Biochimica et Biophysica Acta 1829 (2013) 296–305

Contents lists available at SciVerse ScienceDirect

Biochimica et Biophysica Acta

journal homepage: www.elsevier.com/locate/bbagrm

Review

Identification of RNA polymerase III-transcribed genes in eukaryotic genomes ☆

Giorgio Dieci [a,*], Anastasia Conti [a], Aldo Pagano [b,c], Davide Carnevali [a]

[a] Dipartimento di Bioscienze, Università degli Studi di Parma, Parco Area delle Scienze 23/A, 43124 Parma, Italy
[b] Dipartimento di Medicina Sperimentale, Università degli Studi di Genova, Italy
[c] IRCCS AOU San Martino, IST, Genova, Italy

ARTICLE INFO

Article history:
Received 30 August 2012
Received in revised form 20 September 2012
Accepted 21 September 2012
Available online 2 October 2012

Keywords:
RNA polymerase III
ncRNA
tRNA
SINE
TFIIIC
ChIP-seq

ABSTRACT

The RNA polymerase (Pol) III transcription system is devoted to the production of short, generally abundant noncoding (nc) RNAs in all eukaryotic cells. Previously thought to be restricted to a few housekeeping genes easily detectable in genome sequences, the set of known Pol III-transcribed genes (class III genes) has been expanding in the last ten years, and the issue of their detection, annotation and actual expression has been stimulated and revived by the results of recent high-resolution genome-wide location analyses of the mammalian Pol III machinery, together with those of Pol III-centered computational studies and of ncRNA-focused transcriptomic approaches. In this article, we provide an outline of distinctive features of Pol III-transcribed genes that have allowed and currently allow for their detection in genome sequences, we critically review the currently practiced strategies for the identification of novel class III genes and transcripts, and we discuss emerging themes in Pol III transcription regulation which might orient future transcriptomic studies. This article is part of a Special Issue entitled: Transcription by Odd Pols.

1. Introduction

The RNA polymerase (Pol) III transcription machinery is devoted to the production of non-protein coding (nc) RNAs of small size, whose transcription units are frequently present in multiple copies in eukaryotic genomes. About 400 transcription units are targeted by the RNA polymerase (Pol) III transcription machinery in the *Saccharomyces cerevisiae* genome, a number that approaches 1000 in the human and mouse genomes [1]. The most abundant products of Pol III-dependent transcription are the different species of tRNAs, functionally differing from each other for the ability to charge different amino acids corresponding to different anticodons, and the 5S rRNA, generally encoded by one or a few hundreds of identical transcription units. These RNAs, together with the Pol I-synthesized 5.8S, 28S and 18S rRNAs, are fundamental components of the protein synthesis machinery.

In addition to the abundant ncRNAs involved in translation, whose synthesis represents the major contribution to Pol III workload in eukaryotic cells, Pol III has long been known to synthesize a small, heterogeneous set of ncRNAs, that are generally abundant and are involved in different cellular processes, from protein translocation to rRNA and tRNA processing. The known set of non-tRNA/non-rRNA genes transcribed by Pol III has remarkably expanded during the last decade, especially thanks to the results of transcriptome analyses, genome-wide location studies of transcription factors and computational searches for Pol III regulatory elements in eukaryotic genomes. Recent widening of the known Pol III transcriptome has been previously reviewed [2], but a wealth of significant studies published in the last three years (in particular studies based on chromatin immunoprecipitation followed by sequencing (ChIP-seq) applied to the Pol III machinery in mammalian cells) have further expanded it, allowing not only to identify novel class III genes, but also to shed light on unexpected features of Pol III-targeted loci that complicate the view of their transcriptional regulation. Recent reviews have addressed in detail the novel Pol III-related issues revealed by ChIP-seq studies in mammals, including the cell type-specific variation of Pol III occupancy of tRNA genes in spite of their sharing the same core promoters, the overlap between Pol III and Pol II occupancy at class III gene loci, and the widespread occurrence of TFIIIC-only-associated loci not corresponding to Pol III transcription units [1,3]. In this review, we will provide an outline of the computational and the empirical strategies that, historically and up to the most recent studies, have allowed to establish inventories of Pol III-transcribed genes and corresponding transcripts in eukaryotic genomes, pointing out advantages and limitations of the different approaches, as well as critical issues in the structural annotation of the different types of class III genes. We will also provide an update of newly identified Pol III transcripts whose knowledge further contributes to our understanding of the Pol III-dependent gene expression network.

