# GENES V

## Benjamin Lewin

OXFORD UNIVERSITY PRESS
Oxford New York Tokyo

Tecan Ex. 2080, Page 1 of 28
IDT, Inc. v. Tecan Genomics - IPR2025-00016

*Oxford University Press, Walton Street, Oxford OX2 6DP*

*Oxford New York*
*Athens Auckland Bangkok Bombay*
*Calcutta Cape Town Dar es Salaam Delhi*
*Florence Hong Kong Istanbul Karachi*
*Kuala Lumpur Madras Madrid Melbourne*
*Mexico City Nairobi Paris Singapore*
*Taipei Tokyo Toronto*

*Oxford is a trade mark of Oxford University Press*

*Published in the United States*
*by Oxford University Press Inc., New York*

*© Oxford University Press and Cell Press, 1994*
*First published 1994*
*Reprinted 1994*

*All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, without the prior permission in writing of Oxford University Press. Within the UK, exceptions are allowed in respect of any fair dealing for the purpose of research or private study, or criticism or review, as permitted under the Copyright, Designs and Patents Act, 1988, or in the case of reprographic reproduction in accordance with the terms of licences issued by the Copyright Licensing Agency. Enquiries concerning reproduction outside those terms and in other countries should be sent to the Rights Department, Oxford University Press, at the address above.*

*This book is sold subject to the condition that it shall not, by way of trade or otherwise, be lent, re-sold, hired out, or otherwise circulated without the publisher's prior consent in any form of binding or cover other than that in which it is published and without a similar condition including this condition being imposed on the purchaser.*

*A catalogue record for this book is available from the British Library*

*Library of Congress Cataloging in Publication Data*

*Printed in U.S.A.*

# CHAPTER 23

# The eukaryotic gene: conserved exons and unique introns

A major change in our view of the gene at the molecular level came in 1977 as a result of the discovery that eukaryotic genes may be interrupted. For some years previously, there had been uneasy suspicions that eukaryotic genes might be unusual, that they might differ in some fundamental way from bacterial genes. The roots of this idea lay in the apparently excessive size of eukaryotic genomes described by the C-value paradox.

One escape from this dilemma is to suppose that a large proportion (even a majority) of DNA sequences are not part of the structural genes, but have some other (unknown) function. Another idea is that the structural gene *is much larger than the sequence represented in mRNA.* The additional length must be removed when the RNA is **processed** for transport to the cytoplasm. This idea was supported by the fact that nuclear RNA is much longer than mRNA (see Chapter 31).

The original concept for a large eukaryotic gene supposed that the transcription unit might include extensive sequences on one or the other side of the region represented in mRNA. In addition, if extensive nontranscribed regions were needed for regulation, the unit of gene expression might be rather large relative to the size of the mRNA.

Indeed it turns out that many eukaryotic genes are much longer than their mRNAs. But it was a great surprise that the cause of the discrepancy proved to be the existence of *interruptions that separate different parts of the coding region in DNA.* It is also the case that there may be quite extensive regulatory regions upstream of the startpoint for transcription, but these generally are measured in less than hundreds of base pairs, compared with the thousands of base pairs in the transcription units.

The existence of interrupted genes was revealed by experiments to identify directly the DNA of the gene. The intention of this approach was to proceed back from known mRNAs to their genes, to recover intact transcription units, including flanking sequences such as promoters and other regulatory elements not necessarily represented in the mRNA.

On the basis of experience with bacteria, it was assumed that eukaryotic mRNA would have the same sequence as the DNA from which it is transcribed. As a control to show that an isolated genomic DNA sequence did indeed correspond with the mRNA used to isolate it, their sequences were compared, either by electron microscopy or by restriction mapping. But what this revealed was a discrepancy between the sequences, in the form of *additional regions present in the genomic DNA but absent from mRNA.*

The presence of additional sequences explains the size discrepancy between nuclear RNA and mRNA. The nucleus contains primary transcripts that correspond to the sequence of the gene itself.

Tecan Ex. 2080, Page 3 of 28
IDT, Inc. v. Tecan Genomics - IPR2025-00016

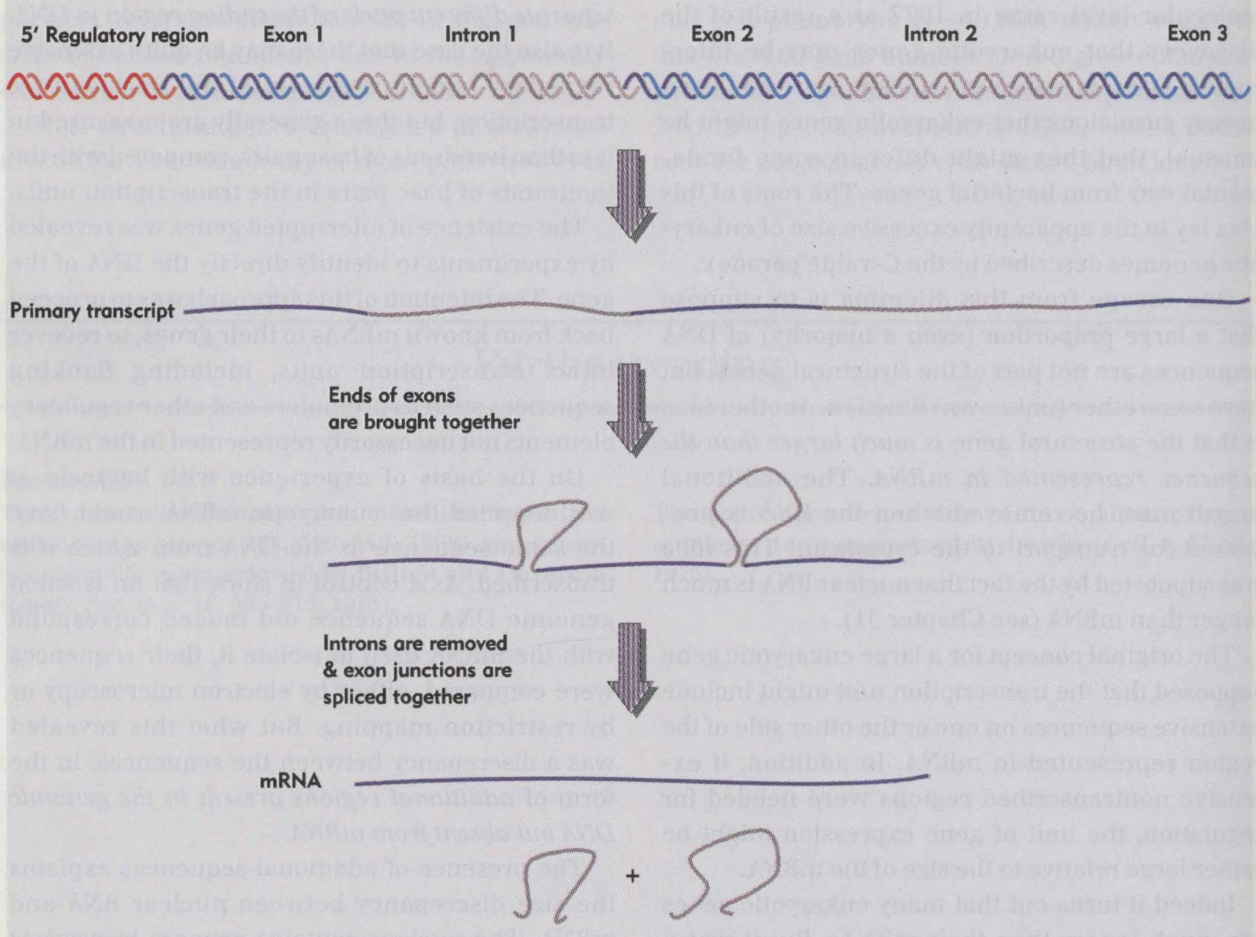
Those parts of the sequence not found in mRNA are removed from the primary transcript before the RNA is exported to the cytoplasm to serve as a messenger.

Recall the terminology for describing the relationship between a gene and its RNA product (see Chapter 6). **Figure 23.1** shows that an interrupted gene consists of an alternating series of **exons** and **introns**:

- The exons are the sequences represented in the mature RNA. By definition, a gene starts and ends with exons, corresponding to the 5′ and 3′ ends of the RNA.
- The introns are the intervening sequences that are removed when the primary transcript is processed to give the mature RNA.

**Figure 23.1**

*Overview:* an interrupted gene consists of alternating exons and introns; the introns are removed by splicing of the RNA transcript, generating an mRNA consisting only of exon sequences. The regulatory region includes the promoter at the 5′ end.



The process by which introns are removed from the primary transcript is called **RNA splicing**. Essentially it involves a precise deletion of an intron from the primary transcript; the ends of the RNA on either side are joined together to form a covalently intact molecule. We discuss the mechanisms and regulation of splicing in Chapter 31.

We define the structural gene as comprising the region in the genome between points corresponding to the 5′ and 3′ terminal bases of mature mRNA. We know that transcription starts at the 5′ end of the mRNA, but probably it extends beyond the 3′ end, which is generated by cleavage of the RNA (see Chapter 29). The definition of the gene can be expanded to include associated regulatory regions: a transcription unit includes a promoter, other regulatory regions in the upstream region, the gene itself, and (sometimes) a terminator.

Proceeding from the structure of the gene itself, the next question is to determine its context. How much material on either side of a gene is involved in its function? How far is it to the next gene, and how much of the genome comprises sequences that lie between transcription units? Do adjacent genes tend to be related and is there any mechanism for regulating regions of the chromosome longer than single genes?

## Organization of interrupted genes may be conserved

When a gene is uninterrupted, the restriction map of its DNA corresponds exactly with the map of its mRNA (obtained by characterizing a cDNA reverse transcript).

When a gene possesses an intron, the map at each end of the gene corresponds with the map at each end of the message sequence. But within the gene, the maps diverge, because additional regions are found in the gene, but are not represented in the message. Each such region corresponds to an intron. The example of **Figure 23.2** compares the restriction maps of a $\beta$-globin gene and mRNA. Two introns can be recognized. Each intron contains a series of restriction sites that are absent from the cDNA. The exons can be recognized because the pattern of restriction sites coincides with that observed in the cDNA.

It is difficult to work directly with genomic DNA, so genes are characterized by using cloned genomic sequences. Once a cDNA is available, genomic clones can be obtained by the techniques described in Chapter 21. They can be compared with the cDNA by restriction mapping. The only practical limitation of this approach is that it may be impossible to isolate very large genes in the form of single cloned fragments, in which case clones representing parts of the gene must be characterized individually. In such cases, it is essential to use clones whose genomic sequences overlap, to ensure that we do not miss parts of a gene lying between the cloned sequences.

