Protists and the Wild, Wild West of Gene Expression: New Frontiers, Lawlessness, and Misfits

David Roy Smith1 and Patrick J. Keeling2

1Department of Biology, University of Western Ontario, London, Ontario, Canada N6A 5B7; email: dsmrt242@uwo.ca
2Canadian Institute for Advanced Research, Department of Botany, University of British Columbia, Vancouver, British Columbia, Canada V6T 1Z4; email: pkeeling@mail.ubc.ca

Abstract

The DNA double helix has been called one of life’s most elegant structures, largely because of its universality, simplicity, and symmetry. The expression of information encoded within DNA, however, can be far from simple or symmetric and is sometimes surprisingly variable, convoluted, and wantonly inefficient. Although exceptions to the rules exist in certain model systems, the true extent to which life has stretched the limits of gene expression is made clear by nonmodel systems, particularly protists (microbial eukaryotes). The nuclear and organelle genomes of protists are subject to the most tangled forms of gene expression yet identified. The complicated and extravagant picture of the underlying genetics of eukaryotic microbial life changes how we think about the flow of genetic information and the evolutionary processes shaping it. Here, we discuss the origins, diversity, and growing interest in noncanonical protist gene expression and its relationship to genomic architecture.

Keywords

constructive neutral evolution, mitochondrial transcription, plastid transcription, posttranscriptional processing, RNA editing, trans-splicing
INTRODUCTION
The storage and expression of genetic information can exemplify the use of simple tools efficiently and effectively. One only needs to think of the beautiful and paradigmatic lactose operon or the lambda phage genetic switch, which are foremost models for teaching gene regulation. But as with many processes in biology, elegance is often the exception; closer inspection, broader sampling, and better understanding of the evolutionary processes that shape these systems can reveal a biological world that is more chaotic than elegant. Genomic architecture and gene expression typify this progression, and nowhere is the chaotic nature of evolutionary complexity more evident than in the genomes of microbial eukaryotes (protists), including their mitochondrial, plastid, and nuclear DNAs (mtDNAs, ptDNAs, and nucDNAs).

Protists are abundant and ubiquitous members of nearly all known ecosystems, and together they account for a large proportion of eukaryotic biodiversity (8, 91, 129). In fact, most major groups of eukaryotes are strictly composed of microbial species, and animals, fungi, and land plants evolved independently from protist ancestors (8). For many protists, genomic architecture is highly variable, especially in the mitochondrion and plastid (110), and protists can also harbor complicated transcriptional and translational jigsaw puzzles (5, 85). Beyond the need for the transcription to RNA and translation to protein, many genes require gratuitous RNA editing; trans-splicing of fragmented, scrambled exons; removal of introns within introns; and/or deciphering via nonstandard genetic codes. In some species, the levels of posttranscriptional processing are so extensive that given the DNA sequence alone it is not possible to distinguish coding from noncoding DNA or to deduce the resulting gene products (102). Taken together, protists can be veritable genetic circus acts, consistently breaking the rules of what was once thought to be axiomatic and generating questions and debate about the evolution and function of such extravagant expression systems (42, 72, 114). Although protists have long been models for investigating the expression of genes and proteins, their propensity toward unconventional transcriptional architectures and the fact that most microbial life is not maintained in culture collections have resulted
in barriers to their study (11, 22). Consequently, some of the most unusual modes of eukaryotic gene expression remain poorly understood and undercharacterized. But that is quickly changing. Recently, there has been increased interest in protists, spurred on by a wider appreciation for their pivotal role in global biogeochemical cycles (11, 129) as well as by the introduction of high-throughput molecular sequencing technologies (83); highly sensitive proteomic methods (7); and sophisticated, user-friendly bioinformatics software (107). The last five years have seen large, international initiatives devoted to comprehending protist transcription, such as the Marine Microbial Eukaryotic Transcriptome Sequencing Project (MMETSP), which assembled, annotated, and made publicly available the transcriptomes from hundreds of marine protists (55). Massive environmental RNA sequencing (metatranscriptomics) is also aiding research into microbial gene expression (76), as is single-cell transcriptomics (61), allowing for the acquisition of RNA data from species that are uncultivated or in complex culture. Proteomics, too, is disentangling protist gene expression, providing large-scale categorizations of proteins in a wide range of species and organelles, from the eyespot of Chlamydomonas reinhardtii (99) to the mitosome of Giardia intestinalis (51).

Ultimately, our current understanding of protist gene expression has come from the combination of many disciplines and approaches, particularly the blending of classical methods with next-generation technologies. The outcome is a complicated, labyrinthine picture of the underlying genetics of microbial life. Below, we highlight compelling examples of noncanonical transcription and translation in protist mitochondria, plastids, and nuclei. We discuss the roles of evolutionary ratchets in shaping convoluted expression systems and explore how high-throughput sequencing technologies have provided unprecedented amounts of data but have also resulted in a departure from more direct analyses of RNA and protein, which remain crucial for accurately characterizing gene expression.

