20 Cudmore, S. *et al.* (1997) Viral manipulations of the actin cytoskeleton. *Trends Microbiol.* 5, 142–148
21 Marques, A.C. *et al.* (2005) Emergence of young human genes after a burst of retroposition in primates. *PLoS Biol.* 3, e357
22 Lander, E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921
23 Brandt, J. *et al.* (2005) Transposable elements as a source of genetic innovation: expression and evolution of a family of retrotransposon-derived neogenes in mammals. *Gene* 345, 101–111

0168-9525/$ – see front matter
doi:10.1016/j.tig.2006.09.006

Chromosomal clustering of nuclear genes encoding mitochondrial and chloroplast proteins in *Arabidopsis*

Andrey Alexeyenko[1], A. Harvey Millar[2], James Whelan[2] and Erik L.L. Sonnhammer[1]

[1] Center for Genomics and Bioinformatics, Karolinska Institutet, S-17177, Stockholm, Sweden
[2] ARC Centre of Excellence in Plant Energy Biology, M316, The University of Western Australia, 35 Stirling Highway, Crawley, WA 6009, Australia

We present a statistical analysis of chromosomal clustering among nuclear genes encoding mitochondrial or chloroplast proteins in *Arabidopsis*. For both organelles, the clustering was significantly increased above the expectation, but the clustering effect was weak, and most clusters were small and dispersed. Clustered genes showed coexpression but not more than expected, and no substantial synteny was detected in other eukaryotic genomes. We propose that the unexpected clustering results from continuous selection favoring chromosomal proximity of genes acting in the same organelle.

Introduction

Mitochondria and chloroplasts both originated from endosymbiotic events. During the course of evolutionary history, most of the essential genes required for mitochondrial and chloroplast function were transferred to the nuclear genome [1]. The organelle protein sets encoded in the nucleus today are not simply the genes originally transferred from the ancient endosymbiont but have a much more complicated genetic history. Evidence for proteins derived from the original endosymbiont, from the host genome, and even originating from other endosymbionts is found in both mitochondrial and chloroplast protein sets [2,3,4]. Due to the discrete metabolic roles of the two plant energy organelles, their origins, the mechanisms of gene transfer and the need for coordination of nuclear and organelle function in plants, it is possible that chromosomal organization of nuclear genes according to function could be an important aspect of regulation. Genes with various kinds of functional relationships are known to be in clusters on the chromosomes of a number of organisms [5–10]. Combining the experimental organelle sets [11–13] with clustering of their location and their expression in *Arabidopsis*, we have attempted to determine if any of the complex factors noted above link physical location with function or origin in this model plant.

Corresponding author: Sonnhammer, E.L.L. (Erik.Sonnhammer@sbc.su.se)
Available online 18 September 2006.

Results

Chromosomal clustering of genes encoding targeted organelle proteins

Clusters of neighboring genes were built by finding stretches of organelle genes closer than 10 kb to another organelle gene along the five chromosomes of *Arabidopsis*. To resolve the problem of tandemly duplicated genes, clustered homologs (those with a unidirectional BLAST similarity score >100 bits) were counted as one gene. To calculate the statistical significance of the number of clustered genes in the experimental sets of 473 mitochondrial and 664 chloroplast genes, we picked the same number of genes randomly from the genome 5000 times to generate a probability distribution (Figure 1 and Table 1). The *P*-values for the observed clustered genes in the mitochondrial and chloroplast sets were 0.0034 and 0.0004, which are greatly significant.

No large clusters were found using a 10-kbp cutoff. Most chloroplast and mitochondrial clusters contained two genes. Five clusters had three organelle genes, and one contained four genes. Variable numbers of other genes were found in these clusters – the largest total cluster size was observed with three mitochondrial genes and six other genes. The importance of the cutoff distance was assessed by using a number of cutoffs between 5 and 80 kbp, as summarized in Figure 2. Approximately 50% of all *Arabidopsis* genes have a neighbor within 80 kbp, hence it was not meaningful to investigate higher cutoffs. For the chloroplast genes, the observed clustering was significant ($P < 0.05$) at all cutoffs except 80 kbp, whereas the mitochondrial genes are generally less significantly clustered, particularly at the 5-kbp cutoff.

