Available online at www.sciencedirect.com
SCIENCE DIRECT
Gene 365 (2006) 41–47
GENE
SECTION
EVOLUTIONARY GENOMICS
www.elsevier.com/locate/gene

The evolution of introns in human duplicated genes

Edda Rayko [a,1], Kamel Jabbari [a,1], Giorgio Bernardi [b,*]

[a] *Laboratoire de Génétique Moléculaire, Institut Jacques Monod, 2 Place Jussieu, F-75005 Paris, France*
[b] *Laboratorio di Evoluzione Molecolare, Stazione Zoologica Anton Dohrn, Villa Comunale, I-80121 Naples, Italy*

Received 22 April 2005; received in revised form 7 July 2005; accepted 7 September 2005
Available online 13 December 2005
Received by T. Gojobori

Abstract

In previous work [Jabbari, K., Rayko, E., Bernardi, G., 2003. The major shifts of human duplicated genes. Gene 317, 203–208], we investigated the fate of ancient duplicated genes after the compositional transitions that occurred between the genomes of cold- and warm-blooded vertebrates. We found that the majority of duplicated copies were transposed to the "ancestral genome core", the gene-dense genome compartment that underwent a GC enrichment at the compositional transitions.

Here, we studied the consequences of the events just outlined on the introns of duplicated genes. We found that, while intron number was highly conserved, total intron size (the sum of intron sizes within any given gene) was smaller in the GC-rich copies compared to the GC-poor copies, especially in dispersed copies (i.e., copies located on different chromosomes or chromosome arms). GC-rich copies also showed higher densities of CpG islands and Alus, whereas GC-poor copies were characterized by higher densities of LINEs. The features of the copies that underwent the compositional transition and became GC-richer are suggestive of, or related to, functional changes.

Keywords: Genome; Isochores; CpG islands; Alus; LINEs

1. Introduction

Since many gene duplications present in the human genome are ancient duplications going back to the origin of vertebrates (see Postlethwait et al., 2004, for a review), a question may be asked about the fate of such duplicated genes after the compositional transitions that occurred between the genomes of cold- and warm-blooded vertebrates (see Bernardi, 2004, for a review). Indeed, at those transitions, the gene-dense "ancestral genome core" of cold-blooded vertebrates underwent a GC enrichment to become the "genome core" of warm-blooded vertebrates (see Fig. 1).

We could show that, by far and large, one copy of the duplicated gene underwent a GC enrichment, the other copy keeping the original GC level (Jabbari et al., 2003). We hypothesized that the former copy was preferentially translocated into the gene-dense compartment of the genome, the "ancestral genome core", namely the gene space which underwent the compositional transition (GC enrichment) at the emergence of warm-blooded vertebrates (see Fig. 2). This hypothesis was based on three assumptions: (i) that duplication occurred more frequently in the gene-poor compartment of the genome of the cold-blooded ancestors, the "ancestral genome desert", as suggested by the abundance of gene duplications in GC-poor pericentromeric regions (see Jabbari et al., 2003, for references); (ii) that one copy acquired a new function; and (iii) that transposition of the duplicated copy into the "ancestral genome core" rather than into the "genome desert" (the gene-poor part of the genome; see Fig. 1) was generally preferred. The latter assumption was based on the observation that the chromatin of the "ancestral genome core" is in an open configuration (as shown by the results of Federico et al., in press, on the genomes of *Rana esculenta* and *Podarcis sicula*), and that integration of retroviral sequences preferentially occurs into open chromatin (see Rynditch et al., 1998; Tsyba et al., 2004). This assumption

* Corresponding author. Tel.: +39 081 5833402; fax: +39 081 2455807.
E-mail addresses: erayko@biologie.ens.fr (E. Rayko), kjabbari@biologie.ens.fr (K. Jabbari), bernardi@szn.it (G. Bernardi).
[1] Present address: Département de Biologie, Ecole Normale Supérieure, 46 Rue d'Ulm, F-75230 Paris, France.

0378-1119/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2005.09.038

was also justified by the fact that it could account for the higher gene density of the "ancestral genome core" compared to the "genome desert" (see Fig. 1).

As far as structural and functional consequences of the major shift of duplicated genes are concerned, the fact that the majority of duplications are ancient duplications, suggested that the copy that experienced a GC shift might exhibit the same properties as any GC-rich gene of the human genome (see Bernardi, 2004 for a review), namely an enrichment in Alus (Jabbari and Bernardi, 1998) and flanking CpG-islands (Aïssani and Bernardi, 1991a,b), a decrease in LINEs (Pavlicek et al., 2001) and a shortening of introns (Duret et al., 1995).