NONCANONICAL GENE EXPRESSION IN MITOCHONDRIA AND PLASTIDS

At various points since their endosymbiotic origins over a billion years ago, mitochondria and plastids acquired transcriptional and translational quirks. In certain species these quirks are severe and multifaceted; in others they are minor or absent entirely. Sometimes they arose in parallel in diverse lineages and different genetic compartments; in other instances they have been restricted to a specific group or genome. But wherever these embellishments have popped up, they often appear to bestow no obvious selective benefit and instead seem to place a sizeable burden on the recipient, much like the unfettered expansion of government bureaucratic complexity. In the following paragraphs, and in Figures 1 and 2, we summarize some of the many bizarre forms of organelle gene expression exposed over the last three decades.

Multilayered Complexity of Organelle Transcription in Euglenozoans

Euglenozoan organelle genomes can contain layer upon layer of transcriptional encryption, the resolution of which requires extensive downstream processing, including RNA editing, unscrambling and rejoining of gene segments, and stepwise progressive splicing. Indeed, the journey from mtDNA to functional protein in the mitochondria of kinetoplastids—the archetype of bizarre gene expression—requires a complex interplay between many chromosomes, transcripts, proteins, and genetic compartments, and up to 90% of the codons within a mature mitochondrial transcript are established through RNA editing (70, 102). In Trypanosoma brucei, for example, the cox3 primary transcript is gibberish until an RNA editing system, involving dozens of nucleus-encoded,
mitochondrion-targeted proteins and over 20 guide RNAs (gRNAs), mediates the insertion and deletion of about 550 and 40 uridine residues, respectively (10, 18, 30). The entire system requires hundreds of proteins and hundreds of gRNAs to express fewer than 20 genes and results in a genome that is fragmented into thousands of pieces and requires an elaborate filing system for accurate replication and segregation, all simply to produce proteins not unlike their ancestral form (70).

Trypanosome mitochondria are extreme but not unique. The mitochondrial genome of Diplonema papillatum, a sister euglenozoan to trypanosomes, comprises more than 75 miniature chromosomes, each with a small coding module that is joined with its partnered modules from neighboring chromosomes through trans-splicing (75, 125). This processing is likely facilitated through antisense RNAs and in some cases requires U-insertion RNA editing (124). The cox1 coding modules from D. papillatum are dispersed over nine chromosomes and need to be transcribed independently before being linked together via eight splicing events into a single molecule (125). The D. papillatum mitochondrion also holds the current record for the largest number of uridines added at a single editing site: 26 (124).

Other euglenozoans break records for subtracting genetic elements. Expression of the Euglena gracilis plastid genome requires the removal of approximately 160 introns (one for every 900 nucleotides of ptDNA), including 15 twintrons (introns within introns), which need to be subtracted sequentially for accurate splicing (43), and some of these transcripts are also polyadenylated (130). As if that weren’t enough, the plastid rps18 gene has a multipartite twintron—an external intron interrupted by an internal intron containing two additional introns (25). The mitochondrial compartment of E. gracilis offers no return to normalcy with split ribosomal RNAs (rRNAs; described below) and what might be the evolutionary precursors for the gRNAs that direct post-transcriptional editing in kinetoplastids and diplonemids (32, 116).

Dinoflagellate Organelle Gene Expression: Not to Be Outdone by Euglenozoans

Gene expression in euglenozoans is odd on many levels, so it is surprising that it is matched almost point for point by gene expression in an unrelated lineage, the dinoflagellates. They, too, experience extensive organelle RNA editing, but of a nonhomologous substitutional type, affecting both the mitochondria (48, 126) and the plastids (5, 21), and where 11 of the 12 possible types of nucleotide substitution (A to C, A to G, etc.) have been identified (86, 96). Plastid transcripts can also sport 3′ polyuridylated tails and are likely expressed by rolling-circle transcription of miniature circular chromosomes (4, 21, 133). Remarkably, the RNA editing and polyuridylation pathways in dinoflagellate plastids have infected replacement plastids, as displayed by Karolodinium veneficum and its close relative Karenia mikimotoi. These species substituted their ancestral
Figure 2

Fragmented large- and small-subunit (LSU and SSU) rRNAs from the mitochondrial, plastid, and nuclear genomes of various protists. Numbers of LSU and SSU rRNA-coding fragments are shown in purple and blue, respectively. Intervening genes (gray boxes), chromosome size (in kilobases; not to scale), and transcriptional polarity (dashed arrows) are shown. Transcribed rRNA fragments can come together to form a functional rRNA via secondary pairing interactions (green checkmarks) or through trans-splicing (Diplonema papillatum). For Symbiodinium sp. clade C3, the complete number of SSU rRNA fragments and how they are ultimately joined are unknown. For the nuclear genome, the LSU rRNA gene includes the 5.8S region.
peridinin-containing plastid for a plastid from haptophyte algae. This group displays neither RNA editing nor polyuridylation, but once housed in a dinoflagellate, the haptophyte-derived organelles acquired both characteristics (24, 96).

