Review

CellPress

Exposing synonymous mutations

Ryan C. Hunt*, Vijaya L. Simhadri*, Matthew Iandoli, Zuben E. Sauna, and Chava Kimchi-Sarfaty

Division of Hematology, Center for Biologics Evaluation and Research, Food and Drug Administration, Bethesda, MD, USA

Synonymous codon changes, which do not alter protein sequence, were previously thought to have no functional consequence. Although this concept has been overturned in recent years, there is no unique mechanism by which these changes exert biological effects. A large repertoire of both experimental and bioinformatic methods has been developed to understand the effects of synonymous variants. Results from this body of work have provided global insights into how biological systems exploit the degeneracy of the genetic code to control gene expression, protein folding efficiency, and the coordinated expression of functionally related gene families. Although it is now clear that synonymous variants are important in a variety of contexts, from human disease to the safety and efficacy of therapeutic proteins, there is no clear consensus on the approaches to identify and validate these changes. Here, we review the diverse methods to understand the effects of synonymous mutations.

An expanding biological footprint of synonymous nucleotide variants

Synonymous nucleotide substitutions in coding regions were historically thought to be of little significance, but they are now the subject of increasing interest to geneticists and pharmacologists [1–4]. Over 50 human diseases have been associated with synonymous mutations [5] and, in a recent survey of 21 429 polymorphisms associated with human disease, nonsynonymous and synonymous variations were determined to have a similar probability of disease association (1.46% versus 1.26%, respectively) in addition to a statistically equivalent effect size, suggesting that the list of disease-causing synonymous mutations will grow [6]. Nonetheless, molecular evolution unmistakably illustrates that most genes tolerate nonsynonymous mutations at a lower rate than their synonymous counterparts [7]. Nature itself may be the best 'experimental' system for engendering fundamental biological principles through comparative genomics, but individual observations continue to highlight the nontrivial nature of this class of genetic variants, including the finding that synonymous and nonsynonymous substitutions introduced within *Salmonella* ribosomal proteins results in remarkably similar distributions of fitness [8]. Hundreds of regions of extreme codon conservation can be found across mammalian genes, areas that are relatively depleted of synonymous but not nonsynonymous substitutions [9]. Thus, purifying selection for synonymous codons can confound classic measurements in evolutionary biology that assume an inflated ratio of nonsynonymous substitutions relative to synonymous changes to be evidence of positive selection [10]. These observations are broadly consistent with the concept that synonymous codons affect the expression and function of the translated protein and, therefore, are under selective pressure [11]. In recent years, there have been significant advances in our understanding of the biochemical, biophysical, and genetic mechanisms by which codon bias is exploited by the translation machinery of cells to control gene expression, the efficiency or speed of protein translation, and the accuracy of protein folding [11–16]. The methodologies that have arisen from these studies are the subject of this review (Figure 1 presents an overview of mechanisms and methods). Given the scope of mechanisms and methodologies applied, we have largely focused on understanding synonymous variants within coding regions of human genes.

Codon usage bias: Nature's cue to the consequence of synonymous codon substitutions

One can study synonymous variants individually or en masse, each approach benefiting from a distinct set of methodologies. Synonymous codons appear globally at nonrandom frequencies, a phenomenon termed 'codon usage bias' (CUB), a subject of intensive investigation. Codon usage can differ across species, within a genome, or even along a single gene. In addition to neutral factors, such as mutational biases and genomic GC content, codon bias has arisen through the optimization of fundamental cellular processes, including the speed and fidelity of translation. The relative contribution of neutral and selective explanations for CUB remains a topic of debate. Evidence for selection is strong in prokaryotes, where isoaccepting tRNA abundance and codon usage frequencies have coevolved for optimal fitness [17]. Although selective forces may account for a smaller share of observed CUB in mammalian genes, a critical mass of evidence has accumulated in recent years arguing for various forms of selective pressure contributing to codon bias in higher eukaryotes. A somewhat unexpected finding is higher constraint at fourfold degenerate sites within hominid genomes relative to murids, despite their discrepant population sizes [18]. Moreover, when careful consideration is

Corresponding authors: Hunt, R.C. (ryan.hunt@fda.hhs.gov); Sauna, Z.E. (zuben.sauna@fda.hhs.gov); Kimchi-Sarfaty, C. (chava.kimchi-sarfaty@fda.hhs.gov).
Keywords: synonymous mutation; synonymous single nucleotide polymorphism; codon usage; mRNA structure; translation speed.
*These authors contributed equally to this work.

0168-9525/
Published by Elsevier Ltd. http://dx.doi.org/10.1016/j.tig.2014.04.006