The clustering analysis was also made with a gene-based distance cutoff, that is, counting the number of intermediate genes rather than the number of base pairs. The cutoffs 0, 1, 2, 4 and 9 in-between genes were used. In this case, we found the clustering tendency much weaker, and, again, less significant in the mitochondrial set (only one *P*-value was <0.05). The optimum was seen at two genes in between for both groups. This indicates that the genes clustered by absolute distance tend to reside

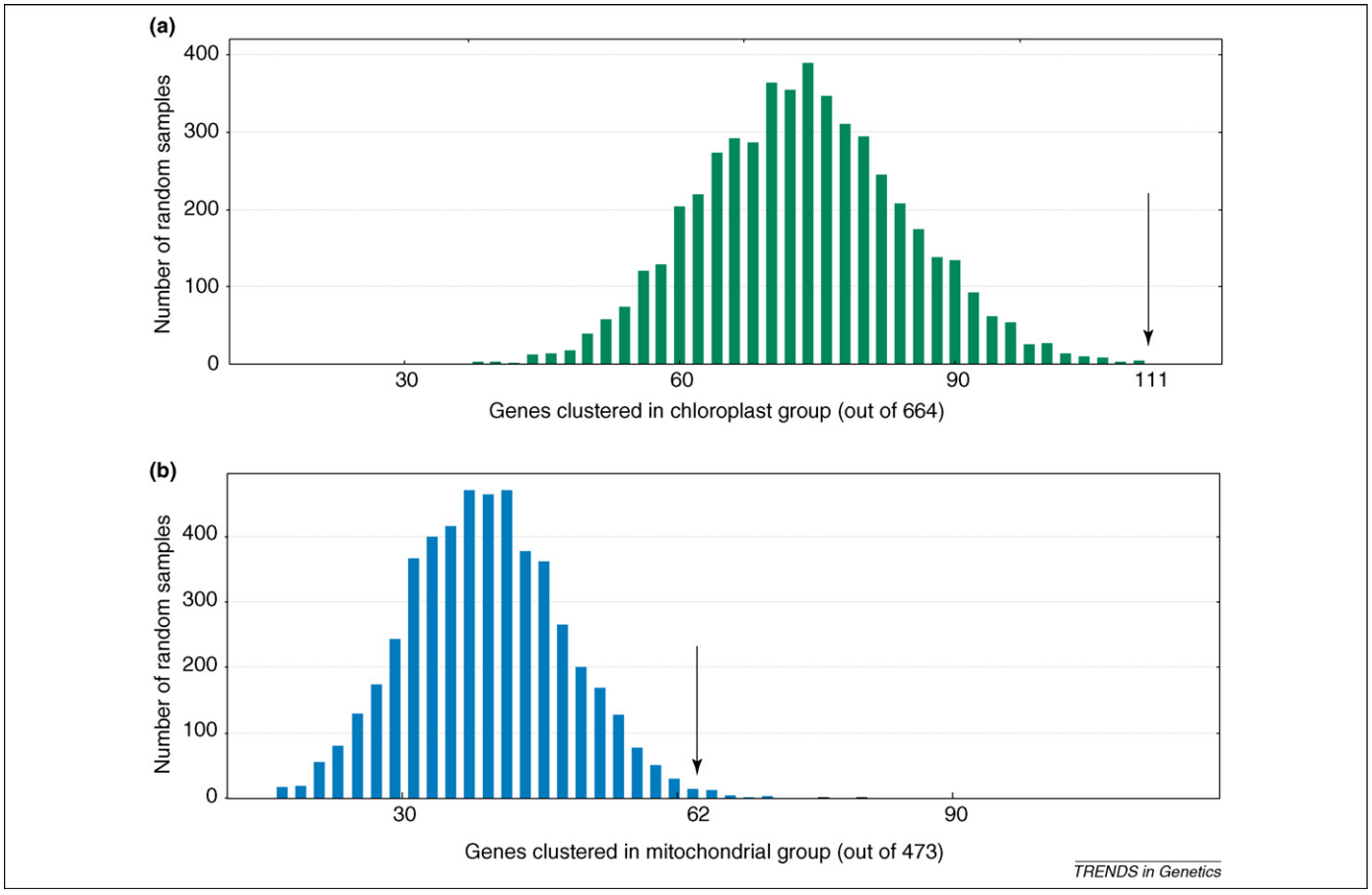

Figure 1. Chromosomal clustering of genes for mitochondrial and chloroplast proteins. Distribution of cluster sizes in 5000 random samples of **(a)** chloroplast and **(b)** mitochondrial gene groups using a 10-kbp cutoff. The observed real number of clustered genes is shown with black arrows.

