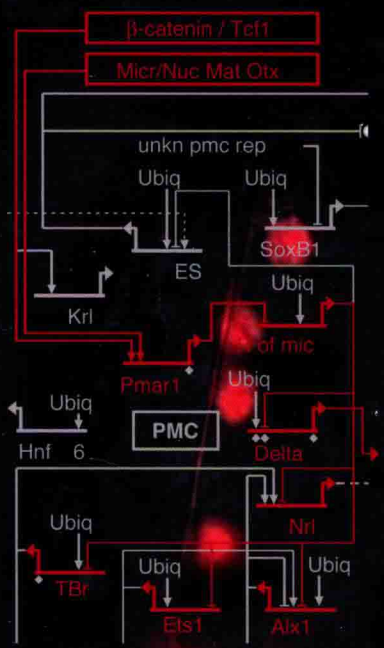
THE REGULATORY GENOME

Gene Regulatory Networks In Development and Evolution



ERIC H. DAVIDSON

THE REGULATORY GENOME

Gene Regulatory Networks in Development and Evolution

Eric H. Davidson
Division of Biology
California Institute of Technology
Pasadena, California

ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO

Academic Press is an imprint of Elsevier

Academic Press is an imprint of Elsevier
30 Corporate Drive, Suite 400, Burlington, MA 01803, USA
525 B Street, Suite 1900, San Diego, California 92101-4495, USA
84 Theobald's Road, London WC1X 8RR, UK

This book is printed on acid-free paper. ♾

Library of Congress Cataloging-in-Publication Data
Application submitted

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

ISBN 13: 978-0-12-088563-3
ISBN 10: 0-12-088563-8

For information on all Academic Press publications
visit our Web site at www.books.elsevier.com

Printed in Canada
06 07 08 09 10 9 8 7 6 5 4 3 2 1

THE REGULATORY GENOME

PREFACE

This book came about in the following way. In the winter of 2004, Jasna Markovac, the publisher of my monograph "Genomic Regulatory Systems" (2001), asked me to think about doing a second edition of that work. But no sooner had I opened that door and begun to survey the landscape than, as in an old legend of magic, everything seemed to have been transformed to another landscape. The conceptual peaks that were the landmarks of "Genomic Regulatory Systems" were still there, but now new mountains towered over them, and the whole domain appeared to be much more brightly illuminated, and to extend farther toward the horizon. It was beyond possibility to "update" a vista so changed.

"Genomic Regulatory Systems" sought as its main objective to crystallize the incontrovertible evidence that causality in development resides ultimately in *cis*-regulatory control of spatial gene expression. But development is the output of regulatory systems comprising large numbers of regulatory genes. Though in that work I often referred to the gene networks that would someday represent developmental programs, the few examples were anecdotal, and their general properties remained entirely obscure. Then in 2002 we published the first real scale gene regulatory network (including about 50 genes) explanatory of a major piece of development, specification of the endomesoderm of the sea urchin embryo. As the ancients used to say, the scales fell from our eyes. A whole field of developmental gene regulatory networks has now sprung forth, that encompasses many different animal systems. The structure/function properties that emerge from the architecture of these networks are a large part of what has transformed the conceptual terrain of this large area of bioscience. There follow more new things: a different way to think comparatively about various forms of development; a different way to think about the process of evolution; hence the subtitle of this book, "Gene Networks in Development and Evolution". Exploration of these new pathways toward scientific explanation of the developmental and evolutionary phenomena of biology is the central object of this volume. As with its predecessor, the approach I have taken here is that of demonstration by example: the points to be made rest upon powerful, exemplary experimental demonstrations, detailed, for those who desire experimental substance, in the figure captions. However, in no way have I attempted to be encyclopedic. So, willy nilly, there is much that could equally well have been included but was not, and my apologies en masse to the authors of these works.

I have not shied away from what are sometimes pejoratively been termed "big ideas," nor taken the view of an anonymous reviewer of a paper of mine who recently amused me with the complaint "But the original ideas in this paper are speculative!" This book includes many diagrams in which concepts are set forth

in specific form, just so they can be subjected to precise tests of falsification, and just so they can be used in precise ways to generate predictions I may not have thought of. One such idea, which underlies everything in this book, is the concept of genomically encoded information processing. To return to my metaphor above, this is like the geological basis of the landscape. In my view, *cis*-regulatory information processing, and information processing at the gene regulatory network circuit level, are the real secret of animal development. Probably the appearance of genomic regulatory systems capable of information processing is what made animal evolution possible.

