if we consider that grass genomes have a high rate of chimeric gene origination by retroposition (Wang *et al.* 2006).

The choice of an outgroup that is highly distant from monocots may have produced results that do not reflect the evolution of chimeric genes that evolved more recently. However analyses of trios of monocots confirmed the ratio of intron losses versus gains (see Table 3a). Moreover closer inspection of data did not identify clear cases of adjacent unique introns which can be expected in cases of retroposition (see Table S2, Table S2).

Gene structure evolution in dicots

The availability of several sequenced genomes of dicot species allowed the analysis of gene structure at different evolutionary stages. Multiple comparisons involving orthologous genes from *Physcomitrella. patens*, and the trios of the dicots *Populus trichocarpa*, *Eucalyptus grandis* and *Vitis vinifera* shed light at very early stages of core eudicots gene structure evolution (see Table 2c).

The ratio between intron losses and gains was unbalanced in favour of losses as already observed for grasses genomes but in this case the losses were less abundant. An interesting exception to a such behaviour was represented by *P. trichocarpa* orthologous genes which showed 33 species specific losses and only one gain (see Table 2c). Table S4 reports the locus names and intron positions which were classified as gains in this analysis

c				(a)				(b)				(c)
	ZM	OS	BD	Nº introns	PT	VV	OS	Nº introns	PT	EG	OS	Nº introns
Conserved	+	+	+	5198	+	+	+	20907	+	+	+	16806
Lost	-	+	+	10	-	+	+	71	-	+	+	52
	+	-	+	8	+	-	+	13	+	-	+	23
	+	+	-	23	+	+	-	286	+	+	-	248
Gain	+	+	+	7	+	+	+	11	+	+	+	3
	-	+	-	2	-	+	-	10	-	+	-	23
	-	-	+	2	-	-	+	34	-	-	+	36
Analyzed introns				5242				21333				17192
				(d)				(e)				(f)
	PT	AT	AL	Nº introns	VV	AT	AL	Nº introns	TH	AT	AL	Nº introns
Conserved	+	+	+	8554	+	+	+	12392	+	+	+	14896
Lost	-	+	+	35	-	+	+	123	-	+	+	31
	+	-	+	20	+	-	+	14	+	-	+	22
	+	+	-	1	+	+	-	6	+	+	-	13
Gain	+	-	-	153	+	-	-	253	+	-	-	43
	-	+	-	0	-	+	-	0	-	+	-	11
	-	-	+	0	-	-	+	0	-	-	+	4
Analized introns				8765				12789				14896

Table 3: Intron gains or losses between trios of reciprocal orthologous. "+" indicates the presence of intron, and "-" absence of introns
(a) 917 reciprocal orthologous between *Z. mays* (ZM) *O. sativa* (OS) and *B. distachyon* (BD)
(b) 6083 reciprocal orthologous between *P. trichocarpa* (PT), *V. vinifera* (VV) and *O. sativa* (OS)
(c) 5261 reciprocal orthologous between *P. trichocarpa* (PT), *E. grandis* (EG) end *O. sativa* (OS)
(d) 3268 reciprocal orthologous between *P. trichocarpa* (PT), *A. thaliana* (AT) and *A. lyrata* (AL)
(e) 2128 reciprocal orthologous between *V. vinifera* (VV), *A. thaliana* (AT) and *A. lyrata* (AL)
(f) 2173 reciprocal orthologous between *T. halophila* (TH)), *A. thaliana* (AT) and *A. lyrata* (AL)

The same picture emerged from analysis of trios involving *Populus trichocarpa* and *Oryza sativa* as outgroup. As shown in Table 3b and 3c in *Populus trichocarpa* there was a higher abundance of cases of intron losses (71 to 11 and 52 to 3). Such a finding was not reproduced for *Vitis vinifera* and *Eucalyptus grandis* suggesting that the reduction of intron number and possibly of intron size may be a distinctive feature of Populus evolutive history (see Table 3b and 3c).

