Research Article

How complex an intron may be? The example of the first intron of the *CTP synthase* gene of *Drosophila melanogaster*

Roberto Piergentili

Istituto di Biologia e Patologia Molecolari del CNR; Dipartimento di Biologia e Biotecnologie, Sapienza Università di Roma; Rome, Italy.

Received on August 6, 2012; Accepted on December 31, 2012; Published on February 20, 2013

Correspondence should be addressed to Roberto Piergentili; Phone: +390649912873, Fax: +39064456866, Email: roberto.piergentili@uniroma1.it

Abstract

In eukaryotes, maturation of primary transcripts into mature messenger RNAs involves the elimination of parts of the gene called 'introns'. The biological significance of introns is not yet completely understood. It has been demonstrated that introns may contain other genes, or regulatory sequences that may be involved in transcriptional control, or also being involved in alternative splicing mechanisms. However, these functions explain the role of only a small number of them, and it is very difficult to formulate any generalization. The CTP synthase gene of *Drosophila melanogaster* is characterized by the presence of a long first intron (approximately 7.2 kilobases) whose role is currently unknown. In the present report we analyzed *in silico* the content of this intron, and found that it contains at least three interesting sub-sequences. Two of them are homologous to the CTP synthase itself and to a putative nucleotide pyrophosphatase, respectively. The third is a short stretch of DNA able to fold into a thermodynamically stable hairpin and showing homology with other 19 sequences from 21 genes inside the *D. melanogaster* genome. These findings suggest a complex yet very accurate way of controlling gene expression inside the fruit fly.

Introduction

There are two main types of nucleotides inside the cell, ribonucleotides (used for RNA) and deoxyribonucleotides (used for DNA). Ribonucleotides may be converted into deoxyribonucleotides through two enzymatic reactions, the reduction of the ribose ring in position 2' and the conversion of uridine into thymine. The first reaction is catalyzed by the ribonucleotide reductase (RNR), that also plays a central role in maintaining their relative abundance (Hofer *et al.* 2012); the second reaction is under thymidylate synthase control (Costi *et al.* 2005). There are two ways to maintain ribonucleotide pool balance inside cells, the *salvage* pathway and the *de novo* pathway. The former allows the recovery of nucleotides from intracellular nucleic acids (such as degraded RNAs) or from free (poly) nucleotides taken from the environment through specific membrane channels; the latter lets the cell assemble new nucleotides starting from simpler compounds present or built in the cytoplasm, such as ribose, phosphate, and amino acids. There are two distinct *de novo* pathways, one specific for purine biosynthesis, the other specific for pyrimidine biosynthesis (Figure 1). Many steps in the *de novo* biosynthesis are reversible, though some are not; the step controlled by the enzyme CTP synthase is an irreversible one, allowing the conversion of UTP into CTP (Figure 1). Thus, CTP synthase is the rate-limiting enzyme for the synthesis of cytosine nucleotides from both the *de novo* and uridine -mediated salvage pathways (van Kuilenburg *et al.* 2000). This enzyme catalyzes the synthesis of CTP through an ATP-dependent reaction between UTP and an ammonia donor, usually a glutamine; products of the reaction are ADP, a phosphate group, glutamate and, of course, CTP.

It has been previously demonstrated that in many organisms, from the yeast *Saccharomices cerevisiae* through *Homo sapiens*, there are two distinct CTP synthase genes. In *S. cerevisiae* the knock out of either of these genes, named *ura7* (Ozier-Kalogeropoulos *et al.* 1991) and *ura8* (Ozier-Kalogeropoulos *et al.* 1994), is not a cause of lethality, indicating that none of them is essential for yeast survival. However, the knock out of both genes causes lethality (Ozier-Kalogeropoulos *et al.* 1994) if no cytidine is supplied to the yeast medium. It is not completely clear why two polypeptides are needed, although it is known that they are controlled differently

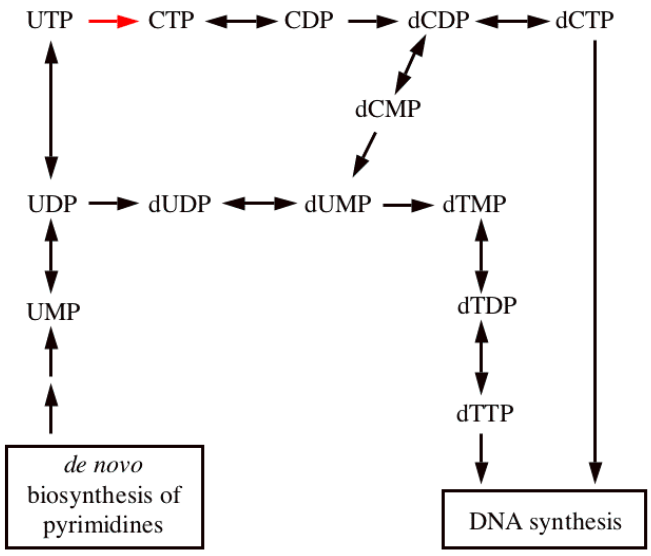

Figure 1. De novo synthesis of pyrimidines. Only metabolic intermediates are illustrated. Double-headed arrows indicate reversible reactions; one-headed arrows indicate irreversible reactions. CTP synthase controls the irreversible transformation of UTP into CTP (red arrow).

at the cellular level (Nadkarni *et al.* 1995). In general, CTP synthase is allosterically regulated by intracellular concentrations of CTP and UTP and shows its highest activity in the presence of physiological concentrations of ATP, GTP and glutamine (Kassel *et al.* 2010). CTP synthase function is regulated at the post-traductional level by phosphorylation (Carman *et al.* 2004; Higgins *et al.* 2007; Kassel *et al.* 2010) and by allosteric interactions with GTP (positive feedback) (Lunn *et al.* 2007), CTP itself (negative feedback) (Endrizzi *et al.* 2005) and 6-diazo-5-oxo-L-norleucine (DON), a non-standard amino acid which is a glutamine antagonist (inhibition) (Ahluwalia *et al.* 1990). Likely, CTP synthase function is somehow controlled also by its ability to create polymers of itself. At low enzyme concentrations and in absence of ATP and UTP, CTP synthase exists in the cell as an inactive monomer. The raising of enzyme concentration promotes the formation of a (still inactive) homodimer and then, in the presence of high concentrations of UTP, ATP and enzyme, it folds up as an active homotetramer (Anderson 1983, Goto *et al.* 2004, Robertson 1995, von der Saal *et al.* 1985). In 2010, three reports (Liu 2010, Noree *et al.* 2010, Ingerson-Mahar *et al.* 2010) showed that in both prokaryotes and eukaryotes, CTP synthase can create much bigger, mainly needle-shaped structures. They are visible at the optic microscope, and were named *cytoophidia* ('cellular snakes' in Greek). So far, cytoophidia were found in bacteria, yeasts, fruit flies, mammalian and human cells, indicating that this kind of organization is evolutionary conserved (Liu 2011). Interestingly, DON and azaserine (another glutamine analog) are both able to promote cytoophidia formation (Chen *et al.* 2011). However, it is not yet clear if the enzyme is the only component of these filaments, nor if it is functionally active inside them (Liu 2011).

Noteworthy, only few data are available about the control of CTP synthase at the transcriptional level. In *Lactococcus lactis* (Jørgensen *et al.* 2003) and *Bacillus subtilis* (Meng *et al.* 2004) the control occurs through attenuation. A similar mechanism acts also in *S. cerevisiae*, at least in the *ura8* gene (Kwapisz *et al.* 2008); cues about gene activation and/or control in higher eukaryotes are still largely missing. In *Droso-*

Table 1. Conservation of the first 53 amino acids of CTP synthase, polypeptide B. NP_648747.1 (line 1) corresponds to CTP synthase, CG6854 polypeptide B; NP_730024.1 (line 3) corresponds to the transcription factor CG6854, polypeptide A.

