The discovery of “split” genes: a scientific revolution
Jan A. Witkowski, Cold Spring Harbor Laboratory, Cold Spring Harbor NY 11724
(Published in Trends in Biochemical Sciences 13: 110-113, 1988)

In 1962 T. S. Kuhn published a seminal book entitled The Structure of Scientific Revolutions in which he suggested that changes in scientific thought come about as revolutions in which an existing theory is replaced by a more or less radical theory, rather than by a gradual cumulative process. Kuhn’s work itself sparked something of a revolution in studies of the history of science and the general theme of his work is largely accepted. However, it is striking that most examples of revolutions have been drawn from the physical sciences, a peculiarity recognized by Kuhn himself and by Cohen in his magisterial book Revolution in Science. I think that there have been many revolutions in the biological sciences and in the early part of 1977 a quite stunning example occurred, the discovery of “split” genes.

Papers describing phenomena attributable to the presence of noncoding, intervening sequences had been published as early as 1974, but it was not until the first half of 1977 that splicing was proposed as the mechanism to account for these phenomena. It was a remarkable feature of this revolution that once the phenomenon had been recognized, results obtained in a wide variety of systems were understood. In alphabetical order these included the adenovirus late region, the chick ovalbumin gene, Drosophila ribosomal genes, mouse, and rabbit globin genes, Simian Virus 40 (SV40) and yeast tRNA. I shall describe these various discoveries, but some disclaimers are required. I have used only published materials as my sources so that this is not a definitive account of the discovery of “split” genes. I am not concerned with trying to establish a precise chronology for these events or with trying to assign priority for the discovery. Instead I want to use the discovery of intervening sequences to illustrate the various aspects of a scientific revolution as defined by Kuhn and Cohen.

A scientific revolution
What constitutes a scientific revolution? Kuhn emphasized that scientists work at solving puzzles within the framework of the existing theory or “paradigm” and that this “normal” science is what occupies most scientists most of the time. As research progresses, an increasing number of findings cannot be accounted for by the current theory and eventually a scientist or group of scientists proposes a radical modification or even a complete replacement of the current theory. The new theory encompasses the older findings as well as those that had been anomalies under the old theory, and in addition the new theory suggests lines of research that were formally undreamt of. The new theory is accepted with varying degrees of enthusiasm but, if successful, it eventually becomes the new paradigm until it in its turn is replaced by further developments.

How does one recognize a scientific revolution? Cohen suggests four tests that can be used to determine if a change in scientific thought counts as a revolution. Firstly, there is what observers of the scene thought at the time; secondly, there is the later documentary evidence such as is found in reviews and textbooks; thirdly, there is the judgement of historians of science; and finally there is the opinion of scientists working in the field today.

I shall treat the topic system by system, but there is the difficulty that I cannot cover all the work that was done and I shall have to be selective. I shall deal with papers published up to and including 1977, and choose papers illustrating Kuhn’s and Cohen’s ideas of a scientific revolution.

Messenger RNA
The prediction of the existence of mRNA is generally regarded as one of the intellectual breakthroughs in the history of molecular biology. The classic experiment was performed by Brenner, Jacob and Meselson\(^3\) although previously others, in particular Volkin and Astrachan\(^,\) had obtained results that were highly suggestive of the existence of mRNA. In their discussion, Brenner et al. remarked that: “It is a prediction that the messenger RNA should be a simple copy of the gene, and its nucleotide composition should therefore correspond to that of the DNA”. This was the unquestioned paradigm that had guided the way people thought about the relationship between the nucleotide sequences of a gene and its mRNA, and the amino acid sequence of the protein. (The discovery of mRNA qualifies in its own right as a scientific revolution and exciting descriptions by Brenner, Crick and Jacob will be found in Ref. 5). How was the paradigm overthrown?