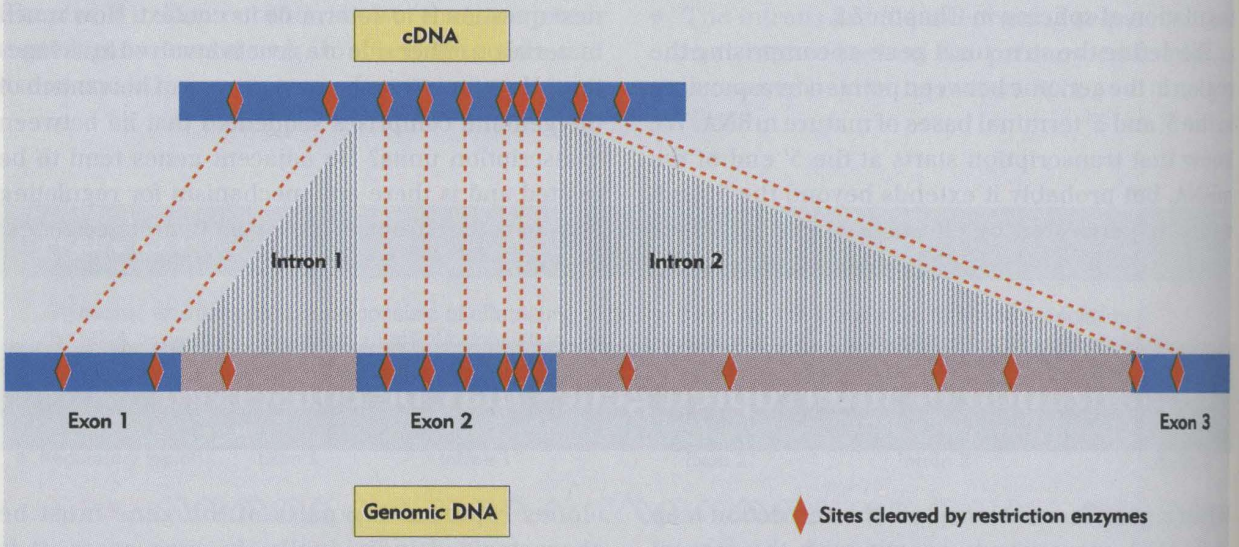
Ultimately a comparison of the nucleotide sequences of the genomic and cDNA clones precisely defines the introns. Resolution at the sequence level is necessary before we can be sure that all the segments of the gene have been identified. Short introns or exons can be missed in restriction maps if they do not happen to contain an appropriate restriction site. (An intron may pass unnoticed if it lies within a long exon, and an exon that is <50 bp long may fail to hybridize with the cDNA probe, and can therefore pass unnoticed within the introns that flank it.) But a sequence comparison is unambiguous. As indicated in **Figure 23.3**, an intron that lies within a coding region usually interrupts the integrity of the reading frame, but an intact reading frame is found in the cDNA sequence.

No particular rhyme or reason yet has been discerned in the extremely varied structures of eukaryotic genes. Some genes are uninterrupted, so that the genomic sequence is colinear with that

Tecan Ex. 2080, Page 5 of 28
IDT, Inc. v. Tecan Genomics - IPR2025-00016

**Figure 23.2**

Comparison of the restriction maps of cDNA and genomic DNA for mouse β-globin shows that the gene has two additional regions not present in the cDNA. The other regions can be aligned exactly between cDNA and gene.




of the mRNA. Most higher eukaryotic genes are interrupted, but the introns vary enormously in both number and size.

Some important features are common to all interrupted genes:

- The *order* of the parts of an interrupted gene is the same in the genome as in its mature RNA product. Genes are thus split rather than dispersed.
- An interrupted gene retains the *same structure* in all tissues, including the germ line and somatic tissues in which it is or is not expressed. Thus the presence of an intron is an invariant feature.
- Introns of nuclear genes generally have termination codons in all reading frames, and have *no coding function.*

All classes of genes may be interrupted: nuclear genes coding for proteins, nucleolar genes coding for rRNA, and genes coding for tRNA. Interruptions also are found in mitochondrial genes in lower eukaryotes, and in chloroplast genes. Interrupted genes do not appear to be excluded from any class of eukaryotes, and have even been found in an archebacterium and a phage of *E. coli.* They appear to be entirely absent only from eubacterial genomes.

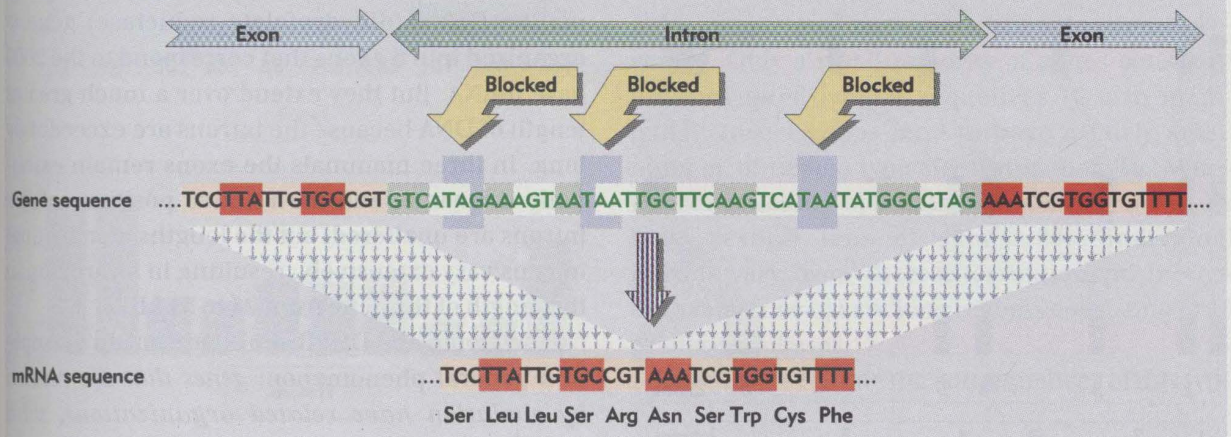
In the case of genes coding for mRNA, the terminal exons include the nontranslated leader and trailer, and the internal exons are likely to carry coding sequences. When interrupted genes code for rRNA or tRNA, the exons do not have a protein-coding function.

Some interrupted genes possess only one or a few introns. The globin genes provide an extensively studied example (see Chapter 24). The two general types of globin gene, $\alpha$ and $\beta$, share a common type of structure. The consistency of the organization of mammalian globin genes is evident from

Tecan Ex. 2080, Page 6 of 28
IDT, Inc. v. Tecan Genomics - IPR2025-00016

**Figure 23.3**

An intron is a sequence present in the gene but absent from the mRNA (here shown in terms of the cDNA sequence). The reading frame is indicated by the alternating open and shaded blocks; note that all three possible reading frames are blocked by termination codons in the intron.




the structure of the 'generic' globin gene summarized in **Figure 23.4**.

Interruptions occur at homologous positions (relative to the coding sequence) in all known active globin genes, including those of mammals, birds, and frogs. The first intron occurs between amino acids 30 and 31 (in a standard mammalian globin sequence), and the second intron is located between amino acids 104 and 105. The first intron is always fairly short, and the second is usually longer, but the actual lengths can vary.

Most of the variation in overall lengths between different globin genes results from the variation in the second intron. In the mouse, the second intron in the $\alpha$-globin gene is only 150 bp long, so the

**Figure 23.4**

All functional globin genes have an interrupted structure with three exons. The lengths indicated in the figure apply to the mammalian β-globin genes.

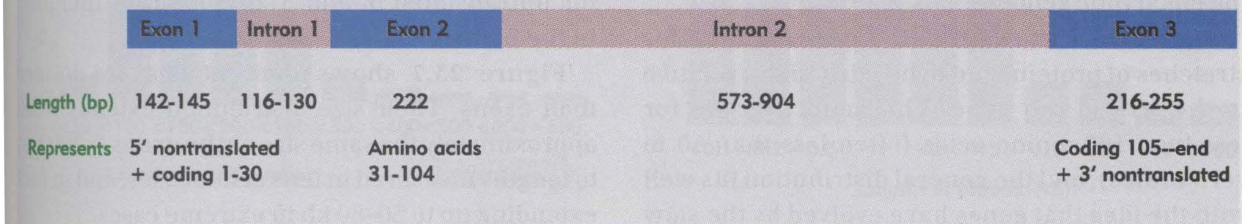
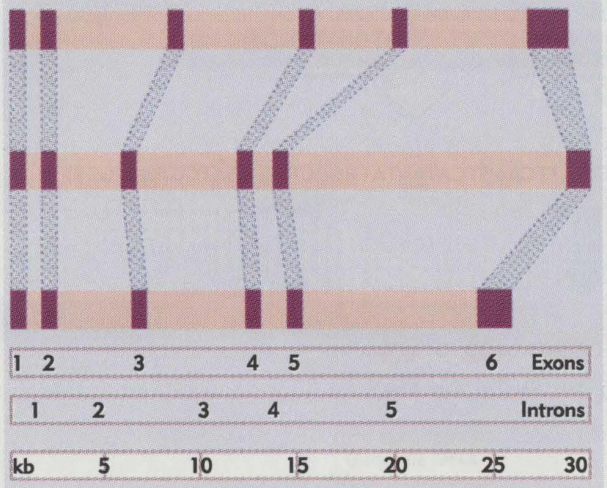


**Figure 23.5**

Mammalian genes for DHFR have the same relative organization of rather short exons and very long introns, but vary extensively in the lengths of corresponding introns.




overall length of the gene is 850 bp, compared with the 1382 bp of the major β-globin gene. Thus the variation in length of the genes is much greater than the range of lengths of the mRNAs (α-globin mRNA is 585 bases, β-globin mRNA is 620 bases).

The example of DHFR, which is a somewhat larger gene, is shown in **Figure 23.5**. The mammalian DHFR (dihydrofolate reductase) gene is organized into 6 exons that correspond to the 2000 base mRNA. But they extend over a much greater length of DNA because the introns are exceedingly long. In three mammals the exons remain essentially the same, and the relative positions of the introns are unaltered, but the lengths of individual introns vary extensively, resulting in a variation in the length of the gene from 25 to 31 kb.

The globin and DHFR genes present an example of a general phenomenon: *genes that are related by evolution have related organizations, with conservation of the positions of at least some of the introns. Variations in the lengths of the genes are primarily determined by the lengths of the introns.*

## Genes show a wide distribution of sizes

The existence of interrupted genes makes it evident that the gene can be much larger than the unit that codes for protein. With several hundred genes either sequenced or at least compared with their mRNA products in outline, we can explore how far the structure of the gene can account for the size of the eukaryotic genome.

**Figure 23.6** shows that the exons coding for stretches of protein tend to be fairly small relative to the size of the gene. The majority code for less than 100 amino acids (often less than 50 in vertebrates), and the general distribution fits well with the idea that genes have evolved by the slow addition of units that code for small, individual domains of proteins (see later). There is no very significant difference in the sizes of exons in different types of organisms, except perhaps for an apparent absence of larger exons in the vertebrates. There are some much larger exons coding for untranslated 5′ and 3′ regions (not included in the figure).

**Figure 23.7** shows that introns are longer than exons. Their size distribution extends from approximately the same size as the exons (<200 bp) to lengths measured in tens of kilobases, and in fact extending up to 50–60 kb in extreme cases.

Tecan Ex. 2080, Page 8 of 28
IDT, Inc. v. Tecan Genomics - IPR2025-00016

**Figure 23.6**

Exons coding for proteins tend to be short.