Accession number	Identifier	Species	Score	Query coverage	E-value	Maximum identity
NP_648747.1	CG6854	*D. melanogaster*	120	100%	1e-30	100%
XP_002030555.1	GM25504	*D. sechellia*	114	96%	2e-29	100%
NP_730024.1	CG6854	*D. melanogaster*	114	96%	2e-29	100%
XP_002094855.1	GE22048	*D. yakuba*	114	96%	2e-29	100%
XP_002134750.1	GA23623	*D. pseudoobscura*	117	96%	2e-29	100%
XP_001972796.1	GG15717	*D. erecta*	114	96%	3e-29	100%
XP_002022231.1	GL24720	*D. persimilis*	116	96%	6e-29	100%
XP_002084926.1	GD14523	*D. simulans*	115	96%	2e-28	100%
XP_001958098.1	GF23684	*D. ananassae*	115	96%	2e-28	100%
XP_001984797.1	GH14829	*D. grimshawi*	111	96%	2e-27	98%
XP_002047983.1	GJ13723	*D. virilis*	111	96%	2e-27	98%
XP_002009248.1	GI13933	*D. mojavensis*	108	96%	2e-26	96%

phila melanogaster, CTP synthase is encoded by only one gene mapping inside the CG6854 locus (Figure 2) (McQuilton *et al.* 2012).

Although CTP synthase is encoded by a single gene, it produces two mRNA isoforms by alternative splicing, that are translated into two polypeptides (B and C, respectively) that are slightly different in length (627 amino acids for polypeptide C vs. 623 amino acids for B) and composition (the C-terminal 570 amino acids are common) (Figure 1B). Interestingly, BLAST search reveals that only the first 53 amino acids in polypeptide B are specific to Drosophilids (Table 1), while the first 57 amino acids of polypeptide C (and similarly, the common portion of polypeptides B and C) are conserved among various eukaryotes, including (but not limited to) *Homo sapiens*, *Mus musculus* (mammal), *Gallus gallus* (bird), *Anolis carolinensis* (reptile), *Xenopus laevis* (amphibian), *Danio rerio* (fish), *Branchiostoma floridae* (cephalochordatum), *Daphnia pulex* (crustacean), *Caenorhabditis elegans* (worm) and *Saccharomyces cerevisiae* (yeast) (data not shown). Thus, also in *D. melanogaster* there are two different polypeptides showing CTP synthase activity as in most eukaryotes, but the fact that one of them is exclusive for Drosophilids suggests that the second form might have a specific role inside the cell, probably related to the biology of these insects. Polypeptide A, encoded by CG6854 mRNA isoforms A and D (the coding sequence being identical), is the product of another gene mapping inside the same locus; it is a transcription factor showing homology to Adf-1, Stonewall, and Dip3 transcription factors (Bhaskar & Courey 2002). Polypeptide A is involved in the expression of the Wingless signalling pathway (Song *et al.* 2010), it interacts with the SAGA complex (Weake *et al.* 2011), it is involved in neural stem cells self-renewal (Neumüller *et al.* 2011) and, likely, in embryogenesis (Michaut *et al.* 2011); its localization is intranuclear, as expected (Buszczak *et al.* 2007). Interestingly, according to the Drosophila database (McQuilton *et al.* 2012) Release FB2012_06, the coding exon 5 is shared between polypeptides A and B, and some parts of the 5'-UTR of the four transcripts are shared as well (Figure 2B).

Another interesting feature of the *D. melanogaster* CTP synthase coding gene is the presence of two long introns at the 5' end of the gene, spanning approximately 7.2 and 2.7 Kb (Figure 2). Insertional mutagenesis performed in different laboratories worldwide allowed the isolation of 34 fly lines having a transposon inserted inside the CG6854 locus (FlyBase Release FB2012_06 reports 33 mutations, and we have another one called RP5, obtained in our laboratory, illustrated in Figure 2 but not yet reported in FlyBase); of them, 31 map inside the first intron. Interestingly, the analysis of a small number of these 31 mutations (Figure 2) revealed that some insertions result in a viable and fertile phenotype, while others induce lethality at the third larval instar of development. Indirectly, this suggests that it is not the mere presence of a transposon to induce lethality: the existence of viable stocks shows that, in these lines, the RNA polymerase is able to transcribe such long stretches of RNA (1.9 Kb of coding sequence plus

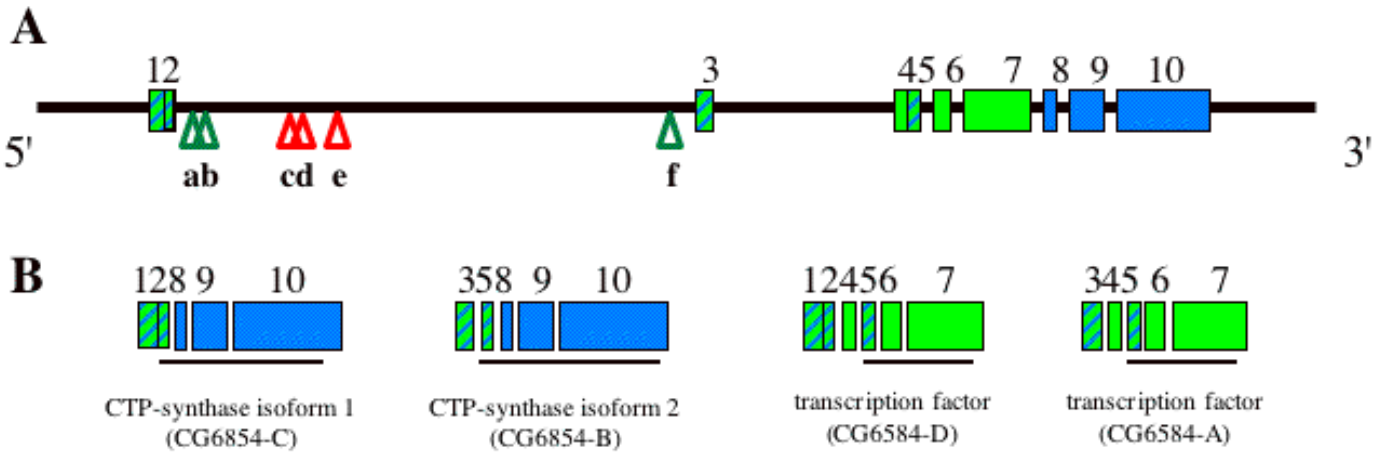

Figure 2. Molecular organization of the CG6854 locus containing the *CTP synthase* gene. Blue boxes: exons of the cytidine synthase coding gene. Green boxes: exons of the transcription factor coding gene. Striped boxes: shared exons; all data are presented according to FlyBase Release FB2012_06. Exons have the same numeration in both parts of the figure. (2A) Black thick line: introns and regions flanking the locus; note the first two long introns inside the *CTP synthase* gene, 7.2 and 2.7 Kb respectively. Triangles: *Drosophila melanogaster* lines with insertions of transposable elements, analysed for viability/lethality. Green triangles: viable and fertile transposon insertions; a: BG01116; b: 5HA1071; f: EY01546. Red triangles: lethal transposon insertions; c: EP1185; d: SH105; e: RP5 (this transposition was induced in our laboratory). (2B) Schematic representation of the four mRNA transcripts of the locus. Black thin lines indicate the extension of the coding sequences.

