MOLECULAR BIOLOGY INTELLIGENCE UNIT

Eugene D. Sverdlov

Retroviruses and Primate Genome Evolution

CRC Press
Taylor & Francis Group

MOLECULAR BIOLOGY
INTELLIGENCE
UNIT

Retroviruses and Primate Genome Evolution

Eugene D. Sverdlov
Institute of Molecular Genetics
of Russian Academy of Sciences
Moscow, Russia

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

CONTENTS

EDITOR

Eugene D. Sverdlov
Institute of Molecular Genetics of Russian Academy of Sciences
Moscow, Russia
e-mail:sverd@humgen.siobc.rus.ru
Chapters 1, 2, 4, 5

CONTRIBUTORS

Jonas Blomberg
Section of Virology
Department of Medical Science
Uppsala University
Uppsala, Sweden
email: Jonas.Blomberg@medsci.uu.se
Chapter 11

Natalia E. Broude
Center for Advanced Biotechnology and Department of Biomedical Engineering
Boston University
Boston, Massachusetts, U.S.A.
e-mail: nebroude@bu.edu
Chapter 2

Gerhard Hunsmann
Department of Virology and Immunology
Deutsches Primatenzentrum
Goettingen, Germany
e-mail: ghunsma@gwdg.de
Chapter 3

Patric Jern
Section of Virology
Department of Medical Science
Uppsala University
Uppsala, Sweden
e-mail: patric.jern@medsci.uu.se
Chapter 11

Aris Katzourakis
Department of Biological Sciences
Imperial College
Ascot, Berkshire, U.K.
e-mail:a.katzourakis@imperial.ac.uk
Chapter 10

Yuri B. Lebedev
Laboratory of Structure and Functions of Human Genes
Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry
Russian Academy of Science
Moscow, Russia
e-mail: yuri@humgen siobc.ras.ru
Chapter 8

Christine Leib-Mösch
GSF-National Research Center for Environment and Health
Institute of Molecular Virology
Oberschleissheim, Germany
e-mail: leib@gsf.de
Chapter 7

Dixie L. Mager
Terry Fox Laboratory
B.C. Cancer Agency
Department of Medical Genetics
University of British Columbia
Vancouver, British Columbia, Canada
e-mail: dmager@bccrc.ca
Chapter 6

Patrik Medstrand
Department of Cell and Molecular Biology
Lund University
Lund, Sweden
e-mail: patrik.medstrand@medkem.lu.se
Chapter 6

Eugene V. Nadezhdin
Laboratory of Structure and Functions of Human Genes
Institute of Bioorganic Chemistry
Russian Academy of Sciences
Moscow, Russia
e-mails: neugene@humgen.siobc.ras.ru
eugene_nadezhdin@mail.ru
Chapter 4

Christian Roos
Working Group Primate Genetics
Deutsches Primatenzentrum
Goettingen, Germany
e-mail: croos@dpz.gwdg.de
Chapter 3

Ulrike Schön
Alopex GmbH
Kulmbach, Germany
e-mail: uschoen@alopexgmbh.de
Chapter 7

Wolfgang Seifarth
Medical Clinic III
Faculty of Clinical Medicine Mannheim
University of Heidelberg
Mannheim, Germany
e-mail: seifarth@rumms.uni-mannheim.de
Chapter 7

Michael Tristem
Department of Biological
Imperial College
Ascot, Berkshire, U.K.
e-mail: a.katzourakis@ic.ac.uk
Chapter 10

Dmitrijs Ushameckis
Section of Virology
Department of Medical Science
Uppsala University
Uppsala, Sweden
e-mail: Jonas.Blomberg@medsci.uu.se
Chapter 11

Louie N. van de Lagemaat
Terry Fox Laboratory
B.C. Cancer Agency
Vancouver, British Columbia, Canada
e-mail: lvandela@bccrc.ca
Chapter 6

Tatiana V. Vinogradova
Laboratory of Structure and Functions of Human Genes
Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry
Russian Academy of Science
Moscow, Russia
e-mail: tv@humgen.siobc.ras.ru
Chapter 9

Hans Zischler
Institute for Anthropology
University of Mainz
Mainz, Germany
e-mail: zischler@mail.uni-mainz.de
Chapter 3

PREFACE

Three famous questions: "Who we are?" "Where are we from?" and "Where are we going?" together with a more general one "What is life?" were asked by people of every culture. When it became clear that certain genes are responsible for certain phenotypic traits, one more question was added to these four: "Which genes make us human?". One could hope to find the answers by studying primates, their communities and differences from humans, as well as their evolutionary relations, succession of appearance and accumulation of differences after the divergence of the human lineage from lineages of the extant human close relatives, great apes (chimpanzees, gorillas, and orangutans). Humans and our closest living relatives, the apes, including both "great apes" and "lesser apes" (gibbons and siamangs) form the *Hominoidea* superfamily. These *Hominoidea* are remarkably similar and at the same time dramatically different. They are different not only in their appearance but also in such characteristics as behavior and resistance to various diseases, including cancer and AIDS. Many lines of evidence indicate that all of them originated from a common ancestor about 17 Mya (million years ago), and that the last common ancestor of human and great apes, i.e., of human and chimpanzee, extincted about 5 Mya. It is a great challenge to reconstruct its genetic architecture and then to understand the ways of its transformation into two closely related, but different architectures of human and chimpanzee. What events caused their divergence in evolution? What genes and regulatory systems were involved in branching hominids off from their closest relatives, chimpanzees and bonobo and then in their proceeding to the *Homo* genus crowned with extant *Homo sapiens*, that is humans with their brain size of at least 600 cubic centimeters, extended period of childhood growth and development, possession of language and many other human specific traits? And what processes step by step shaped the modern human, *Homo sapiens sapiens* during 5 Myr (million years) of its progress after the divergence from chimpanzee?