**Drosophila ribosomal DNA**

In early 1977* Cell* devoted 57 pages to reports from three laboratories on the organization of the ribosomal genes of *Drosophila*. The studies had been carried out using restriction mapping but principally by exploring the R loop technique of White and Hogness\(^6\). In this technique, single-stranded RNA is hybridized to double-stranded DNA under conditions that favor RNA-DNA hybrids. The RNA displaces one of the DNA strands of the duplex which forms a loop that indicates the site of transcription of the RNA. The conclusions of three groups\(^6-9\) were the same. In *Drosophila*, rDNA exists in two forms of repeating units, 17 kb and 11.4 kb in length. What was curious was that the 28S rDNA was in two parts, separated by an insert of variable length ranging from 0.5kb to 6.5kb. However, some 55% of the 28S genes were not interrupted in this way and it could be argued that the interrupted genes were not transcribed, or if they were transcribed they were not functional; there was no evidence that the presence of inserted sequences was a functional feature of these genes.

**Adenovirus**

It was hoped that study of the relatively small adenovirus (Ad2) genome and its transcription would shed light on the processes of gene regulation in eukaryotes. It was the subject of intense investigation in the 1960s and 1970s (and remains so today) and work on adenovirus is particularly well documented. By the mid-1970s, anomalies were observed in three separate investigations using adenovirus.

One approach was to map the transcriptional units by isolating mRNAs and determining what they coded for by in vitro translation and polyacrylamide gel electrophoresis. The first results\(^11\) using mRNAs fractionated on the basis of size showed that some viral mRNAs were much longer than predicted from the amino acid sequences of their protein products. This approach was refined by using EcoRI restriction fragments of the adenovirus late genes to select specific RNAs; as the order of the EcoRI fragments was known, the gene products could be assigned to particular regions of the adenovirus genome\(^12\). There were two puzzling features. Firstly, there were minor protein bands that were explained away by the presence of contaminating non-homologous mRNA (a justifiable assumption at that time and one that became questionable only with hindsight). Secondly, it was found that significant quantities of P-VIII were synthesized by RNAs selected by EcoRI fragments A, and F and D that were widely separated in the adenovirus genome. It was suggested that “The information for the two polypeptides is contained in separate mRNA molecules, but these molecules contain some common sequences”\(^12\).

Another curious result had been obtained by Gelinas and Roberts\(^13\) who analysed the sequences at the 5’ ends of the adenovirus mRNAs and then mapped these in the genome. Remarkably, they found that there was a single 5’ capped oligonucleotide. This was “unexpected” because many adenovirus mRNAs were present late after infection, and Gelinas and Roberts had “... anticipated one 5’ capped oligonucleotide DNA for each mRNA species”. There was one further puzzling feature; Gelinas and Roberts were “surprised” to find that the 5’ oligonucleotide was very sensitive to nuclease digestion in mRNA-DNA hybrids when it should have been protected from attack. They suggested that the 5’ terminus of mRNA or of its DNA complement was able to make a hair-pin loop. The 5’ oligonucleotide might have some regulatory function for which it was distributed at a number of sites in a large transcript that was subsequently “processed” to give individual mRNAs.
Finally in this pre-revolution period, groups at Cold Spring Harbor Laboratory (Chow et al. \(^{14}\)) and at the National Institute of Health (Westphal et al. \(^{15}\)) were using electron microscopy and R loop mapping to determine the sites of transcription of late mRNAs. Chow et al. noticed that about 25\% of loops made with late mRNAs had short free 5’ ends that did not hybridize to the DNA. They referred to the results of Gelinas and Roberts, remarking that the 5’ tails might be due to “... the presence of sequences on the mRNA which are not encoded in the region flanking the hybridized transcript”. Westphal’s group observed similar tails but suggested that they represented 3’poly(A) tails or were artefacts resulting from displacement of RNA by re-annealing DNA.

These various anomalies were resolved in the early part of 1977 when Berget et al. \(^{16}\) and the Cold Spring Harbor Laboratory groups\(^{17-20}\) submitted papers to *Proceedings of the National Academy of Sciences* and *Cell* respectively.

