Scientific Serendipity Initiates an Intron Odyssey

Published, JBC Papers in Press, August 26, 2009, DOI 10.1074/jbc.X109.055400

Marlene Belfort
From the Wadsworth Center, New York State Department of Health, and University at Albany, The State University of New York, Center for Medical Science, Albany, New York 12208

In the fields of observation, chance favors only the prepared mind—Louis Pasteur, 1854

As the child of German immigrants to South Africa and disillusioned by the Apartheid regime, I made my way to the United States for graduate school. After earning my Ph.D. from the University of California at Irvine and a postdoctoral degree at the Hebrew University in Jerusalem, both in phage \( \text{H9261} \) genetics, I faced a five-body problem: two careers and three children. Two positions and a family-friendly environment brought us to the Capital District in New York. For me, it was a research scientist position, working on thymidylate synthase (TS). Even in the 1970s, TS was a well-studied enzyme, highly regulated in each organism in which it had been studied, and above all, TS was a geneticist’s dream: positive selections for both gene function and dysfunction in bacteria. So, I set to work with respected biochemists Frank and Gladys Maley, and almost 30 years ago, we published our first paper together on a single functional arginine involved in TS catalysis (1). This arginine was to reappear on my research radar in a completely different context more than 20 years later. Meanwhile, I developed the TS genetic system for studying the enzyme’s function and regulation (2) and taught my colleague Fred Chu how to sequence the TS coding sequence (the \( \text{td} \) gene) of phage T4. Sequencing was a big deal in those days; entire papers were dedicated to decoding a single gene, and I had taken a week-long course to learn how to sequence DNA. Chu did a superb job, but he presented us with a conundrum. The \( \text{td} \) gene contained a chunk of sequence that did not correspond to TS. Without the benefit of GenBankTM and BLAST searches, we were left to ponder for ourselves what that intervening sequence might be: an artifact? An intron? But dogma insisted that introns existed only in eukaryotic genes.

An Intron in a Phage Gene!

The combination of Chu’s sequencing, the Maleys’ protein chemistry, and my genetics eventually gave us the courage to publish our observation of an “intervening sequence” in the \( \text{td} \) gene, in the TS coding sequence (Fig. 1A) (3). This was my first foray into interdisciplinarity, and from then on, crisscrossing fields became a key feature in my scientific journey. Nevertheless, at the time, we had no clue about how the intact TS enzyme was generated: protein ligation? RNA polymerase hopping? RNA splicing? Over the next few years, we showed that, indeed, processing of the \( \text{td} \) transcript occurred at the RNA level (4–8). In the ensuing period, we warmed up to the words “intron” and “splicing” when referring to the phage T4 \( \text{td} \) gene. My intron odyssey had begun.

Splicing Mechanism as the Means to More Introns

Through a combination of genetics, biochemistry, and electron microscopy, we figured out that the \( \text{td} \) intron is a group I intron that self-splices by the guanosine-initiated pathway (Fig. 1B) (7, 8). This mechanism, elaborated by Tom Cech and colleagues for the renowned \( \text{Tetrahymena} \) nuclear rRNA intron (9), suggested to us a strategy for searching for other introns in...
T4. We used T4-infected cells and labeled the RNA with $^{32}$P]GTP in an autocatalytic reaction in vitro (Fig. 1B). Three labeled bands were apparent, suggesting multiple introns. In collaboration with my colleague David Shub, we identified two additional group I self-splicing introns, one in the nrdB ribonucleotide reductase gene and the other in the nrdD anaerobic ribonucleotide reductase gene (originally called sunY) (10). It is a tantalizing but as yet unexplained observation that all three T4 introns are in genes involved in nucleotide metabolism.

**Homing Sweet Homing: Discovery of Intron Mobility in Phage**

The variable occurrence of the three T4 introns in the T-even phage family, comprising T2, T4, and T6, sug-
gested to us that these introns were mobile (11, 12). Indeed, we were able to demonstrate hopping of the td and nrdD introns from intron-containing donor replicons to cognate intronless recipients containing the td or nrdD homing sites (13). Endonucleases encoded by open reading frames within the mobile td and nrdD introns were responsible for initiating the homing process (Fig. 1, A and C). As we later discovered, the T4 nrdB intron carries a disabled endonuclease and is therefore not mobile.