It is widely believed that the evolutionary history of a species is reflected in its genome sequence, and therefore the most straightforward way to study primates is sequence comparisons of various primate genomes. The sequencing of the human genome has already contributed a great deal to such an analysis, and as soon as the sequencing of the chimpanzee genome is finished, we will have enormous information to work on. Multiple differences of various kinds that can be envisioned between the two genomes will inevitably puzzle researchers trying to find the genetic reasons for human speciation and rapid phenotypic evolution. How can one single out a relatively small number of the differences that have been actually or most possibly involved in speciation from the multitude of just random neutral mutations accumulated during millions of years?

One way is to try to identify the regions positively selected in evolution. In the case of coding stretches of the genome, an enhanced rate of nonsynonymous substitutions compared to synonymous ones is a widely accepted indication of positive selection. However, the situation with regulatory regions or regions which encode noncoding RNAs is much more tangled. Their conservation indicates that these regions were important for a sufficiently long period(s) of evolution, but generally speaking could be of no importance in other periods. Clearly, this criterion may not be applicable to the regulatory units that appeared after the human-chimp lineages divergence. But precisely these units might be the acquisitions that played a major role in shaping our human phenotype.

Therefore, it seems inevitable to resort to rather traditional hypothesis-driven approaches, when the research starts from the hypotheses aimed at explaining why and how the most significant human features, such as language or cognitive capacities, could emerge. In this approach, only particular loci will be taken for interspecies comparison. This "last line of research" will undoubtedly be stimulated by new information on genome-wide comparisons.

Obviously, the chances to reconstruct the succession of the genetic events which had occurred during millions of years of evolution are negligible. But what we can hope to gain as a result of such a comparative research is a deeper understanding of the mechanisms governing the modern genome and the role of particular elements in the networks responsible for the functional integrity of the genome. We will also certainly be able to reveal differences in spatial-temporal networks of the events determining the development of different species and thus to form a basis for the second order hypotheses related to the genetic basis of differences in the phenotypes of extant species. The achievement of this goal will be a great step towards the understanding of what life is in general, and what its peculiarities are regarding our presently prospering, but still endangered, species as well as how these peculiarities could evolve.

What kind of differences might promote speciation? Since a classical work of King and Wilson, who in 1975 undertook a thorough comparison of the molecular data available on chimpanzee (*Pan troglodytes*) and human (*Homo sapiens*), it is widely accepted that, as put by the authors: "...a relatively small number of genetic changes in systems controlling the expression of genes may account for major organizational differences between human and chimpanzees". It is now known as the regulatory hypothesis. Later, it became a major constituent of the Evo-Devo concept suggesting that evolution depends on heritable changes in the development and, according to Duboule and Wilkins (1998), "...the primary source of developmental differences... will prove to be not unique gene products but rather the way that comparable, or the same, gene functions are differentially deployed in their development.... Many so called heterochronic shifts altering developmental programs and morphologies involve no more than alteration in the times and cellular site at which particular regulatory molecules are expressed rather than alteration in those molecules themselves". In metazoan evolution, these processes have been brought under intercellular control regarding the time, place, and conditions of functioning. It seems logical to propose that such developmental functional shifts could be caused by changes in gene regulation, which in turn could result from addition of a new regulatory module(s) to the pre-existing gene regulatory system.

The title of this volume, *Retroviruses and Primate Genome Evolution,* reflects its goal to conceive the role of the obligate inhabitants of all vertebrate genomes—endogenous retroviruses, especially those emerged in genomes rather recently, during primate evolution. Although a special focus in the volume is put on human endogenous retroviruses (HERVs), some attention is also paid to other retroelements (REs), like LINEs and Alu to give a more comprehensive view of the evolutionary potential of these perpetually mobile entities now occupying almost a half of the human genome.

Keeping in mind that REs are jumping carriers of the regulatory *cis*-elements adapted for RNA-polymerase II or III transcription regulation, it is quite reasonable to put them on the list of highly probable candidates for evolutionarily significant

changes, capable of affecting the regulation of the genes in the vicinity of which they were inserted. Such changes could quite possibly occur in the developmental regulatory machinery thus causing the above mentioned developmental shifts.

Indeed, the data obtained for different species clearly demonstrate that REs insertions can change not only the structure of genes, and hence their products, but also their regulation. Moreover, transposable elements can have their own genes and thus enrich the genome with new genetic information, like genes of reverse transcriptase or viral resistance. Although the newly inserted elements are known to mostly cause deleterious effects including hereditary diseases, the host cells sometimes exploit the ability of REs to generate variations for their own benefit. Among other REs, HERVs are considered to be the most sophisticated. ERV-related sequences are believed to represent footprints of ancient germ-cell retroviral infections which now occupy up to 8% of the human genome. They have excitingly diverse tools of affecting the human genome functioning originated from exogenous retrovirus systems of successful life cycle. They can change the host genome function through expression of retroviral genes, human genome loci rearrangements due to retropositions of HERVs, or by the ability of their long terminal repeats (LTRs) to regulate nearby genes. A multitude of solitary LTRs comprise a variety of transcription regulatory elements, such as promoters, enhancers, hormone-responsive elements, and polyadenylation signals. This feature makes LTRs potentially able to strongly affect the expression patterns of neighboring genes. It can be imagined that the appearance of such invaders in the genome can change some functions relevant to development and thus provide new traits for subsequent natural selection. They can therefore be considered prime suspects for being a major class of causative agents in speciation.

