Genome Analysis

Inversions and the dynamics of eukaryotic gene order

Martijn A. Huynen, Berend Snel and Peer Bork

Comparisons of the gene order in closely related genomes reveal a major role for inversions in the genome shuffling process. In contrast to prokaryotes, where the inversions are predominantly large, half of the inversions between *Saccharomyces cerevisiae* and *Candida albicans* appear to be small, often encompassing only a single gene. Overall the genome rearrangement rate appears higher in eukaryotes than in prokaryotes, and the current genome data do not indicate that functional constraints on the co-expression of neighboring genes have a large role in conserving eukaryotic gene order. Nevertheless, qualitatively interesting examples of conservation of gene order in eukaryotes can be observed.

Beyond the counting of shared genes, the comparative analysis of whole genomes only took off after a substantial number of genomes at varying evolutionary distances became available. We already have a variety of so-called 'gene context' analyses (Refs 1 and 2, and references therein) that shed light on evolutionary and functional aspects of the interactions between genes in prokaryotes. However, despite a number of comparative genome studies in eukaryotes, the age of gene context analysis in eukaryotes has only just begun. An elegant paper by Seoighe *et al.*[3] describes the comparison between the genome of *Saccharomyces cerevisiae* and the almost-complete genome of *Candida albicans*, with a focus on the evolution of gene order.

The current set of complete eukaryotic genomes (human, fly, worm and yeast) are too divergent to reveal the dynamics of gene-order evolution. Therefore, Seoighe *et al.*[3] compared *S. cerevisiae* and *C. albicans* to study eukaryotic genome dynamics (most of the *C. albicans* genome is available in the public domain http://www-sequence.stanford.edu/group/candida). It is the first large-scale study that documents the important role of local inversions in shuffling the eukaryotic genome. The authors model various types of genome rearrangement using differential equations and estimate that local inversions (containing less than ten genes) disrupt gene order as frequently as inter-chromosomal or long-distance transpositions.

Illustrating the small size of the inversions is the fact that the relative orientation of genes has been reversed in 103 of the 298 pairs of genes that occur as neighbors both in *C. albicans* and *S. cerevisiae*. The authors estimate that 1100 of such single-gene inversions occurred after the divergence of the two species (140–330 Myr ago)[3]. Another large-scale comparison of *S. cerevisiae* with a number of other hemiascomycetous yeast species, including *Candida tropicalis*, also points out the large frequency of inversions between *Saccharomyces* and some *Candida* species[4]. The authors argue, however, that the rate of inversions is not as constant as modeled by Seoighe *et al.*[3] Their results indicate that gene inversions have not played a large role within the *Saccharomyces* taxon (see also Ref. 5) and have mainly occurred at larger evolutionary distances, in the lineage leading to the *Candida* species and to *Yarrowia lipolytica*[4].

Previous studies document the importance of inversions in the evolution of eukaryotic gene order (Ref. 3 and references therein). A recent example is the comparison of the right arm of chromosome 3 of *Drosophila melanogaster* with its homolog in *Drosophila repleta*[6], in which 114 inversions are estimated to have occurred since the divergence of these species (40–62 Myr ago). It is the prevalence of local inversions that make the results of Seoighe *et al.* so interesting. They indicate that our view of the colinearity between chromosomal regions of closely related eukaryotes, suggested by low-resolution studies, might be refined by more high-resolution studies.

The role of inversion in genome rearrangements

Inversions appear also to be a major component in genome rearrangements in the evolution of prokaryotes[7–9]. They tend to be predominantly large scale, and to be centered around the terminus or origin of replication[8,9]. Local, small-scale inversions seem to be rare in prokaryotic genomes and have not been reported in large-scale analyses of gene-order conservation, although they can incidentally be observed in comparative gene-order plots (e.g. in the Chlamydiae[9]). Finally, there are lineages (e.g. Mycoplasmas) in which inversions have not been observed at all[10].

These results raise the question whether the rates of genome shuffling in Eukaryotes are comparable to those in Prokaryotes. We examined this question by comparing the gene-order conservation between *S. cerevisiae* and *C. albicans* with that between the bacteria *Haemophilus influenzae* and *Escherichia coli*. The evolutionary distance between *C. albicans* and *S. cerevisiae* as measured by protein-sequence divergence is similar to that between *H. influenzae* and *E. coli*. However, between *H. influenzae* and *E. coli* 36% of gene pairs are conserved, whereas between *C. albicans* and *S. cerevisiae* 9% of the gene pairs are conserved[3] (Table 1). Here, a conserved gene pair is defined as two adjacent genes in species A that are also adjacent in species B.