The extraordinarily efficient phage intron homing reactions paralleled those that had been demonstrated for group I introns in yeast mitochondria and protist nuclei (reviewed in Ref. 14). There was so much interest in RNA catalysis and intron splicing and mobility that Shub and I decided to bring together the community of giants on whose shoulders we were standing. We organized a conference in 1988 entitled "RNA: Catalysis, Splicing, Evolution" (Fig. 2A). Among the 24 speakers, 11 are now members of the National Academy of Sciences, and three have won Nobel prizes (Fig. 2B). Their collective contributions to the conference have been published (15). The odyssey has indeed been one in which we have been accompanied by many visionary colleagues, all with an interest in demystifying the RNA world.

**Dreaded Recombination-dependent Mobility Pathways**

In a previous incarnation, I had vowed never to work on recombination, but intron homing proved irresistibly seductive. Soon after discovering mobility of the phage introns, "gene conversion," "flanking-marker inheritance," "Holliday junction resolution," and "crossover versus non-crossover products" became part of my lexicon. The genetics of the phage and bacterial systems were simply so facile that we dove into dissecting the intron mobility pathways. After all, molecular biology had its origins in phage crosses, and I was not about to break with tradition. How could I, when my mobile-intron colleagues, who performed their studies in mitochondria and chloroplasts, were faced with similar mechanistic questions but needed to conduct ballistics to move DNA into these organelles to get their questions answered?

In short, we demonstrated that the intron-encoded endonuclease was responsible for making a double-strand break.

---

**FIGURE 2. RNA: Catalysis, Splicing, Evolution.** A, a conference by this name was held in 1988 in Rensselaerville, NY. Articles were published in a special issue of Gene and in a corresponding book in 1989. B, some speakers at the conference were as follows: 1, Bernard Dujon; 2, John Belote; 3, Thomas Cech; 4, Christine Guthrie; 5, Olke Uhlenbeck; 6, Norman Pace; 7, John Abelson; 8, Phillip Sharp; 9, Sidney Altman; 10, Hugh Robertson; 11, Alan Weiner; 12, John Boothroyd; 13, Nancy Maizels; 14, Francois Michel; 15, Michael Green; 16, Susan Wessler; 17, Michael Rosbash; 18, Marlene Belfort; 19, George Bruening; 20, David Shub; and 21, Philip Perlman.
(DSB) at the intron homing site (13, 16). DNA cleavage paralleled the situation for the fungal mitochondrial introns (14). We were then able to take the plunge into recombination and showed that the td intron mobilizes via homology-dependent DSB repair (DSBR) and synthesis-dependent strand annealing (SDSA) pathways (Fig. 1C) (17, 18). These pathways differ in that DSBR involves recombination per se, generating both crossover and non-crossover products, whereas SDSA is replication-based, yielding only non-crossover products. As it turns out, both pathways are embedded in the context of recombination-dependent replication of phage T4, usurping replication, recombination, and repair functions to complete the mobility process (17–20). It became clear that we had been correct in viewing these parasitic introns as cunning scavengers of functions of their hosts (21).

**Intron-encoded Endonucleases: Organisms unto Themselves**

Shortly after the discovery of the td intron, the endonuclease encoded by that intron, called I-TevI, was shown to share a sequence motif with a protein of filamentous fungi (22). This motif, GIY-YIG, is characteristic of an entire family of endonucleases that nick or
cleave DNA. We now know that this superfamily includes restriction enzymes, recombination and repair proteins, and retrotransposons in all three domains of life (23). Clearly, GIY-YIG-containing nuclease modules have been shuffled with other protein domains throughout their evolutionary history.