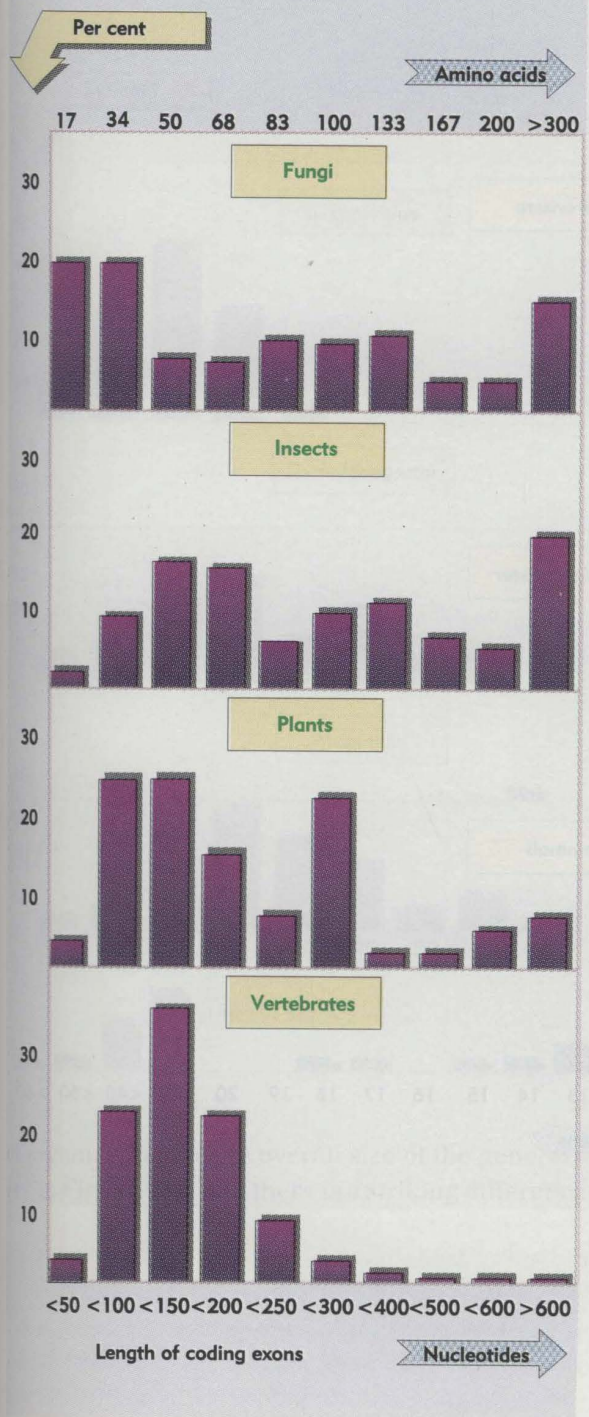


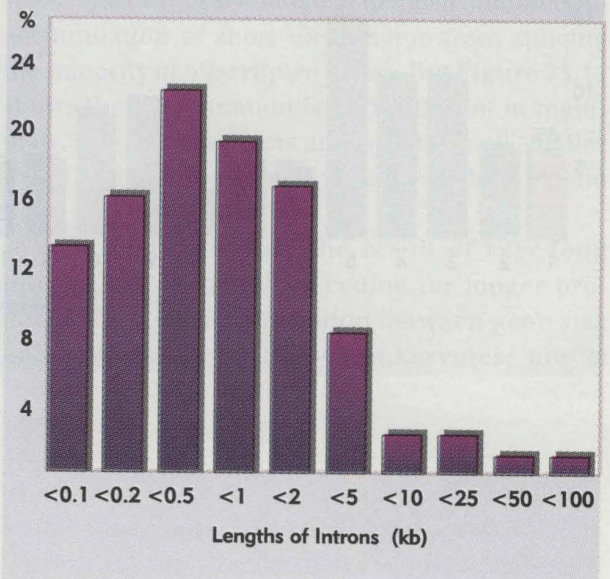

**Figure 23.8** shows the overall organization of genes in yeasts, insects and mammals. In *S. cerevisiae*, the great majority of genes in fact are not interrupted, and those that have exons usually remain reasonably compact. There are virtually no *S. cerevisiae* genes with more than four exons.

In insects and mammals, the situation is reversed. Only a small minority of genes consist of uninterrupted coding sequences (6% in mammals). Insect genes tend to have a fairly small number of exons, typically fewer than 10. Mammalian genes are split into more pieces, and some have several tens of exons. The individual introns vary from short interruptions in the coding sequence to very long regions consisting of 10 or more kilobases.

If we now examine the consequences of this type

**Figure 23.7**

Introns in vertebrate genes range from short to long.



## Figure 23.8

Most genes are uninterrupted in yeast, but most genes are interrupted in flies and mammals. (Uninterrupted genes have only 1 exon, and are totalled in the leftmost column.)

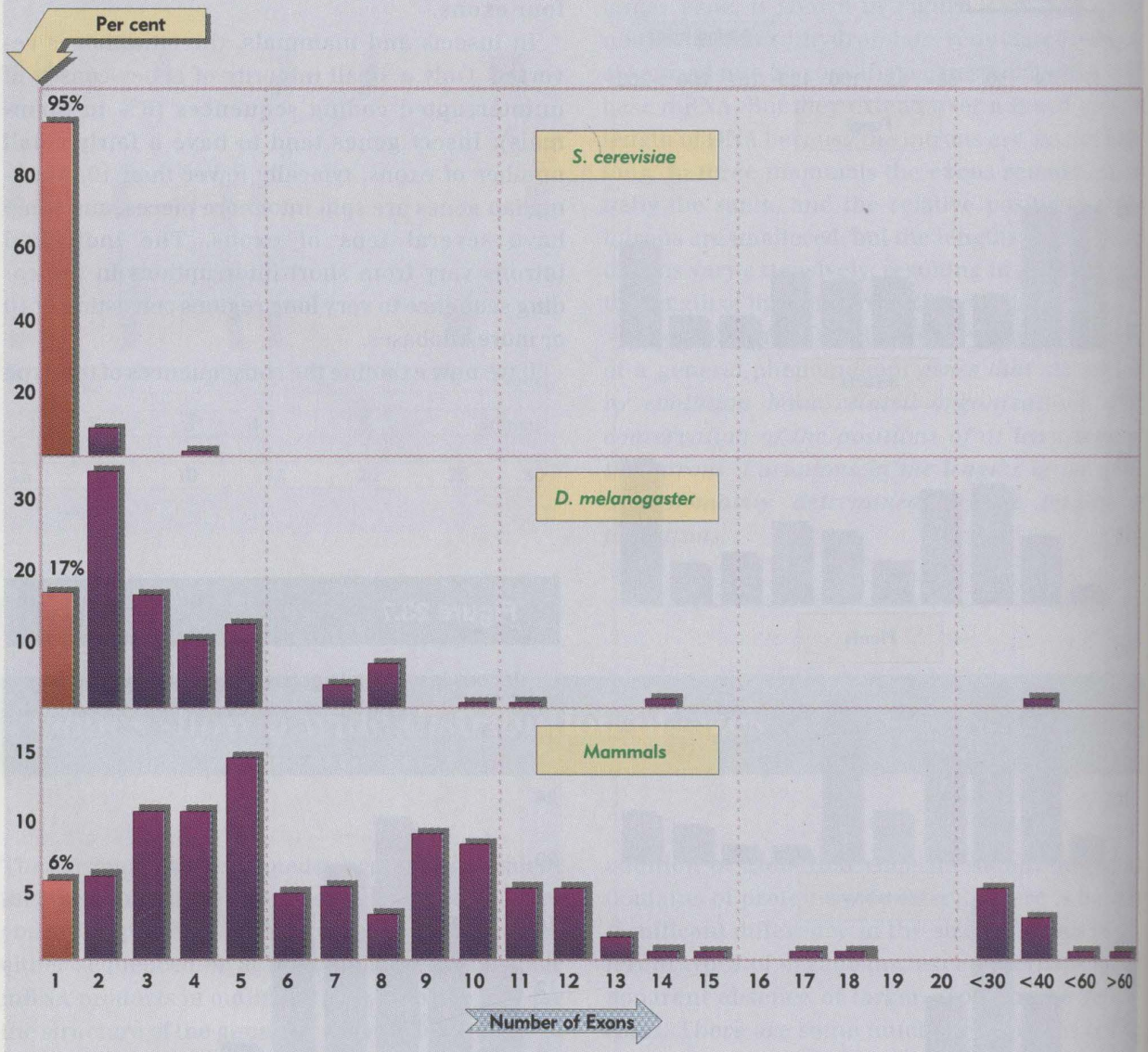
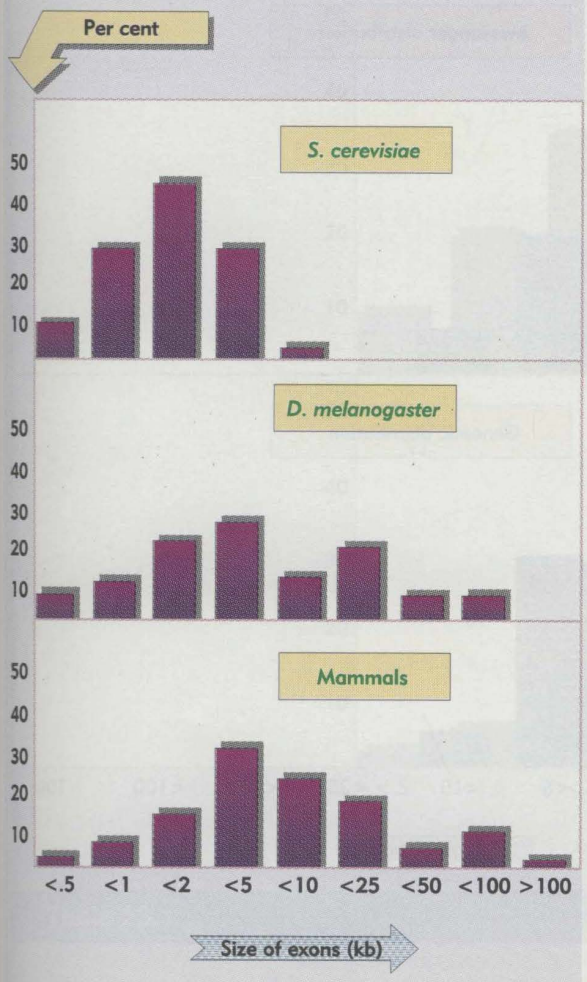


**Figure 23.9**

Yeast genes are fairly small, but genes in flies and mammals have a dispersed distribution extending to very large sizes.




of organization for the overall size of the gene, we see in **Figure 23.9** that there is a striking difference between yeast and the higher eukaryotes. Most yeast genes are shorter than 2 kb, and very few are longer than 5 kb. By contrast, relatively few genes in flies or mammals are shorter than 2 kb, and many have lengths between 5 and 100 kb.

The switch from largely uninterrupted to largely interrupted genes occurs in the lower eukaryotes. In fungi (excepting the yeasts), the majority of genes are interrupted, but they have a relatively small number of exons (<6) and are fairly short (<5 kb). The switch to long genes occurs within the higher eukaryotes, and genes become significantly larger in the insects. Perhaps genes become large at the same point where the relationship between genome complexity and organism complexity is lost (see Figure 22.1).

We still do not know the cause of the variations in DNA size within a phylum; for example, *D. melanogaster* is an example of a fly with a very small genome, and we may ask whether the 6-fold increase in genome size seen in the house fly represents an increase in lengths of individual genes, more space between the genes, or even an increased number of genes.

