CellPress

Review

Genes in Hiding

Gertraud Burger,[1,*] Sandrine Moreira,[1] and Matus Valach[1]

Unrecognizable genes are an unsettling problem in genomics. Here, we survey the various types of cryptic genes and the corresponding deciphering strategies employed by cells. Encryption that renders genes substantially different from homologs in other species includes sequence substitution, insertion, deletion, fragmentation plus scrambling, and invasion by mobile genetic elements. Cells decode cryptic genes at the DNA, RNA or protein level. We will focus on a recently discovered case of unparalleled encryption involving massive gene fragmentation and nucleotide deletions and substitutions, occurring in the mitochondrial genome of a poorly understood protist group, the diplonemids. This example illustrates that comprehensive gene detection requires not only auxiliary sequence information – transcriptome and proteome data – but also knowledge about a cell's deciphering arsenal.

The Dilemma of the Genomics Era

Current genomics technologies allow rapid sequencing of entire genomes, but detecting genes in the 'haystack' of DNA sequences remains a challenging task, because gene finding primarily relies on sequence similarity to known genes from a limited set of (model) species. An implication is that certain genes are overlooked, especially when they contain numerous **introns**, short exons, and atypical splice-junctions. The puzzle becomes even more complex when genes are disrupted by foreign DNA inserts, carry massive mutations, or are broken up in pieces. While genomicists struggle with incognito genes, these genes are accurately decoded in the cell. Understanding cellular deciphering strategies is an important asset in assigning function to yet 'vacant' genome regions.

In this review, we summarize scenarios that render genes cryptic and briefly examine cellular decoding schemes that ensure functional gene products. We then focus on a recently discovered system whose genetic information is hidden in an unparalleled way: by massive fragmentation and scrambling, as well as nucleotide substitutions and deletions. The molecular processes involved in deciphering these extremely 'mutilated' genes are intriguingly inventive, raising questions about how such a complex system may have emerged and why it has persisted during evolution. This review aims to make genomicists aware of unconventional gene structures encountered in nature, underscoring the importance of transcriptome and proteome data for a full exploration of genome information.

Types of Gene Encryption and Corresponding Decoding Strategies

A gene is called 'cryptic' when its sequence deviates substantially from that of its product. Deviations can take various forms (Table 1, Figure 1). Nucleotide substitutions have the potential to obscure a gene if replacements are abundant. The largest number (approx. 10%) and combinatorial diversity of substitutions are reported in dinoflagellate mitochondrial and chloroplast genes [1–3]. Similar high rates, but mostly cytidine to uridine replacements (C-to-U), occur in plant organelles [4]. In both cases, genes appear highly unusual, but are still recognizable. Nucleotide substitutions in these genes are corrected after transcription, a process known as RNA editing. Substitution RNA editing of organellar pre-mRNAs and pre-tRNAs is most studied in plants, certain amoebas, and fungi, showing that the molecular processes and enzymes

Trends

Omics technologies facilitate research into organisms beyond model systems, tapping into an underexploited wealth of information.

Combination of genomics, transcriptomics, and proteomics is a potent means for uncovering hidden genes and genetic elements, assigning function to genomic regions thought to be 'junk DNA'.

Unconventional gene structures promise to reveal innovative strategies and novel molecular mechanisms in gene expression, and thus to expand our toolbox for synthetic biology, genetic engineering, and molecular therapy.

[1]Department of Biochemistry and Robert-Cedergren Centre for Bioinformatics and Genomics, Université de Montréal, Montreal, Canada

*Correspondence: gertraud.burger@umontreal.ca (G. Burger).

employed by these groups are radically different from each other [5]. Note that, likewise, many animal nuclear transcripts undergo nucleotide substitutions [predominantly adenosine to **inosine** (A-to-I), but also C-to-U], but positions fall outside coding regions [6,7] or, alternatively, generate tissue-dependent protein isoforms [8], a subject area covered extensively in other reviews [9]. Finally, rectification of substitutions also occurs during translation. For instance, illegitimate in-frame stop codons in viral genes are reinterpreted as amino-acid codons through suppressor tRNAs [10].

Another encryption type is nucleotide insertion and deletion. **Indels** can severely conceal genes that encode proteins, because conceptual translation will shift the reading frame at indel sites, changing the downstream protein sequence, and usually introducing premature stop codons. Encryption by indels was first discovered in trypanosome mitochondria, where thymidines (Ts) are either missing or superfluous at certain gene positions. These 'errors' are fixed by RNA editing through an enzymatic machinery (editosome) that adds uridines (Us) in some places and removes Us in other places from the precursor transcript [11]. A much different system of indel elimination operates in mitochondria of the slime mold *Physarum*, in which repair coincides with RNA synthesis without producing pre-edited RNA [12]. Again, other indels are corrected at the translational level by programmed frameshifting, as frequently seen in viral genes. Frameshifting requires 'slippery' codons, RNA pseudoknot structures, frameshift-promoting tRNAs, and probably trans-acting factors [13–15].