Through a productive partnership with colleagues Vicky Derbyshire and Patrick Van Roey, we characterized the structure and function of the 28-kDa I-TevI. The GIY-YIG motif is in the ~90-amino acid catalytic domain (24–26), which is joined by a 75-amino acid linker (27) to an elongated 80-residue DNA-binding domain (Fig. 3A) (28–30). Strikingly, the I-TevI recognition sequence, the td homing site, spans three turns of the DNA helix (28, 31, 32). I-TevI binds this site as a monomer (33), winding around the primary binding region of ~20 bp that is centered on the intron insertion site (28).

The catalytic module, joined to the DNA-binding domain via the linker, uses both sequence and distance determinants to select the remote cleavage site. The DSB comprises two nicks, one at 23 nucleotides (nt) upstream of the intron insertion site and the other at 25 nt upstream, on the top and bottom strands, respectively (16). If the preferred cleavage site (CXXXG) is displaced from the optimal distance of 23 and 25 nt, the enzyme will search within a finite window upstream and downstream, and if an optimal site is not found, cleavage will occur at reduced efficiency at the normal distance (Fig. 3B) (29, 34, 35). This ruler function resides within the linker, which includes a zinc finger (27, 29). We postulate that such a level of complexity evolved because I-TevI has a second function, namely to regulate its own synthesis by acting as an autorepressor (35, 36), as will be described below.

In as yet unpublished findings, we have discovered that the zinc finger in the linker is redox-sensitive. Because I-TevI cleavage is regulated by the linker, we are tempted to speculate that mobility of the intron responds to oxidative stress of the organism (J. Robbins, D. Smith, and M. Belfort, unpublished data). We have shown recently that a different kind of self-splicing mobile intron, a group II intron, disperses to new sites under conditions of nutritional stress (37).

**Regulation of td Intron Mobility and I-TevI Self-control**

Remarkably, although I-TevI is transcribed along with the td gene early in phage infection, it is not translated until late in the infectious cycle, being under the control of a T4 late promoter (P_L). Protein synthesis from the primary transcript is inhibited by a stem-loop structure that involves the ribosome-binding site (Fig. 3C). Only when a transcript is initiated proximal to the I-TevI sequence at P_L are ribosomes able to bind, and the RNA can then be translated (38). A second level of control is exerted by I-TevI itself, which, as noted above, can act as an autorepressor by binding to an operator that blocks P_L (Fig. 3D) (35). Thus, when I-TevI binds the operator, which shares sequence similarity with the primary I-TevI binding sequence, repression occurs because the preferred cleavage sequence (CXXXG) is not within range. In contrast, when the protein binds the homing site, in which the CXXXG sequence is at the optimal distance, cleavage ensues.

Delays in expression until late in infection and autorepression appear to pertain similarly to I-TevII and I-TevIII, the respective nrdD and nrdD intron endonucleases (35, 38). Why, one might ask, is there such strict down-regulation? Spurious DNA breaks early in infection may be deleterious to the infectious cycle. Indeed, timing of both ectopic and directed breaks, the latter to promote homing late in infection, when the DNA repair and replication machinery is plentiful, seems highly expedient.

**Evolution**

We hypothesized that mobile introns, like the td intron, arose because the self-splicing element played host to an invasive endonuclease gene, which entered the genome by virtue of the DNA break created by its product. Indeed, the capacity for recognition of intron sequences by an endonuclease supports this hypothesis (39). Genomic shuffling is thought to extend beyond endonuclease invasion to the genesis of the I-TevI-like modular nucleases that make the DSBs. These enzymes have rapidly evolved a broad range of binding specificities through exchanging GIY-