Individual chapters in this book are devoted to specific areas of research into human genome evolution and possible involvement of REs in the processes related to evolution and includes REs own evolution which was prime interdependent with the host genome evolution.

The book opens with four chapters giving a general insight into human genome structure and function analysis and ideas on the genome evolution. Chapter 1, "A glance at evolution through the genomic window" (by E. Sverdlov), describes the status of whole genome sequence comparisons which, for the first time, opened a possibility to analyze evolutionary changes at a whole genome level. The intra- and interspecies comparisons of the sequenced genomes demonstrated that the genome complexities did not directly correlate with the number of genes and suggested the importance of combinatorial interactions in the cells and organisms as a major player in the complexity of live systems. The whole genome comparisons allowed one to elucidate the role of gene duplications, gross genome rearrangements, transposable elements and other genomic changes in divergence of genomes, thus forming a solid basis for understanding genetic mechanisms of evolution. The whole genome analyses developed in parallel and interdependently with the development of new concepts of evolution, such as evolutionary developmental biology (Evo-Devo), aimed at explaining how developmental processes and mechanisms become modified in evolution, and how these modifications produce changes in animal morphology. This review considers new data and trends and supports the idea that transposable elements play a role of a major pacemaker in evolution being a "depot" of evolvability factors.

Chapter 2 entitled "Complex Genome Comparisons: Problems and Approaches" by N. Broude and myself provides a brief outline of the experimental approaches to genome-wide interindividual, interpopulation, and interspecies comparisons. Such comparisons form a unique background for deciphering spatial and temporal genomic regulatory networks and their changes during evolution. They are also indispensable for understanding genetic and environmental contributions to complex diseases afflicting modern society. The chapter describes also a range of modern approaches to genome-wide complex genome comparisons with their advantages and disadvantages.

Chapter 3 by H. Zischler and his colleagues is devoted to primate evolution. Information on evolutionary events and relations of different primate extant species is indispensable for understanding the role of particular genomic elements in evolution. The authors review the modern status of investigations on primate phylogeny with all its problems and contradictions. An emphasis is made on the divergence of non-human primates, relevant interpretations of the fossil record and molecular evidence, including retropositional evidence. Whereas a congruent view is emerging concerning phylogenetic relationships among primate taxa at a higher taxonomic level, e.g., primate infraorders, there is still considerable debate on primate origins or very recent splits in primate evolution. Obtaining more clarity about primate origins is to a large degree hampered by the sparseness of the critical fossil record. If both molecular and fossil evidence is available for a certain splitting, many interpretations based on these two completely different approaches seem to be remarkably compatible. Attention is also paid to problems of modern human evolution.

Chapter 4 "How different is the human genome from genomes of the great apes?" by E. Nadezdin and E. Sverdlov gives an account on sequence and chromosomal organization differences between highly related genomes of humans and the African great apes that were accumulated during Hominoid evolution. Some of them certainly form a genetic basis for recently evolved, specifically human traits. Human genome sequencing revealed its characteristic features, and the ongoing sequencing of the chimpanzee genome continuously widens the possibilities of large-scale systematic comparison of the two genomes. Now it is more and more apparent that most probably hundreds and thousands of genes were involved in the divergence of even the most closely related species of human and chimpanzee. The divergence might be caused by changes in gene regulation and by modifications of protein biochemical functions, gene duplications, losses and acquisitions. A great challenge will be to single out functionally significant differences from the mess of all changes accumulated during evolution.

These four general chapters are followed by reviews devoted to various aspects of evolution, interactions with the host genome, and involvement of REs in various human diseases. This series opens with a very brief introduction, Chapter 5, "Endogenous Retroviruses and other retroinsiders", which I wrote to give general information about retroviruses, their endogenous counterparts, and other retroelements in the human genome. I hope it will make the reading of the following more specialized reviews easier.

The next two chapters, by Dixie Mager et al (Chapter 6, "Genomic Distributions of Human Retroelements") and by Christine Leib-Mosch et al (Chapter 7, "Influence of human endogenous retroviruses on cellular gene expression") focus on distribution and function of REs in the human genome. Chapter 6 reviews the studies performed in the last 20 years on chromosomal arrangements of human

retroelements including endogenous retroviruses. Biological mechanisms or evolutionary forces that might influence their modern distribution patterns are also discussed. Chapter 7 discusses a variety of effects of newly inserted REs on adjacent genes. These effects include not only impairment of gene function, but also enhancement of transcription, changes in tissue specificity of gene expression, and creation of new gene products with modified functions, e.g., via alternative splicing. The conclusion is that retrotransposable elements may have served as catalysts of genomic evolution and possibly played a role in primate speciation and adaptation.