Genes also become unrecognizable by insertion of mobile genetic elements that are either 'selfish' mobile endonucleases/transposases or elements of viral or bacteriophage origin. The most studied insert elimination is (posttranscriptional) intron splicing, operating through various mechanisms specific to the particular intron type (spliceosomal, Group I, Group II, and tRNA-introns). Removal at the protein level exists as well and applies to insertion elements referred to as inteins, which are transcribed and translated together with their host gene, and subsequently eliminated by protein splicing [16]. Typically, inteins do not obscure gene detection, but rather impede functional assignment, as they are usually inserted in highly conserved protein domains. In contrast, 'hops' (for hopping) and 'byps' (for bypassing) are a class of insertion elements that do conceal genes. The hop element resides in gene 6 of bacteriophage T4 and was described as a 'persistently untranslated sequence' [17], while byps are inserted *en masse* in mitochondrial protein-coding genes of the yeast *Magnusiomyces capitatus* [18]. Unlike introns, hops and byps are retained in mRNA, yet 'ignored' during translation, by ribosomes that have 'learnt' to bypass these elements [19].

Fragmented and scrambled genes are most difficult to detect in genome sequences. Genes in pieces prevail in the germline nucleus of certain ciliates, with fragments assembled at the DNA level in the working copy of the genome, the somatic nucleus. For example, the germline nucleus of *Oxytricha* contains thousands of scrambled genes; reordering and assembly of gene segments is guided by epigenetically inherited antisense RNAs [20]. Fragment joining at the RNA level is more widespread. Typically, gene breakpoints are within introns. Gene fragments are transcribed separately and fragment transcripts associate via RNA base-pairing that allows splicing by the cognate machinery in trans (reviewed in [21]). An unorthodox case of **trans-splicing** was reported recently in certain dinoflagellates, where one of the three mitochondrial protein-coding genes (*cox3*) was split into two pieces that were separately transcribed and then combined to a full-length mRNA by an unknown, intron-independent mechanism [22]. Another RNA-level strategy occurs in retroviruses, whose split genes are transcribed directly into contiguous mRNA via template switching of the RNA polymerase [23]. Finally, there are cases of fragmented genes that produce fragmented products, which, in turn, associate noncovalently. Examples include mitochondrial rRNA genes of diverse taxa (e.g., [24–26]), but also protein-coding genes [27,28], with pieces engaging in RNA–RNA, RNA–protein, or protein–protein interactions [29].

Glossary

Adenosine deaminase acting on RNA/tRNA (ADAR/ADAT): adenosine deaminase acting on (double-stranded) RNA/tRNAs, respectively.
Cassette: a short sequence region that is unique to a given chromosome in diplonemid mtDNA. The remaining sequence of the chromosome is shared with the other chromosomes of the multipartite genome.
Cis element: a sequence or secondary structure motif (e.g., of DNA or RNA) that acts on another region of the molecule by playing a regulatory or auxiliary role in a given molecular process (transcription, splicing, etc.). The counterpart is a trans factor (see later).
Constructive neutral evolution: a ratchet-like process that describes the evolution of complex systems by nonadaptive forces.
Diplonemids: a monophyletic group of unicellular, flagellated eukaryotes that are generally free-living. Diplonemids, their sister-group kinetoplastids, and euglenids form the Euglenozoa.
Deep sea pelagic diplonemids clade I (DSPDI) and DSPDII: recently discovered large clades of deep sea pelagic diplonemids, so far including exclusively noncultured species.
Genetic drift: a neutral force in species evolution proceeding by random segregation of genetic variants within a population.
Indels: insertions and deletions of nucleotides in a sequence.
Inosine: a nucleoside, commonly found in tRNAs, which is composed of hypoxanthine (i.e., hydrolytically deaminated adenine) and ribose.
Introns: Group I and Group II introns are ribozymes, characterized by distinctive 3D structures. Splicing involves transesterification. Splice-site consensus sequences are present but weak in Group I introns. Spliceosomal introns have a highly conserved splice-site consensus and are removed from pre-mRNA by several hydrolytic and trans-esterification steps that are performed by a large RNA–protein complex, the spliceosome. 'Archaeal' or 'tRNA' introns are characterized by a distinct secondary structure context. They are eliminated from the