---

**FIGURE 3. Structure and regulation of I-TevI.** A, structure of I-TevI. A globular catalytic domain is joined by a linker to an extended DNA-binding domain, which wraps around the DNA homing site (see cross-section). ZF, zinc finger; HTH, helix-turn-helix. Cleavage occurs at a fixed distance from the intron insertion site (red outline). B, I-TevI has both sequence and distance determinants. I-TevI cleaves at a distance of 23 nt on the top strand at a preferred sequence (CXXXG). When the cleavage site is moved downstream or upstream (e.g. to 18 or 28 nt), the wild-type (WT) enzyme defaults to distance (solid red). However, linker mutants have lost the ruler mechanism and can either reach out or retract to cleave at the preferred sequence (red outline). C, sequence; D, distance; HS, homing site. C, I-TevI is subject to transcriptional and translational control. I-TevI cannot be translated from the primary td transcript because a stem-loop structure sequesters the ribosome-binding site (RBS). A late transcript originates from P_L, which has a free ribosome-binding site for translation late in infection. D, I-TevI moonlights as an autorepressor. I-TevI can bind the P_L sequence, but, because the preferred cleavage sequence is not within range, repression rather than RNA cleavage occurs. In contrast, when binding to the homing site, cleavage at CXXXG and homing ensue.
YIG-like catalytic cassettes with a variety of DNA-binding domains (23).

It remains perplexing, however, as to why introns are maintained in streamlined phage genomes. At first, we considered regulatory rationales related to the host genes involved in DNA metabolism. We then pondered whether recombination, which is known to occur among the phage introns, provides a long-term advantage (40). We also speculated whether these mobile introns are simply parasitic elements, with their invasiveness accounting for their persistence (21). It is indeed likely that intron selfishness accounts for their presence in streamlined genomes. However, the host organism might eventually adapt to their presence in useful ways. This “lemonade-from-lemons” scenario is one that is responsible for the persistence of mobile elements in all three domains of life (41).

There has been a satisfying convergence of discoveries made 25 years apart. One observation resulted from study of the TS enzyme that hosts the td intron; the other arose from analyzing the cleavage specificity of I-TevI, the homing endonuclease encoded by the td intron. The favored cleavage sequence of I-TevI within the td gene corresponds precisely to the coding sequence of that invariant active-site arginine of TS mentioned early in this Reflections article (1, 42). How cunning of a homing endonuclease to select nucleotides that are conserved throughout the TS family to not only ensure homing targets and intron spread but to also guard against haphazard intron loss by occupying a site that has no margin of error!

**Lessons from 30 Years Spent with a Single Kilobase of DNA**

Although other kinds of introns, as well as protein splicing elements known as inteins, are also research foci in our laboratory, the group I td intron has provided a persistent thread over 30 years. The initial serendipitous discovery of an intron in a bacteriophage gene, at a time when only eukaryotes were known to contain introns, was followed by several other surprises: guanosine-initiated group I splicing, intron mobility, homing endonuclease regulation, distance measurement for nucleolytic cleavage, and DNA endonuclease moonlighting as an autorepressor. Additionally, there were multiple evolutionary insights along the way. The journey was a mix of following one’s nose and formulating hypotheses to discover the beauty and complexity of an intron. Eventually, a depth perception emerged, allowing us to understand in great detail a parasitic element (the intron), its host enzyme (TS), and the relationship between the two. The work has taught us both to focus and to think broadly and to repeatedly zoom in and out. There were times when a single nucleotide or amino acid became the object of investigation. At other times, our attention swept across species, genus, even kingdom lines, when horizontal gene transfer of mobile introns was the issue.

There were valuable lessons to be learned during these intellectual travels. First, from early on as a microbial geneticist, I realized that, for my studies to become more definitive, I would need to cross disciplines and to apply biochemistry, then structural biology, and later bioinformatics to my work. Fine collaborators have been invaluable guides and companions in these excursions. Second, the value of basic research became ever clearer, both to enhance our fundamental understanding and for application. As a clear example of the latter, the use of site-specific homing endonucleases, like I-TevI and its molecular cousins, has become a cottage industry in DNA manipulation for the biomedical research enterprise and for gene therapy (43). Finally, I am struck by how science has changed over these 30 years. Back then, it was about a problem of confined scope; today, the span is enormous, from interrogating single molecules to querying entire systems. This range of approaches continues to unearth the rich properties, dynamics, and evolutionary potential of a single kilobase of DNA, its RNA and protein products.