This book begins with an overview of the regulatory genome and the concept of information processing in gene regulation (Chapter 1). It proceeds to an in-depth analysis of modular cis-regulatory designs for generation of spatial patterns of gene expression, and consideration of how they generate regulatory output (Chapter 2); thence to a comparative treatment of developmental pathways in terms of transient regulatory states (Chapter 3); to gene regulatory network theory and the character of diverse real developmental regulatory networks (Chapter 4); and finally to the application of network structure/function relations to some unsolved problems of animal evolution (Chapter 5). The image of a genomically encoded information processing system that throughout the life cycle responds conditionally to incident regulatory inputs can never lie far from the surface of any of these subject areas.

Science is made by scientists, whose creations deeply affect each others' progress. For me there have been certain scientists in each period of my own progress whose work and ideas have particularly illuminated the world: among them I must mention as of particular importance in this present period, and for what is included herein, Mike Levine, Ellen Rothenberg, Doug Erwin, Sorin Istrail, Bill McGinnis, and Lee Hood. This book would not have whatever worth it does were it not for the generosity of these people and also of Paola Oliveri and Joel Smith, postdoctoral colleagues in my laboratory, in reading, criticizing, and improving drafts of various parts, and in some cases all, of the manuscript. I have been extremely fortunate to have had the very expert services of a superb illustrator, Tania Dugatkin. In my own domain Deanna Thomas has provided invaluable assistance with figures, references, and everything else; and my graduate student Pei-yun Lee not only helped with technical research but also with figure attributions. Nor would this project have ever reached fruition were it not for the continued encouragement of Jasna Markovac, and of the careful, obsessive work of the production manager Paul Gottehrer at Academic Press/Elsevier. I also wish to say that since so much of what follows is linked to our expanding experimental invasion of gene regulatory networks, the support we have had for that research has been indirectly essential for this book as well: mainly this support has come from the National Institute of Child Health and Human Development and from the Genomes to Life Program of DOE, but also from NIGMS, NIRR, NIHGRI, NSF, NASA, Caltech's Beckman Institute, and Applied Biosystems, Inc.

Finally, I would like to dedicate this book to the person who has worked most closely with me on it, good days and bad, and that is Jane Rigg. She has been my editor, judge, administrator, research aide, and advisor throughout, as also on three other books I have written in the more than 35 years that we have worked together. Only my first book, "Gene Activity in Early Development" (1968) preceded the Jane Rigg era, but that was a very long time ago indeed.

Eric Davidson
April 2006

CONTENTS

CHAPTER 1

The "Regulatory Genome" for Animal Development

THE FRAMEWORK

Animal body plans, their structures and the functions with which their morphology endows them, are the integrals over time and space of their successive developmental processes. In abstract terms the mechanism of development has many layers, expanding in the diversity of its parts the farther removed from its core. At the outside, development is mediated by the spatial and temporal regulation of expression of thousands and thousands of genes that encode the diverse proteins of the organism, and that catalyze the creation of its nonprotein constituents. Deeper in is a dynamic progression of regulatory states, defined by the presence and state of activity in the cell nuclei of particular sets of DNA-recognizing regulatory proteins (transcription factors), which determine gene expression. At the core is the genomic apparatus that encodes the interpretation of these regulatory states. Physically, the core apparatus consists of the sum of the modular DNA sequence elements that interact with transcription factors. These regulatory sequences "read" the information conveyed by the regulatory state of the cell, "process" that information, and enable it to be transduced into instructions that

can be utilized by the biochemical machines for expressing genes that all cells possess. The sequence content, arrangement, and other aspects of the organization of these modular control elements are the heritage of each species. They contain the sequence-specific code for development; and they determine the particular outcome of developmental processes, and thus the form of the animal produced by every embryo. In evolution, the alteration of body plans is caused by changes in the organization of this core genomic code for developmental gene regulation.