2. Computational search algorithms for Pol III-transcribed genes

2.1. tRNA genes

The most striking and easily recognizable DNA sequence feature of class III genes is the presence, in a significant subset of them – in particular in all tRNA genes – of two internal control regions,

☆ This article is part of a Special Issue entitled: Transcription by Odd Pols.
* Corresponding author. Tel.: +39 0521 905649; fax: +39 0521 905151.
E-mail address: giorgio.dieci@unipr.it (G. Dieci).

1874-9399/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.bbagrm.2012.09.010

known as A box and B box, forming together a bipartite binding site for the multisubunit, basal transcription factor TFIIIC on the DNA, and also contributing, at the RNA level, to the universally conserved D- and T-loops in the tRNA structure (reviewed in [4]; see Fig. 1). However, given the strongly conserved structural organization of tRNA genes, and the occurrence at particular positions of bases that are common to all or most tRNAs, searching for such conserved tRNA structural features and/or invariant positions, independently from knowledge of A box and B box promoter elements, was initially found to be a satisfactory criterion for the efficient computational identification of tRNA genes in early DNA sequence databases [5–7]. A complementary, transcription-oriented tRNA search algorithm was later proposed, relying on the weight matrix-based recognition of correctly positioned transcriptional control elements (A box, B box, poly(dT) termination signal) [8]. A modified version of this algorithm (Pol3scan, including an additional step checking amino acid stem secondary structure to exclude tRNA-like elements) was successfully employed for the definition of the first complete tRNA gene inventory of an eukaryotic organism, *S. cerevisiae* [9]. Moreover, being independent from tRNA secondary structure parameters, promoter-based tRNA search also allowed the identification, through specifically developed sub-routines, of genes for the structurally unusual selenocysteine tRNAs, and of tRNA genes bearing additional B box downstream of the terminator [8], whose role is still largely unexplored. The complementarity of promoter-based and structure-based heuristic tRNA search algorithms was later exploited, in combination with a highly sensitive and selective tRNA covariance model [10], for the setting up of a highly reliable and fast-operating program, tRNAscan-SE [11], which has become the most widely used computational tool for tRNA gene detection and annotation in newly sequenced eukaryotic genomes. Later developed programs have been proposed as useful computational tools whose employment, in parallel with tRNAscan-SE, could give a higher probability of correct and comprehensive tRNA gene identification [12,13]. A widely used genomic tRNA database (GtRNAdb) has been developed as a repository for all identifications made by tRNAscan-SE [14]. In 2008 this database, which provides direct links to UCSC eukaryotic and microbial genome browsers, included over 74,000 tRNA genes from 740 bacterial, archaeal and eukaryotic species. GtRNAdb also contains questionable tRNA gene identifications, due to tRNA-derived SINEs, pseudogenes and other tRNA-like elements, that are often individually documented. While tRNA misidentification may be regarded as a problem when genuine tRNA gene complements are the focus of interest, nevertheless tRNA-derived elements, that are very abundant in the mammalian and other genomes, are a potential source of novel Pol III-synthesized tRNA-like ncRNAs whose exploration might be interesting (see below, Sections 5 and 6). In this respect the Pol3scan search algorithm, mainly based on the detection of Pol III transcriptional control elements and thus more commonly identifying pseudogene tRNA fragments, SINE-like repetitive elements or other tRNA-like sequences containing A and B boxes [8,9,15], may be regarded as a useful tool for the computational identification of novel tRNA-like ncRNAs.