UTRs, 7 Kb of transposon and the length of the intron, either 7.2 or 2.7 Kb according to the isoform) and that the splicing machinery is still able to recognize this very long transcript and perform its job, allowing the final production of a functional CTP synthase coding mRNA. In fact, it has been found by rt-PCR that both mRNA isoforms are present inside the BG01116 mutant line (Figure 2A) (Ceprani 2004). Consequently, it is possible that the first intron might have a regulatory function on gene expression in some parts (identified by lethal insertions) of its sequence, but not in all of it (viable insertions). The aim of this report was to investigate *in silico* the content of the first intron (the 7.2 Kb long one), to discuss the data available from FlyBase and to integrate these data with present, original findings, in order to suggest possible ways of gene control at the transcriptional and/or translational level.

Materials and Methods

All experiments and data mining were performed using free software and databases available in the world wide web. In particular, we took advantage of the Drosophila database (FlyBase) (http://flybase.org/), which contains genomic data about several drosophilids, for the description of the CG6854 locus in *D. melanogaster* and the analysis of the corresponding locus in other *Drosophila* species; in the same web site (http://flybase.org/blast/), we also used the Basic Local Alignment Search Tool (BLAST) engine (at default settings) for the alignment of the two, high-complexity sequences found inside the first intron and for the analysis of the hairpin matches inside the *D. melanogaster* genome. The sequence of the hairpin-forming region is the following: 5'-actaaataTATGTACATACATATGTATGTACATAgatatagt-3', with the capitalized letters representing the central, 26 nt long, perfect inverted repeat. The analysis of the evolutionarily conservation of the first 53 amino acids of the CTP synthase was achieved using the National Center for Biotechnology Information (NCBI) BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi). The analysis of the stability of RNA secondary structures was performed on line as well (www.bioinfo.rpi.edu/applications/mfold/old/rna/) (Brennecke *et al.* 2003). The dot-plot analysis was carried out using various software packages retrieved from the Internet (http://molbiol-tools.ca/Alignments.htm).

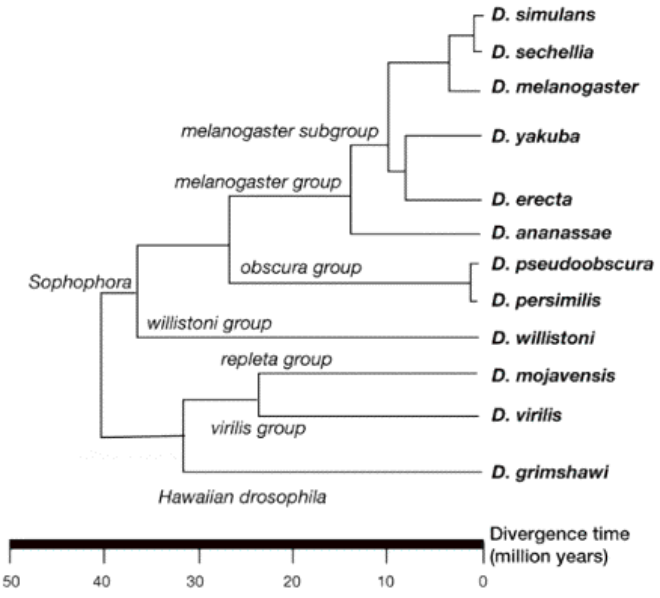

Figure 3. Phylogenetic relationships of Drosophilids belonging to the *Sophophora* group.

Results

The general organization of the CTP synthase locus is shared among drosophilids

First, it was investigated whether the complex organization of the CG6854 locus, containing a CTP synthase coding gene, another gene coding for a transcription factor and harbouring long introns, is a peculiar feature of *D. melanogaster* only, or if it is conserved among drosophilids. To verify this, we analyzed the genomes of other *Drosophila* species available in FlyBase, whose phylogenetic relationships are illustrated in Figure 3.

In most cases, the database search for CTP synthase retrieved two genes instead of one as in *D. melanogaster*. However, in all these situations, the two genes (i) are in the same chromosomal region of ~20 Kb of length; (ii) they are in the same orientation; (iii) one, usually the 5'-most, is much smaller than the other; (iv) they are separated by a long DNA spacer (the genomic region containing the two putative CTP synthase coding genes being long in all species approximately 15 Kb, thus similar to *D. melanogaster*); (v) between the two identified CTP encoding genes there is always another coding sequence showing homology with transcription factors, or the CTP synthase putative gene shows homologies with transcription factors. The only exceptions to these rules are for *Drosophila pseudoobscura* (no evidence of the presence of a transcription factor) and *Drosophila willistoni* (only one gene, and no evidence of a transcription factor). We believe that the differences between *D. melanogaste*r and the other Drosophilids are mainly caused by the lower quality of genome annotation for the latter. The fact that a CTP synthase protein shows a domain of a transcription factor may be interpreted as an error of the automated software analysis, which

joined two different genes into one unit; similarly, the transcription factor might have not been recognized by the software in some species. Thus, taken together, these data support the hypothesis that in all these cases the two CTP synthase genes are probably just the two parts of the same gene as it happens in *D. melanogaster*. In conclusion, it is likely that all species reported in Figure 3 show a local molecular organization similar to the CG6854 locus, therefore including (i) a CTP synthase coding gene; (ii) a transcription factor coding gene; (iii) at least one long intron at the 5' end of the CTP synthase coding sequence.

Identification of new homologies inside the intron. As described in the Introduction, in both bacteria and yeast the CTP synthase gene expression is controlled through RNA secondary structures, which are able to interact with RNA polymerase II, altering its processivity (Jørgensen *et al.* 2003, Kwapisz *et al.* 2008, Meng *et al.* 2004). Thus, a first approach to identify 'interesting' regions inside the first intron was to evaluate the stability of the hypothetical RNA produced during the transcription, with the rationale that non-repetitive, high-complexity sequences might fold into stable double-stranded structures, allowing for their identification. To perform this task, the intronic sequence was analysed in blocks of 800 nucleotides with a 160 nucleotides overlap. In other words, calling +1 the first nucleotide after the first exon/intron junction, the stability of the sequences +1/+800, +640/+1440, +1280/+2080, +1920/+2720 and so on, plus regions –640/+160 and +7280/+8080 (that include part of the flanking exons), has been evaluated. For each sequence, the best score in terms of D G value was