The reviews by Yu. Lebedev, "Genome-wide search for human specific retroelements" (Chapter 8), and Tatyana Vinogradova, "Approaches to genome wide analysis of human gene expression: application to analysis of expression of human endogenous retroviruses in normal and cancerous tissues" (Chapter 9), provide an insight into experimental techniques used for revealing species specific REs and analysis of their functional status.

Chapter 10 by A. Katzourakis and M. Tristem, "Phylogeny of human endogenous and exogenous retroviruses", is somewhat different in its style from other reviews in this book. And although it is rather a research article than a review, this chapter successfully demonstrates the state-of-the art for attempts to reconstruct the correct phylogeny of endogenous retroviruses with all their problems and difficulties. Quite a number of assumptions made to smooth evident contradictions of grouping based on just the level of sequence identity make the resulting tree appreciably dependent on the researcher's intuition and prejudice. Similarly, the results obtained with even more sophisticated tools of modern computer-based phylogeny analysis are by no means final or indisputable. Unfortunately, our past seems to be almost as cloudy as our future. With all their assumptions, Tristem and his colleagues found 31 HERV families in the human genome. Currently, it is probably the most extensive survey of HERVs diversity. I think that sequencing of other primate genomes will reveal even more HERV families.

Finally, the title of the last chapter by J. Blomberg et al, "Evolutionary aspects of human endogenous retroviral sequences (HERVs) and diseases", precisely reflects its content. Also discussed is the impact of retroviruses on a variety of human diseases.

Taken together, these partially overlapping chapters hopefully provide a balanced and accurate overview of our current knowledge of the complex interplay of the human genome with its mobile inhabitants, retrotransposons.

To conclude the Preface, I would like to stress that hardly a particular genomic constituent or even a numerous group of constituents like REs has caused such a pronounced phenotypic difference between human and chimpanzee. Undoubtedly, hundreds or even thousands of changes occurred during 5 million years after the two species diverged, and that eventually made us humans. However, the chain of events leading to human might well be initiated by some very first genomic perturbations, and this initiation could be caused by retroviral integration(s) and/or by other RE retropositions into a critical site of the ancestor genome.

Exact dating of RE integrations and comparative functional analysis of the genes in the species which sustained the integrations and those retaining the same but native genes will help us to better understand our own evolution.

Eugene D. Sverdlov

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

ABBREVIATIONS

aa	amino acid
ADHD	attention deficit/hyperactivity disorder
AFLP	amplified fragment length polymorfism
ALV	Avian leukosis virus
ATM	ataxia telangiectasia mutated, gene responsible for ataxia telangiectasia
BAC	bacterial artificial chromosome
BES	bacterial artificial chromosome end sequence
BLAST	basic local alignment search tool
BLV	bovine leukemia virus
bp	base pair
BRCA1 and 2	genes responsible for hereditary breast cancer
CA	capsid protein surrounding the RNAs bound to nucleocapsid (NC) proteins
CGAP	Cancer Genome Anatomy Project
CGH	comparative genomic hybridization
CNS	central nervous system
COX	cytochrome *c* oxidase genes
CSF	cerebrospinal fluid
DD	differential display
DDRT-PCR	differential display reverse transcription PCR
DLBCL	diffuse large B-cell lymphoma
dUTPase	deoxyuridinetriphosphatase
EBV	Epstein-Barr virus
EM	electron microscopy
ERV	endogenous retroviral sequence
ES	embryonic stem cells
EST	expressed sequence tag
FeLV	feline leukemia virus
FISH	flourescence In Situ Hybridization
FV	friend virus
GaLV	gibbon ape leukemia virus
GDB	genome database
GC	germ cell
GCT	germ cell tumour
GEF	gene expression fingerprinting
GMS	genomic mismatch scanning;
GRE	glucocorticoid responsive element
HERV	human endogenous retroviral sequence, human endogenous retrovirus
HHV6	human herpesvirus 6
HHV7	human herpesvirus 7
HIV	Human Immunodeficiency Virus
HML	human MMTV-like sequence

Hsp90	heat-shock protein 90 in *Drosophila*
HSV	Herpes Simplex virus
HTDV	human teratocarcinoma derived virus
HTLV	human T-lymphotropic virus, human T-cell leukemia virus
HVRI, HVRII	hypervariable regions I and II of mtDNA (sequences)
HYAL2	hyaluronidase type 2
IAP	intracisternal type A particle
IDDM	insulin dependent diabetes mellitus
IF	immunofluorescence
IFN	interferon
IgG	immunoglobulin G
IgM	immunoglobulin M
IN	integrase
Inr	initiator element
IRS-PCR	interspersed repetitive sequence PCR
ISU	immunosuppressive unit, a conserved sequence derived from p15E
JSRV	Jaagsiekte retrovirus
kb	kilobase pairs
L1	LINE1, the most abundant LINE in mammalian genomes
LCR	locus control region
LINE	long interspersed nuclear element non-LTR retroelements encoding their own reverse transcriptase
LTR	long terminal repeat
MA	matrix protein,. matrix
MaLR	mammalian LTR retroelements. A large heterogeneous group of LTR retroelements found in mammals
Mb	megabase, million base pairs
MDV	Marek disease virus, a tumourigenic herpesvirus of turkeys
MHC	major histocompatibility complex
MHR	major homology region
MIR	mammalian interspersed repeat, class of SINEs
MLV	mouse (murine) leukemia virus
MMTV	mouse mammary tumour virus
MPMV	Mason-Pfizer monkey virus
MRCA	most recent common ancestor
MS	Multiple sclerosis
MSRV	Multiple sclerosis – associated retrovirus (retroviral element)
mtDNA	mitochondrial DNA
Mya	million years ago
Myr	million of years
MZ	monozygotic twins
NC	nucleocapsid proteins
ncRNAs	noncoding RNAs
Neu5Ac	N-acetyl-neuraminic acid