Expression in dinoflagellate mitochondria is, if anything, stranger still. In addition to undergoing widespread substitutive editing, mitochondrial transcripts are regularly 5’ polyuridyalted, 3’ oligo-adenylated, and trans-spliced, and most lack canonical start and stop codons. The mitochondrial cox3 transcript, for example, is fused with cob in some species (104), whereas in others it is fragmented and trans-spliced so that a species-specific number of bases from its polyadenylated tail are incorporated at the splice site (47, 48) and are essential for expression (Figure 1). The polyA tail is also used by several species to form a stop codon on cox3, whereas other genes appear to lack stop codons altogether and simply translate the tail as oligo-lysine (47, 104).

Translational Slippage in Mitochondria

The oyster parasite Perkinsus marinus is closely related to dinoflagellates, but its mode of mitochondrial expression has taken a different but equally perplexing route. Although RNA editing does not occur in P. marinus mitochondria, the cox1 and cob genes contain multiple frameshifts that would ultimately lead to nonfunctional proteins if not amended (79, 131). These irregularities are retained in the corresponding mature transcripts and eventually corrected by programmed ribosomal slippage (79). Every time an in-frame AGG or CCC appears in the mRNA, the reading frame moves forward by one or two bases, respectively. The result is that AGGY codes for glycine and CCCCC for proline. How these frameshifts are introduced during translation is poorly understood, but it is hypothesized that stalled ribosomes skip the first bases of these codons or that specialized tRNAs recognize nontriplet codons (79). Whatever the mechanism, frameshifts have to be introduced 10 times within the cox1 mRNA alone to yield a functional protein. In Perkinsus chesapeaki, the same system was found but with slight variation in the number of codon slippages that occur (131). Perkinsus spp. also appear to harbor a relic, nonphotosynthetic plastid, but it has lost its genome and gene expression system (79), which may be viewed as the ultimate in noncanonical expression diversity (see sidebar “Gene Expression–Less Organelles”).

A parallel form of translational slippage in the mitochondria of a fungus has been described, but here the ribosome does not really skip but jumps a considerable distance. At over 80 sites in the Magnusiomyces capitatus mitochondrial genome, the ribosome encounters unassigned codons and secondary-structure-rich sequences similar to mobile elements in other fungi. From these cues, the ribosome recognizes takeoff and landing sites and uses these sites to bypass large segments of mRNA, resulting in intact and functional polypeptides (64).

GENE EXPRESSION–LESS ORGANELLES

Among the most radical changes to gene expression is to lose it altogether. In various lineages, mitochondria have been subverted into anaerobic organelles, called hydrogenosomes or mitosomes (45), and plastids have lost photosynthetic capabilities (50, 112). In some cases, these vestigial organelles retain a genome and gene expression (e.g., 82), but in other instances the genome has been completely lost (109). This has been known for some time for relic mitochondria (9, 45, 90), and more recently genome-lacking plastids have been identified (50, 112). These organelles are entirely dependent on genes in the nucleus, which is a radical change from their original state as free-living bacteria.
Noncanonical Genetic Codes

The universal genetic code is undeniably ancestral and highly conserved, but it is not quite universal, particularly within mitochondria (60). In the mtDNAs of kinetoplastids, diplomemids, and many red algae, for example, UGA (normally a stop codon) specifies tryptophan (121, 125), and dinoflagellate mitochondria regularly employ start codons that differ from the standard AUG (104, 126), as do some apicomplexan and ciliate mitochondrial genes (26). Novel code alterations are still being uncovered: The green alga *Pseudomuriella schumacherensis* was recently found to use UCG as a mitochondrial stop codon (35). In fact, most mitochondria, including those of animals, fungi, and plants, have experienced at least one mtDNA codon alteration, resulting in over a dozen mitochondrial codes and departures from the universal code. In contrast, the universal code still dominates in plastids, with sporadic examples of noncanonical codes in some apicomplexans and dinoflagellates (65, 80).

Connecting the Transcriptional Dots of Fragmented Genes

Introns are sometimes described as splitting genes into pieces, but most intron-containing genes are expressed as contiguous RNA molecules. However, many genes really are split in that they are physically separated in the genome (sometimes on different DNA molecules), resulting in disjointed, incomplete transcripts that need to be strung together after transcription (or after translation; 49). Mitochondria exhibit a proclivity for gene fragmentation, especially in their rRNAs, which have independently evolved fragmented architectures in diverse groups, including chlorophycean green algae, euglenozoa, and alveolates (6, 31, 116) (Figure 2).