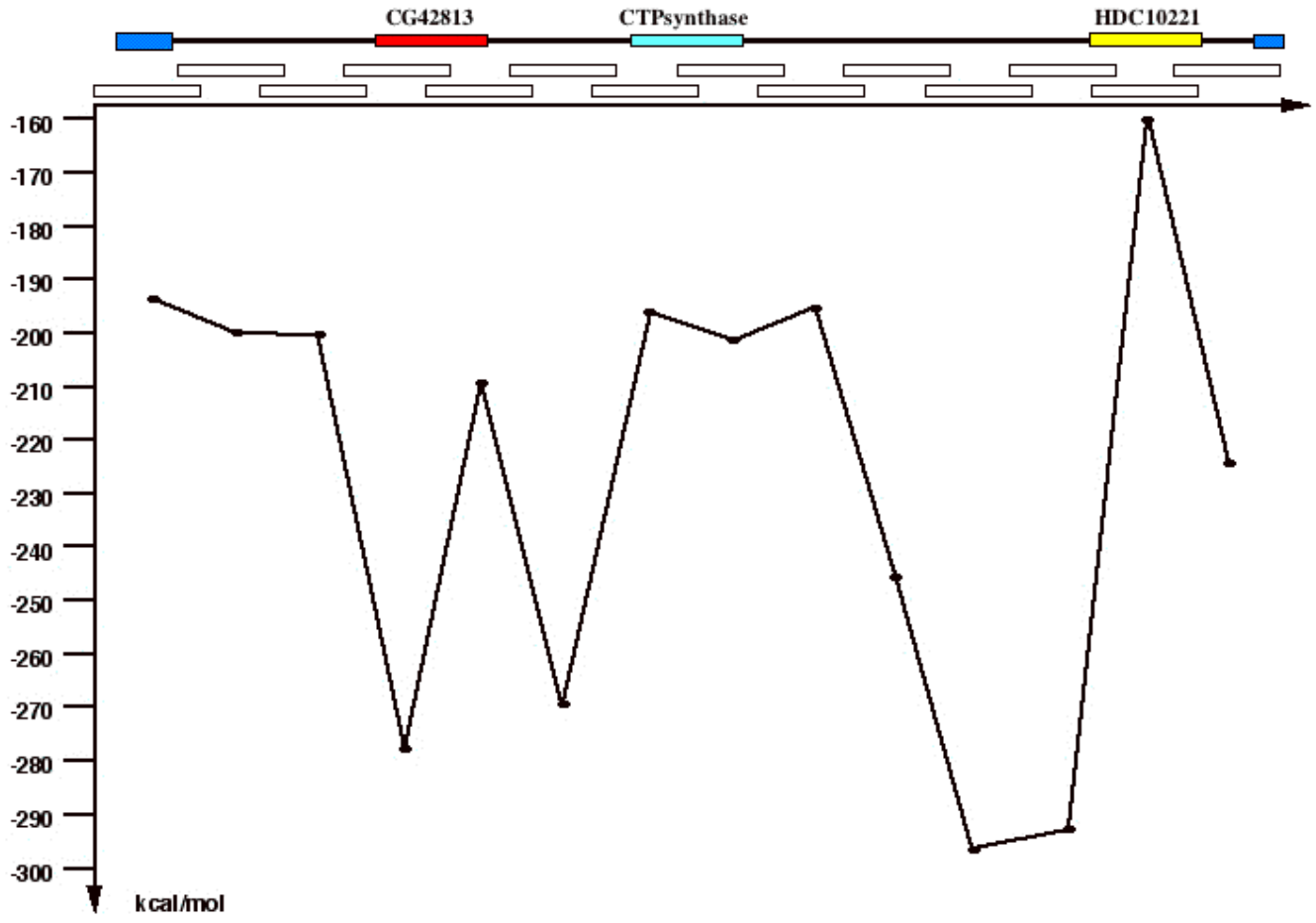

Figure 4. Analysis of RNA secondary structure stability of the CG6854 first intron. Open rectangles: the position of 800 nucleotides stretches with 160 nucleotides overlap, used for the stability analysis. For each region of 800 nucleotides, the best score in terms of DG value (kcal/mol) was plotted in the lower diagram, which was then paired to the sequence of the first intron of the *GC6854* locus (top line, in colors; the arrow in the plot indicates the 5'-3' gene orientation). Each colour represents different parts of the genomic region. Blue: exons flanking the first intron. Red: the region homologous to *CG42813*. Pale blue: the region homologous to the terminal part of the CTP synthase gene. Yellow: the region corresponding to *HDC10221* putative open reading frame. Black thick line: other parts of the intron.

then considered (Figure 4). This allowed the detection of two regions with particularly low values of free energy, identifying putative high complexity sequences.
A deeper analysis of accessible data in FlyBase allowed identifying, next to the low energy region at the 3' end of the intron, a sequence called *HDC10221* (GenBank: BK002148.1); an inferred open reading frame containing one small intron and potentially coding a polypeptide of 201 amino acids showing no evident homologies with other known proteins. Since there is at least one viable and fertile *D. melanogaster* mutant (namely EY01546) with a transposable element inserted inside the putative coding region, this sequence was not investigated any further, since it was considered unnecessary for gene expression and fly viability. As for the low energy region located at the 5' end of the intron, no coding sequences are reported in FlyBase, thus it was aligned against the *D. melanogaster* genome, to verify the presence of external homologies. Quite interestingly, this search allowed identifying a homology with the 3'-UTR region (plus part of the following DNA spacer) of the gene *CG42813*, of unknown function but coding for a protein containing a double NUDIX hydrolase domain (Lin *et al.* 2009) suggesting a nucleoside diphosphate pyrophosphatase activity (McLennan 2006, Mildvan *et al.* 2005). Comparing the whole sequence, the homology spans 620 nucleotides (excluding an internal region without homology), with 86% identity (533/620 nucleotides); limiting the analysis to the CG42813 locus alone (without spacer DNA), the homology spans 418 nucleotides, with 83% identity (347/418 nucleotides) (Figure 5).

Aligning the gene vs. itself reveals new internal homologies.

To verify the presence of other homology regions, a dot-plot analysis of the CG6854 locus against itself was performed. This analysis led to the discovery that the first intron contains a duplication of the 3' end of the *CTP synthase* gene itself, but in reverse orientation. This sequence is located between the other two, abovementioned, high complexity regions (Figure 4) and spans a length of 401 nucleotides (85% identity, 341/401 nucleotides) partly overlapping the 3'UTR of the gene (90% identity, 213/236 nucleotides) and the following intergenic spacer (Figure 6).

Besides other shorter regions and repetitive sequences, this analysis also allowed the identification of a stretch of 26 nucleotides representing a perfect inverted repeat able to create a hairpin structure, surrounded by other 16 nucleotides with a lower homology but still able to take part in this structure (Figure 7).

Analysis of the entire 42 nucleotides long sequence using the BLAST software available in FlyBase web site, revealed that in the *D. melanogaster* genome, there are a total of 412 different target sequences (excluding partial, duplicated overlaps due to the inverted repeat, and excluding false positives due

Figure 5. Alignment of the first intron of the *CTP synthase* gene with the 3'-UTR region and the following DNA spacer of the *CG42813* gene. Vertical lines: perfect matches; colon: conserved pyrimidine/purine; dots: non conserved positions; dashes: gaps. Upper line: CTP synthase sequence; lower line: CG42813 sequence.

to reverse orientations) having at least 20 matches with the 42 nucleotide query; the choice of this threshold length is based on known data from literature about siRNA. Among these 412 targets, 19 of them map inside exonic sequences, for a total of 21 genes involved (the different values are due to shared targets i.e. overlapping genes) (Table 2).

Interestingly, looking at the temporal expression of these genes, two thirds of them (14/21, 66.7%) are expressed during embryogenesis; of the remaining seven, just two (9.5%) are expressed in adults only, and for the remaining five no data are available. Moreover, looking at their biological function, genes involved in morphogenesis (tissue differentiation and/or cell differentiation, shape and motility) are 8/21 (38%) and raise up to 9/22 (40.9%) if the CG6854 locus is included. More specifically, half of them (i.e., 4/21 genes or 19%), are involved in neurogenesis and neuronal function. Another interesting fact is that 5/21 genes (23.8%) are involved in post-translational modifications of target proteins (three kinases and two peptidases).