To investigate whether the high number of intron losses in Populus genes is common to other eurosids we analyzed the genes of *A. thaliana*, *A. lyrata* and *T. halophila* using *P. patens* as outgroup (see Table 9). These three dicots belong to the Brassicaceae family and are eurosids II. The dominance of intron losses over gains already reported for the trios of more ancient eudicots was confirmed also in these trios. On average intron losses outnumbered tenfold the gains (see Table 2d). Fawcett *et al.* (2011) have recently reported that the genome of *Arabidopsis thaliana* is subjected to a strong selection for size reduction. The reduction of intron numbers observed in *A. thaliana* genes when compared to *A. lyrata* orthologous is an evidence of the genomic response to such a selective pressure (see Table 3d and 3e). To confirm a such proposal we analyzed trios including the two *Arabidopsis* species and either *Vitis vinifera* or *Populus trichocarpa* as outgroups (Table 10 and Table 11). In both analyses we confirmed that intron losses in *A. thaliana* genes outnumbered the cases of intron losses in *A. lyrata* (see Table 3d and 3e). However it was interesting to note that a such tendency toward a decrease in intron number was attenuated in analyses involving *Thelluginella halophila* as outgroup (Table 3f).

Conclusions

In the present work, we analyzed the evolution of genic structures in twelve species spanning the evolutionary range from green algae to extant dicots. In agreement with previous findings (Roy 2006; Fedorov *et al.* 2002; Lin *et al.* 2006) our data confirm that during plant evolution the cases of intron losses outnumbered the gains. However, we found two cases indicating that the rate of intron losses can be subjected to significant variations in specific lineages. In particular we report that both in *Populus tricocharpa* and *Arabidopsis thaliana* genes the rate of intron loss was significantly higher then in closely related species. Importantly these conclusions were confirmed by analyses which used several species as outgroups. For example we compared Populus genes with orthologous from either the monocot *Oryza sativa* or the moss *Physcomitrella patens*. The evidence for the high rate of intron losses was confimed in both analyses suggesting that the force responsible for intron depletion is active also on very ancient introns. As said before a very similar case to Populus was recorded for the dicot *Arabidopsis thaliana*. In this latter case we could demonstrate that in the close relative *Arabidopsis lyrata* intron lossed occured at lower rate. A similar finding was recently reported by Fawcett and coworkers (2012) in a thorough analysis of intron dynamics in *A. thaliana* and *A. lyrata*. These authors suggested that *A. thaliana* genome is under strong selective pressure for size miniaturization which involves also a high rate of intron losses. In analogy with a such hypothesis we would propose that a similar tendency toward size miniaturization is also at work on Populus genome. Close inspection of data, ruled out the hypothesis that the increase of intron losses was due to high rate of retroposition events leaving room for hypotesizing that the mechanism of intron depletion acts with a genome wide breadth. Further analyses are needed to understand the dominant mechanism responsible for intron loss. Moreover comparative analyses of expression data are needed to understand whether the changes in gene models associated to intron depletions have an impact on gene expression or function. Another remarkable exception to the general view was the high rate of intron gais that occured in angiospems after their divergence from their common ancestor with moss.

These conclusions were drawn by data obtained by comparing *Physcomitrella patens* genes with the orthologous from *Arabidopsis thaliana* and *Oryza sativa* and using *C. reinhardti* and *V. carteri* as outgroups. Unfortunately due to the unavailability of

sequenced gymnosperm species we could not identify with more precision the point at which this rate increase occured.

In conclusion we propose that intron exon dynamic is a complex trait that played a major role during angiosperm evolution. Further analyses on the mechanisms of intron loss/ gain at work in different genomic context will reveal important information on the mechanism of gene architecture evolution. All this information will be of particular relevance for those interested in deciphering the evolutionary trajectories of the various plant lineages.

Acknowledgments:

We wish to thank Alexei Fedorov for encouragment and suggestions in exon intron structure analysis.