The most extreme example of rRNA fragmentation yet described is in the malaria parasite *Plasmodium falciparum*, where the large- and small-subunit (LSU and SSU) mitochondrial rRNA genes have splintered into numerous coding modules randomly distributed across both strands of the genome and interspersed with protein-coding genes (31). Given their small sizes (23–190 nucleotides) and disorganization, it has taken years to identify the more than 25 rRNA modules, which are expressed via cleavage of long precursor polycistronic transcripts, undergo oligoadenylation, and come together through secondary pairing interactions to form functional rRNAs (31, 52). A similar situation exists in the mitochondria of ciliates (44), chlamydomonadalean algae (6, 27), and dinoflagellates (126); moreover, the latter group can also have split plastid rRNAs (21). As in *P. falciparum*, the fragmented rRNAs from these different groups and compartments are processed, sometimes from long polycistronic transcripts, and then joined via base-pairing into complete rRNAs (6, 21, 27, 44).

In other systems, discontinuous rRNA genes are reassembled into a single covalently continuous molecule by trans-splicing. In *D. papillatum* this affects most genes, including the LSU rRNA, and appears to be mediated by antisense RNAs (124). Intron-mediated trans-splicing of organelle mRNAs is well documented for many different groups (39), particularly the plastid-encoded *psaA* gene, but it has rarely been observed for bridging discontinuous organelle rRNAs (39). This is particularly surprising for dinoflagellate mitochondria, where a trans-splicing mechanism exists for ligating *cox3* mRNAs, but the fragmented mitochondrial rRNAs from this group are not trans-spliced (48).

SIMILAR PHENOMENA IN NUCLEAR GENE EXPRESSION: ALTERED CODES, FRAGMENTATION, TRANS-SPlicing, AND SLIPPAGE

Mitochondria and plastids have stretched the limits of genetic expression, but many of the same peculiarities are also found in nuclear genes, albeit at a lower frequency and to a lesser extent than
that typically seen in organelles (110). Other features of organelle expression are near absent from nuclei, such as RNA editing, which can occur in nuclear genomes (15) but not in the widespread or gratuitous form observed in organelles.

Noncanonical genetic codes have evolved many times in nuclear genomes—including those of yeasts, green algae, diplomonads, oxymonads, and especially ciliates where several code changes have occurred independently in different subgroups (17, 39, 46, 56, 57). The nonstandard code whereby the canonical stop codons UAG and UAA encode glutamine is the most common change in nuclei, but it has surprisingly not yet been found in organelles or bacteria, likely reflecting differences in translation (57).

Fragmented genes are rare but present in nuclear systems. Euglenozoans can have fragmented rRNAs, broken into as many as 14 pieces that function as a noncovalent network (Figure 2) (100, 127). In the diplomonad parasite *G. intestinalis*, some proteins are encoded by discontinuous pre-mRNAs, which are independently transcribed from disparate chromosomal regions and joined into a contiguous mRNA by spliceosome-mediated *trans*-splicing (54, 98). Nuclear tRNAs can be similarly disjointed. The 5′ halves of several tRNA genes in the red alga *Cyanidioschyzon merolae* are located downstream of their 3′ portions and are expressed via circular RNA intermediates (113); permuted tRNA genes have also been spotted in the nucDNAs of prasinophyte green algae and the nucleomorph genome of *Bigelowiella natans* (78)—nucleomorphs as a whole have bizarre forms of gene expression (see sidebar “Nucleomorph Gene Expression”). Lastly, spliceosomal introns can also form progressively spliced “stwintrons,” like the group II twintrons of organelles (33).

**Spliced-Leader *Trans*-Splicing and Polycistronic mRNA**

Euglenozoans and dinoflagellates have idiosyncratic organelle gene expression systems, but they can undergo strange forms of nuclear expression as well. Trypanosome nuclear genomes, for instance, have large tracks of genes oriented on the same strand lacking canonical sequence-based promoters. Instead, RNA polymerase II initiates transcription at regions of modified histones upstream of gene tracts, often resulting in extremely long polycistronic transcripts, which can contain tens to hundreds of genes and cover large sections of the genome (77). These polycistronic transcripts are processed into gene-sized mRNAs by *trans*-splicing: The spliceosome initiates two *trans*-esterification reactions that cleave the genes from one another and add a short spliced-leader (SL) to the 5′ end of each gene (66, 118). Consequently, there is little or no control of expression at the level of transcription in trypanosomes, which is overcome by more strict control of transcript degradation and by genome organizational features (59). Other euglenozoans also appear to cap mRNAs with an SL (31, 34, 62), but apparently these systems have not reached the same extremes as those of trypanosomes.
Widespread use of SL trans-splicing also occurs in dinoflagellates (132). Like in trypanosomes, these systems have evolved polycistronic transcription for at least some genes, but the situation in dinoflagellates is poorly understood because their genomes are so large. The existence of SL trans-splicing likely preconditions expression systems for the emergence of polycistronic transcription because it allows for efficient and effective processing of multigene transcripts into functional mRNAs (71), although the presence of one by no means necessitates the presence of the other.