Discussion

The CG6854 locus, as described to date in FlyBase, shows a complex organization, as it harbors two genes: a CTP synthase coding gene with two splicing forms, and a transcription factor with two splicing forms and part of its sequence shared with CTP synthase isoforms (Figure 2). The locus is also characterized by the presence of a long first intron (approximately 7.2 Kb). In *D. melanogaster* there are two types of introns, according to their size: short (less than 86 bp, with an average length of 61±10 bp) and long (more that 86 bp). The short introns are characterized by splicing mechanisms different from those used for the long ones (Mount *et al.* 1992; Yu *et al.* 2002). Although short introns are more numerous inside the fly genome, they only represent a small fraction of total intronic DNA, since long introns may span several kilobases in length. Ten years ago it was hypothesized that long introns should be negatively selected by evolution since the transcription of unnecessary long sequences is costly (Castillo-Davis *et al.* 2002); indeed, it was also demonstrated that, in general, long introns are negatively selected in active chromosomal domains (Marais *et al.* 2005, Prachumwat *et al.* 2004). Apparently, these features do not fit with our data: the overall organization of the locus is conserved among drosophilids (present report) and the CTP synthase is clearly an essential protein, thus the gene is functional in all tissues during the whole fly lifetime, especially in those having actively replicating cells. As a consequence, the first intron of CG6854 cannot be considered "unnecessary", and this is also supported by the presence of transposable element insertions causing lethality and mapping inside it. This contradiction may be overcome recalling that "first introns" indeed behave differently from the rest. They are usually longer than other long introns (on average, 2.7x longer) (Bradnam & Korf 2008) and they probably harbor sequences necessary for transcription regulation (Bradnam & Korf 2008, Duret 2001, Marais *et al.* 2005, Parsch 2003). In fact, separated analysis of first and non-first introns revealed that the former are positively correlated to gene expression (Marais *et al.* 2005). Moreover, long first introns also are under evolutionarily constraints, since they evolve more slowly than both non-first long introns and short ones, with a direct correlation between length and conservation (Haddrill *et al.* 2005).

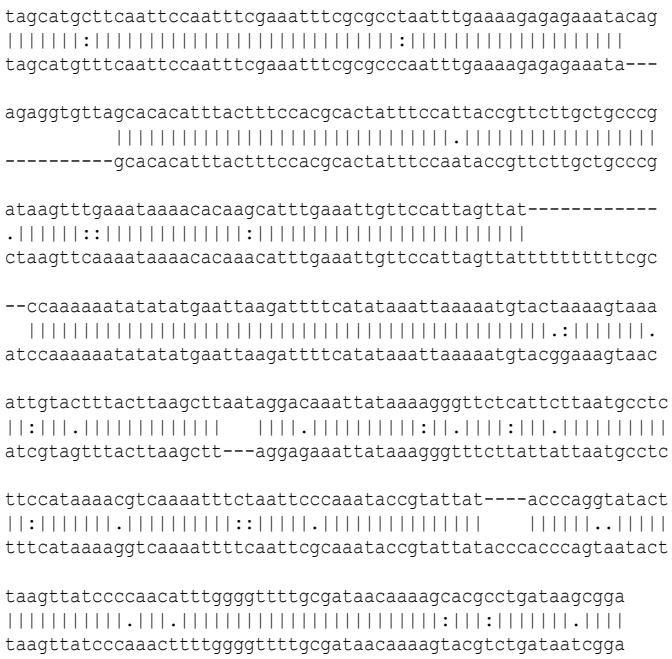

Figure 6. Alignment of the first intron of the *CTP synthase* gene with the 3'-UTR region and following DNA spacer of the CG6854 locus. Vertical lines: perfect matches; colon: conserved pyrimidine/purine; dots: non conserved positions; dashes: gaps. Upper line: first intron sequence; lower line: 3'-UTR plus following the DNA spacer of the CTP synthase sequence.

An in-depth analysis of this intron in the present report actually revealed that it harbors at least three sub-sequences of interest, which are not yet reported in the annotated genome. Starting from the 5'-end of the gene, the first sequence is a partial copy of another gene, namely *CG42813*, probably a nucleoside diphosphate pyrophosphatase. This fact is quite interesting not only *per se*, but also because both *CTP synthase* and *CG42813* genes are involved in nucleotide

Table 2. Exon targets of the 42nt hairpin-forming region in the genome of *D. melanogaster*. Lines in the table are listed according to descending BLAST score values. The first line represents the BLAST query, using the intronic sequence inside the *CTP synthase* gene (locus: CG6854). Identity: if numbers are different, internal mismatches are present. Overlap: positions are referred to the 42 nucleotides query from the *CTP synthase* intron. Genes: those listed consecutively and marked with an asterisk share the same target sequence (overlapping genes). Notes: (i) if two genes overlap, they are reported separately; (ii) if the same target belongs to both an intron and an UTR (splicing alternatives), the latter is considered for the target position. Molecular functions, biological process and temporal expression are reported according to FlyBase, Release FB2012_03.

Score	Identity	Overlap	Chromosome – gene	Position	Molecular function	Biological process	Temporal expression
83.7518	42/42	1-42	3L-CG6854	intron	CTP synthase	nucleotide biosynthesis; neurogenesis	unknown
52.0341	26/26	6-31	3L-bab2	3'-UTR	DNA binding; transcription	development; morphogenesis	embryo and early pupa
52.0341	26/26	9-34	2L-CG7227	3'-UTR	scavenger receptor	defense response	embryo
44.1047	25/26	1-26	2R-gprs	3'-UTR	unknown	unknown	embryo and early larva
42.1223	21/21	14-34	X-SK	3'-UTR	unknown	copulation (morphogenesis?)	unknown
40.1400	29/32	8-39	3L-CR43470*	exon	unknown	unknown; non-coding RNA	unknown
40.1400	29/32	8-39	3L-CG14830*	3'-UTR	unknown	unknown	embryo and late pupa
40.1400	26-28	6-33	3L-Mes2	5'-UTR	unknown	embryo and larva development	early embryo and adult female
40.1400	26/28	8-35	2R-CG11163	5'-UTR	zinc ion transmembrane transporter	transmembrane cation transport	embryo, late larva and early pupa
40.1400	23/24	17-40	2R-Pkn	5'-UTR	protein kinase	embryo dorsal closure; wing development	embryo, late larva, pupa, adult female
38.1576	25/27	7-33	3L-fax	3'-UTR	unknown	axonogenesis; neurogenesis	embryo, late larva and pupa
38.1576	22/23	6-28	X-Edem1	3'-UTR	mannosyl-oligosaccharide 1,2 -alpha-mannosidase	determination of adult lifespan	early embryo
36.1753	21/22	7-28	X-Fur2	5'-UTR	serine-type endopeptidase	proteolysis	embryo, adult female
34.1929	20/21	13-33	3L-Wnk	5'-UTR	protein serine/ threonine kinase	axon guidance	early embryo, late larva, pupa, adult
34.1929	23/25	6-30	X-Rph	5'-UTR	protein transporter	synaptic vesicle exo- and endo-cytosis	early embryo
34.1929	20/21	3-23	X-Rbp2	3'-UTR	mRNA binding; translation initiation factor	translational initiation	unknown
34.1929	23/25	6-30	2L-CG4629	3'-UTR	serine/threonine kinase	regulation of cell shape; cell adhesion	late pupa and adult male
34.1929	23/25	6-30	2L-Sur	5'-UTR	unknown	central nervous system development	unknown
34.1929	23/25	5-29	3R-CG2006*	3'-UTR	unknown	unknown	early embryo and early larva
34.1929	23/25	5-29	3R-Spase12*	3'-UTR	peptidase	signal peptide processing	unknown
32.2105	22/24	7-30	X-Zw	3'-UTR	glucose-6-phosphate dehydrogenase	pentose-phosphate shunt	embryo, late pupa, adult
32.2105	22/24	6-29	X-Sdic1	3'-UTR	microtubule motor	microtubule-based movement	adult male