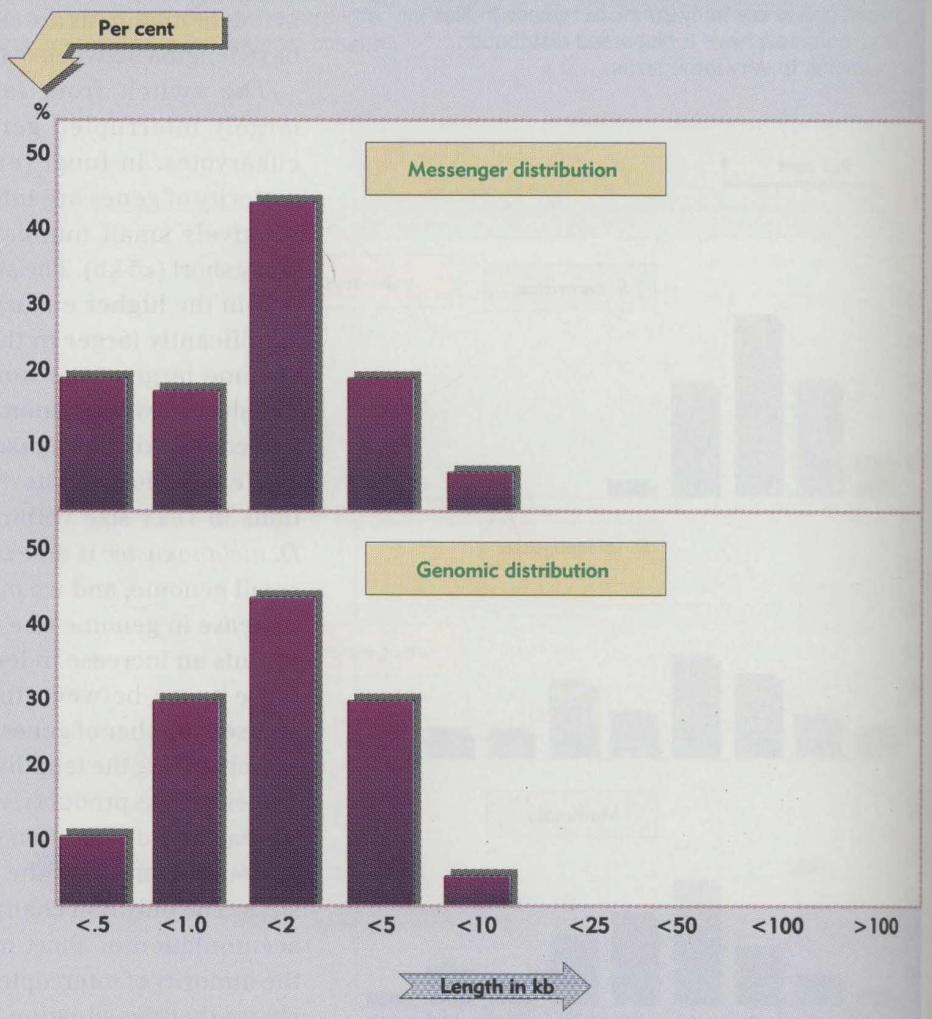
Comparing the lengths of genes with the lengths of their mRNA products, we see in **Figure 23.10** that in yeast the distribution of messenger sizes is not very different from the distribution of genomic sizes. The only real change in the distribution is an accumulation of short messengers from splicing the minority of interrupted genes. But **Figure 23.11** shows that the situation is very different in mammals, where messengers are essentially all smaller than 10 kb, contrasted with the genes, which extend up to ~100 kb.

Very long genes are the result of very long introns, not the result of coding for longer products. There is no correlation between gene size and mRNA size in higher eukaryotes; nor is

Tecan Ex. 2080, Page 11 of 28
IDT, Inc. v. Tecan Genomics - IPR2025-00016

**Figure 23.10**

The size distribution of mRNAs and genes corresponds in yeast.

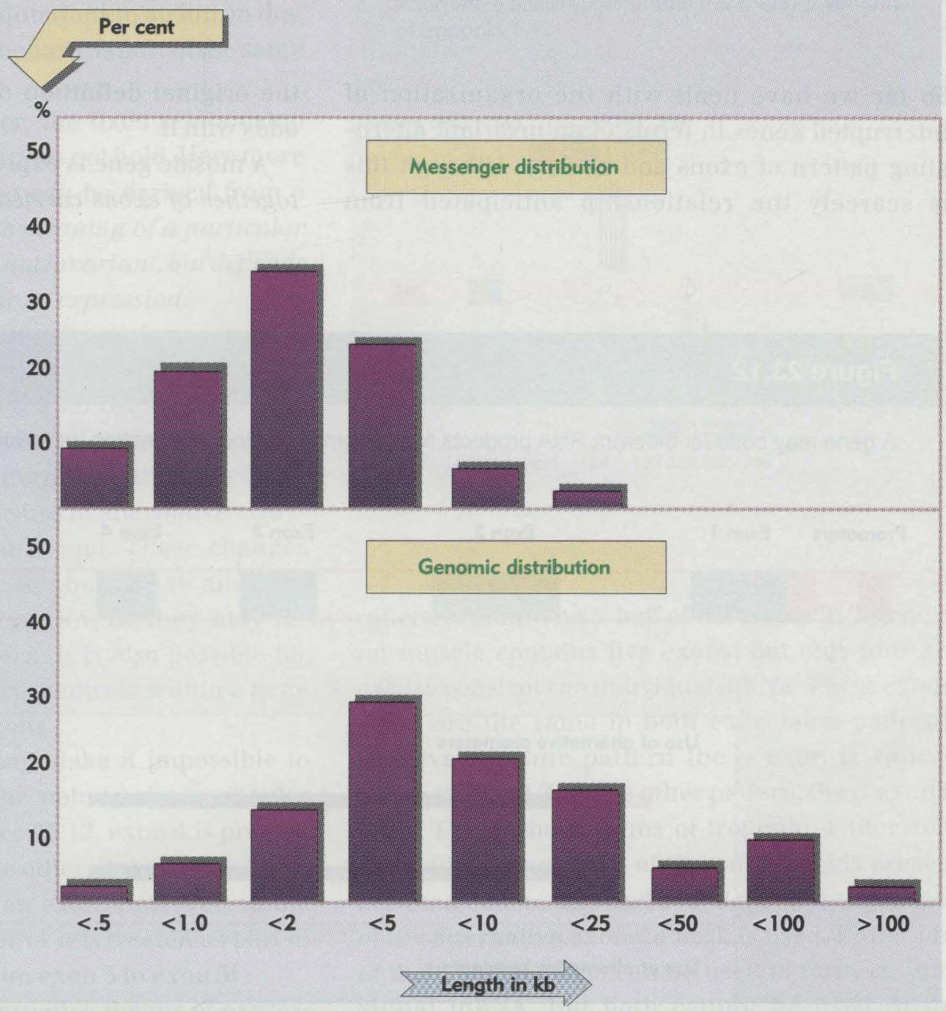



there a good correlation between gene size and the number of exons. The size of a gene therefore depends primarily on the lengths of its individual introns.

**Table 23.1** summarizes the average gene sizes in a range of organisms. When we compare the average sizes of mRNAs and genes in various species, we see that there is a small increase proceeding from the lower to higher eukaryotes. A large increase in the average number of exons is evident by the start of the higher eukaryotes, and the genes become much larger with the insects. In mammals, insects, and birds, the 'average' gene is approximately 5 times the length of its mRNA.

## Figure 23.11

Mammalian genes are much larger than the corresponding mRNAs.




## Table 23.1

The average lower eukaryotic gene is similar in size to the average mRNA, but the average size of the higher eukaryotic gene is about 5 times the length of its mRNA product.

| Species | Average exon number | Average gene length (kb) | Average mRNA length (kb) |
|---|---|---|---|
| *S. cerevisiae* | 1 | 1.6 | 1.6 |
| Fungi | 3 | 1.5 | 1.5 |
| *C. elegans* | 4 | 4.0 | 3.0 |
| *D. melanogaster* | 4 | 11.3 | 2.7 |
| Chicken | 9 | 13.9 | 2.4 |
| Mammals | 7 | 16.6 | 2.2 |

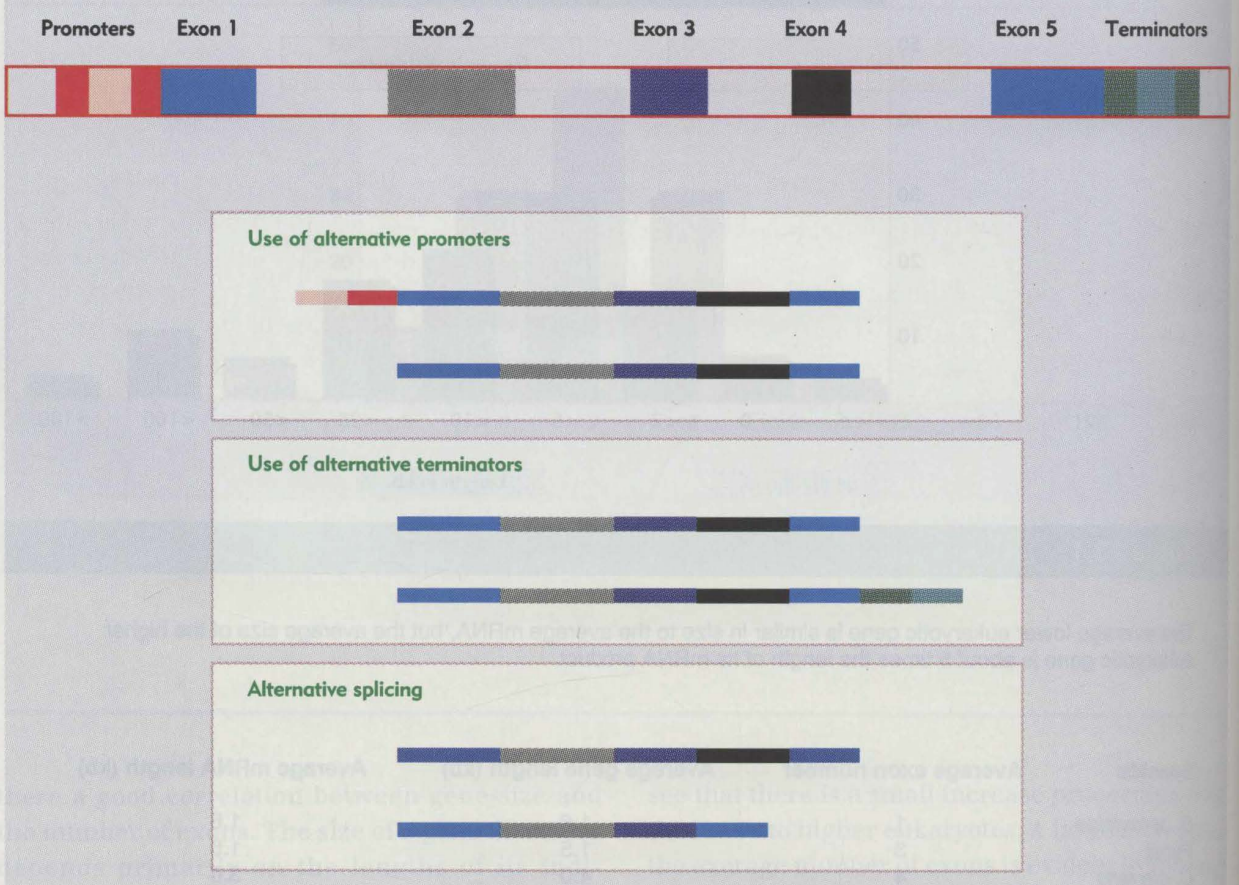
## One DNA sequence may code for multiple proteins

So far we have dealt with the organization of interrupted genes in terms of an invariant alternating pattern of exons and introns. Although this is scarcely the relationship anticipated from the original definition of the cistron, it is not at odds with it.

A mosaic gene is expressed *only via the splicing together of exons carried by one molecule of RNA,*

**Figure 23.12**

A gene may code for different RNA products by changing initiation, termination, or splicing.



thus conforming to the original concept of the *cis/trans* test (see Chapter 3). Mutations in the exons of a gene therefore fail to complement one another (see Chapter 6). Mutations in an intron that affect splicing of the RNA behave as part of the same complementation group.