Neu5Gc	N-glycolyl-neuraminic acid
NK cells	natural killer cells
NMD	nonsense mediated decay
nr and htgs	nonredundant and high-throughput subset of GDB, correspondingly
NRE	negative regulatory sequence
nt	nucleotide
NWM	New World monkey
ODD	ordered differential display
ORF	open reading frame
OWM	Old World monkey
PA	polyacrylamide
PAGE	PA gel electrophoresis
PBL	human peripheral blood leukocytes
PBS	primer binding site
PCR	polymerase chain reaction
PERV	porcine endogenous retrovirus
PHA	phytohemagglutinin
PLZF	promyelocytic leukemia zinc finger
PML	promyelocytic leukemia
PPT	polypurine tract
PR	protease
PS	PCR suppression
PTN	pleiotrophin
RAPD	randomly amplified polymorphic DNA
RDA	representational difference analysis
RE(s)	retroelement(s) A transposable element that transposes via an RNA intermediate
READS	restriction endonucleolytic analysis of differentlly expressed sequences
REs	retroelements
RFLP	restriction fragment length polymorfism;
RIDGE	regions of increased gene expression
RLGS	restriction landmark genome scanning;
RNAi	RNA interference
RSV	rous sarcoma virus
RT	reverse transcriptase
RT-PCR	reverse transcription - PCR
SAGE	serial analysis of gene expression
SBH	sequencing by hybridization
SDD	systematic differential display
SDDIR	selective differential display of RNAs containing interspersed repeats
SH	subtractive hybridization
SINE	short interspersed nuclear element; non autonomous retroelement typically derived from a small functional RNA that has amplified in the genome by retrotransposition.

SLE	systemic lupus erythematosus
SNP	single nucleotide polymorphism
SRV	simian retrovirus
SSH	suppression subtractive hybridization
STR	short tandem repeat
SU	envelope surface glycoprotein, surface unit; outer portion of retroviral Env proteins
TDD	targeted differential display
TE	transposable element
TGDD	targeted genomic DD
TGF-β	transforming growth factor beta
TK	thymidin kinase
TM	transmembrane glycoprotein, inner portion of retroviral Env proteins
TPRT	target-site primed reverse transcription
3'UTR	3' untranslated region
5'UTR	5' untranslated region
VDA	variant detection array
VNTR	variable number of tandem repeats
VZV	Varicella-Zoster virus (mumps virus)
Vβ7TCR	T cell receptor containing a beta 7 variable chain
WGD	whole genome duplication
XRV	exogenous retrovirus

CHAPTER 1

A Glance at Evolution through the Genomic Window

Eugene D. Sverdlov

"Nothing in biology makes sense except in the light of evolution."
Th. Dobzhansky

Abstract

A large number of various genomes sequenced recently for the first time make it possible to analyze evolutionary changes at a whole genome level, unlike a single gene level. Intra- and interspecies comparisons of the sequenced genomes demonstrated that the organism's complexities did not directly correlate with the number of genes and suggested the importance of combinatorial interactions in cells and organisms as a major player in the complexity of live systems. They made it possible to reveal conserved and variable elements of the genomes and to suppose that tens of thousands of proteins are made of just about 1,500-2,000 discrete structural protein units called domains or modules. Different modular proteins are formed from these modules taken in different combinations, and this shuffling might play an extremely important role in the genesis of evolutionary novelties. The new domain architectures (defined as the linear arrangement of domains within a polypeptide) have emerged in evolution by shuffling, adding or deleting domains, resulting in new proteins composed of old parts. More complex organisms seem to contain more various protein architectures than the simpler ones. Whole genome comparisons allowed one to elucidate the role of gene duplications, gross genome rearrangements, transposable elements and other genomic changes in genome divergency, thus forming a solid basis for understanding genetic mechanisms of evolution. Whole genome analyses developed in parallel and interdependently with the development of new concepts of evolution such as evolutionary developmental biology (Evo-Devo) aimed at explaining how developmental processes and mechanisms become modified in evolution, and how these modifications produce changes in animal morphology and body plans. Among these new concepts can be found such fruitful notions as (i) a universal principle of modular organization at various levels of living systems, particular modules being changed and co-opted into new functions without affecting other modules, (ii) a concept of network-like organization of cellular regulatory systems with cis-regulatory elements of the genome functioning as major nodes of the networks, and a crucial evolutionary role of changes in the regulatory systems, (iii) an assumption of increase in functional load per regulatory gene with increasing the complexity of the organism, (iv) an idea of evolvability as a universal feature of the living entities, and a very important concept (v) that not only natural selection, but also internal developmental biases can form the basis for evolutionary changes.

Retroviruses and Primate Genome Evolution, edited by Eugene D. Sverdlov.

This chapter considers new data and trends and supports the idea that transposable elements play a role of a major pacemaker in evolution being a "depot" of evolvability factors.