More dramatic than the difference in the overall level of gene-order conservation is the conservation of the relative orientation of the genes within conserved gene pairs. Between *C. albicans* and *S. cerevisiae* the relative orientation is conserved in 189 (64%) of 294 gene pairs, however between *E. coli* and *H. influenzae*, it is conserved in 476 (99%) out of 481 gene pairs (Table 1). Thus, small-scale inversions that maintain the gene order but not the relative orientation of the genes are much rarer in prokaryotes than in *C. albicans* and *S. cerevisiae*.

Co-regulation and conservation of gene pairs

Analysis of *S. cerevisiae* expression data revealed that, although adjacent genes do have a significantly higher chance of being co-regulated than non-adjacent genes, less than 10% of adjacent gene pairs appear to be co-regulated[11]. The gene pairs that display the highest level of co-regulation are transcribed either divergently (← →) or in the same direction (→ →)[11]. Surprisingly, the relative orientation of genes conserved in pairs between *C. albicans* and *S. cerevisiae*

 0168–9525/01/$ – see front matter PII: S0168-9525(01)02302-2

do not indicate that functional constraints on gene order have a role in gene-order conservation[3]. In fact, gene pairs that are transcribed convergently (→ ←) are more often conserved than gene pairs that are transcribed divergently and are conserved as often as genes that are transcribed in the same direction (Table 1). By contrast, in prokaryotes, the conservation of gene order is strongly dominated by genes transcribed in the same direction, probably reflecting operons, with divergently transcribed genes (hinting at conserved divergent promoters) as a distant second (Table 1)[12,13].

Nevertheless, a direct comparison of the gene-order conservation data[3] with the expression data[11] indicates a small, but significant, effect of co-regulation on the conservation of gene-order between *S. cerevisiae* and *C. albicans*. Out of 18 co-regulated gene pairs from Ref. 11 having both genes present in *C. albicans*, four (22%) are conserved as pairs in *S. cerevisiae*. This exceeds the 6% of gene pairs that are conserved, including their relative direction of transcription, between *S. cerevisiae* and *C. albicans* and indicates that the selective constraints imposed by co-regulation have slowed down the genome rearrangements, if only slightly. Note also that in yeast mitochondrial genomes incidental cases of gene-order conservation are linked to co-regulation of the genes[14].

Table 1. A comparison of gene-order conservation between *C. albicans* and *S. cerevisiae* and between *H. influenzae* and *E. coli*

	C. albicans–S. cerevisiae	*H. influenzae–E. coli*
No. genes (genome A / genome B)	9168/5800	1709/4289
Elongation factor 1α identity[a]	91%	93%
No. shared orthologs	3960 (68%)	1330 (78%)
Conserved pairs[b]	9%	36.2%
Conserved pairs including gene orientation	6%	35.7%
Gene orientation of conserved pairs[c] (→→/←→/→←)	1/0.76/0.99	1/0.11/0.0

[a]The protein sequence conservation within the pairs of species, as measured by the sequence identity of elongation factor 1α, is similar.
[b]A conserved gene pair is defined as an adjacent pair of genes in genome A that is both present and adjacent in genome B. In determining whether two genes are adjacent, genes that are not shared between the two species are ignored[3]. Genome data are from GenBank, except *C. albicans* (http://www-sequence.stanford.edu/group/candida).
[c]The largest fraction of conserved pairs with conserved direction of transcription was set to one, the other fractions are relative to this. The data clearly show that prokaryotes show a higher degree of gene-order conservation in general than *C. albican*s and *S. cerevisiae*, specifically regarding the conservation of the orientation of genes in conserved pairs.

Operons in Eukaryotes

Operon-like structures that allow co-transcription of adjacent genes and that contain functionally related genes have been described in *C. elegans* and other nematodes (reviewed in Ref. 15). The nematode operons probably evolved independently from the prokaryotic ones. A comparison of the gene order in the prokaryotes and *C. elegans* revealed only three cases of functionally interacting, neighboring genes that are present in both *C. elegans* and at least one prokaryote: BO272.3 (3-hydroxyacyl-CoA dehydrogenase) and BO272.4 (enoyl CoA hydratase/isomerase), K07E3.3 (methylene tetrahydrofolate dehydrogenase) and K07E3.4B (tetrahydrofolate synthase), and Y38F2AR.A and Y38F2AR.B (subunits of 5-oxoprolinase). Using gene-order conservation for the detection of operons and functionally interacting genes in *C. elegans*, as was done for the prokaryotes, will therefore mainly have to rely on the sequencing of other nematodes like *Caenorhabditis briggsae*, but also more distantly related ones.