In some cases, however, the fixed relationship between gene and protein does not hold. Here more than one mRNA sequence can be derived from a single stretch of DNA. *The meaning of a particular region of DNA therefore is not invariant, but depends on the pathway selected for its expression.*

**Figure 23.12** summarizes the various mechanisms that are used for alternative expression of genes. They include differential use of promoters or of termination sites as well as use of alternative exons. The use of alternative promoters may change the 5′ end, and the use of alternative cleavage sites may change the 3′ end. These changes may be restricted just to the initial and terminal sequences of the mRNA, or they may influence the splicing pattern. It is also possible for alternative splicing to occur entirely within a gene to change the usage of exons.

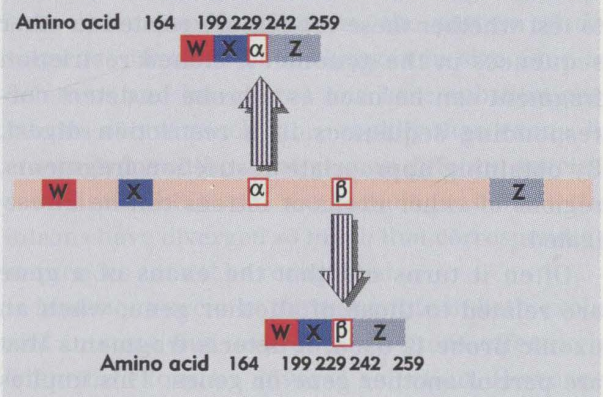
Alternative splicing may make it impossible to define unambiguously the nature of a particular segment of DNA. In Figure 23.12, exon 4 is present in one mRNA but not in the other. Its presence in the first mRNA defines it as an exon. However, in the splicing of the second mRNA it is treated as part of an intron that extends from exon 3 to exon 5!

In some cases, the alternative means of expression do not affect the sequence of the protein; for example, changes that affect the 5′ nontranslated leader or the 3′ nontranslated trailer may have regulatory consequences, but the same protein is made. In other cases, the substitution of one exon for another, or the addition or deletion of exons, may change the protein sequence as indicated in **Figure 23.13**.

In this example, the proteins produced by the two mRNAs contain sequences that overlap extensively, but that are different within the alternatively

**Figure 23.13**

Alternative splicing generates the α and β variants of troponin T.




spliced region. The 3′ half of the troponin T gene of rat muscle contains five exons, but only four are used to construct an individual mRNA. Three exons, *WXZ*, are the same in both expression patterns. However, in one pattern the $\alpha$ exon is spliced between *X* and *Z*; in the other pattern, the $\beta$ exon is used. The $\alpha$ and $\beta$ forms of troponin T therefore differ in the sequence of the amino acids present between codons 229 and 242, depending on which of the alternative exons, $\alpha$ or $\beta$, is used. Either one of the $\alpha$ and $\beta$ exons can be used to form an individual mRNA, but both cannot be used in the same mRNA.

Thus alternative (or differential) splicing can generate proteins with overlapping sequences from a single stretch of DNA. It is curious that the higher eukaryotic genome is extremely spacious in having large genes that are often quite dispersed, but at the same time it may make multiple products from an individual locus. It is really not possible to say how many genes have alternative modes of expression, but anecdotally, the number would seem to be a few per cent.

Tecan Ex. 2080, Page 15 of 28
IDT, Inc. v. Tecan Genomics - IPR2025-00016

## Exon sequences are conserved but introns vary

By using restriction fragments that correspond to specific regions of a structural gene, it is possible to test whether these regions are related to other sequences in the genome. A cloned restriction fragment can be used as a probe to detect corresponding sequences in a restriction digest. By obtaining appropriate restriction fragments, regions of either exons or introns can be investigated.

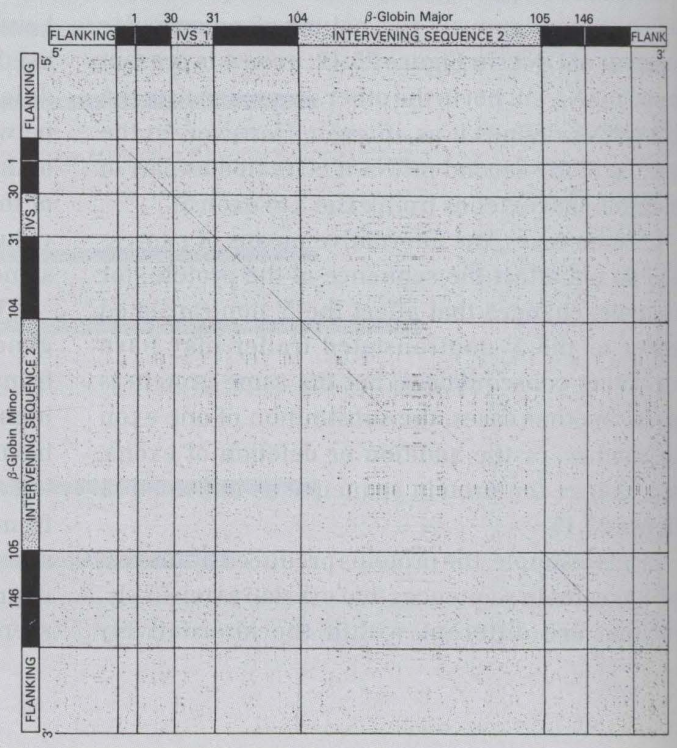
Often it turns out that the exons of a gene are related to those of another gene; when an exonic probe is used, it detects fragments that are part of another gene or genes. This implies that the two genes originated by a duplication of some common ancestral gene, after which differences accumulated between the copies.

Usually the introns are not related to other sequences. And when two genes have related exons, the relationship between their introns is more distant than the relationship between the exons.

Our original description considered nonrepetitive DNA to represent sequences that are unique in the genome. Introns come closer than exons to fitting this definition. In asking whether structural genes are nonrepetitive, we see therefore that the answer is ambiguous: the entire length of the gene is unique as such, but its exons often are related to those of other genes. At least for some genes, the exons constitute slightly repetitive sequences embedded in a unique context of introns.

**Figure 23.14**

The sequences of the mouse *αmaj* and *αmin* globin genes are closely related in coding regions, but differ in the flanking regions and large intron. Data kindly provided by Philip Leder.



The relationship between two genes can be plotted in the form of the dot matrix comparison of **Figure 23.14**. A dot is placed to indicate each position at which the same sequence is found in each gene. The dots form a line at an angle of 45° if two sequences are identical. The line is broken by regions that lack similarity, and it is displaced laterally or vertically by deletions or insertions in one sequence relative to the other.

When the two β-globin genes of the mouse are compared, such a line extends through the three exons and through the small intron. The line peters out in the flanking regions and in the large intron. This is a typical pattern, in which coding sequences are well related, the relationship can extend beyond the boundaries of the exons, but it is lost in longer introns and the regions on either side of the gene.

The overall degree of divergence between two exons is related to the differences between the proteins. It is caused mostly by base substitutions. In the translated regions, the exons are under the constraint of needing to code for amino acid sequences, so they are limited in their potential to change sequence. Thus many of the changes do not affect codon meanings, because they lie in third-base positions. Changes occur more freely in nontranslated regions (corresponding to the 5′ leader and 3′ trailer of the mRNA).

In corresponding introns, the pattern of divergence involves both changes in size (due to deletions and insertions) and base substitutions. Introns evolve much more rapidly than exons; in comparisons of the same gene in different species, sometimes the exons are homologous while the introns have diverged so much that corresponding sequences cannot be recognized.

Mutations occur at the same rate in both exons and introns, but are removed more effectively from the exons by adverse selection. However, in the absence of the constraints imposed by a coding function, an intron is able quite freely to accumulate point substitutions and other changes. These changes imply that the intron does not have a sequence-specific function. Whether its presence is at all necessary for gene function is not clear.

## Genes can be isolated by the conservation of exons

A useful consequence follows from the contrast between the conservation of exons and the variation of introns. In a region containing a gene whose function has been conserved among a range of species, the sequence actually representing the protein should have two distinctive properties: it must of course have an open reading frame; and it is likely to have a related sequence in other species. These features can be used to isolate genes whose functions we suspect, but of which we otherwise have insufficient knowledge.

Suppose we know by genetic data that a particular genetic trait is located in a given chromosomal region, and we have walked along the chromosome to that region. If we lack knowledge about the nature of the gene product, how are we to identify the gene in a region that may be (for example) >100 kb?

A heroic approach that has proved successful with some genes of medical importance is to screen relatively short fragments from the region for the two properties expected of a conserved gene. First we seek to identify fragments that cross-hybridize with the genomes of other species. Then we examine these fragments for open reading frames.

The first criterion is applied by performing a **zoo blot**. We clone a short fragment from the region and use it as a (radioactive) probe to test for related DNA from a variety of species by Southern blotting. If we find hybridizing fragments in several species

Tecan Ex. 2080, Page 17 of 28
IDT, Inc. v. Tecan Genomics - IPR2025-00016

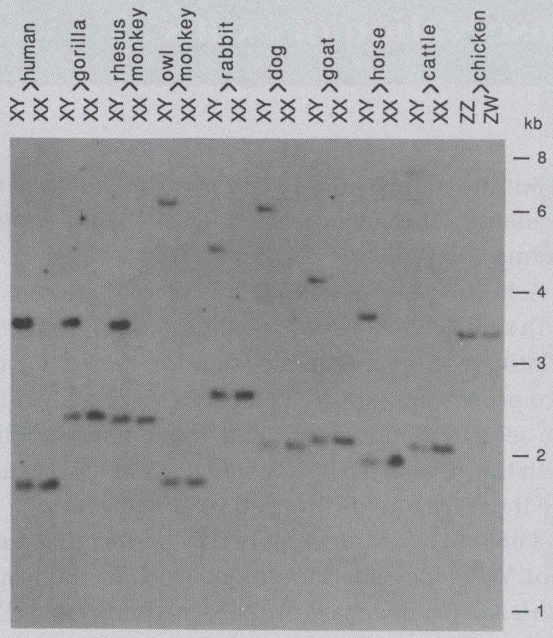
related to that of the probe—the probe is usually human—the probe becomes a candidate for an exon of the gene.

Such candidates are sequenced, and if they contain open reading frames, are used to isolate surrounding genomic regions. If these appear to be part of an exon, we may then seek to identify the entire gene, to isolate the corresponding cDNA or mRNA, and ultimately to identify the protein.