Acknowledgments—This work was originally supported by the National Science Foundation and has since been supported by two longstanding grants from the National Institutes of Health (Grants GM39422 and GM44844). I thank Maryellen Carl for expert handling of the manuscript and John Damsareau for professional rendering of the figures for this publication and many that have gone before. This chapter is dedicated first, to my husband, Georges, and my three sons, David, Gabi, and Yona, who have made all my work possible, and second, to those students, postdoctoral graduates, technicians, and collaborators whose names appear only in the reference list. They did the work and stimulated much of the thinking.

Address correspondence to: belfort@wadsworth.org.

**REFERENCES**

1. Belfort, M., Maley, G. F., and Maley, F. (1980) Arch. Biochem. Biophys. **204**, 340–349
2. Belfort, M., and Pedersen-Lane, J. (1984) J. Bacteriol. **160**, 371–378
3. Chu, F. K., Maley, G. F., Maley, F., and Belfort, M. (1984) Proc. Natl. Acad. Sci. USA. **81**, 3049–3053
4. Belfort, M., Pedersen-Lane, J., West, D., Ehrenman, K., Maley, G., Chu, F., and Maley, F. (1985) Cell **41**, 375–382
5. Chu, F., Maley, G. F., Belfort, M., and Maley, F. (1985) J. Biol. Chem. **260**, 10680–10688
6. Belfort, M., Pedersen-Lane, J., Ehrenman, K., Chu, F. K., Maley, G. F., Maley, F., McPheeters, D. S., and Gold, L. (1986) Gene **41**, 93–102
7. Chu, F. K., Maley, G. F., West, D. K., Belfort, M., and Maley, F. (1986) Cell **45**, 157–166
8. Ehrenman, K., Pedersen-Lane, J., West, D., Herrman, R., Maley, F., and Belfort, M. (1986) Proc. Natl. Acad. Sci. USA. **83**, 5875–5879
9. Kruger, K., Grabowski, P. J., Zaugg, A. J., Sands, J., Gottschling, D. E., and Cecchini, T. R. (1982) Cell **31**, 147–157
10. Gott, J. M., Shub, D. A., and Belfort, M. (1986) Cell **47**, 81–87
11. Pedersen-Lane, J., and Belfort, M. (1987) Science **237**, 182–184
12. Quirk, S. M., Bell-Pedersen, D., Tomaschewski, J., Rüger, W., and Belfort, M. (1989) *Nucleic Acids Res.* 17, 301–315
13. Quirk, S. M., Bell-Pedersen, D., and Belfort, M. (1989) *Cell* 56, 465–465
14. Lambowitz, A. M. (1989) *Cell* 56, 323–326
15. Belfort, M., and Shub, D. A. (1989) RNA: Catalysis, Splicing, Evolution, Elsevier Science Publishing Co., Inc., New York
16. Bell-Pedersen, D., Quirk, S. M., Aubrey, M., and Belfort, M. (1989) *Gene* 82, 119–126
17. Mueller, J. E., Clyman, J., Huang, Y. J., Parker, M. M., and Belfort, M. (1996) *Genes Dev.* 10, 351–364
18. Baylin, S. B., and Belfort, M. (1989) *Genes Dev.* 3, 1386–1394
19. Arnold, D. J., and Belfort, M. (1989) *Proc. Natl. Acad. Sci. U.S.A.* 86, 11983–11987
20. Bryk, M., and Belfort, M. (1990) *Nature* 346, 394–396
21. Beauregard, A., Curcio, M. J., and Belfort, M. (2008) *Annu. Rev. Genet.* 42, 587–617
22. Edgell, D. R., Stanger, M. J., and Belfort, M. (2004) *J. Mol. Biol.* 343, 1231–1241
23. Belfort, M. (2005) in *Homing Endonucleases and Inteins* (Belfort, M., Derbyshire, V., Stoddard, B. L., and Wood, D. W., eds) pp. 1–10, Springer-Verlag, Heidelberg, Germany