in gene-rich regions. We confirmed this by comparing the neighbor distance distributions of the two organelle sets with genes from none of the sets. Both the chloroplast and mitochondrial sets were significantly ($P < 0.00001$ by Fisher's exact test on four bins) enriched in genes having a close neighbor (data not shown).

For comparison, the same procedure was applied to the set of 750 proteins found in yeast (*Saccharomyces cerevisiae*) mitochondria [14]. The base-pair cutoffs were reduced in accordance with the more compact (2.15 times shorter intergenic distances) yeast genome. This displayed a similar, but weaker, pattern of gene clustering, with a minimum *P*-value of 0.0066 at 4 kbp. The clustering tendency practically disappeared when using the gene-based distance cutoffs (minimal *P*-value = 0.065 at two genes in between). The enrichment of genes with a close neighbor was also significant ($P < 0.05$).

It is striking that the yeast results were so similar to *Arabidopsis* despite the fact that a much larger set of the proteome encodes experimentally confirmed mitochondrial proteins (12% compared with 1.7% and 2.4% in the *Arabidopsis* organelle sets).

Function and coexpression of the genes in organelle groups and clusters

The clustered genes were found to encode proteins with a wide set of biological functions, and covered a broad range of expression levels based on the available EST information (see Supplementary Material Table 1). We found no evidence for greater enrichment of targeting presequences in the clustered genes compared with the organelle sets as a whole. Also, no increased coexpression among genes in the same cluster compared with the average coexpression among all genes in the same organelle sets was observed. The lack of elevated coexpression in chromosomal clusters might be caused by the fact that the nuclear organelle genes are often coexpressed [mean Pearson correlation coefficient (r) = 0.35 and 0.11 for chloroplast and mitochondria, compared with 0.01 for random gene pairs and 0.03 for close (<10 kbp) chromosomal neighbors]. Extracting

Table 1. Gene and clustering statistics using a 10-kbp cutoff, and *P*-values of the observed number of clustered genes relative to the random distribution for a range of distance cutoffs

Organelle set	Total number of genes	Observed number of clusters (10 kbp)	Number of clustered genes (10 kbp)		*P*-value at different cutoffs (kbp)				
			Observed	Above expected	5	10	20	40	80
Mitochondrial	473	29	62	22.7	0.0602	0.0034	0.0223	0.0346	0.1085
Chloroplast	664	55	111	37.2	0.0005	0.0004	0.0031	0.0281	0.0717