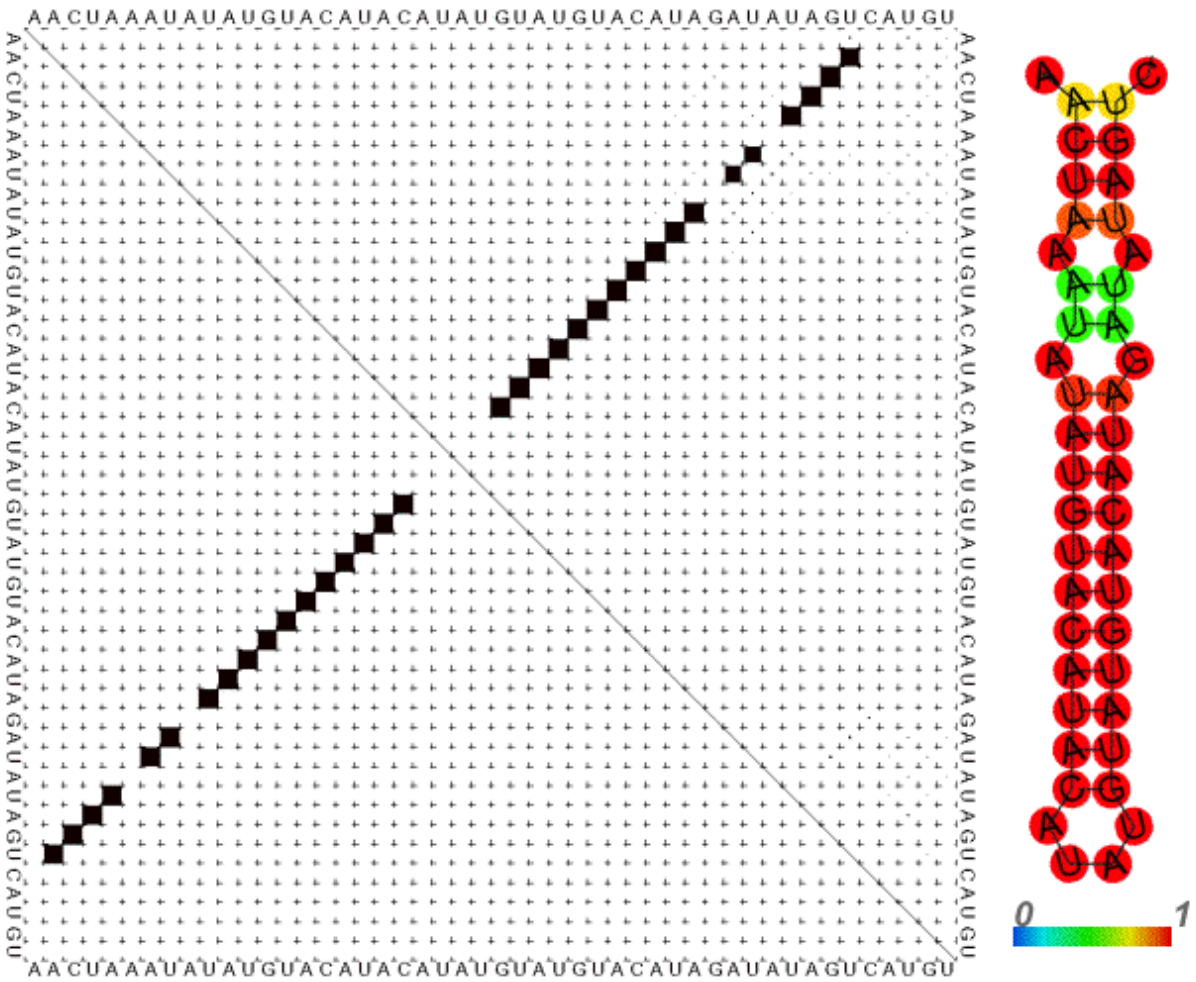

Figure 7. The first intron harbours a sequence able to form a hairpin. Left: dot plot analysis; right: structure of the hairpin. Colours in the hairpin reflect the base-pairing probability of each nucleotide position against the surrounding nucleotides (see colour scale at the bottom of the hairpin structure). In particular, for paired regions the colour denotes the probability of being paired; for unpaired regions the colour denotes the probability of being unpaired. In both cases, the red colour marks the highest probability and the blue colour marks the lowest probability.

metabolism, suggesting that the presence of this duplication is not casual. The second sequence is a copy of the *CTP synthase* gene itself, with homology overlapping both isoforms, but in reverse orientation, compared to the main transcript. The third sequence is a perfect inverted repeat of 26 nucleotides, surrounded by 16 other nucleotides, able to fold into a complex hairpin; interestingly, the same sequence partially matches the exons of a group of genes mostly involved in embryogenesis and morphogenesis, with an enrichment (4 targets) in genes involved in neuronal formation/function, a task in which the transcription factor mapping inside CG6854 is involved as well. Some questions arise. Do these sequences have a biological meaning? If so, how do they exert their function? And is there a reason why they are inside "this" locus?

In *D. melanogaster*, the same locus encodes two CTP synthase polypeptides and the cDNA analysis from FlyBase (Figure 8) reveals that they are formed by alternative splicing; indeed, both mRNA isoforms are transcribed in the wild type (Ceprani 2004).

This implies that, when the RNA polymerase transcribes the 5'-most isoform (encoding polypeptide C), the intron is also transcribed, and, consequently, the same applies to the antisense strand of the *CTP synthase* gene, present inside the cell. Since both isoforms would be affected by it (both share the 3'-end, Figure 2), in theory the gene might be non-functional because of the presence of both sense and antisense RNA strands. Of course, this is not true – the gene works fine in the wild type. Data presented here allow only for a complex explanation: it is possible to hypothesize the presence of some other regulatory element that (i) might be able to block the antisense RNA and allow the sense RNA to be regularly translated in case of necessity, but also (ii) allow antisense formation if CTP synthase is not required (for example, in the presence of a high CTP concentration or in the ab-

sence of cell/DNA replication). The easiest way to block an antisense RNA is to transcribe an anti-antisense sequence, targeting it. A specific search in FlyBase for all ESTs mapping inside the CG6854 region reveals the presence of a putative transcribed antisense RNA inside the intron (Figure 8, red sequences indicated by blue arrows). As shown, there are at least four such sequences: one upstream the 5' end of the first exon of isoform C, and three inside the first intron itself. Moreover, the last FlyBase update (Release FB2012_06, November 2012) also indicates the presence of two putative long non-coding RNAs (lncRNA) inside the first intron, namely CR43972 and CR43973, which are transcribed in reverse orientation compared to the main transcript (Figure 8). It is thus tempting to imagine that these ESTs are part of the same, longer, antisense transcript, likely spanning the entire first intron and maybe even more (another putative lncRNA named CR43971 is partly located inside the second intron, and shows the same orientation to the other two) (Figure 8). If these sequences are validated to be part of a longer transcription unit and not a mere computational error, it would be possible to envisage a genetic system potentially able to transcribe a gene (*CTP synthase*), its antisense, and its anti-antisense from the same locus, providing a complex yet very accurate way to control CTP synthase concentration inside the cell.

At the same time, the intron also allows the transcription of a sequence partially homologous to the *CG42813* gene (again, the homology is inside the 3'-end of the gene plus the DNA spacer, similarly to CTP synthase). Since this sequence is transcribed in the same orientation of the original gene, this should not interfere with its function. But if the reverse strand is also transcribed, then also *CG42813* might be under the control of an antisense transcript, in a way that when CTP synthase levels are high, levels of CG42813 protein are low, and vice versa. Therefore, this genetic system might control with the same mechanism, but with opposite effects, two different steps of nucleotide metabolism. Why should these two proteins have negatively related levels? At the moment there are no clues to an answer, the identification of the function of *CG42813* will be necessary for the comprehension of this relationship.