This book is about the system level organization of the core genomic regulatory apparatus, and how this is the locus of causality underlying the twin phenomena of animal development and animal evolution. Because the sequence of the DNA regulatory elements is the same in every cell of each organism, the regulatory genome can be thought of as hardwired, and genomic sequence may be the only thing in the cell that is. Indeed that is a required property of gene regulatory elements, for they must endow each gene with the information-receiving capacity that enables it to respond properly to every conditional regulatory state to which it might be exposed during all phases of the life cycle, and in all cell types. For development, and therefore for major aspects of evolution, the most important part of the core control system is that which determines the spatial and temporal expression of regulatory genes. As used here, "regulatory genes" are those encoding the transcription factors that interact with the specific DNA sequence elements of the genomic control apparatus. The reason that the regulation of genes encoding transcription factors is central to the whole core system is, of course, that these genes generate the determinant regulatory states of development.

There follow several important and general principles of organization of the developmental regulatory apparatus, that is, of the control machinery directing expression of the regulatory genes themselves. First, signaling affects regulatory gene expression: The intercellular signals upon which spatial patterning of gene expression commonly depends in development must affect transcription of regulatory genes, or else they could not affect regulatory state. Therefore, the transcriptional termini of the intracellular signal transduction pathways required in development are located in the genomic regulatory elements that determine expression of genes encoding transcription factors. Second, developmental control systems have the form of gene regulatory networks: Since when they are expressed given transcription factors always affect multiple target genes, and since the control elements of each regulatory gene respond to multiple kinds of incident regulatory factor, the core system has the form of a gene regulatory network. That is, each regulatory gene has both multiple inputs (from other regulatory genes) and multiple outputs (to other regulatory genes), so each can be conceived as a node of the network. Third, the nodes of these gene regulatory networks are unique: Though it is not *a priori* obvious, each network node performs a unique job in contributing to overall regulatory state, in that its inputs are a distinct set, just as the factor it produces has a distinct set of target genes.

Fourth, regulatory genes perform multiple roles in development: The repertoire of regulatory genes is evolutionarily limited, and all animals use more or less the same assemblage of DNA binding domains, which define the classes of transcription factor. However, given factors are frequently required for different processes in different forms of development, and they are often used for multiple unrelated purposes within the life cycle. Thus, both within and among animal species, many regulatory genes must be able to respond to diverse regulatory inputs that are presented in various space/time places in the developing organism.

THE REGULATORY APPARATUS ENCODED IN THE DNA

Genomes, Genes, and Genomic "Space"

Viewing the animal genome as a whole, we may ask how much sequence information is required for the regulatory apparatus, compared to the amount encoding proteins. The question is confounded at the outset by the great variation among animal species in the overall amount of DNA per haploid genome, even within given phylogenetic clades. Examples are the greater than tenfold differences in genome size seen among insects, among fish, and among amphibians. This was already known by the end of the 1960s, from measurements carried out on dozens of species (reviewed by Britten and Davidson, 1971). On the other hand, estimations of the amount of genetic information read out into the mRNA populations of organisms of diverse genome size indicated early on that the large differences in genome size are not reflected quantitatively as differences in expressed mRNA complexity ("complexity" is here total mRNA sequence length in nucleotides if single molecules of each of the different mRNA species represented in a population were laid end-to-end). Two direct sets of measurements led to this conclusion (reviewed in Davidson, 1986). One was a comparison of maternal RNA complexities in eggs of various species of animal, the genomes of which range more than 100-fold in size. The results boiled down to the conclusion that the egg RNAs are all of roughly the same complexity, give or take a small variation. This is of course reasonable, since animal eggs have essentially similar jobs to do with their stored maternal mRNAs. The second set of measurements consisted of cytological and molecular analyses of the number of transcription units active in the extended "lampbrush" chromosomes in the oocytes of two species of amphibian that differ in genome size by a factor of about 10. About the same number of diverse genes is transcribed in the oocytes of these species, though the size of the individual transcription units appears to scale with genome size. The general implication from both data sets was that the complexity of given phases of gene expression is tightly constrained and independent of genome size across species. On the other hand, the amount of transcribed noncoding sequence, i.e., mainly intronic sequence, and of nontranscribed intergenic sequence, seems to have been relatively a "free variable" in animal evolution.