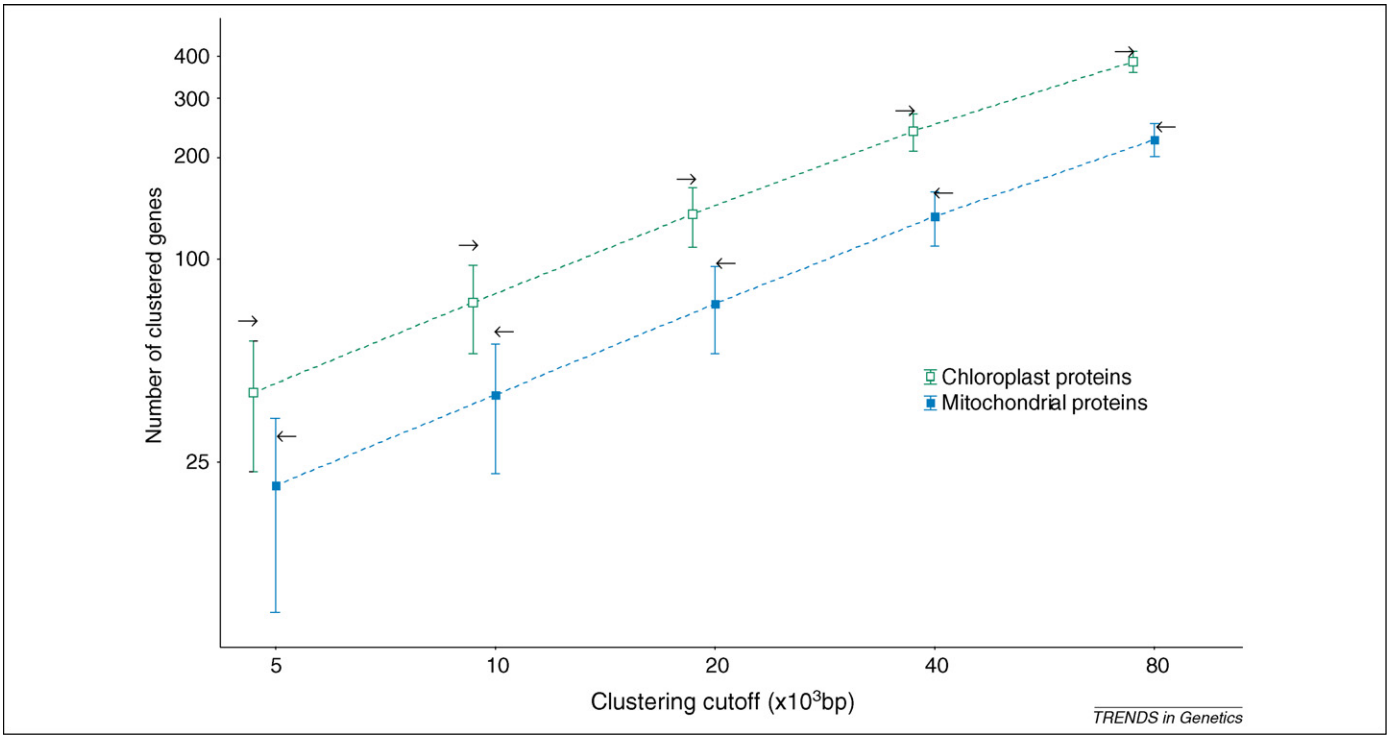

Figure 2. Number of clustered genes in chloroplasts and mitochondria as a function of the clustering cutoff. The chloroplast groups were clustered under the same cutoffs as the mitochondrial ones, but their points were shifted to the left for visibility. Observed real data points for the groups are denoted with arrows. The bars denote 95% confidence intervals.

only greatly coexpressed ($r > 0.77$) clustered gene pairs gave fractions also statistically indistinguishable from the complete organelle sets. The occurrence of elevated coexpression in the clusters thus seems to correspond to the general degree of functional coupling, but does not exceed it.

Clustering of orthologous genes between *Arabidopsis* and rice

We analyzed the degree of clustering among orthologs to the mitochondrial and chloroplast genes in *Oryza sativa* (rice), using the InParanoid program [15]. The fraction of *Arabidopsis* nuclear organelle genes with orthologs was great: 52% for mitochondrial genes and 69% for chloroplast ones, whereas the whole-genome average was 26%. However, conservation of the gene clusters was scant – we extracted all possible gene pairs in each cluster in each species, and looked for such pairs that were found clustered in both species. Only 1 of the 131 clustered *Arabidopsis* gene pairs had conserved their nearby genomic position in *O. sativa* (Table 2). Even when adopting a much looser limit (*Arabidopsis* pairs at <60 kbp and *O. sativa* orthologs at <1 Mbp) we found conservation of only two (of 179) mitochondrial and five (of 287) chloroplast pairs (Table 2). This is approximately the same as the fraction of such neighbors in the whole genome (~1.0%); hence we found no evidence of increased conservation of organelle clusters between rice and *Arabidopsis*. No clustered mitochondrial pairs defined in *Arabidopsis* were found conserved in four other eukaryotes (*S. cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster* and *Homo sapiens*).