Literature Cited

Cenik C., Chua H.N., Zhang H. Tarnawsky S.P., Akef A., Derti A., Tasan M., Moore MJ, Palazzo A.F., Rooth F.P. 2011 Genome analysis reveals interplay between 5' UTR introns and nuclear mRNA export for secretory and mithocondrial genes. *Plos Genet.* **7:**e1001366

EdgarR. C. 2004 MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. **32:** 1792-1797

Fawcett J.A., Rouzé P. and Van de Peer Y. 2012 Higher Intron Loss Rate in *Arabidopsis thaliana* than *A. lyrata* is consistent with stronger selection for a smaller genome. *Mol Biol Evol* **29:**849-859

Fedorov A., Merican A.F. and Gilbert W 2002 large-scale comparison of intron position among animal, plant and fungal genes *National Acad Science* **99:** 16128–16133

Goodstein D. M., Shu S., Howson R., Neupane R., Hayes R. D., Fazo J., Mitros T., Dirks W., Hellsten U., Putnam N. And Rokhsar D. S.2012 Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res*. **40**: D1178-D1186

Jeffares D.C., Mourier T., Penny D., Vinter J. Gracia-Fernandez J. and Roy S.W. 2008 Origin of introns by intronization of exonic sequences. *Trends Genet* **24:**378-381.

Kreitman M. and Comeron J. M. 1999 Coding sequence evolution. *Curr Opin Genet Dev*. **9**: 637-641

Le Hir H., Nott A., Moore M.J., 2003 How introns influence and enhance eukaryiotic gene expression. *Trens Biochem Scie* **28:**215-220.

Lin H., Zhu W., Silva J. C., Gu X. and Buell C. R. 2006 Intron gain and loss in segmentally duplicated genes in rice. *Genome Biol,* **7**: R41

Nott A., Hir H. L. and Moore M. J. 2004 Splicing enhances translation in mammalian cells: an additional function of the exon junction complex. *Genes Dev*. **18**: 210-222

Rogozin I. B., Wolf Y. I., Sorokin A. V., Mirkin B. G. And Koonin E. V. 2003 Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr Biol*. 2003, **13**: 1512-1517

Roy S.W 2006 Intron-rich ancestor *Trends in Genetics* **22** (9):468-71

Roy S. W. and Gilbert W. 2006 The evolution of spliceosomal introns: patterns, puzzles and progress. *Nat Rev Genet*. **7:** 211-221

Roy S W., Penny D. 2007 Patterns of intron loss and gains in plants: Intron loss dominated evolution and genome wide comparison of O. sativa and A thaliana. *Mol. Biol. Evol* **24:** 171-181

Wang W., Zheng H., Fan C., Li J., Shi J., Cai Z., Zhang G., Liu D., Zhang J., Vang S., Lu Z., Wong G.K.S., Long M. And Wang J., 2006 High Rate of Chimeric Gene Originatio by Retroposition in Plant Genomes. *The Plant Cell* **18:** 1791-1802