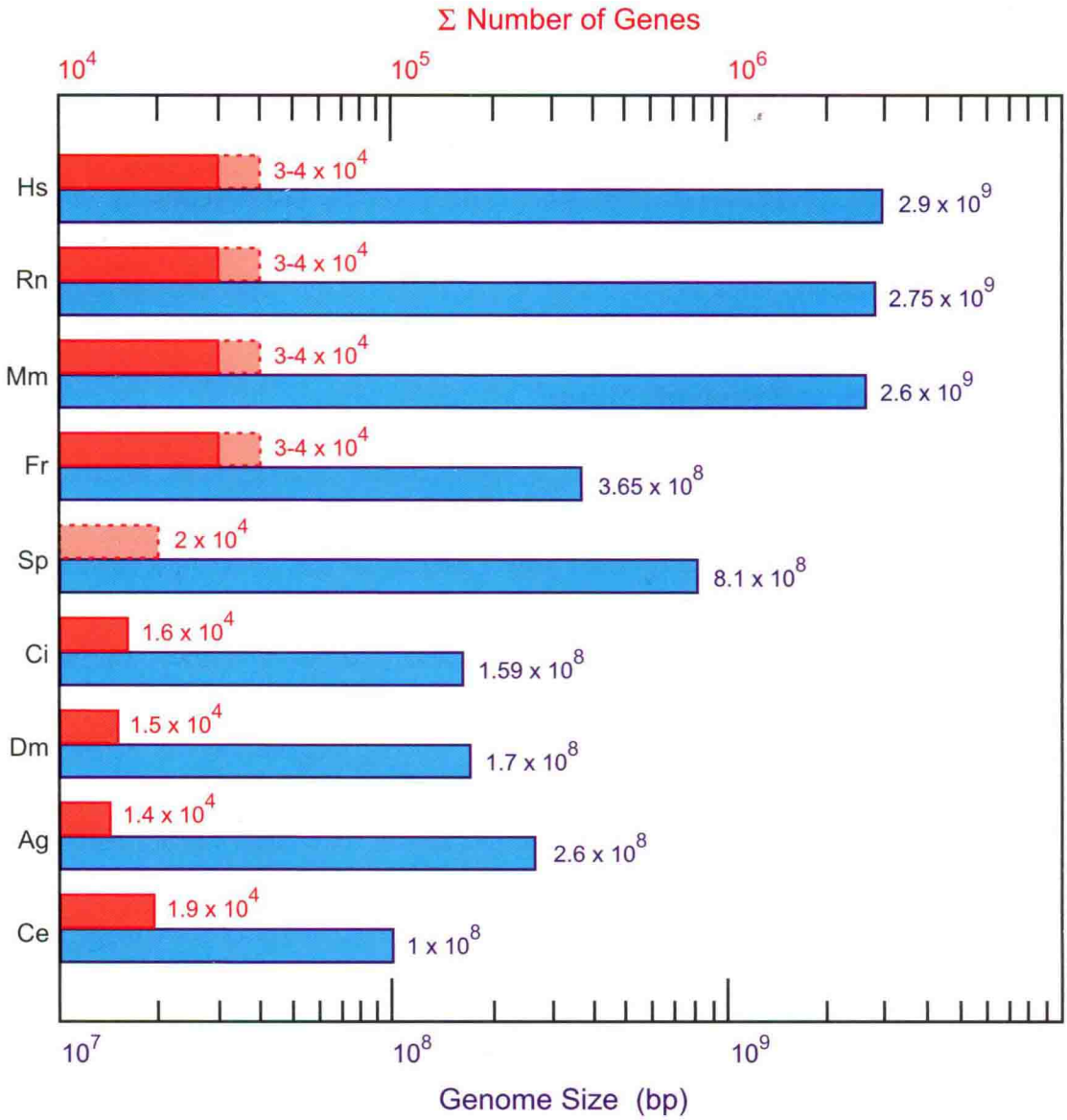


FIGURE 1.1. Representative animal genome sizes and gene numbers. Data are from genome sequencing. Dashed lines indicate larger alternative possible estimates of gene number also consistent with current data, or uncertainties. Genome sizes, indicated by blue bars (bottom scale), are given in