Here we considered, therefore, the consequences of the events just outlined on the size of the introns and on the frequencies of CpG islands, Alus and LINEs located in the introns of duplicated genes. The changes that we found in the GC-rich copies compared to the GC-poor copies further support the idea that translocations preferentially took place from the gene-poor "ancestral genome desert" to the gene-rich "ancestral genome core". Moreover, they are suggestive of functional changes in the copies that underwent the compositional change.

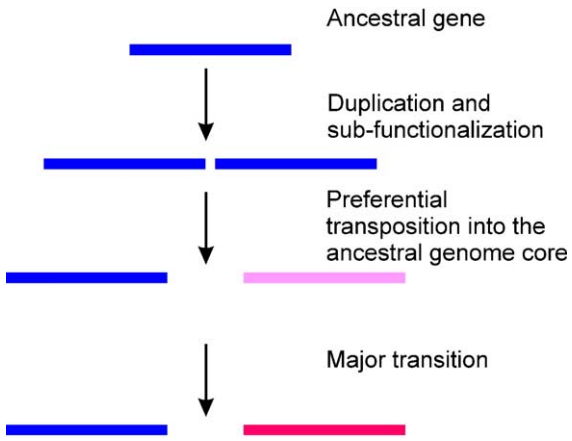

Fig. 2. A scheme of the most frequent pathway following ancient gene duplication (blue bars). One copy is supposed to be preferentially transposed into the ancestral genome core (pink bar), which then undergoes the major compositional transition (red bar) (modified from Jabbari et al., 2003).

2. Materials and methods

We retrieved 3947 complete human coding sequences from HOVERGEN (Duret et al., 1994), with gene family annotation. We collected pairs of coding sequences (CDS) with similar size (the difference in size being lower than, or equal to, 7%) from each family. Redundant and alternatively spliced CDS were disregarded. This led to a data set of 412 pairs of duplicated human genes representing 214 gene families and including 596 individual CDS. Exon and intron sizes were calculated using their coordinates along the gene sequences, as indicated in EMBL annotations. We excluded a few pseudogenes that were present in the group of intronless genes. We should notice that only the exons containing the initiator methionine were considered as first exons, last exons being those ending with a stop codon (Supplementary Table 1).

CpG islands were searched using the program cpgplot http://www.hgmp.mrc.ac.uk/Software/EMBOSS/Apps/cpgplot.html. This program defines, by default, a CpG island as a region where the GC level is over 50%, the calculated Observed/Expected (O/E) CpG ratio is over 0.6, these conditions holding for a minimum of 200 bases. Alu and LINE sequences located within introns were identified using RepeatMasker (A. Smit and P. Green, unpublished, http://ftp.genome.washington.edu/RM/RepeatMasker.html).

χ^2-test for means comparisons with a two-sided hypothesis was used to test the statistical significance; 1% standard errors are indicated (see figures). We used correlation analysis to test the statistical significance of the association between the

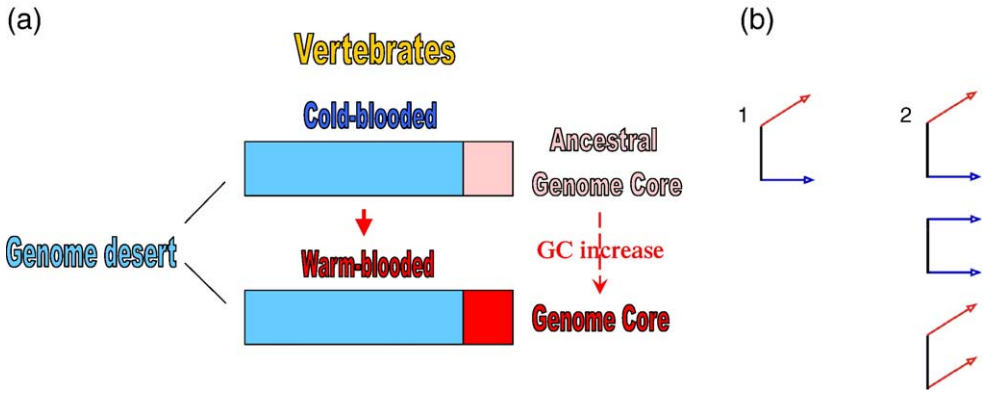

Fig. 1. (a) Scheme of the compositional genome transitions that took place between cold- and warm-blooded vertebrates. The "genome desert" of cold-blooded vertebrates is GC-poor and gene-poor (blue box) and essentially did not undergo any compositional change. In contrast, the gene-dense, moderately GC-rich "ancestral genome core" (pink box) underwent a compositional change into a gene-dense, GC-rich "genome core" (red box). (b) A model of the two situations hypothesized for duplicated genes at the transitions from cold- to warm-blooded vertebrates. (1) One copy of each pair preferentially underwent the transition (red arrow), the other copy maintaining its original low GC level (blue arrow). (2) In addition to situation 1, both copies underwent the transition or maintained their original low GC level (modified from Jabbari et al., 2003).