Discussion

If a biological mechanism is driving the statistically significant coupling of genes in the two organelle groups

Table 2. Pairs of genes that have intergenic distance <60 kbp in *Arabidopsis* and orthologs in rice (*O. sativa*) at a distance <1 Mbp

Arabidopsis				Rice			
Gene 1	**Gene 2**	**Distance (bp)**	**Genes in between**	**Gene 1**	**Gene 2**	**Distance (bp)[a]**	**Genes in between[a]**
Mitochondrion							
At2g43780	At2g43750	4122	2	Os12g41760	Os12g42980	823 323	121
At5g50850	At5g50810	13 166	3	Os08g42410	Os08g42380	27 227	2
Chloroplast							
At1g21750	At1g21650	41 615	9	Os11g09280	Os11g08980	200 075	29
At3g56940	At3g56910	6337	2	Os01g17170	Os01g17150	13 631	1
At4g25100	At4g25080	6521	1	Os06g02500	Os06g04150	867 202	164
At5g35170	At5g35100	58283	6	Os08g19140	Os08g19610	274 570	46
At5g42765	At5g42650	51 146	11	Os03g56320	Os03g55800	340 053	51

[a]Minimal value of alternative orthologs.

to be chromosomal neighbors in the *Arabidopsis* nuclear genome, it seems to be operating mostly at the intergenic distance of ~10 000–20 000 base pairs (Figure 2). Open chromatin domains extending >2–5 genes in *C. elegans* chromosomes have been identified that contain coexpressed genes responsible for tissue-specific functions in the worm muscle [7]. Also, genes that are strongly expressed in a variety of human tissues or that are coexpressed in yeast are known to be clustered along the chromosomes across similarly sized intergenic regions [6,10,16].

However, we show here that coexpression can be largely ruled out as the driving force for the observed clustering. There have been many other reports that physical clustering is not required for coexpression or regulation of genes for organelle proteins. Although insertion into the vicinity of other genes encoding organelle proteins could initially provide signals required for expression or targeting (or both), coexpression of genes encoding subunits of multisubunit complexes can be readily achieved without physical clustering [17,18]. A study analyzing expression of 3292 genes enriched for nuclear-encoded plastid genes defined coexpression of only three gene pairs, when strongly homologous genes were excluded [19].

Could the insertion sites of the original gene transfer be the cause of some of the clusters? In a range of studied cases of gene transfers the genes are located adjacent to or inserted into a gene encoding a mitochondrial protein, for example *RPS11* (a gene encoding a mitochondrial protein of the small ribosomal subunit) in rice [20], *RPS10* in carrot and fuchsia [21], succinate dehydrogenase 3 (*sdh3*) in *Arabidopsis* and cotton [22], and *RPL15* (a gene encoding a mitochondrial protein of the large ribosomal subunit) in wheat [23]. This is most strikingly seen with the *RPS14* gene in maize and rice, which is inserted into a gene encoding succinate dehydrogenase subunit 2 – differential processing results in both functional transcripts [24,25]. In *Arabidopsis, RPS10* (At3g22300) has undergone recent gene transfer [21] and we show that it is encoded in a mitochondrial cluster in *Arabidopsis* (see Table 1 in Supplementary Material). However, because we could find no elevated synteny between *Arabidopsis* and rice for the observed clusters, we had to reject the hypothesis that the clusters are relicts of original insertion groups.

Co-inheritance of specific allele combinations has been suggested to explain clustering of genes for organelle function in plants. Elo *et al.* [26] noted that a large region of chromosome 3 in *Arabidopsis* contained a raft of genes for chloroplast and mitochondrial DNA and RNA maintenance. This 6–10-Mbp region is a different order of magnitude to the clustering investigated here, but some conservation was noted with regions of close clustering here on chromosome 3 (see Table 1 in Supplementary Material).

It therefore seems most likely that the observed clustering is the result of sampling of functionally related genes and has been ongoing in the course of evolution. Although hardly having an impact on organelle expression and function across the whole plant kingdom, this phenomenon would enable important gene or allele groups to survive both recombination and catastrophic genome-scale rearrangements. Reduced recombination rate has, for example, been demonstrated in clusters of essential yeast genes [27].