We now have genomic sequence for a number of animal species, and a huge accumulation of gene expression data. Comparison of results suggests that indeed, similarity in complexities of given expressed gene sets, more or less irrespective of genome size, is a basic fact of animal life. A limitation is that the genomic sequences and expression databases available are still clustered in only a few phylogenetic clades, and for obvious practical reasons are somewhat biased toward animals which have the smaller genomes in these clades. In Fig. 1.1 is shown a summary of total gene numbers estimated for different animal species, in comparison to their genome sizes (in this book we are concerned almost exclusively with bilaterally organized animals, the "bilaterians," and occasionally with their cousins the cnidarians, i.e., jellyfish and sea anemones; this excludes protozoans, sponges, and ctenophores). A quick glance at Fig. 1.1 conveys the main import: The basic "package" of genes needed in the genome of a bilaterian is about 15,000, and even the most complex vertebrates, i.e., in Fig. 1.1, rats, mice, and us, do not appear to have more than twice this complement. Compared to this the genome sizes of the animals included in Fig. 1.1 differ enormously (see legend for references). However, a caveat must be noted. This is that the definition of "gene" is in practice not trivial, even considering only protein coding genes. The values in Fig. 1.1 are certainly not overestimates, but particularly in vertebrates they could be to some extent underestimates of the actual gene complement (dashed bars). Estimates of gene number based on expressed cDNA sequences are often higher than those based only on exon sequence homology and computational prediction from the DNA sequence. In amniote vertebrates alternative splicing, and splicing over huge distances, are very common (Johnson *et al.*, 2003), so that determining what exons belong to what gene can be problematical. An important point is that while a large number of diverse exon combinations may be generated by alternative splicing, in regulatory terms given

base pairs in blue numerals; gene number estimates are given in red numerals. Sources: Hs, human: (International Human Genome Sequencing Consortium, 2001; Venter *et al.*, 2001; Johnson *et al.*, 2003; http://maple.lsd.ornl.gov/cgi-bin/GCat/GetOrg.cgi?org = human); Rn, *Rattus norvegicus* (rat): (Rat Genome Sequencing Project Consortium, 2004); Mm, *Mus musculus* (mouse): (Mouse Genome Sequencing Consortium, 2002; The FANTOM Consortium and the RIKEN Genome Exploration Research Group Phase I & II Team, 2002); Fr, *Fugu rubipres* (puffer fish): (Aparicio *et al.*, 2002); Sp, *Strongylocentrotus purpuratus* (sea urchin): (Hinegardner, 1974; Cameron *et al.*, 2000; Sea Urchin Sequencing Consortium, 2006); Ci, *Ciona intestinalis* (ascidian): (Dehal *et al.*, 2002); Dm, *Drosophila melanogaster* (fly): (Ashburner, 1989; Adams *et al.*, 2000); Ag, *Anopheles gambiae* (mosquito): (Holt *et al.*, 2002); Ce, *Caenorhabditis elegans* (nematode): (The *C. elegans* Sequencing Consortium, 1998; Stein *et al.*, 2001).

protein coding genes are tightly defined. Thus there come to mind no examples of genes that have more than one, two, or three alternative transcriptional basal promoters, i.e., locations where productive transcription of mRNA is initiated, and for a given gene, one or another of the exons beginning at these sites are present in all splice variants.