Outlook

Among the eukaryotes, there are well-known functionally interacting genes such as the histone genes or Hox genes that are conserved in clusters. Quantitatively, however, only the histone genes have a significant role in the amount of gene-order conservation between the sequenced eukaryotes of yeast, worm, fly and human. When we compared the gene order among these eukaryotes, excluding the histone genes from the analysis, the amount of gene-order conservation did not exceed the expected level for randomly shuffled genomes (M.A. Huynen, unpublished).

Thus, it appears that, even though there are qualitatively interesting examples of conservation of gene order among the eukaryotes and between eukaryotes and prokaryotes, with the present genome data these examples will not have a significant role for the prediction of functional interactions between proteins. It is possible that with the sequencing of more species such as the nematodes or yeast species in which neighboring genes do have an above-average probability of being co-regulated, the importance of gene neighborhood for the prediction of functional interactions will increase. Furthermore, less strict, but relevant, forms of neighborhood (allowing a larger distance between genes) or less strict forms of conservation[16] might become apparent when more eukaryotic genomes are sequenced. In any case the large-scale, high-resolution comparisons of closely related eukaryotic genomes[3] and the explicit modeling of the various processes that rearrange the genome[3] will be necessary to detect patterns of gene-order conservation and to judge their significance.

Acknowledgements

This work was supported by Bundesministerium fuer Bildung und Forschung. M.A.H. and P.B. also carry out research at Max Delbruck Centrum for Molecular Medicine, 131122 Berlin-Buch, Germany. M.H. thanks Kenneth Wolfe for useful discussions and Richard Copley and Elia Stupka for providing human genome data.

References

1 Huynen, M.A. *et al.* (2000) Prediction protein function from genomic context: Quantitative evaluation and qualitative inferences. *Genome Res.* 10, 1204–1210
2 Marcotte, E.M. (2000) Computational genetics: finding protein function by nonhomology methods. *Curr. Opin. Struct. Biol.* 10, 359–365
3 Seoighe, C. *et al.* (2000) Prevalence of small inversions in yeast gene order evolution. *Proc. Natl. Acad. Sci. U. S. A.* 97, 14433–14437
4 Llorente, B. *et al.* (2001) Genomic exploration of the Hemiascomycetous yeasts: 18. Comparative analysis of the chromosome maps and synteny with *Saccharomyces cerevisiae*. *FEBS Lett.* 487, 101–112
5 Langkjaer, R.B. *et al.* (2000) Yeast chromosomes have been significantly reshaped during their evolutionary history. *J. Mol. Biol.* 304, 271–288
6 Ranz, J.M. *et al.* (2001) How malleable is the eukaryotic genome? Extreme rate of chromosomal rearrangement in the genus *Drosophila*. *Genome Res.* 11, 230–239
7 Tillier, E.R.M and Collins, R.A. (2000) Genome rearrangement by replication-directed translocation. *Nat. Genet.* 26, 195–197
8 Eisen, J.A. *et al.* (2000) Evidence of symmetrical chromosomal inversions around the replication origin in bacteria. *Genome Biol.* 1, 1–9
9 Suyama, M. and Bork, P. (2001) Evolution of prokaryotic gene order: genome rearrangements in closely related species. *Trends Genet.* 17, 10–13
10 Himmelreich, R. *et al.* (1997) Comparative analysis of the genomes of the bacteria *Mycoplasma pneumoniae* and *Mycoplasma genitalium*. *Nucleic Acids Res.* 25, 701–712

11 Kruglyak, S. and Tang, H. (2000) Regulation of adjacent yeast genes. *Trends Genet.* 16, 109–111
12 Huynen, M.A. and Snel, B. (2000) Gene and context: integrative approaches to genome analysis. In *Analysis of Amino Acid Sequences* (*Adv. Prot. Chem.* Vol. 54) (Bork, P., ed.), pp. 345–379, Academic Press
13 Bork, P. *et al.* (2000) Comparative genome analysis: exploiting the context of genes to infer evolution and predict function. In *Comparative Genomics* (*Computational Biology*) (Sankoff, D. and Nadeau, J.H., eds), pp. 281–294, Kluwer
14 Groth, C. *et al.* (2000) Diversity in organization and the origin of gene orders in the mitochondrial DNA molecules of the genus *Saccharomyces*. *Mol. Biol. Evol.* 17, 1833–1841
15 Blumenthal, T. (1998) Gene clusters and polycistronic transcription in eukaryotes. *BioEssays* 20, 480–487
16 Lathe, W. *et al.* (2000) Gene context conservation of a higher order than operons. *Trends Biochem. Sci.* 25, 474–479