Very recently, a tRNA gene database manually curated by experts (tRNADB-CE) has been developed, which contains almost 300,000 tRNA genes. It is based on a combination of different complementary tRNA search programs and mainly features prokaryotic sequences and sequences obtained from metagenome analyses of environmental samples, but also the complete genome sequences of a few eukaryotic genomes [16].

Relevant issues, which mainly concern tRNA genes as the most numerous family of Pol III-transcribed genes, are their redundancy (number of gene copies for each isoacceptor tRNA [9]), their genomic organization (which was recently subjected to a comprehensive analysis in 74 different eukaryotic genomes, revealing complex lineage-specific patterns [17]) and their sequence diversity within supposedly homogeneous families. The latter issue has been raised by the unexpected observation that the number of tRNA genes having the same anticodon but different sequences elsewhere in the tRNA body (defined as tRNA isodecoder genes) varies significantly in different eukaryotic genomes [18], suggesting a previously under-appreciated diversity in tRNA function. Such a diversity is further supported by the tissue-specific expression of individual human tRNA species [19], and by the highly variable composition of Pol III-occupied tRNA gene subsets recently revealed by genome-wide location analyses [1] (see Section 4).

2.2. 5S rRNA genes

As is the case for tRNA genes, transcription of eukaryotic 5S rRNA genes also relies on *cis*-acting control elements (A box, intermediate element and C box) located within the transcribed region (Fig. 1). The most downstream of these element, the C box, is recognized by the sequence-specific DNA binding protein TFIIIA, while the two other elements help the TFIIIA-dependent recruitment of TFIIIC [4]. In contrast with tRNA genes, however, the internal control regions have generally not been exploited for the computational identification of 5S rRNA genes in genomic sequences. Instead, 5S rRNA gene detection (as well as the detection of Pol I-transcribed rDNA) has been based for a long time on generalized sequence similarity search programs. While such approaches may be successful in identifying highly conserved sequence tracts in the core regions of the genes, they may be sub-optimal for complete annotation purposes. A useful, 5S rRNA-specific search algorithm was thus developed, based on general weight matrix method [20]. More general and versatile ncRNA detection algorithms, exploiting either hidden Markov models (HMM) methods [21] or covariance models [22] have later proven to be highly reliable for detection of ncRNA, among which the rRNAs, including 5S rRNA [23].

2.3. Other Pol III-synthesized ncRNAs

Even though tRNA and 5S rRNA genes, given their biological (and historical) centrality, have been discussed separately, the issue of their annotation is part of the more general problem of annotating ncRNA in complete genome sequences. This aspect of genome annotation has become increasingly important in the last ten years, inasmuch as the number of genes that do not encode for proteins has proven to be much larger than expected. A useful, comprehensive discussion about problems and tools for ncRNA annotation in genomes, including “classical” Pol III-transcribed ncRNA gene families (tRNAs, 5S rRNA, U6 snRNA, SRP (7SL) RNA, RNase P and RNase MRP RNAs, vault RNAs, Y RNAs, 7SK RNA), has been published recently [24]. As outlined in that review, the most widely used general tool for annotating ncRNA genes is the Rfam database, which maintains alignments, consensus secondary structures and covariance model of known ncRNA families, with the aim of favoring automated, accurate annotation of ncRNAs in genomic sequences [25]. A common problem encountered with automated annotation of Pol III-transcribed genes, however, is the presence of large numbers of repeats and pseudogenes. All active SINE repeats in mammalian genomes (including human Alu and rodent B1 and B2 elements) are derived from Pol III-transcribed genes, and tRNA, U6 RNA, 7SK RNA, and Y RNA gene families all have hundreds of predicted homologs in the human genome, most of which are likely to be pseudogenes. Pseudogene generation is thought to be favored in part by internal promoter sequences of class III genes, that confer transcriptional activity to duplicated copies. For this reason, manual annotation is necessary in order for reliable inventories of putatively active genes to be established. For example, ~300 5S rRNA sequences are identified by Rfam in the human genome that, at variance with genuine gene copies that tend to be clustered in tandem repeats, are dispersed throughout the genome and thus likely to be pseudogenes: only a minority of them are classified as authentic rRNAs by manual annotation [24]. A