As for the inverted repeat able to form a hairpin structure, its presence also makes sense in this context. Present data show that a group of 21 genes (i) have an exonic sequence (mostly inside the UTRs) partly matching it; (ii) are mostly active during embryogenesis and morphogenesis; (iii) are enriched in

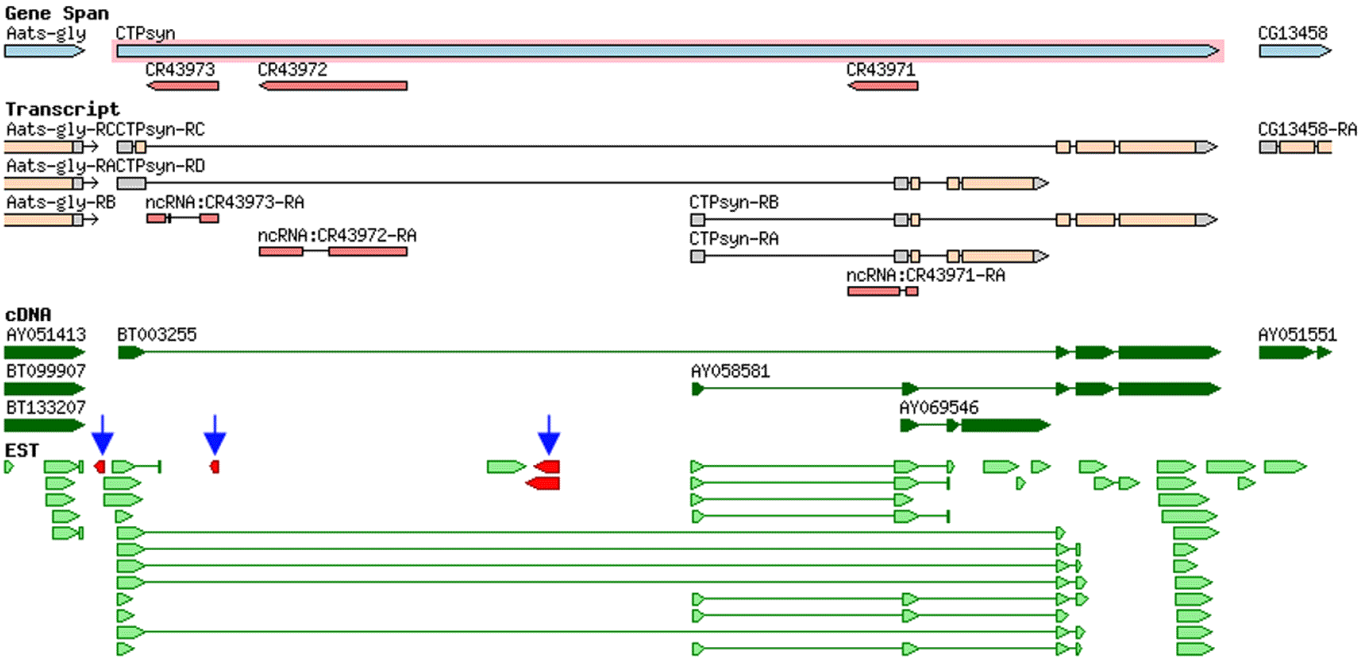

Figure 8. Antisense sequences transcribed inside the first intron of the CG6854 locus. The picture represents an elaboration of a partial snapshot of the FlyBase web page showing the genomic region containing the CG6854 locus (FlyBase Release FB2012_06). From top to bottom: genes mapping in the region (CTP synthase, blue, and three putative long non-coding RNAs, pink); mRNAs transcribed inside the CG6854 locus (two CTP synthase isoforms, RB and RC, orange; two transcription factor isoforms, RA and RD, orange; three long non-coding RNAs, pink); cDNAs, two for the CTP synthase (encoding polypeptides B and C), one for the transcription factor (encoding polypeptide A), all in dark green; ESTs (brilliant green: sense sequences supporting the mRNA models illustrated above; red: antisense sequences; blue arrows highlight these putative antisense sequences).

neurogenesis and neuron function. During embryogenesis there is intense cell duplication, and consequently fast DNA replication, requiring a high nucleotide concentration. This creates a hypothetical link between the CTP synthase and them. Moreover, the CG6854 locus encodes a transcription factor that is involved in neurogenesis and morphogenesis. This creates a link between this protein (polypeptide A) and the latter. Thus, this sequence might also play a specific function during embryogenesis, for DNA replication and neural system development, and is likely not inside this locus just by chance. Its effects might also be amplified, recalling that five targets fall inside genes coding for proteins involved in post-translational modifications. Further analyses are required to verify if this regulation indeed occurs, and if the mechanism involved is gene silencing, activation, or both, since in almost all cases they are inside UTR regions (Thomson *et al.* 2011). However, it is noteworthy that, being an inverted repeat, this sequence should not be influenced by sense or antisense transcription. In conclusion, this intron might be in the center of a complex network of interacting genetic functions regulated by complex relationships among cell status, protein levels, mRNA levels and the presence/absence of regulatory non-coding RNAs.

Acknowledgements

I am deeply grateful to Prof. A. Chabes (Umeå University, Sweden) for his invaluable help during the writing of this report, and for his critical reading of the manuscript. The present work was performed partly in Prof. Chabes' laboratory, and partly in Prof. Gatti's laboratory (Sapienza University, Rome); I thank both of them for the possibility to use their IT resources.

Conflicts of Interest

None declared.

References

Ahluwalia GS, Grem JL, Hao Z & Cooney DA 1990 Metabolism and action of amino acid analog anticancer agents. *Pharmacol Ther* **46** 243-271.

Anderson PM 1983 CTP synthetase from *Escherichia coli*: an improved purification procedure and characterization of hysteretic and enzyme concentration effects on kinetic properties. *Biochemistry* **22** 3285-3292.

Bhaskar V & Courey AJ 2002 The MADF-BESS domain factor Dip3 potentiates synergistic activation by *Dorsal* and *Twist*. *Gene* **299** 173-184.

Bradnam KR & Korf I 2008 Longer first introns are a general property of eukaryotic gene structure. *PLoS ONE* **3** e3093.

Brennecke J, Hipfner DR, Stark A, Russell RB & Cohen SM 2003 *bantam* encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene *hid* in *Drosophila*. *Cell* **113** 25-36.

Buszczak M, Paterno S, Lighthouse D, Bachman J, Planck J, Owen S, Skora AD, Nystul TG, Ohlstein B, Allen A, Wilhelm JE, Murphy TD, Levis RW, Matunis E, Srivali N, Hoskins RA & Spradling AC 2007 The Carnegie protein trap library: a versatile tool for *Drosophila* developmental studies. *Genetics* **175** 1505-1531.

Carman GM & Kersting MC 2004 Phospholipid synthesis in yeast: regulation by phosphorylation. *Biochem Cell Biol* **82** 62-70.

Castillo-Davis CI, Mekhedov SL, Hartl DL, Koonin EV & Kondrashov FA 2002 Selection for short introns in highly expressed genes. *Nat Genet* **31** 415-418.

Ceprani F 2004 Identificazione e caratterizzazione di geni necessari per la stabilità cromosomica in *Drosophila melanogaster*. PhD Thesis, Sapienza - Università di Roma, pp. 48-49.

Chen K, Zhang J, Tastan ÖY, Deussen ZA, Siswick MY & Liu JL 2011 Glutamine analogs promote cytoophidium assembly in human and *Drosophila* cells. *J Genet Genomics* **38** 391-402.