**Juggling Gene Expression with Two Distinct Nuclear Genomes**

The oddities of nuclear gene expression are mostly poor reflections of their organelle counterparts. Protist nuclear genomes do, however, exhibit some eccentricities not found in mitochondria or plastids. Ciliates, for instance, have two distinct nuclear genomes: a dormant germline micronucleus (MIC) and a transcriptionally active somatic macronucleus (MAC) (92). The MIC genome architecture resembles a canonical nuclear genome in that it comprises a modest number of large chromosomes containing thousands of genes interspersed with long stretches of noncoding DNA (16). But MIC genes are mostly transcriptionally silent and are interrupted by nonintronic non-coding sequences called internal eliminated sequences (IES), which generally disrupt the reading frame (16). The MAC genome, on the other hand, has functional genes, but they are arranged on thousands of tiny, multicopy chromosomes (1, 119). One of the most extreme examples of this occurs in *Oxytricha trifallax*, where there is essentially one gene per MAC chromosome (119). The MAC is generated after sexual conjugation from a copy of the MIC, so functional genes must be generated by the programmed elimination of IES DNA, massive genomic reorganization, and the shattering and amplification of the chromosomes into small, multicopy pieces (87).

Functional MAC genes are reconstructed in ciliates epigenetically, but the mechanisms used to accomplish this can differ among species (20, 87, 101). In *Tetrahymena thermophila* and *Paramecium tetraurelia*, short noncoding RNAs are transcribed from across the entire MIC genome and transported to the parental MAC, where they can then bind to the pool of transcripts generated from the whole genome. The short RNAs that fail to bind to the parental MAC transcripts are sequestered to the developing MAC, where they single out genomic DNA (corresponding to IES) for excision and deletion, thus epigenetically passing along the genome content of the MAC (20, 101).

*O. trifallax* also employs RNA to epigenetically recreate the MAC genome, but the strategy differs from those of *T. thermophila* and *P. tetraurelia*. Both short and long noncoding RNAs are generated from the *O. trifallax* parental MAC genome, and they are transported to the developing MAC, where they target sequences not for deletion but for retention (28). These noncoding RNAs not only help identify and remove IES but also provide information for gene rearrangements, resulting in the unscrambling of MIC coding segments into coherent MAC gene sequences (88). This type of unscrambling mechanism would not be possible using only short RNAs, such as those found in *T. thermophila* and *P. tetraurelia*. Despite their differences, all of these systems use mechanisms derived from transposon defense to perform the actual elimination and rearrangement of DNA (14, 20, 120). Indeed, in *P. tetraurelia* some of the IES are clearly homologous to transposons (3).

Although the MIC genome has typically been considered transcriptionally inactive (except during sexual reproduction), this now appears to be an oversimplification: a number of germline-restricted protein-coding genes reside within IES and are not transmitted to the MAC but are clearly expressed and regulated (16). Further departures from conventional ciliate nuclear gene expression (12, 36) are now showing how truly dynamic the storage and decoding of genetic information can be.
MOLECULAR RUBE GOLDBERG MACHINES

Even when we begin to see how systems like the ciliate nuclear dualism have evolved, the question remains: Why did they evolve? Biologists have a propensity for seeing the world in an adaptive light, perhaps because at the levels of organization that we are used to dealing with, this is mostly reasonable. But there is increasing evidence that complexity and variation at deeper levels in the molecular world may also be a consequence of nonadaptive processes (42, 72, 73). Seeing the hand of natural selection in shaping the spear-like snout of a marlin or the streamlined feathers of a hawk is easy; explaining the origins of a noncanonical genetic code or jumbled rRNA is not so straightforward.

What advantage can there be to taking an intact, unadorned coding region and shattering it into dozens of unordered and independently transcribed pieces, only to undo all of this encryption and stitch the segments back together again at the RNA or amino acid level? What benefit can come from the addition of an intron into an intron that is already inside of an intron? Or from becoming dependent on an extensive and costly RNA-editing infrastructure that returns mRNAs to their ancestral state? Why alter a genetic code that has existed since the last universal common ancestor?

The evolution of unconventional and elaborate modes of gene expression has puzzled researchers for decades (37) and divided them along adaptive versus nonadaptive lines (72). For certain genetic eccentricities, there are step-by-step hypotheses explaining their evolution. There are multiple, well-regarded neutral hypotheses for reassigning codons in the genetic code (73, 89). But there are also adaptive alternatives, such as the genome-streamlining hypothesis, which argues that variation in the code is the result of selection for a reduced translational apparatus (2). Similarly, a variety of adaptationist explanations for RNA editing have been suggested, such as gene regulation, generating genetic variation, and/or mutational buffering (29, 40, 53). The existence of the MAC nuclear genome and its associated RNA infrastructure in *O. trifallax* has been suggested to provide an adaptive store of heritable variation (in addition to the MIC nucleus), which contributes to the evolutionary success of ciliates (87). The appearance and proliferation of introns has long been considered an adaptation for generating organismal complexity by exon shuffling and alternative splicing (37, 97).