M.A. Huynen*
B. Snel†
P. Bork‡
EMBL, Biocomputing,
Meyerhofstrasse 1,
69117 Heidelberg, Germany.
*e-mail: huynen@embl-heidelberg.de
†e-mail: snel@embl-heidelberg.de
‡e-mail: Bork@embl-heidelberg.de

Paramecium genome survey: a pilot project

Philippe Dessen, Marek Zagulski, Robert Gromadka, Helmut Plattner, Roland Kissmehl, Eric Meyer, Mireille Bétermier, Joachim E. Schultz, Jürgen U. Linder, Ronald E. Pearlman, Ching Kung, Jim Forney, Birgit H. Satir, Judith L. Van Houten, Anne-Marie Keller, Marine Froissard, Linda Sperling and Jean Cohen

A consortium of laboratories undertook a pilot sequencing project to gain insight into the genome of *Paramecium*. Plasmid-end sequencing of DNA fragments from the somatic nucleus together with similarity searches identified 722 potential protein-coding genes. High gene density and uniform small intron size make random sequencing of somatic chromosomes a cost-effective strategy for gene discovery in this organism.

The ciliated protozoan *Paramecium* was one of the first microorganisms discovered by the early microscopists in the 18th century and has been extensively studied since then. These studies made important discoveries such as microbial sexuality and the occurrence of mating types[1], surface antigens[2], cytoplasmic inheritance[3] and an epigenetic phenomenon not mediated by DNA, called structural heredity[4]. More recently, *Paramecium* has become a powerful model unicell in various fields including membrane excitability[5] and signal transduction[6,7], regulated secretion[8], cellular morphogenesis[9,10], surface antigen variation[11], developmental genome rearrangements[12,13], and homology-dependent epigenetic regulation of both gene expression[14] and developmental genome rearrangements[15]. The recent availability of DNA-mediated transformation[16] allowed complementation cloning of genes identified by mutation[17–19] and gene inactivation by homology-dependent gene silencing through a mechanism related to RNA interference[14,20].

Paramecium and the other ciliates are located at a key position in the terminal crown of the eukaryotic phylogenetic tree, together with fungi, plants and metazoa. Moreover, ciliates display a unique feature in the unicellular world: the differentiation of germ and somatic lines in the form of nuclei, not cells. The somatic nucleus (macronucleus) and the germinal nucleus (micronucleus) both derive from the zygotic nucleus, itself derived from parental micronuclei through meiosis and fertilization. During macronuclear development, programmed DNA rearrangements affect the entire genome through amplification to a high ploidy level, chromosome fragmentation and telomere addition, and internal sequence elimination. Many sexual and developmental processes present in metazoa therefore also exist in ciliates, which could serve as pertinent models for their study.

For the moment, no full-scale ciliate genome project has been funded. The community working with *Tetrahymena* has mobilized great ingenuity in genome mapping and development of other tools, including sequencing of expressed sequence tags (ESTs) (J. Fillingham *et al.*, unpublished), with the objective of the complete sequencing of the genome of *Tetrahymena thermophila*[21], a ciliate whose evolutionary distance from *Paramecium tetraurelia* is estimated at greater than 100 Myr.

The pilot sequencing study

All these considerations stimulated the *Paramecium* community to undertake a genome project. Before being able to establish the full 100–200-megabase genome sequence, *Paramecium* scientists present at the FASEB Ciliate Molecular

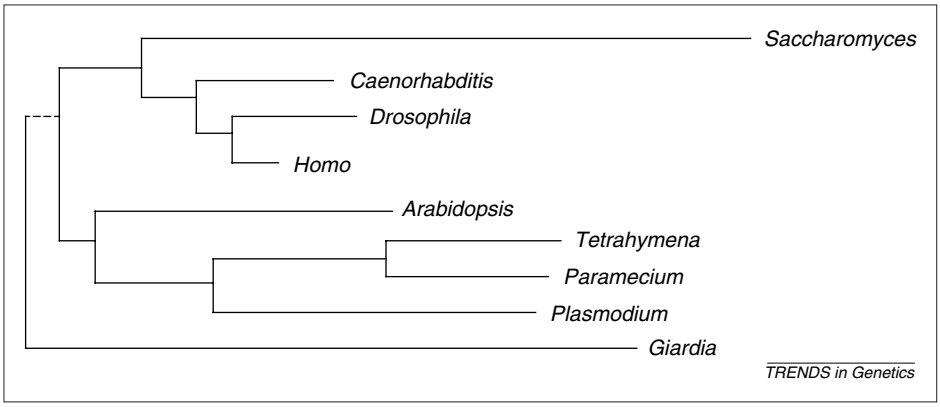

Fig. 1. Eukaryotic phylogeny simplified from Ref. 26.

 0168–9525/01/$ – see front matter PII: S0168-9525(01)02307-1