Berget et al.\(^{16}\) had been using the R looping technique to map the hexon mRNA to the Ad2 genome. When hexon mRNA was hybridized to the HindIII A fragment, an R loop was formed, but there were also short 5’ and 3’ tails. A similar result was obtained when the mRNA was hybridized to single-stranded HindIII-digested DNA; double-stranded hybrids were formed, again with short single-stranded 5’ and 3’ tails. The most spectacular result came from the hybridization of hexon mRNA to single-stranded EcoRI A fragment. No fewer than three loops of single-stranded DNA formed at the 5’ end of the hybrid molecule, indicating that the 5’ end of the mRNA was transcribed from map positions 16.8, 19.8 and 26.9. These sequences, Berget et al. suggested, were “spliced” to the main body of the mRNA during processing of a long precursor mRNA.

Chow et al.\(^{17}\) continued their studies by identifying the origins of the free 5’ tails that they had observed using single-stranded DNA restriction fragments as probes. Their results were identical to those of Berget et al.; sequences from map positions 16.6, 19.6 and 26.6 were complementary to the 5’ tails of mRNA that were transcribed to the right of position 36. With a degree of understatement, Chow et al. remarked that their results were “... not directly consistent with any mechanism previously suggested for the biosynthesis of mRNA in eukaryotic cells”.

What of the anomalies in the earlier studies using in vitro translation of mRNA selected with restriction fragments? Lewis et al. \(^{18}\) used other enzymes to produce smaller restriction fragments and found that the HpaI C fragment selected mRNA for IVa2, 15k and IX polypeptides, but it and the neighbouring fragment selected for almost all late mRNA at levels that were substantially higher than those obtained with A DNA used as a negative control. These findings “... were initially difficult to explain” but interpretation had been “... greatly facilitated recently by several different investigations ... “. In short, it was now clear that several late mRNAs had sequences that were transcribed from a common site away from the main transcription region of each mRNA.

Finally, Klessig\(^{19}\) performed a refined analysis of the 5’ oligonucleotide studied by Gelinas and Roberts\(^{13}\), concentrating on producing highly purified fiber and 100K mRNA. The purified mRNAs had the same 5’ oligonucleotide, confirming the results of Gelinas and Roberts. Klessig tried to purify the mRNAs by hybridizing them to the EcoRI DNA fragments and then treating with RNase. The DNA fragments used coded for the 5’ ends of the mRNAs but failed to protect the 5’ oligonucleotide from digestion. Furthermore, the 5’ oligonucleotide was protected in DNA-RNA hybrids using the BamHI (0-29.1) and HindIII (7.5-17.00) DNA fragments. Klessig wrote: “The scheme that emerges for the biosynthesis of Ad2 mRNAs is unlike any hitherto described or postulated”.

These studies followed on from previous lines of investigation, but Dunn and Hassell\(^{20}\) presented new data from an analysis of the transcription of SV 40- Ad2 hybrids. Their results showed that the transcripts included sequences that were distant from the site of insertion of the SV40 DNA in the Ad2 genome. They suggested that what was needed was a direct sequence comparison between a gene and the mRNA transcribed from it.

**Yeast tRNA genes**
The first sequence comparisons seem to have been carried out not using an mRNA but rather using yeast tRNA\textsuperscript{tyr} genes. Goodman, Olson and Hall\textsuperscript{21} identified an EcoRI fragment that contained the SUP4 locus and used this as a probe to screen an ochre suppressor SUP4-o yeast \lambda library. They then determined the sequences of the mutant gene and three wild-type genes and found that the genes contained a stretch of 14 nucleotides to the 3’ side of the anticodon triplet that was not present in tRNA\textsuperscript{tyr}. The insert was present in functional genes and it was unlikely to be a cloning artefact. As they remarked then, “...it is difficult to muster any real evidence on how such genes or gene systems might work”.

The rabbit and mouse β-globin genes

The experiment suggested by Dunn and Hassell had already been done in part by Jeffreys and Flavell\textsuperscript{22} using the rabbit β-globin mRNA, and their report appeared in the same historic issue of Cell as did the Ad2 work. In fact Jeffreys and Flavell had compared restriction maps rather than sequence data but the results were convincing nevertheless. Two examples of their anomalous results must suffice here: (1) the enzyme HaeIII liberated a 333 bp fragment from the cDNA but an 800 bp fragment from the genomic DNA; (2) a BamHI site known to be 67 bp from an EcoRI site was found to be in a ‘strange’ position some 700 bp from the EcoRI site in the genomic DNA. Jeffreys and Flavell thought that it was remarkable that there was a 600 bp fragment within the gene that did not code for globin. With considerable understatement, echoing a paper published 23 years earlier, they went on that if this was a general phenomenon, “…then there are certain implications for recombinant DNA research”.