This approach is valuable for genes whose existence is implied by genetics, but whose nature is unknown. One example is the gene *zfy* located on the human Y chromosome. **Figure 23.15** shows a zoo blot using a probe from this region. It hybridizes specifically with sex chromosomes of mammals

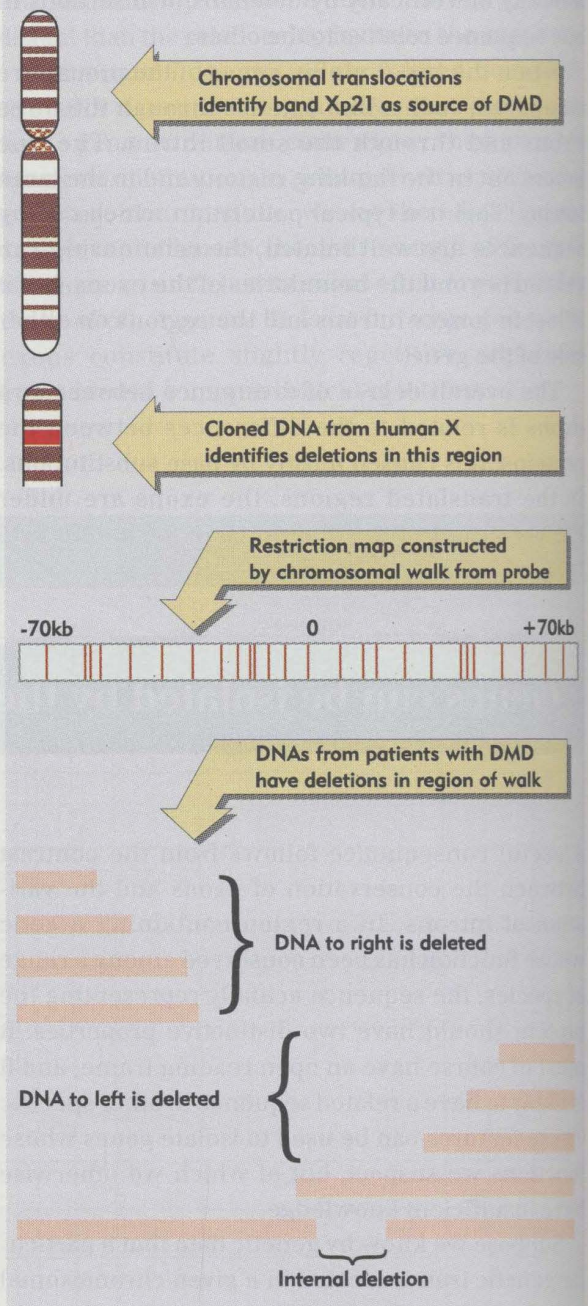
## Figure 23.15

A zoo blot with a probe from the human Y chromosomal gene *zfy* identifies cross-hybridizing fragments on the sex chromosomes of other mammals and birds. There is one reacting fragment on the Y chromosome and another on the X chromosome. Data kindly provided by David Page.




## Figure 23.16

The gene involved in Duchenne muscular dystrophy has been tracked down by chromosome mapping and walking to a region in which deletions can be identified with the occurrence of the disease.



and also with other species. It contains an open reading frame, which identifies a conserved gene.

This approach is especially important when the target gene is spread out because it has many large introns. This proved to be the case with Duchenne muscular dystrophy (DMD), a degenerative disorder of muscle, which is X-linked and affects 1 in 3,500 of human male births. The steps in identifying the gene are summarized in **Figure 23.16**.

Linkage analysis localized the DMD locus to chromosomal band Xp21. Patients with the disease often have chromosomal rearrangements involving this band. By comparing the ability of X-linked DNA probes to hybridize with DNA from patients and with normal DNA, cloned fragments were obtained that correspond to the region that was rearranged or deleted in patients' DNA.

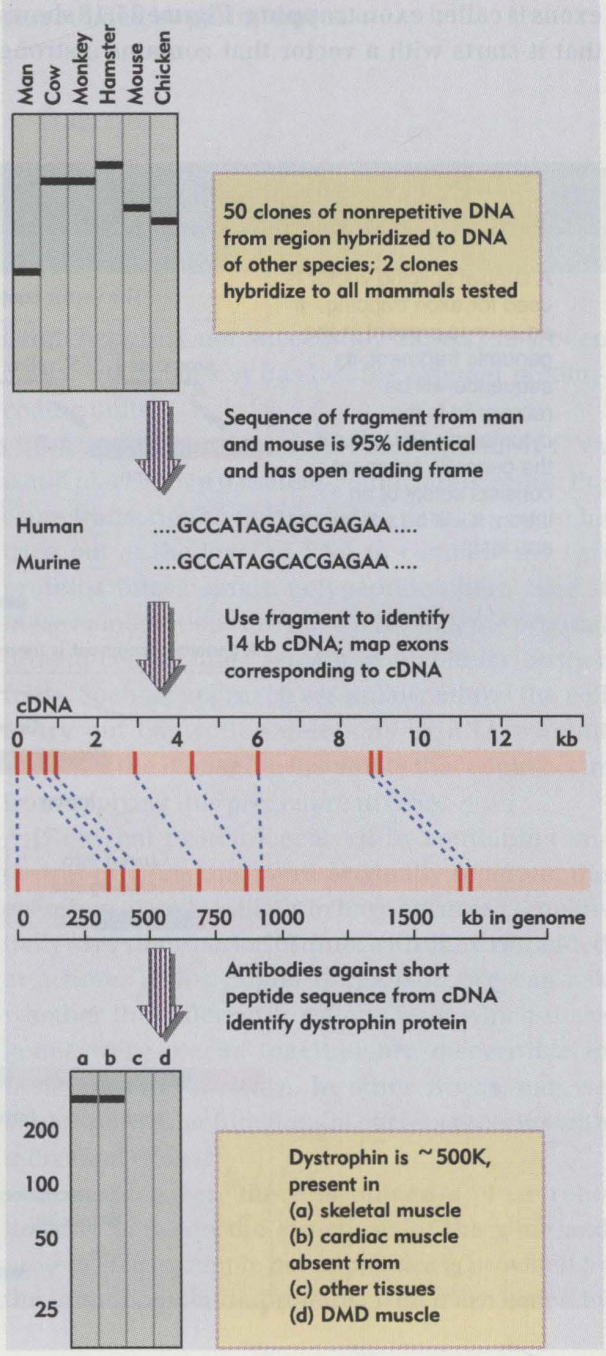
A chromosomal walk was used to construct a restriction map of the region on either side of the probe, covering a region of >100 kb. Analysis of the DNA from a series of patients identified large deletions in this region, extending in either direction. The most telling deletion is one contained entirely within the region, since this delineates a segment that must be important in gene function and indicates that the gene, or at least part of it, lies in this region.

Having now come into the region of the gene, we need to identify its exons and introns. A zoo blot identified fragments that cross-hybridize with the mouse X chromosome and with other mammalian DNAs. As summarized in **Figure 23.17**, these were scrutinized for open reading frames and the sequences typical of exon–intron junctions. Fragments that met these criteria were used as probes to identify homologous sequences in a cDNA library prepared from muscle mRNA.

The cDNA corresponding to the gene identifies an unusually large mRNA, ~14 kb. Hybridization back to the genome shows that the mRNA is represented in >60 exons, which are spread over ~2,000 kb of DNA. This makes DMD the longest gene identified; in fact, it is 10× longer than any other known gene.

**Figure 23.17**

The Duchene muscular dystrophy gene has been characterized by zoo blotting, cDNA hybridization, genomic hybridization, and identification of the protein.

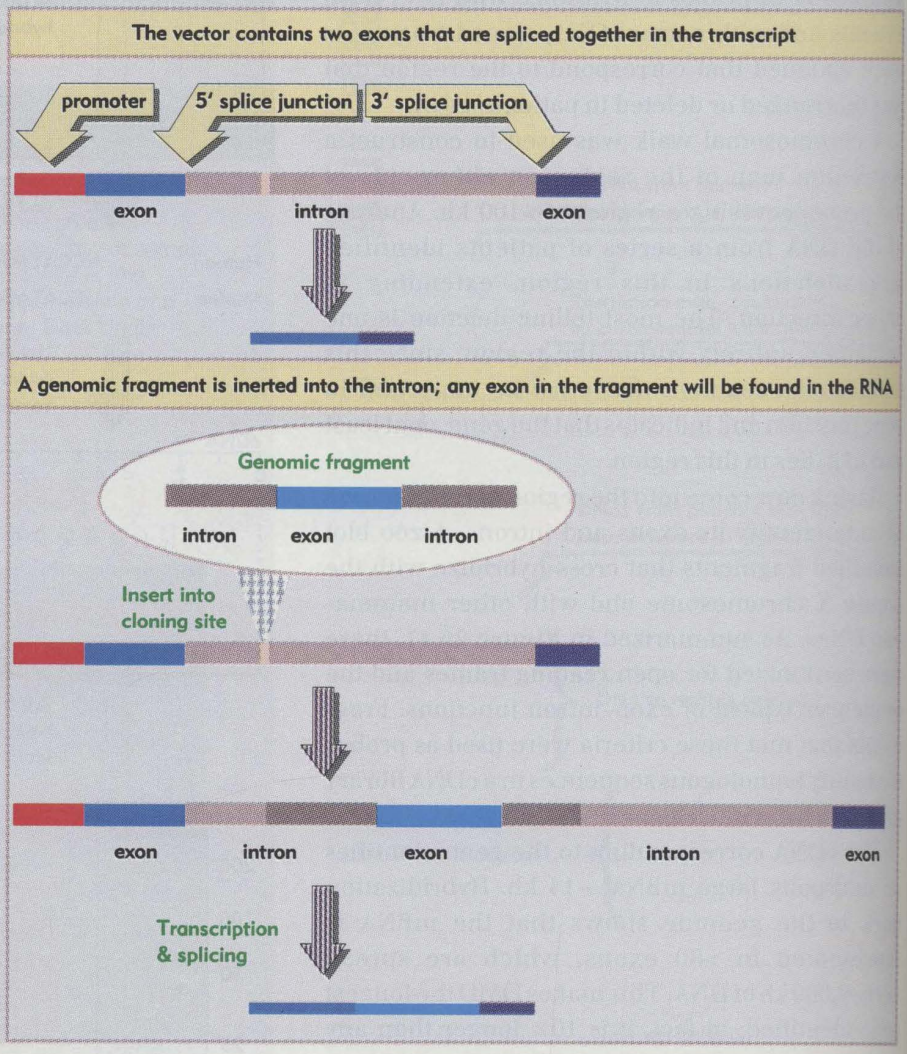


The gene codes for a protein of ~500,000 daltons which has been named dystrophin. It is a component of muscle, present in rather low amounts. All patients with the disease have deletions at this locus, and lack (or have defective) dystrophin.

Another technique that allows genomic fragments to be scanned rapidly for the presence of exons is called **exon trapping**. **Figure 23.18** shows that it starts with a vector that contains a strong promoter, and has a single intron interrupting the two exons. When this vector is transfected into cells, its transcription generates large amounts of an RNA containing the sequences of the two exons. A restriction cloning site lies within the intron, and is used to insert genomic fragments from a region of interest. If a fragment does not contain an exon, there is no change in the splicing pattern, and the RNA contains only the same sequences as the

**Figure 23.18**

A special splicing vector is used for exon trapping. If an exon is present in the genomic fragment, its sequence will be recovered in the cytoplasmic RNA, but if the genomic fragment consists solely of an intron, it will be spliced out and lost.



parental vector. But if the genomic fragment contains an exon flanked by two partial intron sequences, the splicing sites on either side of this exon are recognized, and the sequence of the exon is inserted into the RNA between the two exons of the vector. This can be detected readily by reverse transcribing the cytoplasmic RNA into cDNA, and using PCR to amplify the sequences between the two exons of the vector. Thus the appearance in the amplified population of sequences from the genomic fragment indicates that an exon has been trapped. Because introns are usually large and exons are small in animal cells, there is a high probability that a large random piece of genomic DNA will contain the required structure of an exon surrounded by partial introns.

## How do interrupted genes evolve?

What was the original form of genes that today are interrupted?

- Did the ancestral protein-coding units consist of uninterrupted sequences of DNA, into which introns were subsequently inserted?
- Or did these genes initially arise as interrupted structures, which since have been maintained in this form?