These and other adaptationist arguments for genetic complexity might be valid, but most remain completely untested, several are inconsistent with vital aspects of the system, and in many cases cause is mistaken for effect. For example, suggesting that RNA editing evolved to buffer against deleterious mutations implies that the editing apparatus, which is tailored for specific sites, was put in place after the deleterious mutation(s) that it now corrects. However, this requires the deleterious mutation to persist in a population during the (probably lengthy) time required to evolve the elaborately complex and expensive compensatory mechanism. An alternative view is that the RNA editing apparatus already existed, potentially serving a different purpose altogether, and was ultimately fixed within the cell through fortuitous events (19, 42). Once the editing machinery for a single edit is fixed in the population (neutrally), the conditions are set for additional editing sites to arise more easily. Indeed, RNA editing in trypanosomes illustrates this most effectively because the gRNAs encode the ancestral sequences, so based on parsimony they should have existed before the mutations they now edit.

More and more attention is being given to nonadaptive models for the origins of cellular and molecular complexity (42, 68, 73). One general model is called constructive neutral evolution (CNE) (19, 42, 117): a ratchet-like process whereby neutral (or slightly deleterious) mutations that result in increased complexity are fixed by random processes, such as genetic drift. The ratchet in this model comes from the idea that the gains in complexity that are fixed neutrally are not easily
unfixed by random events. Although criticized by some (114; but see 58), CNE has been used to explain everything from the origins of RNA editing and intron splicing (19, 117) to the complexity of ribosomes and mitochondrial respiratory complexes (23, 69). The ratcheting of neutral mutations has also been invoked for major transitions in evolution (81), such as the shift from a unicellular to a multicellular existence (68). In essence, CNE and other nonadaptive evolutionary models, such as the mutational hazard hypothesis (73), argue that biological complexity is not necessarily driven by fine-tuning or sophistication, but instead can be the consequence of runaway bureaucracy—“biological parallels of nonsensically complex Rube Goldberg machines that are overengineered to perform a single task” (42).

One of the challenges of evaluating the different models for the evolution of molecular complexity in protists is a lack of population-genetics data. We presently know very little about the effective population sizes and mutation rates of protists. There is also a paucity of information on DNA repair and recombination. We do not even know if most microbial eukaryotes have sex, although most likely do (41, 115), and when sex is characterized, it can be complicated: Some species have many different mating types (13, 123).

Data on protist population dynamics, mutation, recombination, and DNA maintenance are fundamental to understanding the underlying processes shaping their genomes and gene expression systems. One way to gain insights into these processes is by studying genetic diversity within and among populations. Interesting relationships between genomic/transcriptional complexity and genetic diversity have already been observed in protists (73, 74, 111); and, overall, species with a propensity for genetic embellishments, such as RNA editing, have been shown to harbor low levels of silent-site nucleotide diversity, suggesting that they have small effective population sizes (74, 111). Other studies have found that inadequacies or idiosyncrasies in DNA maintenance processes can contribute to increased genomic complexity (110). These are based on a few relatively well-studied taxa, but population-genetics and mutational studies of protists will only get easier as massively parallel sequencing methods improve and our databases grow accordingly.

A FUTURE PAVED WITH RNA-SEQ, BUT LOOK OUT FOR POTHOLES

Besides simplifying and streamlining genomics and transcriptomics, next-generation sequencing (NGS) has provided a huge reserve of unexplored data. Vast quantities of raw sequencing data from diverse eukaryotic groups have been deposited in freely accessible repositories, such as GenBank’s Sequence Read Archive (SRA) (http://www.ncbi.nlm.nih.gov/sra). As of early September 2015, the SRA contains \( \sim 2 \times 10^{15} \) open-access NGS bases, including \( \sim 700,000 \) datasets from eukaryotes, a large fraction of which are RNA-seq. Only about 5% of these datasets come from eukaryotes that are not animals, land plants, or fungi, but that still leaves thousands of NGS projects from protists, including ones for closely related species or distinct isolates of the same species, many of which are transcriptomic surveys (55).

These transcriptomic data are obviously a valuable resource for studying nuclear gene expression, but are data from organelle genomes as readily available? The high copy number per cell and elevated expression levels of organelle DNAs mean their transcripts are well represented in eukaryotic RNA-seq experiments (94, 106), and because mitochondrial and plastid intergenic regions are often transcribed (4, 93), near-complete organelle genome assemblies are available (67, 95). Consequently, the SRA contains billions of unanalyzed reads from hundreds of microbial eukaryotes and includes both nuclear and organelle sequences, making it an excellent, untapped resource for investigating unusual and poorly understood forms of gene expression.