Tilghman et al.\textsuperscript{23} worked on the mouse β-globin gene and obtained similar results. They found that HindIII and SstI which failed to cut the cDNA, readily cleaved the cloned gene (their italics). R looping was also used and ‘surprisingly’ two loops instead of the expected one were found, indicating the presence of an inserted fragment. Tilghman et al. sequenced 40 bp across the suspected junction of the inserted DNA and found that the DNA sequence agreed with the known amino acid sequence for residues 93-104, but that it then diverged and the DNA sequence did not code for at least the next five amino acids. Here was direct evidence for a non-translated region within a eukaryotic gene.

The chicken ovalbumin gene

In the early days of gene cloning it was tremendously advantageous to work with genes that were highly expressed, for example the globin genes and the chicken ovalbumin gene. Breathnach, Mandel and Chambon\textsuperscript{24} “unexpectedly” found that the ovalbumin gene contained inserts. For example, the cDNA contained no HindIII or EcoRI sites, yet the cDNA probe detected several bands in genomic DNA cut with these enzymes. Comparisons of the lengths of fragments obtained with other enzymes produced a ‘striking anomaly”, once again demonstrating the presence of inserts within a gene. It was possible that the unusual arrangement of the ovalbumin gene in these cells was related in some way to its regulation; the cells of the tubular gland of the oviduct respond to hormonal stimulation by synthesizing very large amounts of ovalbumin. However, Breathnach et al. showed that the same insert was present in DNA from erythrocytes and from 5-day embryos.

SV40 transcription

By 1977 the SV40 genome was being mapped and functions were being assigned to its various regions. However, the sequences between 0.67 and 0.76 that were transcribed late in infection had not been ascribed a function. Aloni et al.\textsuperscript{25} set out to remedy this and began by isolating mRNA hybridizing to DNA in the 0.67-0.76 region and also RNA hybridizing to DNA fragments from the region 0.0-0.17 that coded for 16S and 19S RNA. They found that these two sets of RNAs hybridized to both sets of restriction fragments and that analysis on sucrose gradients showed that they all contained 16S and 19S RNA. Mapping the 5’ ends of these RNAs showed that they were derived from the 0.67 region. Aloni et al. concluded that leader sequences (0.67-0.76) of the 16S RNAs were added to transcripts from the 0.95-0.17 region and they speculated that the deleted region might have some regulatory function.

“Split” genes as a Kuhnian scientific revolution

The first stage as described by Kuhn is the accumulation of anomalies that cannot be explained by existing theory. It is clear that the scientists working in this field were all too aware that they were finding anomalies. Words such as ‘surprised”, “unexpected”, ‘strange”, “ambiguity” and “anomaly” recur
repeatedly in papers published in this period. The results were not easy to explain for the old ideas on the relationship between the gene and its mRNA had failed. It had been recognized that hnRNA must be processed but it was not thought that the processing might involve the removal of sequences from within coding regions. (Although studies of hnRNA were beginning to suggest this might occur.) As Weinstock et al. 26 later remarked, their results using the ovalbumin gene were “not predictable from an examination of the restriction map of mRNA”. Crick wrote: “What is remarkable is that the possibility of splicing had not at any time been seriously considered before it was forced upon us by the experimental facts.... Lacking evidence, we have become overconfident in the generality of our ideas.”27. This idea of being “forced” is explicit in some writings. O’Malley and Woo’s laboratory wrote of their work 29 on the ovalbumin gene that “...the existence of intervening sequences within structural genes was unpredicted by our prior understanding of eukaryotic genes”. They were “obligated” to conclude that the coding sequences were non-contiguous in the gene and Dunn and Hassell 20 wrote of their work that this conclusion seemed “unavoidable” (my italics). Yet once it was admitted that mRNAs “…possess a structure-fundamentally different from any mRNAs hitherto described”, the observed anomalies became excellent data to support the new theory.