Another form of this question is to ask whether the difference between eukaryotic and prokaryotic genes is to be accounted for by the acquisition of introns in the eukaryotes or by the loss of introns from the prokaryotes.

Could the mosaic structure be a remnant of an ancient approach to the reconstruction of genes to make novel proteins? Suppose that an early cell had a number of separate protein-coding sequences. One aspect of its evolution is likely to have been the reorganization and juxtaposition of different polypeptide units to build up new proteins.

If the protein-coding unit must be a continuous series of codons, every such reconstruction would require a precise recombination of DNA to place the two protein-coding units in register, end to end in the same reading frame. Furthermore, if this combination is not successful, the cell has been damaged, because it has lost the original protein-coding units.

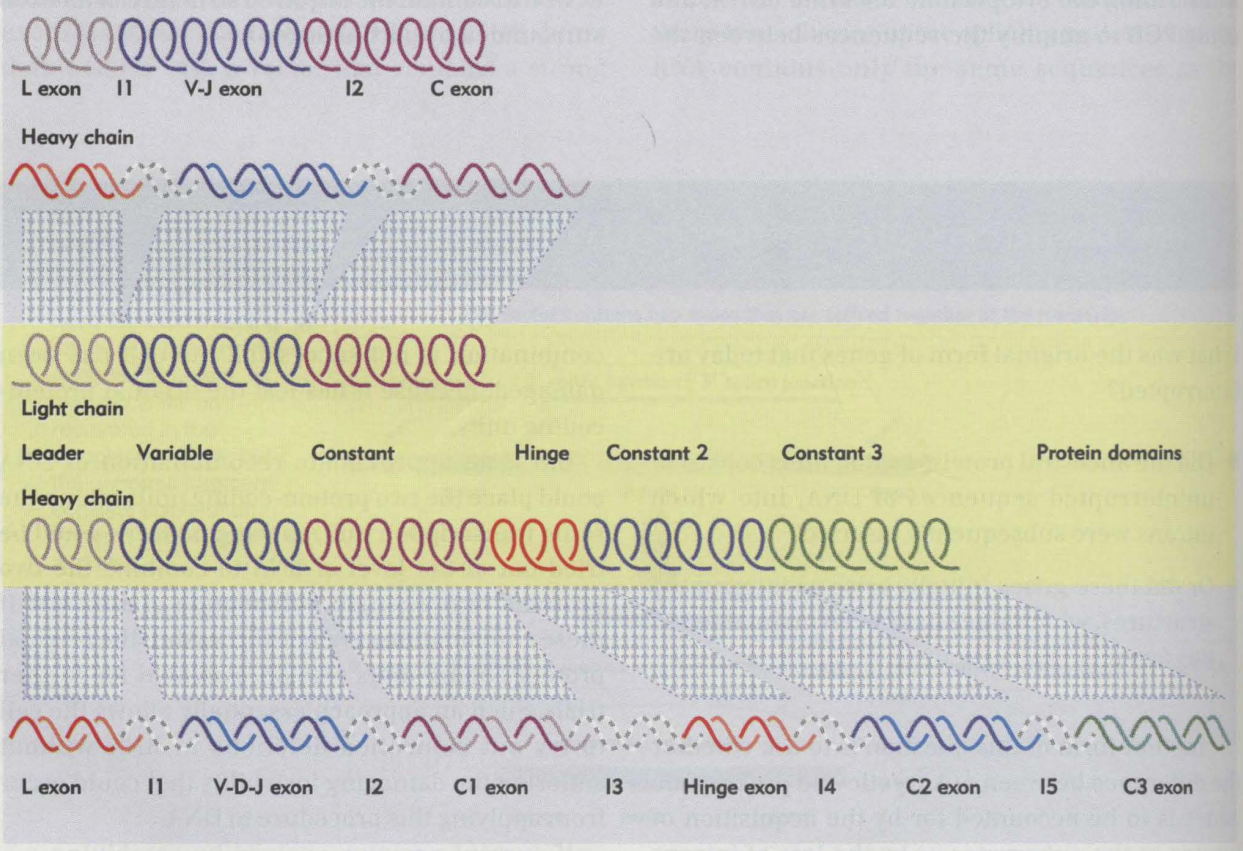
But if an approximate recombination of DNA could place the two protein-coding units within the same transcription unit, splicing patterns could be tried out at the level of RNA to combine the two proteins into a single polypeptide chain. And if these combinations are not successful, the original protein-coding units remain available for further trials. Such an approach essentially allows the cell to try out controlled deletions in RNA without suffering the damaging instability that could occur from applying this procedure to DNA.

If current proteins evolved by combining ancestral proteins that were originally separate, the accretion of units is likely to have occurred sequentially over some period of time, with one exon added at a time. If this model is realistic, we can ask whether the different functions from which these genes were pieced together are discernible in their present structure. In other words, can we equate particular functions of current proteins with individual exons?

In some cases, there is indeed a clear relationship between the structures of the gene and protein. The example *par excellence* is provided by the immunoglobulin proteins, which are coded by

**Figure 23.19**

Immunoglobulin light chains and heavy chains are coded by genes whose structures (in their expressed forms) correspond with the distinct domains in the protein. Each protein domain corresponds to an exon; introns are numbered 1-5.




genes in which every exon corresponds exactly with a known functional domain of the protein. **Figure 23.19** compares the structure of an immunoglobulin with its gene.

An immunoglobulin is a tetramer of two light chains and two heavy chains, which aggregate to generate a protein with several distinct domains. Light chains and heavy chains differ in structure, and there are several types of heavy chain. Each type of chain is expressed from a gene that has a series of exons corresponding with the structural domains of the protein. (Chapter 37 explains how the expressed form of the gene is generated from its components.)

In many instances, some of the exons of a gene can be identified with particular functions. In secretory proteins, the first exon, coding for the N-terminal region of the polypeptide, often specifies the signal sequence involved in membrane secretion. An example is insulin.

Sometimes the evolution of a gene involves the duplication of exons, creating an internally repetitious sequence in the protein. In chicken collagen, a 54 bp exon appears to have been multiplied many times, generating a series of exons that are either 54 bp or multiples of 54 bp in length.

Sequences held in common between genes that

are related only in part may represent exons that have migrated or been recruited between genes. **Figure 23.20** summarizes the relationship between the receptor for human LDL (plasma low density lipoprotein) and other proteins.

In the center of the LDL receptor gene is a series of exons related to the exons of the gene for the precursor for EGF (epidermal growth factor). In the N-terminal part of the protein, a series of exons codes for a sequence related to the blood protein complement factor C9. Thus the LDL receptor gene was created by assembling *modules* for its various functions, these modules also being used in other proteins.

The relationship between exons and protein domains is somewhat erratic in known genes. In some cases there is a clear 1:1 relationship; in others no pattern is to be discerned. One possibility is that removal of introns has fused the adjacent exons. A difficulty in this idea is the need to suppose that the intron removal was precise, not changing the integrity of the coding region. An alternative is that some introns arose by insertion into a coherent domain; here the difficulty is that we must suppose that the intron carried with it the ability to be spliced out.

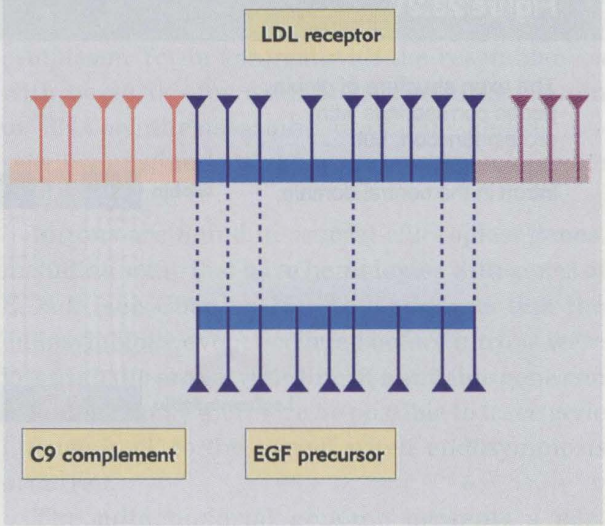
Exons tend to be fairly small (see Figure 23.6), around the size of the smallest polypeptide that can assume a stable folded structure, ~20–40 residues. Perhaps proteins were originally assembled from rather small modules. Each module need not necessarily correspond to a current function; several modules could have combined to generate a function. The number of exons in a gene tends to increase with the length of its protein, which is consistent with the view that proteins acquire multiple functions by successively adding appropriate modules.

This idea might explain another feature of protein structure: it seems that the sites represented at exon-intron boundaries often are located at the surface of a protein. As modules are added to a protein, the connections, at least of the most recently added modules, could tend to lie at the surface.

The equation of at least some exons with protein domains, and the appearance of related exons in different proteins, supports the idea that 'exon-shuffling' has been a fundamental relationship in the evolution of genes. Certainly the duplication and juxtaposition of exons has played an important role in evolution. We cannot trace the actual events involved in the evolution of every gene; many relationships between exons and protein domains do not conform to a simple equation, but they could be accounted for if events such as exon fusions have modified the ancestral structure during the evolution of nuclear genes. Indeed, one school of thought is that the number of ancestral exons, from which all proteins have been derived by duplication, variation, and recombination, could be relatively small (a few thousands or tens of thousands).

**Figure 23.20**

The LDL receptor gene consists of 18 exons, some of which are related to EGF precursor and some to the C9 blood complement gene. Triangles mark the positions of introns. Only some of the introns in the region related to EGF precursor are identical in position to those in the EGF gene.




Tecan Ex. 2080, Page 23 of 28
IDT, Inc. v. Tecan Genomics - IPR2025-00016

A fascinating case of evolutionary conservation is presented by the globins, all of whose genes have three exons (see Figure 23.2). The two introns are located at constant positions relative to the coding sequence. The central exon appears to represent the heme-binding domain of the globin chain. The active protein is a tetramer containing two globin chains of the $\alpha$ type and two of the $\beta$ type.

Another perspective on this structure is provided by the existence of two other types of protein that are related to globin. Myoglobin is a monomeric oxygen-binding protein of animals, whose amino acid sequence suggests a common (though ancient) origin with the globin subunits. Leghemoglobins are oxygen-binding proteins present in the legume class of plants; like myoglobin, they are monomeric. They too share a common origin with the other heme-binding proteins. Together, the globins, myoglobin, and leghemoglobin constitute the globin 'super-family', a set of gene families all descended from some (distant) common ancestor.

Myoglobin is represented by a single gene in the human genome, whose structure is essentially the same as that of the globin genes. The three exon structure therefore predates the evolution of separate myoglobin and globin functions.

Leghemoglobin genes contain three introns, the first and last of which occur at points in the coding sequence that are homologous to the locations of the two introns in the globin genes. This remarkable similarity suggests an exceedingly ancient origin for the heme-binding proteins in the form of a split gene, as illustrated in **Figure 23.21**.

The central intron of leghemoglobin separates two exons that together code for the sequence corresponding to the single central exon in globin. Could the central exon of the globin gene have been derived by a fusion of two central exons in the ancestral gene, bringing together the sequences coding for two parts of the protein chain that together form the heme-binding structure?

Cases in which homologous genes differ in structure may provide information about their evolution. An example is insulin. Mammals and birds have only one gene for insulin, except for the rodents, which have two genes. **Figure 23.22** illustrates the structures of these genes.

The principle we use in comparing the organization of related genes in different species is that *a common feature identifies a structure that predated the evolutionary separation of the two species.* In chicken, the single insulin gene has two introns;

**Figure 23.21**

The exon structure of globin genes corresponds with protein function, but leghemoglobin has an extra intron in the central domain.