But relying too heavily on bulk sequence data for the identification and characterization of novel gene expression systems is rarely sufficient for understanding how complex genomes function.
Accurate and detailed characterizations of gene expression tend to involve a combination of massive sequencing with more demanding methods. But it is these technically challenging techniques that are sometimes lacking from contemporary transcriptome research, particularly that focusing on organelle genetics (108). High-throughput sequencing and bioinformatics have removed barriers to identifying where potentially complex expression systems might exist and understanding how they might operate, but additional information is required to really understand the machinery involved in posttranscriptional processing, modes of transcription, and the underlying DNA/RNA maintenance processes.

Moving forward, studies of protist expression systems need to combine NGS with molecular-biology-focused methods. This in combination with data on the population genetics and mutation rates, as well as a more unified understanding of cytonuclear interactions and coevolution (105), will indeed lead to an exciting synthesis.

**DISCLOSURE STATEMENT**

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

**ACKNOWLEDGMENTS**

This work was supported by Discovery Grants to D.R.S. and P.J.K. from the Natural Sciences and Engineering Research Council (NSERC) of Canada. P.J.K. is a fellow of the Canadian Institute for Advanced Research.

**LITERATURE CITED**


36. Gao F, Roy SW, Katz LA. 2015. Analyses of alternatively processed genes in ciliates provide insights into the origins of scrambled genomes and may provide a mechanism for speciation. *mBio* 6:e01998-14


109. Smith DR, Asmail SR. 2014. Next-generation sequencing data suggest that certain nonphotosynthetic green plants have lost their plastid genomes. *N. PhytoL* 204:7–11
New From Annual Reviews:

Annual Review of Cancer Biology
cancerbio.annualreviews.org • Volume 1 • March 2017

Co-Editors: Tyler Jacks, Massachusetts Institute of Technology
Charles L. Sawyers, Memorial Sloan Kettering Cancer Center

The Annual Review of Cancer Biology reviews a range of subjects representing important and emerging areas in the field of cancer research. The Annual Review of Cancer Biology includes three broad themes: Cancer Cell Biology, Tumorigenesis and Cancer Progression, and Translational Cancer Science.

TABLE OF CONTENTS FOR VOLUME 1:

- How Tumor Virology Evolved into Cancer Biology and Transformed Oncology, Harold Varmus
- The Role of Autophagy in Cancer, Naiara Santana-Codina, Joseph D. Mancias, Alec C. Kimmelman
- Cell Cycle–Targeted Cancer Therapies, Charles J. Sherr, Jiri Bartek
- Ubiquitin in Cell-Cycle Regulation and Dysregulation in Cancer, Natalie A. Borg, Vishva M. Dixit
- The Two Faces of Reactive Oxygen Species in Cancer, Colleen R. Reczek, Navdeep S. Chandel
- Analyzing Tumor Metabolism In Vivo, Brandon Faubert, Ralph J. DeBerardinis
- Stress-Induced Mutagenesis: Implications in Cancer and Drug Resistance, Devon M. Fitzgerald, P.J. Hastings, Susan M. Rosenberg
- Synthetic Lethality in Cancer Therapeutics, Roderick L. Beijersbergen, Lodewyk F.A. Wessels, René Bernards
- Noncoding RNAs in Cancer Development, Chao-Po Lin, Lin He
- p53: Multiple Facets of a Rubik’s Cube, Yun Zhang, Guillemina Lozano
- Resisting Resistance, Ivana Bozic, Martin A. Nowak
- Deciphering Genetic Intratumor Heterogeneity and Its Impact on Cancer Evolution, Rachel Rosenthal, Nicholas McGranahan, Javier Herrero, Charles Swanton
- Immune-Suppressing Cellular Elements of the Tumor Microenvironment, Douglas T. Fearon
- Overcoming On-Target Resistance to Tyrosine Kinase Inhibitors in Lung Cancer, Ibiayi Dagogo-Jack, Jeffrey A. Engelman, Alice T. Shaw
- Apoptosis and Cancer, Anthony Letai
- Chemical Carcinogenesis Models of Cancer: Back to the Future, Melissa Q. McCreery, Allan Balmain
- Extracellular Matrix Remodeling and Stiffening Modulate Tumor Phenotype and Treatment Response, Jennifer L. Leight, Allison P. Drain, Valerie M. Weaver
- Aneuploidy in Cancer: Seq-ing Answers to Old Questions, Kristin A. Knouse, Teresa Davoli, Stephen J. Elledge, Angelika Amon
- The Role of Chromatin-Associated Proteins in Cancer, Kristian Helin, Saverio Minucci
- Targeted Differentiation Therapy with Mutant IDH Inhibitors: Early Experiences and Parallels with Other Differentiation Agents, Eytan Stein, Katharine Yen
- Determinants of Organotropic Metastasis, Heath A. Smith, Yibin Kang
- Multiple Roles for the MLL/COMPASS Family in the Epigenetic Regulation of Gene Expression and in Cancer, Joshua J. Meeks, Ali Shilatifard
- Chimeric Antigen Receptors: A Paradigm Shift in Immunotherapy, Michel Sadelain
Contents