Introduction

Jack London in his novel "The Sun Dog Trail" described the following scene:

Sitka Charley smoked his pipe and gazed thoughtfully at the Police Gazette illustration on the wall. For half an hour he had been steadily regarding it... "Well?" I finally broke the silence.

He took the pipe from his mouth and said simply, "I do not understand."

"That picture-what does it mean? I do not understand."

I looked at the picture. A man, with a preposterously wicked face, his right hand pressed dramatically to his heart, was falling backward to the floor. Confronting him, with a face that was a composite of destroying angel and Adonis, was a man holding a smoking revolver.

"That picture is all end," he said. "It has no beginning."

"It is life," I said.

"Life has beginning," he objected..." Something happens in life. In pictures nothing happens. No, I do not understand pictures."

His disappointment was patent. I felt, also, that there was challenge in his attitude. He was bent upon compelling me to show him the wisdom of pictures. "Pictures are bits of life," I said. "We paint life as we see it. For instance, Charley, you are coming along the trail. It is night. You see a cabin. The window is lighted. You look through the window for one second, or for two seconds, you see something, and you go on your way. You saw maybe a man writing a letter. You saw something without beginning or end. Nothing happened. Yet it was a bit of life you saw."

I think it is an exact description of our efforts to understand evolution. We see it as a momentary picture without beginning and end and try to understand life from the very beginning in all its diversity and in movement. And we try to animate the picture looking at it through different windows: through the window of paleontology, the window of phylogenetics, the window of developmental biology, and through the window of comparative genomics. In this chapter I will try to sketch what we see through the genomic window. It will mainly focus on human genome evolution, whereas the data on other genomes will be used for comparative purposes.

Towards the Understanding of "The Mechanisms That Bring about Evolutionary Changes"

Dobzhansky in *Genetics and the Origin of Species*, first published in 1937 (citation from ref. 1) wrote: "The problem of evolution may be approached in two different ways. First, the sequence of the evolutionary events as they have actually taken place in the past history of various organisms may be traced. Second, the mechanisms that bring about evolutionary changes may be studied... ".

However difficult may appear to be the reconstruction of successive evolutionary events, the unraveling of the mechanisms leading to the morphological changes, which are then fixed due to natural selection and eventually lead to the emergence of new species is a much more difficult task. A fundamental question (further designated as Question 1) is what kind of the genomic changes are transformed into the phenotypic changes subject to natural selection and how these transformations are materialized?

DNA sequence variation is abundant in modern populations, yet the relationship between the phenotypic variation and the genomic variation producing it is extremely complex.[2] The main problems here involve, as a rule, many genes in creation of a function and a great gap in our understanding of the chain of events bringing about the conversion of the genetic information into the phenotype. A great variety of interdependent evolutionary changes could act

cooperatively, sometimes within limited periods of evolution, and then cease to operate. Moreover, certain changes could accumulate in the genome first without any visible effect (dormant changes) and then suddenly manifest themselves due to a mutation in a single (or a few) "capacitor" gene (see below). Finally, an overwhelming majority of changes in the genome structure are most probably just neutral, or almost neutral, and play no role in the selection.

"A Relatively Small Number of Genetic Changes in Systems Controlling the Expression of Genes May Account for Major Organizational Differences between Human and Chimpanzees"[3]

An attempt to answer Question 1 at least partially has been undertaken in the classical work of King and Wilson[3] on comparison of the chimpanzee (*Pan troglodytes*) and human (*Homo sapiens*) gene and protein primary structures. It was demonstrated that human proteins and genes are generally 99% identical to their chimpanzee counterparts. This remarkably low difference seemed to be too small to account for the evident dissimilarity of the organisms. Moreover, King and Wilson indicated that "Since the time that the ancestor of these two species lived the chimpanzee lineage has evolved slowly relative to the human lineage, in terms of anatomy and adaptive strategy". At the same time, the rates of molecular changes in proteins and genes were rather similar in these two species, and even close to the values for anatomically highly conservative species like frogs. The following two remarkable conclusions were drawn by King and Wilson: "The contrasts between organismal and molecular evolution indicate that the two processes are to a large extent independent of one another" and that "a relatively small number of genetic changes in systems controlling the expression of genes may account for major organizational differences between human and chimpanzees". These regulatory mutations may affect either *trans*-acting regulatory proteins, such as transcription factors, participants of the signal transduction pathways etc., or *cis*-acting sequences responsible e.g., for the regulation of the gene expression at the transcriptional or posttranscriptional levels, or even, as we understand now, non-protein regulatory molecules such as noncoding RNAs (ncRNAs, see below).

The regulatory character of the changes responsible for evolutionary progress is now widely accepted and the next question concerning the mechanisms of the appearance of new regulatory proteins and/or new *cis*-acting regulatory sequences is already being discussed.

Spate of Facts and Slow Progress to Real Knowledge of the Mechanisms of Evolutionary Changes

"...our ignorance of the laws of variation is profound."
Darwin[4]

Gabriel Dover starts his seminal review[5] with a citation of the prominent evolutionist Richard Lewontin. "For many years population genetics was an immensely rich and powerful theory with virtually no suitable facts on which to operate. It was like a complex and exquisite machine, designed to process a raw material that no one had succeded in mining...."