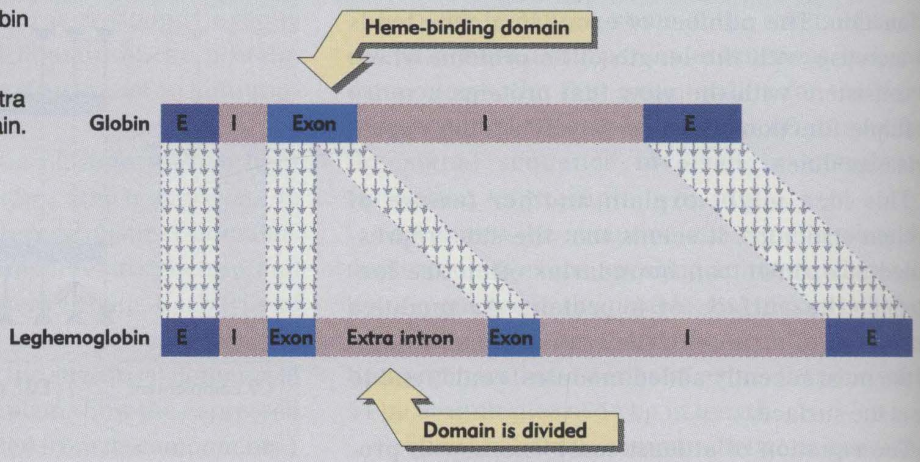
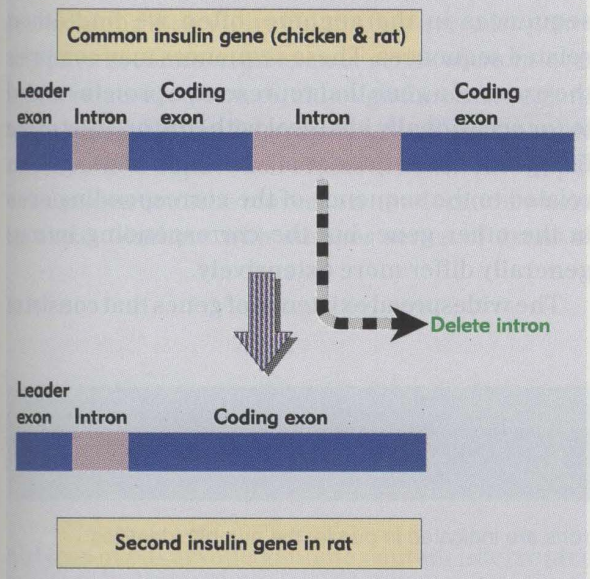


**Figure 23.22**

The rat insulin gene with one intron evolved by losing an intron from an ancestor with two interruptions.




one of the two rat genes has the same structure. The common structure implies that the ancestral insulin gene had two introns. However, the second rat gene has only one intron. It must have evolved from the first by a process in which a gene duplication in rodents was followed by the precise removal of one intron from one of the copies.

The organization of some genes shows extensive discrepancies between species. In these cases, there must have been extensive removal or insertion of introns during evolution.

The best characterized case is represented by the actin genes. The typical actin gene has a nontranslated leader of <100 bases, a coding region of ~1,200 bases, and a trailer of ~200 bases. Most actin gens are interrupted; the positions of the introns can be aligned with regard to the coding sequence (except for a single intron sometimes found in the leader).

**Figure 23.23** shows that almost every actin gene is different in its pattern of interruptions. Taking all the genes together, introns occur at 12 different sites. However, no individual gene has more than six introns; some genes have only one intron, and one is uninterrupted altogether. How did this situation arise? If we suppose that the primordial actin gene was interrupted, and all current actin genes are related to it by loss of introns, different introns have been lost in each evolutionary branch. Probably some introns have been lost entirely, so the primordial gene could well have had 20 or more. The relationships between the intron locations found in different species may be used ultimately to construct a tree for the evolution of the gene.

Polymorphisms seem common in genes for rRNA and tRNA, where alternative forms can often be found, with and without introns. In the case of the tRNAs, where all the molecules conform to the same general structure, it seems unlikely that evolution brought together the two regions of the gene. After all, the different regions are involved in the base pairing that gives significance to the structure. So here it may be that the introns were inserted into continuous genes.

Organelle genomes provide some striking connections between the prokaryotic and eukaryotic worlds. Because of many general similarities between mitochondria or chloroplasts and bacteria, it seems likely that the organelles originated by an **endosymbiosis** in which an early bacterial prototype was inserted into eukaryotic cytoplasm. Yet in contrast with the resemblances with bacteria—for example, as seen in protein or RNA synthesis—some organelle genes possess introns, and therefore resemble eukaryotic nuclear genes.

Introns are found in several chloroplast genes, including some that have homologies with genes of *E. coli* (see Chapter 25). This suggests that the endosymbiotic event occurred before introns were lost from the prokaryotic line. If a suitable gene can be found, it may therefore be possible to trace gene lineage back to the period when endosymbiosis occurred.

The mitochondrial genome presents a particularly striking case. The genes of yeast and

mammalian mitochondria code for virtually identical mitochondrial proteins, in spite of a considerable difference in gene organization. Vertebrate mitochondrial genomes are very small, with an extremely compact organization of continuous genes (see Chapter 25), whereas yeast mitochondrial genomes are larger and have some complex interrupted genes. Which is the ancestral form?

If we accept the endosymbiotic theory, we must conclude that the interrupted genes represent the source. This in turn implies that the ability to splice out introns was lost independently during the evolution of bacteria and vertebrate mitochondria.

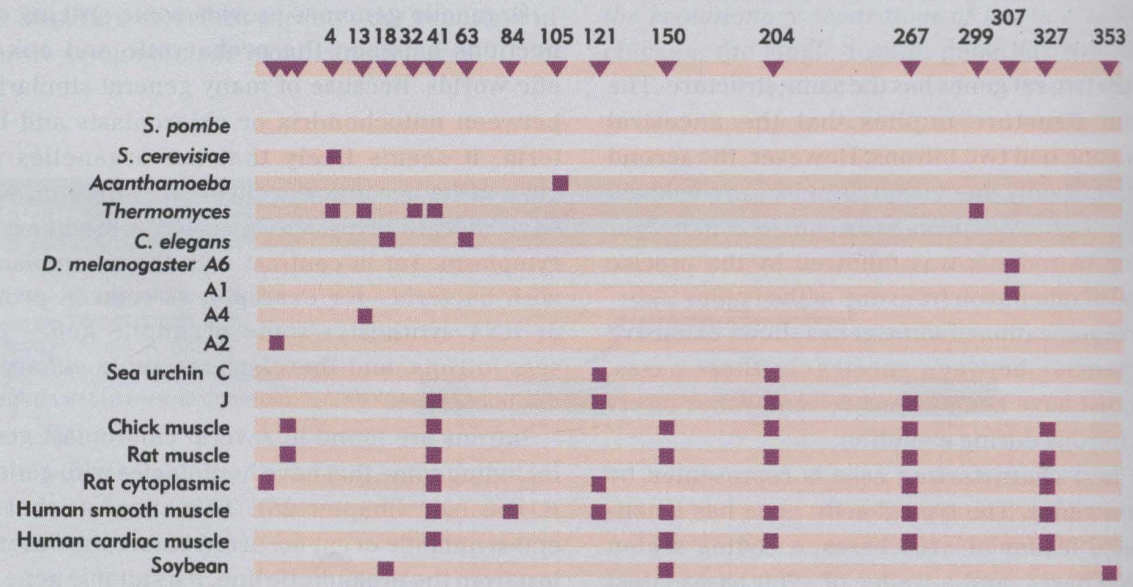
Considering the interrupted, sometimes very extensively spread out structure of eukaryotic genes, we can picture the eukaryotic genome as a sea of introns (mostly but not exclusively unique in sequence), in which islands of exons (sometimes very short) are strung out in individual archipelagoes that constitute genes.

Viewed in terms of their exons, few genes are alone in the eukaryotic genome. When the exons of a gene are used as a probe to identify corresponding sequences in the genome, often we find other, related sequences. These sequences may comprise the exons of a gene that represents a protein related to (or occasionally identical with) the product of the first gene. The sequence of each exon of one gene is related to the sequence of the corresponding exon in the other gene, but the corresponding introns generally differ more extensively.

The widespread existence of genes that consist of

**Figure 23.23**

Actin genes vary widely in their organization. The sites of introns are indicated in purple; the number identifies the codon interrupted by the intron.



corresponding exons emphasizes the importance of gene duplication as a mechanism for generating new genes. Duplication of the entire gene allows one copy to evolve via mutation, while the other retains its original function. Such genes are likely to code for proteins that have related functions.

In addition to genes that code for proteins that are related along their entire length, there are also genes that have in common an exon or series of exons coding for related sequences, although the other regions of the genes do not correspond. Such genes could originate by a duplication of part of a gene; the duplicated region could be moved to a new location where its exons combine with another series of exons to generate a new gene. The proteins coded by these genes share features coded by the related exons, but otherwise have different functions.

The history of related genes that have the same organization is likely to encompass a series of events in which first the component exons were brought together to form a coding region; later the entire series of exons and introns constituting the gene may have been duplicated. The duplication will have been followed by lesser divergence in exon sequences and greater divergence in intron sequences. We can attempt to reconstruct the history of their evolution from the sequences of genes related in this manner.

## Summary

All types of eukaryotic genome contain interrupted genes. The proportion of genes that is interrupted is low in yeasts and increases in the lower eukaryotes; few genes are uninterrupted in higher eukaryotes. The size of a gene is determined primarily by the lengths of its introns. Introns become larger early in the higher eukaryotes, when gene sizes therefore increase significantly. The range of gene sizes in mammals is generally from 1 to 100 kb, but it is possible to have even larger genes; the longest known case is dystrophin at 2,000 kb.

Introns are found in all classes of genes, including those coding for rRNA and tRNA. The structure of the interrupted gene is the same in all tissues, exons are joined together in RNA in the same order as their organization in DNA, and the introns usually have no coding function. Introns are removed from RNA by splicing. Some genes are expressed by alternative splicing patterns, in which a particular sequence is removed as an intron in some situations, but retained as an exon in others.

Positions of introns are conserved when the organization of homologous genes is compared between species. Intron sequences, however, vary, and may even be unrelated, although exon sequences remain well-related. The conservation of exons can be used to isolate related genes in different species.

Some genes share only some of their exons with other genes, suggesting that they have been assembled by addition of exons representing individual modules of the protein. Such modules may have been incorporated into a variety of different proteins. The idea that genes have been assembled by accretion of exons implies that introns were present in genes of primitive organisms. Some of the relationships between homologous genes can be explained by loss of introns from the primordial genes, with different introns being lost in different lines of descent.

# Further reading

## Reviews

The first major review of gene structure following the discovery of interrupted genes was provided by **Breathnach and Chambon** (*Ann. Rev. Biochem.* **50**, 349–383, 1981).

The relationship between gene organization and protein structure has been analyzed by **Blake** (*Int. Rev. Cytol.* **95**, 149–185, 1985).

The general characteristics of gene and protein evolution were reviewed by **Wilson *et al.*** (*Ann. Rev. Biochem.* 46, 573–639, 1977). The evolution of multigene families was reviewed by **Maeda and Smithies** (*Ann. Rev. Genet.* 20, 81–108, 1986).

Tecan Ex. 2080, Page 28 of 28
IDT, Inc. v. Tecan Genomics - IPR2025-00016