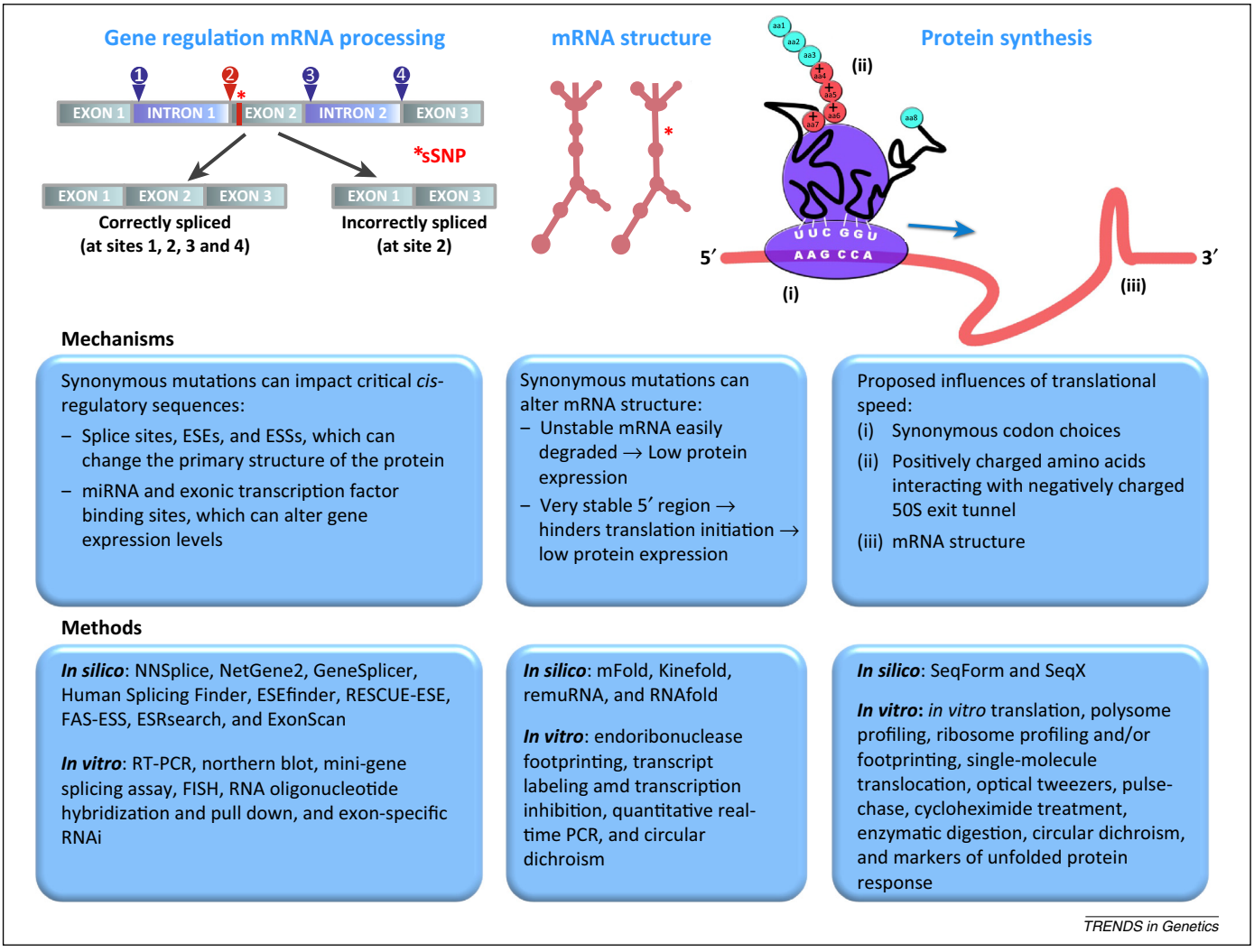

Figure 1. Schematic depiction of the general means by which synonymous codon substitutions may exert a biological effect, and select methods and approaches to delineate their underlying molecular and cellular mechanisms.

given to the nature of human genes analyzed (i.e., using data sets of genes expressed at consistent levels across tissues and physiologic states), selection strengthens for gene characteristics, including CUB, that achieve expression efficiency [19]. A recent report implicated a surprising percentage (approximately 15%) of all human codons in the binding of transcription factors (TF). These conserved exonic TF binding sites impose a selective constraint that partly drives observed codon preferences within mammalian genomes [20]. Of relevance to this discussion, >17% of single nucleotide polymorphisms (SNPs), independent of class of variant (nonsynonymous and synonymous), falling within these regions were determined to alter TF binding. Binding sites for TFs, exonic splice enhances, and other *cis*-regulatory elements can generate local CUB that is distinct from that imposed by translational selection [21], making the notion of universal 'codon optimality' a spurious term. Indeed, codons near intron–exon boundaries in many cases are not thought to be translationally optimal, an evolutionary trade-off to preserve important consensus regulatory sequences [22].

CUB, particularly the existence of selective pressures that help generate codon preferences, underscores the nontrivial nature of synonymous codon choices. There exists a multitude of approaches to quantify CUB. These measurements can be carried out at various levels, from within individual genes or functional domains to whole-genome assessments. Some approaches control for the background genomic nucleotide composition [23,24]. Others seek to understand translational-based impact of codon usage isolated from other influences; such codon bias indices, including the Codon Adaptation Index and Relative Codon Adaptation Index [25], accomplish this latter task. Assessing codon bias may reveal the existence of skewed employment of synonymous codons across a gene or within a given protein domain; these genes or specific loci may warrant particular interest during the investigation of synonymous nucleotide variants.

Identifying candidate synonymous changes for investigation

CUB suggests that synonymous mutations, as a population, have biological consequence, but how are those with particularly high impact identified? A straightforward approach is genotyping of candidate gene(s). Often, however, genetic contributions to disease or phenotypic traits are unknown, a challenge that is being addressed with the advent of increasingly economical high-throughput