Quite suddenly the situation has changed. The mother-lode has been tapped and facts in profusion has been poured into the hoppers of this theory machine. And from the other end has issued nothing..... The machine can not transform into a finished product the great volume of raw material that has been provided... The entire relationship between the theory and facts needs to be reconsidered". Then Dover continues "...unless and until we uncover the 'rules of transformation' that connect 'genotype space' with 'phenotype space' then we can not seriously entertain, or be satisfied with, a gene based theory of evolution. How an individual phenotype emerges and reproduces from a given unique set of genes inherited from its sexual parents is the central question of evolutionary theory: all the rest is subsidiary". Both Lewontin

and Dover describe the exact status of the modern theory of evolution. It operates now with a great number of facts concerning gene and genome structures, it tries to understand changes in the gene content and regulation in various species, and finally comes to a conclusion: 'Genome speaks biochemistry—not phenotype'.[6] This has been said right after the first complete genome sequence of a multicellular organism *C. elegans* has become available. Strong efforts of a consortium formed to inactivate all 19,099 genes of *C. elegans* led to an idea that inactivation of a great many genes of the animal yields either no obvious phenotype or early death. It was clearly understood that there is no quick and easy way to search for gene function even with this excellent model containing only 959 cells, 302 neurons and 97,000,000 bp forming about 19,000 genes.[7] Once again citing G.Dover,[5] we find that modern "Evolutionary genetics...in its current focus on DNA variation reduces phenotypes to symbols. Varying phenotypes, however, are the units of evolution, and if we want comprehensive theory of evolution we need to consider both the internal and external evolutionary forces that shape the development of phenotypes".

But however scanty is the information we have accumulated since Darwin wrote his bitter words cited in the epigraph, today we know much more than he could even dream of. We can be satisfied with the progress in understanding what is going on with the genome structure and, to a lesser extent, expression during evolution. Also, we considerably advanced with concepts on how the genomic information is transformed into phenomes at least at the very first biochemical stages of this process. And we move, though slowly, from traditional molecular biology to system-level understanding[8] and from 'one gene-one product' philosophy to modular cell biology and to genome-wide thinking. However, we do not understand phenotype and we do not understand evolution because we do not understand phenotype. We can only understand phenotype through its evolution and vice versa: Dobzhansky was right saying that "Nothing in biology makes sense except in the light of evolution". Comparison of phenotypic and genomic changes in a great variety of species is the only hope to answer Question 1.

What Is Going on with Genomes during Evolution

Over the last decade massive information has accumulated on complete structures of the genomes starting from bacteria and finishing with an almost complete sequence of the human genome (leaving aside numerous viral genomes sequenced earlier). Some of the organisms with the genomes sequenced are listed in Table 1. These genomes together with the products of their expression form what could be called an integrated genome-information space. It opens enormous opportunities for comparisons aimed to reveal common and different features of various genomes and their functional organizations. In the following sections I will try to briefly outline what is emerging from such comparisons.

The Complexity of the Organisms Does Not Correlate with the Number of Genes

An estimated total size of the human genome is 3.2 billions bp. Over half of the human DNA is occupied by repeated sequences of various types, and only 1.1% of the genome is spanned by exons, whereas 24% is introns, with 75% of the genome being intergenic DNA.

A comparison with some other sequenced genomes shown in Table 1 demonstrates that genome sizes increase with increasing complexity of organisms. It seems quite logical, though we remember the 'C paradox': the observed 40,000-fold variation in eukaryote haploid DNA content ('C value') is unrelated to organism complexity.[13] The problem is with the number of genes. The human genome is believed to contain 26,000[14] to 31,000[15] protein-encoding genes. In addition, several hundreds of genes are known to encode non-protein-coding RNAs. The number of coding genes in the human genome is thus only twice as large as in a worm *C. elegans*[7] (19,000) and approximately the same as in a plant, *Arabidopsis thaliana*[11] (about 26,000)

Table 1. Some of the sequenced genomes with their characteristics*

Organism	Genome Size (1,000 kb)	Gene Number Estimated
Homo sapiens (draft)[14, 15]	2,900 (euchromatic part)	~30 –40,000
Mus musculus[9]	2,500 (euchromatic part)	~30, 000
Fugu rubripes[10]	365	~30 – 40,000
Arabidopsis thaliana[11]	125	25,498
Drosophila melanogaster[12]	120	13,600
Caenorhabditis elegans[7]	97	19,000
Saccharomyces cerevisiae[12a]	12.1	6,034
Escherichia coli K-12[12b]	4.6	4,288

* When this chapter was finished, a complete sequence of the human genome was reported: http://www.sanger.ac.uk/Info/Press/2003/030414.shtml.

and in fish *Fugu*.[10] Moreover, the genome of such a complex organism as *Drosophila melanogaster*[12] contains even fewer genes (~13,000) than a rather primitive worm.[7] If the estimated number of human genes is more or less correct (it is still being debated), then we have a new paradox—lack of correlation between the organism complexity and the gene number. Some authors called this new paradox 'N-paradox'.[16]

It is not so simple to give the definition to the organism complexity.[17-20] A common sense-based definition was suggested, for example, by David Baltimore[17] in his reflections on the appearance of the human genome draft sequence: "Understanding what does give us our complexity—our enormous behavioral repertoire, ability to produce conscious action, remarkable physical coordination (shared with other vertebrates), precisely tuned alterations in response to external variations of the environment, learning, memory...". To roughly evaluate the complexity I will use criteria (Table 2) based on the number of cells comprising the organism, on the number of neurons forming neural network, and on the number of cell types in the organism. The latter estimate has been used, in particular, by Raff and Kaufman.[21]

Table 2 demonstrates an enormous jump in complexity between *C. elegans* and *H. sapiens*. Then arises the question of how a modest increase in number of genes creates such a jump.