Costi MP, Ferrari S, Venturelli A, Calò S, Tondi D & Barlocco D 2005 Thymidylate synthase structure, function and implication in drug discovery. *Curr Med Chem* **12** 2241-2258.

Duret L 2001 Why do genes have introns? Recombination might add a new piece to the puzzle. *Trends Genet* **17** 172-175.

Endrizzi J, Kim H, Anderson PM & Baldwin EP 2005 Mechanisms of product feedback regulation and drug resistance in cytidine triphosphate synthetases from the structure of a CTP-inhibited complex. *Biochemistry* **44** 13491-13499.

Goto M, Omi R, Nakagawa N, Miyahara I & Hirotsu K. 2004 Crystal structures of CTP synthetase reveal ATP, UTP, and glutamine binding sites. *Structure* **12** 1413-1423.

Haddrill PR, Charlesworth B, Halligan DL & Andolfatto P 2005 Patterns of intron sequence evolution in *Drosophila* are dependent upon length and GC content. *Genome Biol* **6** R67.

Higgins ML, Graves PR & Graves LM 2007 Regulation of human cytidine triphosphate synthetase 1 by glycogen synthase kinase 3. *J Biol Chem* **282** 29493-29503.

Hofer A, Crona M, Logan DT & Sjöberg BM 2012

DNA building blocks: keeping control of manufacture. *Crit Rev Biochem Mol Biol* **47** 50-63.

Ingerson-Mahar M, Briegel A, Werner JN, Jensen GJ & Gitai Z 2010 The metabolic enzyme CTP synthase forms cytoskeletal filaments. *Nat Cell Biol* **12** 739-746.

Jørgensen CM, Hammer K & Martinussen J 2003 CTP limitation increases expression of CTP synthase in *Lactococcus lactis*. *J Bacteriol* **185** 6562-6574.

Kassel KM, Au da R, Higgins MJ, Hines M & Graves LM 2010 Regulation of human cytidine triphosphate synthetase 2 by phosphorylation. *J Biol Chem* **285** 33727-33736.

Kwapisz M, Wery M, Després D, Ghavi-Helm Y, Soutourina J, Thuriaux P & Lacroute F 2008 Mutations of RNA polymerase II activate key genes of the nucleoside triphosphate biosynthetic pathways. *EMBO J* **27** 2411-2421.

Lin J, Hu Y, Tian B & Hua Y 2009 Evolution of double MutT/Nudix domain-containing proteins: similar domain architectures from independent gene duplication-fusion events. *J Genet Genomics* **36** 603-610.

Liu JL 2010 Intracellular compartmentation of CTP synthase in *Drosophila*. *J Genet Genomics* **37** 281-296.

Liu JL 2011 The enigmatic cytoophidium: compartmentation of CTP synthase via filament formation. *Bioessays* **33** 159-164.

Lunn FA, MacDonnell JE & Bearne SL 2007 Structural requirements for the activation of *Escherichia coli* CTP synthase by the allosteric effector GTP are stringent, but requirements for inhibition are lax. *J Biol Chem* **283** 2010-2020.

Marais G, Nouvellet P, Keightley PD & Charlesworth B 2005 Intron size and exon evolution in *Drosophila*. *Genetics* **170** 481-485.

McLennan AG 2006 The Nudix hydrolase superfamily. *Cell Mol Life Sci* **63** 123-143.

McQuilton P, St. Pierre SE, Thurmond J & the FlyBase Consortium 2012 FlyBase 101 – The basics of navigating FlyBase. *Nucleic Acids Res* **40** (Database issue) D706-714.

Meng Q, Turnbough CL Jr & Switzer RL 2004 Attenuation control of pyrG expression in *Bacillus subtilis* is mediated by CTP-sensitive reiterative transcription. *Proc Natl Acad Sci USA* **101** 10943-10948.

Michaut L, Jansen HJ, Bardine N, Durston AJ & Gehring WJ 2011 Analyzing the function of a *hox* gene: an evolutionary approach. *Dev Growth Differ* **53** 982-993.

Mildvan AS, Xia Z, Azurmendi HF, Saraswat V, Legler PM, Massiah MA, Gabelli SB, Bianchet MA, Kang LW & Amzel LM 2005 Structures and mechanisms of Nudix hydrolases. *Arch Biochem Biophys* **433** 129-143.

Mount SM, Burks C, Hertz G, Stormo GD, White O & Fields C 1992 Splicing signals in *Drosophila*: intron size, information content and consensus sequences. *Nucleic Acids Res* **20** 4255-4262.

Nadkarni AK, McDonough VM, Yang WL, Stukey JE, Ozier-Kalogeropoulos O & Carman GM 1995 Differential biochemical regulation of the *URA7*- and *URA8*-encoded CTP synthetases from *Saccharomyces cerevisiae*. *J Biol Chem* **270** 24982-24988.

Neumüller RA, Richter C, Fischer A, Novatchkova M, Neumüller KG & Knoblich JA 2011 Genome-wide analysis of self-renewal in *Drosophila* neural stem cells by transgenic RNAi. *Cell Stem Cell* **8** 580-593.

Noree C, Sato BK, Broyer RM & Wilhelm JE 2010 Identification of novel filament-forming proteins in *Saccharomyces cerevisiae* and *Drosophila melanogaster*. *J Cell Biol* **190** 541-551.

Ozier-Kalogeropoulos O, Fasiolo F, Adeline MT, Collin J & Lacroute F 1991 Cloning, sequencing and characterization of the *Saccharomyces cerevisiae URA7* gene encoding CTP synthetase. *Mol Gen Genet* **231** 7-16.

Ozier-Kalogeropoulos O, Adeline MT, Yang WL, Carman GM & Lacroute F 1994 Use of synthetic lethal mutants to clone and characterize a novel CTP synthetase gene in *Saccharomyces cerevisiae*. *Mol Gen Genet* **242** 431-439.

Parsch J 2003 Selective constraints on intron evolution in *Drosophila*. *Genetics* **165** 1843-1851.

Prachumwat A, DeVincentis L & Palopoli MF 2004 Intron size correlates positively with recombination rate in *Caenorhabditis elegans*. *Genetics* **166** 1585-1590.

Robertson JG 1995 Determination of subunit dissociation constants in native and inactivated CTP synthetase by sedimentation equilibrium. *Biochemistry* **34** 7533-7541.

Song H, Goetze S, Bischof J, Spichiger-Haeusermann C, Kuster M, Brunner E & Basler K. 2010 *Coop* functions as a corepressor of *Pangolin* and antagonizes *Wingless* signaling. *Genes Dev* **24** 881-886.

Thomson DW, Bracken CP & Goodall GJ 2011 Experimental strategies for microRNA target identification. *Nucleic Acids Res* **39** 6845-6853.

van Kuilenburg AB, Meinsma R, Vreken P, Waterham HR & van Gennip AH 2000 Isoforms of human CTP synthetase. *Adv Exp Med Biol* **486** 257-261.

von der Saal W, Anderson PM & Villafranca JJ 1985 Mechanistic investigations of *Escherichia coli* cytidine -5'-triphosphate synthetase. Detection of an intermediate by positional isotope exchange

experiments. *J Biol Chem* **260** 14993-14997.
Weake VM, Dyer JO, Seidel C, Box A, Swanson SK, Peak A, Florens L, Washburn MP, Abmayr SM & Workman JL 2011 Post-transcription initiation function of the ubiquitous SAGA complex in tissue-specific gene activation. *Genes Dev* **25** 1499-1509.
Yu J, Yang Z, Kibukawa M, Paddock M, Passey DA & Wong GK-S 2002 Minimal introns are not "junk". *Genome Res* **12** 1185-1189.