The preferential location of the clustered genes in gene-dense regions serves as additional evidence of a post-insertion rearrangement.

A mixture of other factors might also be responsible for the observed chromosomal clustering of genes for organelle proteins in the *Arabidopsis* genome. Further functional genomics studies will be necessary to identify specific patterns of functional coupling in individual species. The discovered clustering could be important in our understanding of the evolutionary history of gene transfer and activation, and the co-inheritance of this material in particular lineages.

Supplementary material

The data and methods for our analysis are described in the supplementary material associated with this article, which can be found online at doi:10.1016/j.tig.2006.09.002.

References

1 Timmis, J.N. *et al.* (2004) Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nat. Rev. Genet.* 5, 123–135
2 Abdallah, F. *et al.* (2000) A prediction of the size and evolutionary origin of the proteome of chloroplasts of *Arabidopsis*. *Trends Plant Sci.* 5, 141–142
3 Andersson, S.G. *et al.* (2003) On the origin of mitochondria: a genomics perspective. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 358, 165–177 discussion 177–169
4 Richly, E. *et al.* (2003) Evolutionary diversification of mitochondrial proteomes: implications for human disease. *Trends Genet.* 19, 356–362
5 Blumenthal, T. (1998) Gene clusters and polycistronic transcription in eukaryotes. *Bioessays* 20, 480–487
6 Caron, H. *et al.* (2001) The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science* 291, 1289–1292
7 Roy, P.J. *et al.* (2002) Chromosomal clustering of muscle-expressed genes in *Caenorhabditis elegans*. *Nature* 418, 975–979
8 Firneisz, G. *et al.* (2003) Identification and quantification of disease-related gene clusters. *Bioinformatics* 19, 1781–1786
9 Lee, J.M. and Sonnhammer, E.L. (2003) Genomic gene clustering analysis of pathways in eukaryotes. *Genome Res.* 13, 875–882
10 Hurst, L.D. *et al.* (2004) The evolutionary dynamics of eukaryotic gene order. *Nature. Rev. Gen.* 5, 299–310
11 Heazlewood, J.L. *et al.* (2004) Experimental analysis of the *Arabidopsis* mitochondrial proteome highlights signaling and regulatory components, provides assessment of targeting prediction programs, and indicates plant-specific mitochondrial proteins. *Plant Cell* 16, 241–256
12 van Wijk, K.J. (2004) Plastid proteomics. *Plant Physiol. Biochem.* 42, 963–977
13 Millar, A.H. *et al.* (2005) The plant mitochondrial proteome. *Trends Plant Sci.* 10, 36–43
14 Sickman, A. *et al.* (2003) The proteome of *Saccharomyces cerevisiae* mitochondria. *Proc. Natl. Acad. Sci. U. S. A.* 100, 13207–13212
15 Remm, M. *et al.* (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* 314, 1041–1052
16 Cohen, B.A. *et al.* (2000) A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nat. Genet.* 26, 183–186
17 Kelly, D.P. and Scarpulla, R.C. (2004) Transcriptional regulatory circuits controlling mitochondrial biogenesis and function. *Genes Dev.* 18, 357–368