My Lifelong Passion for Biochemistry and Anaerobic Microorganisms
  Rudolf Kurt Thauer .......................................................... 1

The Pyromaniac Inside You: *Salmonella* Metabolism in the Host Gut
  Fabian Rivera-Chávez and Andreas J. Bäumler .......................... 31

Regulation of Transcript Elongation
  Georgiy A. Belogurov and Irina Artsimovitch .......................... 49

*Candida albicans* Biofilms and Human Disease
  Clarissa J. Nobile and Alexander D. Johnson .......................... 71

Transcription Factors That Defend Bacteria Against Reactive Oxygen Species
  James A. Imlay ...................................................................... 93

Bacterial Proteasomes
  Jordan B. Jastrab and K. Heran Darwin ................................. 109

The Organellar Genomes of *Chromera* and *Vitrella*, the Phototrophic Relatives of Apicomplexan Parasites
  Miroslav Oborník and Julius Lukeš ........................................ 129

The Gut Microbiota of Termites: Digesting the Diversity in the Light of Ecology and Evolution
  Andreas Brune and Carsten Dietrich ...................................... 145

Microsporidia: Eukaryotic Intracellular Parasites Shaped by Gene Loss and Horizontal Gene Transfers
  Nicolas Corradi ................................................................. 167

Anthrax Pathogenesis
  Mahatab Moayeri, Stephen H. Leppla, Catherine Vrentas, Andrei P. Pomerantsev, and Shibui Liu ........................................... 185

CRISPR-Cas: New Tools for Genetic Manipulations from Bacterial Immunity Systems
  Wenyan Jiang and Luciano A. Marraffini .................................. 209
Perception and Homeostatic Control of Iron in the Rhizobia and Related Bacteria
Mark R. O’Brien ................................................................. 229

Thymineless Death Lives On: New Insights into a Classic Phenomenon
Arkady Khodursky, Elena C. Gazmán, and Philip C. Hanawalt .................. 247

Membrane-Coupled mRNA Trafficking in Fungi
Carl Haag, Benedikt Steinert, and Michael Feldbrügge .................................. 265

Molecular Pathogenesis of Ehrlichia chaffeensis Infection
Yasuko Rikihisa ........................................................................................................ 283

Microbiology Meets Big Data: The Case of Gut Microbiota-Derived Trimethylamine
Gwen Falony, Sara Vieira-Silva, and Jeroen Raes ........................................... 305

The Unique Molecular Choreography of Giant Pore Formation by the Cholesterol-Dependent Cytolysins of Gram-Positive Bacteria
Rodney K. Tweten, Eileen M. Hotze, and Kristin R. Wade ............................. 323

Ion Regulation in the Malaria Parasite
Kiaran Kirk ............................................................................................................. 341

Physics of Intracellular Organization in Bacteria
Ned S. Wingreen and Kerwyn Casey Huang ......................................................... 361

Stochastic Switching of Cell Fate in Microbes
Thomas M. Norman, Nathan D. Lord, Johan Paulson, and Richard Losick .......... 381

Assembly of the Mycobacterial Cell Wall
Monika Jankute, Jonathan A.G. Cox, James Harrison, and Gurdyal S. Besra ........ 405

Mechanisms of Bacterial Colonization of the Respiratory Tract
Steven J. Siegel and Jeffrey N. Weiser ................................................................ 425

Interactions Between the Gastrointestinal Microbiome and Clostridium difficile
Casey M. Theriot and Vincent B. Young ................................................................. 445

Lytic Cycle of Toxoplasma gondii: 15 Years Later
Ira Blader, Bradley Coleman, Chun-Ti Chen, and Marc-Jan Gubbels ................ 463

Septins and Generation of Asymmetries in Fungal Cells
Anum Khan, Molly McQuilken, and Amy S. Gladfelter ....................................... 487

How Is Fe-S Cluster Formation Regulated?
Erin L. Mettert and Patricia J. Kiley ...................................................................... 505
Protein Phosphatases of Pathogenic Bacteria: Role in Physiology and Virulence

Andaleeb Sajid, Gunjan Arora, Anshika Singhal, Vipin C. Kalia, and Yogendra Singh

Index

Cumulative Index of Contributing Authors, Volumes 65–69

Errata

An online log of corrections to *Annual Review of Microbiology* articles may be found at
http://www.annualreviews.org/errata/micro