Supplemental Materials

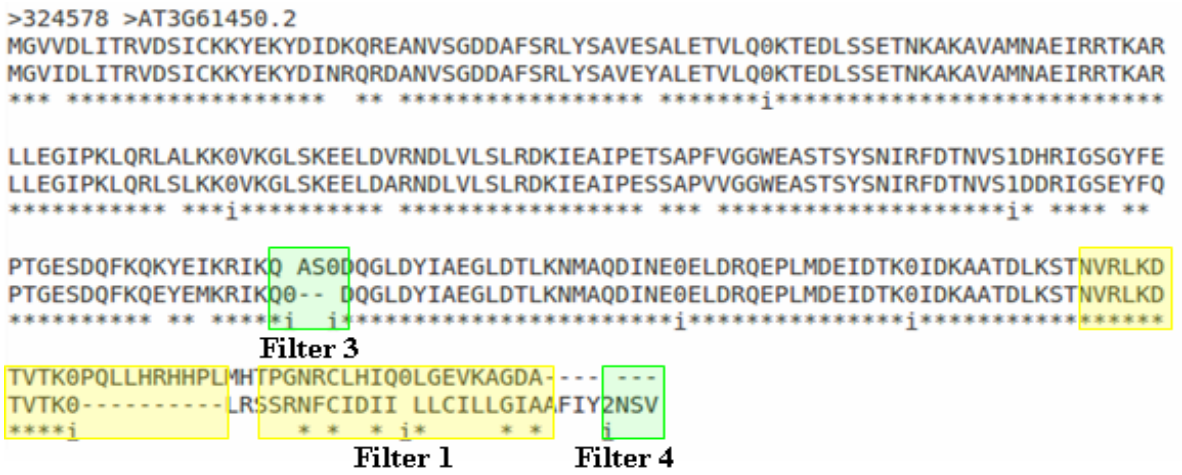


Figure S1: Example of filtred introns

Figure S2: Example of Protein Alignment with intron positions .Yellow boxes indicate intron unique, red box indicate conserved intron in the **phase0** (intron that don't interrupt codon), green box intron conserved in **phase1 (**intron positioned between first and second bases of codon**)** and green box intron conserved in **phase2** (intron collocated between the second and third bases of codon)

Figure S3: Classification of Conserved, gained or Lost Introns.The blue boxes indicates conseved introns, the yellow rectangle higlight intron gain and the box green marks lost intron.

	VC	PP	OS	AT	N° introns
Conserved	+	+	+	+	**283**
Lost	-	+	+	+	**680**
	+	-	+	+	**45**
	+	+	-	+	**2**
	+	+	+	-	**8**
Gain	+	-	-	-	**781**
	-	+	-	-	**31**
	-	-	+	-	**15**
	-	-	-	+	**3**
Others	+	-	-	+	**1**
	+	+	-	-	**17**
	+	-	+	-	**2**
		+	+	-	**16**
	-	+	-	+	**10**
	-	-	+	+	**156**
Analyzed introns					**2049**

Table S1: Intron gains or losses between 928 reciprocal orthologous *V.carteri* (**VC**), *P.patens* (**PP**), *O.sativa* (**OS**) and *A.thaliana* (**AT**); "+" indicates the presence of intron , and "-" absence of introns**:**

Gained Intron	Ortholog	Ortholog	Ortholog
GRMZM2G044011_T01_4	Pp1s227_55V6.1	Bradi5g24820.1	LOC_Os04g56480.1

Table S2: Intron classified as Gains: Intron gains between 917 reciprocal orthologous of *Z. mays, O. sativa*, *B.distachyon* and *P.patens* where *P.patens* was used as outgroup

Gained Intron	Ortholog	Ortholog
LOC_Os12g15470.2_1	Bradi4g39350.1	GRMZM2G118241_T01
LOC_Os01g59630.1_5	Bradi2g53000.1	GRMZM2G139691_T01
Bradi3g56020.1_1	LOC_Os02g57720.1	GRMZM2G081843_T01
Bradi4g32140.1_2	LOC_Os09g28420.1	GRMZM2G138468_T01

Table S3: Intron classified as Gains: Intron gains between 917 reciprocal orthologous of *O. sativa*, *B. distachyon* and *Z. mays* where *Z. mays* was used as outgroup

Gained Intron	Ortholog	Ortholog	Ortholog
GSVIVT01009948001_9	Pp1s313_104V6.1	Eucgr.I02329.1	POPTR_0005s11280.1
Eucgr.A00595.1_1	Pp1s111_85V6.1	GSVIVT01030332001	POPTR_0016s12860.1
Eucgr.H00404.1_1	Pp1s197_25V6.1	GSVIVT01010735001	POPTR_0007s02170.1
Eucgr.F02787.1_2	Pp1s136_3V6.1	GSVIVT01030990001	POPTR_0010s13050.1
Eucgr.E00644.1_1	Pp1s15_161V6.1	GSVIVT01019398001	POPTR_0003s13110.1
POPTR_0010s23600.1_26	Pp1s142_79V6.1	GSVIVT01016183001	Eucgr.G02746.1

Table S4: Intron classified as Gains: Intron gains between 2261 reciprocal orthologous of *V. vinifera*, *E. grandis* and *P. trochocarpa and P. patens* where *P. patens* was used as outgroup