18 Welchen, E. and Gonzalez, D.H. (2005) Differential expression of the *Arabidopsis* cytochrome C genes Cytc-1 and Cytc-2. Evidence for the involvement of TCP-domain protein-binding elements in anther- and meristem-specific expression of the Cytc-1 gene. *Plant Physiol.* 139, 88–100
19 Biehl, A. *et al.* (2005) Analysis of 101 nuclear transcriptomes reveals 23 distinct regulons and their relationship to metabolism, chromosomal gene distribution and co-ordination of nuclear and plastid gene expression. *Gene* 344, 33–41
20 Kadowaki, K. *et al.* (1996) Targeting presequence acquisition after mitochondrial gene transfer to the nucleus occurs by duplication of existing targeting signals. *EMBO J.* 15, 6652–6661
21 Adams, K.L. *et al.* (2000) Repeated, recent and diverse transfers of a mitochondrial gene to the nucleus in flowering plants. *Nature* 408, 354–357
22 Adams, K.L. *et al.* (2001) Multiple losses and transfers to the nucleus of two mitochondrial succinate dehydrogenase genes during angiosperm evolution. *Genetics* 158, 1289–1300
23 Sandoval, P. *et al.* (2004) Transfer of RPS14 and RPL5 from the mitochondrion to the nucleus in grasses. *Gene* 324, 139–147
24 Figueroa, P. *et al.* (1999) Transfer of RPS14 from the mitochondrion to the nucleus in maize implied integration within a gene encoding the iron-sulphur subunit of succinate dehydrogenase and expression by alternative splicing. *Plant J.* 18, 601–609
25 Kubo, N. *et al.* (1999) A single nuclear transcript encoding mitochondrial RPS14 and SDHB of rice is processed by alternative splicing: common use of the same mitochondrial targeting signal for different proteins. *Proc. Natl. Acad. Sci. U. S. A.* 96, 9207–9211
26 Elo, A. *et al.* (2003) Nuclear genes that encode mitochondrial proteins for DNA and RNA metabolism are clustered in the *Arabidopsis* genome. *Plant Cell* 15, 1619–1631
27 Pal, C. and Hurst, L.D. (2003) Evidence for co-evolution of gene order and recombination rate. *Nat. Genet.* 33, 392–395

0168-9525/$ – see front matter
doi:10.1016/j.tig.2006.09.002

Overlapping genes as rare genomic markers: the phylogeny of γ-Proteobacteria as a case study

Yingqin Luo, Cong Fu, Da-Yong Zhang and Kui Lin

MOE Key Laboratory for Biodiversity Science and Ecological Engineering and College of Life Sciences, Beijing Normal University, Beijing 100875, China

Phylogenies can be constructed in many ways, including using shared complex characters known as rare genomic changes (RGCs), such as insertions and deletions (indels), retroposon integrations and intron positions. Here, we demonstrate that distance-based phylogenies, which were determined by shared overlapping genes from 13 completely sequenced γ-Proteobacteria genomes, are consistent with phylogenies based on 16S rRNAs and other robust markers. These findings suggest that overlapping genes could provide interesting additional insights into the phylogenomics of completely sequenced microbial genomes.

Introduction

With more and more completely sequenced genomes available, phylogenies based on whole genomes capture more phylogenetic signatures and are less influenced by anomalous events than those based on single genes. Along with sequence-based methods, which mainly involve the comparison of primary sequences, gene content and gene order are currently used in phylogenetic analyses of completely sequenced genomes; see Delsuc *et al.* [1] for a detailed review. However, gene content might prove to change too little, and gene order might change too much, for adequate analyses to be performed [2]. Genomes can also be compared by looking for shared complex characters known as rare genomic changes (RGCs), such as indels, retroposon integrations and intron positions. Until recently, only a few types of RGC character have been used for inferring phylogenetic relationships among completely sequenced genomes; for example, Gupta *et al.* [3,4] deduced the branching order for bacterial groups using conserved indels in protein sequences, and Yang *et al.* [5] determined prokaryotic phylogeny based on protein domain content (see also Rokas and Holland [6] for a more detailed review).

Orthologous overlapping genes among bacterial genomes

Overlapping genes in bacterial genomes are adjacent genes whose coding sequences partially or entirely overlap. Intuitively, we would assume that this character does not evolve as slowly as gene content because the formation of overlapping genes is more frequently observed than the variation of gene repertoires among completely sequenced genomes, especially in closely related genomes [7,8]. By contrast, overlapping genes might be more conserved than gene order during the course of evolution, because functional constraints might prevent breaking of the linkage of two overlapping genes [9–11]. The rate of evolution is also expected to be slower for stretches of DNA encoding overlapping genes than for similar DNA sequences that encode only one reading frame [11]. Therefore, choosing overlapping genes as an indicator of phylogeny is feasible because mutations generating overlapping genes are acquired at a universal rate across species [7,8].

The complete sequences of 13 γ-Proteobacteria genomes were downloaded from GenBank in August 2004. Open reading frames (ORFs) annotated as 'hypothetical' or 'putative', or with products annotated as 'unknown

Corresponding author: Lin, K. (linkui@bnu.edu.cn)
Available online 12 September 2006.