Splicing was not only compatible with earlier results but clearly it opened up a whole new area for research. Paper after paper closed with references to the consequences of splicing for obtaining transcription of cloned genes, and to speculations on the functions of the inserted sequences and the 5’ leader sequences, and of course on the mechanism of splicing itself. It is not necessary to elaborate how splicing has become the new paradigm and in its turn has led to another revolution, the recognition that the RNA possesses catalytic activity once thought to be the exclusive property of proteins.30. The implications for studies of gene evolution are only just being explored.

Cohen’s criteria
The RNA splicing story seems to fit the Kuhnian model quite well; how well does it satisfy Cohen’s four criteria? Firstly, it is obvious that the participants recognized that they were in the midst of a remarkable scientific event. The language used in the papers expresses astonishment at what was being discovered, exemplified by the title of the paper by Chow et al., “An Amazing Sequence Arrangement at the 5’ ends of the Adenovirus 2 Messenger RNA” 27. Watson 30 wrote in the 1977 Annual Report of Cold Spring Harbor Laboratory that ‘splicing is likely to be the most important observation in molecular genetics since the 1960 discovery of messenger RNA” and with uncanny prescience, the 1977 Cold Spring Harbor Quantitative Symposium was devoted to chromatin. The summaries of the meeting written by Chambron 31 and Sambrook 32 again show that these events were recognized at the time as being revolutionary, and Watson 33 remarked that the participants left “feeling that they had been part of an historic occasion”. Cohen lays great emphasis on contemporary opinion, particularly if the word “revolution” is used. At least three references in the early years referred to splicing as revolutionary; “In the last 2 years there has been a mini-revolution in molecular genetics” (Crick 27); “...revolutionize our ideas of genes and the proteins for which they code” (Watson 30); “Our picture of the organization of genes in higher organisms has recently undergone a revolution” (Gilbert 34).

Secondly, the later documentary evidence from reviews and textbooks is overwhelming. In his 1979 review Damell 35 wrote that it was a “most astonishing discovery”. All undergraduate text books on molecular biology describe splicing and Lewin devotes a chapter of 21 pages in Genes 36 to a discussion of splicing. It is ironic that the third edition of Watson’s classic text The Molecular Biology of the Gene 37 appeared in 1976, just prior to the discovery of splicing, but in the fourth edition 38 splicing is covered extensively, including a chapter dealing with the evolutionary implications.

It is difficult to satisfy Cohen’s third criterion that the opinions of competent historians of science should be consulted because the events of only ten years ago have not yet attracted analysis by professional historians. I hope that there will be an increasing trend for historians to deal with recent events in areas like molecular biology where changes occur with disturbing but exciting speed.

I have discussed this topic with many scientists who were not directly involved in the discovery of splicing, and there is a unanimous opinion that the discovery was indeed a revolution. To a degree, events
of that period have taken on a mythical quality, with stories of how word of the American findings first reached London and were received with frank disbelief, and discussions of whether a Nobel Prize could be awarded for the work. (The discussion does not center on whether the work justifies a prize, but who would receive it!)

Conclusion
The discovery of non-coding, intervening sequences in genes qualifies as a scientific revolution according to the tests of both Kuhn and Cohen. It was perceived as such at the time and it has opened up new areas of research that were undreamt of in the pre-revolutionary period. The paradigm that it overthrew was that there was a direct correspondence at the nucleotide level between a gene and its RNA, and that the coding sequence of a gene was continuous without interruption. As Broker and Chow wrote: “This startling fact [splicing] has abruptly changed our conception of what a eukaryotic gene is and how its expression is controlled”.39

4 Volkin, E. and Astrachan, L. (1956) Virology 2,149
5 Judson, H. F. (1979) The Eighth Day of Creation, Jonathan Cape
6 White, R. L. and Hogness, D. S. (1977) Cell 10, 177-192
9 Pellegrini, M., Manning, J. and Davidson, N. (1977) Cell 10, 213-224
27 Crick, F. (1979) *Science* 204, 264-271