Though the bilaterians have rather similar sized gene toolkits, some of the constituents are specific to each clade, while others are shared with all other bilaterians. As the individual references in the legend to Fig. 1.1 detail, every genome includes sets of genes, often large, paralogous replications of certain gene families, that perform special functions for that kind of organism (common examples are the particular chemosensor and immune function genes that each clade uses). Every one of these genomes also contains a huge set of panbilaterian genes that encode common cytological, enzymatic, and cell type-specific differentiation functions, though these genes are present in diverse numbers in different genomes, and often display strikingly clade specific variation in protein coding sequence. But all these distinctions, as well as the bilaterian gene number constraint shown in Fig. 1.1, are peripheral to an essential fact: If we focus explicitly on the genes encoding transcription factors, and the genes encoding components of signaling systems required for developmental spatial regulation, there is almost no qualitative variation among the genomes of bilaterians. The genetic repertoires of each of these diverse bilaterians include genes encoding every known major family of transcription factor, and components of every known signaling pathway. The point is strengthened by the observation, made anew as each additional genome comes on line, that in each bilaterian clade, though all the regulatory and signaling gene families are represented, these gene families have diversified differently (Ruvkun and Hobert, 1998; Rubin *et al.*, 2000). That is, different bilaterian genomes have different numbers of genes encoding transcription factors belonging, for instance, to the various subfamilies of homeodomain regulators, Ets regulators, T-box regulators, nuclear receptors, or winged helix regulators, or different numbers of genes encoding TGFβ ligands. The replication and diversification of these gene families and subfamilies are always correlated with diversification of their functional roles in development. This major feature can be regarded as direct evidence of the process of reapplication of the same shared gene regulatory toolkit, a process that has occurred endlessly during bilaterian evolution (reviewed by Erwin and Davidson, 2002). Thus we can exclude the proposition that given bilaterian body plans and morphological structures differ from others because each has its own qualitatively specific class of gene regulatory protein and its own set of signaling pathways. Instead, the exact opposite is true.

What of the great majority of the genomic DNA that is not included in genes, here including not only those genes that encode proteins, but also those encoding rRNA and other kinds of RNA that function at posttranscriptional levels? To begin let us think about all forms of known gene regulatory elements in which the genomic sequence is important for function (here and in the following, the term "element" is used broadly, to denote any genomic feature that has a specific

regulatory role). The sign of sequence-dependent function is constraint in its rate of change during evolution, relative to the majority basal rate at which selectively neutral DNA sequence diverges (e.g., most intronic DNA, most third base codon sequence, most intergenic DNA). A very approximate estimation can be made based on those few gene regulatory systems that are relatively well known (for cases, turn the pages of this volume). By extrapolation from these cases, this evidence suggests that there is at least as much sequence-dependent gene regulatory information built into the genome as there is sequence included in mature gene products; that is, than in protein coding mRNAs plus all other kinds of functional transcripts that ever appear in the cytoplasm. There could indeed be twice as much gene regulatory as coding information, or more; but not enough is known to recognize most of it *a priori*. It is an amazing comment on the current predilections of molecular biology and genomics that, relatively speaking, only minute attention has so far been devoted to reading the enormous regulatory code carried in the genomes of animals, compared to reading the protein coding capacity. What this has meant is focus mainly on structure/function relations at the outermost layers of animal life systems, whereas it is only at the innermost layers, where the genomic control apparatus operates, that development and evolution can be explained.

Even assuming high-end estimates for the dimensions of the regulatory genome, it would still be true that most of the DNA in the genomes of Fig. 1.1 has no likely sequence-dependent role. Repetitive elements account for some of the sequence-independent DNA. In larger genomes much of the DNA sequence is repetitive, in smaller genomes less. The repeats occur in tens, hundreds, or thousands of copies of more or less related sequence per genome (early studies based mainly on genome-wide DNA renaturation kinetics reviewed by Davidson *et al.*, 1974; current data for sequenced genomes are summarized in references in legend of Fig. 1.1). Repetitive sequences are mainly due to insertions and replications of transposable elements (Moore *et al.*, 1978; Britten, 1984; Deininger and Batzer, 2002). In terms of both the position of these elements in the genome and their frequencies, they change during evolution many times faster than do the underlying syntenic (chromosomal gene linkage) scaffolds of which animal genomes seem basically to be composed (Aparicio *et al.*, 2002; Bourque *et al.*, 2004). Related animal genomes differ more in their repetitive sequence content than in anything else, so by that definition as well as by provenance, repetitive sequences are charter components of the sequence-independent portion of the genome. Even so, they occasionally transpose into a location carrying a gene regulatory element with them, or they mutate to constitute such (reviewed by Britten, 1996; for example, Zhou *et al.*, 2002). To what extent the transposition of repetitive sequence elements has contributed overall to the generation of novel gene regulatory systems during evolution is a question that remains to be resolved.

In most animal genomes the larger part of the sequence-independent, freely evolving genomic DNA is the single or very low copy sequence that constitutes the major extent of intronic and intergenic sequence. Should we think of this major