The Number and Modifications of Proteins Encoded by Genes Do Not Explain Changes in Complexity

Estimates from the genes analyzed to date suggest that the average number of alternates spliced from the transcript of a single mammalian gene might be in the range of two to three or more.[14,15,23] With an estimate of about 30,000 genes, this would give us about 90,000 or more distinct proteins encoded by the human genome.[16,23,24] It was suggested that the extent of alternative splicing is higher in humans than in worm or *Drosophila*. However, this attractive explanation was recently called in question[23] and the extent of alternative splicing was shown to be likely similar in various animals, including invertebrates.

Another source of complexity can be at the protein level.[23-25] Proteins are much more complicated than nucleic acids: more than 200 different types of post-translational protein modifications are known. In addition, different proteins can be produced from one and the same gene due to alternative splicing (but see above), by varying translation start or stop sites, or by frameshifting due to which a different set of triplet codons in the mRNA is translated. All of these possibilities result in a proteome estimated to be an order of magnitude more complex than the genome.[25] Moreover, proteins respond to altered conditions by

Table 2. Evaluations of the complexity of some organisms

	E. coli	*S. cerevisiae*	*C. elegans*	*H. sapiens*
Cell number	1	1	959	10^{14}
Cell type number	1	3–4*	About 20	200**
Neuron number	0	0	302	10^{10-11}***

*Raff and Kaufmann,[21] **Alberts et al,[22] ***Koch and Laurent.[20]

translocation to different cellular compartments, by getting cleaved into pieces, and by changing their ability to bind other proteins, nucleic acids or low-molecular ligands. Important is also the ability of one protein to be involved in more than one process. Even minor alterations in the nature of protein-protein interactions, protein modifications, and localization can have dramatic effects on cellular physiology.[26] This increased complexity of proteome certainly contributes to organismal complexity but still seems to be not sufficient to be fully responsible for its enormous jump.

Non-Coding RNAs and Epigenetic Mechanisms Might Be Partially Responsible for the Jump in Complexity

Yet another source of the increased complexity might be greater usage of ncRNAs for regulation.[15,27,28] Thousands of non-identified human genes may produce ncRNAs as their ultimate products.[15] Indeed, the analysis of mouse transcriptome[29] indicated that ncRNA is a major component of the transcriptome. The ncRNAs lack translated ORFs, they are often small and not polyadenylated and, accordingly, novel ncRNAs cannot readily be found by computational gene-finding techniques or experimental sequencing of cDNA or EST libraries and need special experiments to be revealed.[16] Their importance is now widely accepted[30] and their involvement in developmental processes has been demonstrated at least in model organisms.[31]

Epigenetic differences might also contribute to the greater complexity of mammalian genomes. Wide involvement of epigenetic mechanisms in gene expression regulation is now a common knowledge. Epigenetic modifications of mammalian DNA, such as methylation, are important for genome functioning in development and in adult organisms. DNA methylation is of central importance to genomic imprinting and other aspects of epigenetic control of gene expression, and methylation patterns are largely maintained during development in somatic lineages. The mammalian genome undergoes major reprogramming of methylation patterns in the germ cells and in the early embryo. Some of the factors that are involved both in maintenance and in reprogramming, such as methyltransferases, are being identified.[32] Epigenetic mechanisms may be quite different in different species. For example, *Drosophila*, *C. elegans* and yeast were long thought not to use methylation for their genome regulation. And although methylation was quite recently detected in *Drosophila*[33] and is suggested to have some regulatory function in development, it is still not known definitely whether DNA methylation has a functional role in *Drosophila*. In any case methylation features are quite distinct in fly and known mammalian systems. Epigenetic effects are known to regulate such important effects as dosage compensation by which the expression levels of sex-linked genes are appropriately altered in one sex to offset a difference in sex-chromosome number between males and females of heterogametic species. It was shown that different species use very different mechanisms to achieve such a

compensation: the male X chromosome is hypertranscribed in drosophilid flies, both hermaphrodite X chromosomes are downregulated in the nematode *C. elegans*, and one of two X chromosomes is inactivated in mammalian females with the participation of ncRNAs. The *trans*-acting factors (proteins and ncRNAs) that have been shown to mediate dosage compensation are unrelated among the three lineages.[34]

Multiple Combinatorial Interactions Can Be a Major Source of the Complexity

Many other possibilities were also suggested to explain the molecular basis of increased complexity, but the most powerful source of the complexity is probably a combinatorial use of the repertoire of regulatory factors. A single gene product can be involved in various processes, in particular during organism development.[35] This multiple involvement can be connected with the ability of modular promoters (enhancers) to interact with various combinations of transcription factors and thus to vary cell compartments and/or time of the gene expression during development.[5] The combinations of a protein with different partners form complexes with different features: a property known at least for different combinations of transcription factors, activators and co-activators capable of interacting with different enhancers or promotor regions switching various genes on and off.[36] Moreover, depending on the particular combination, the function of these factor complexes can be dramatically changed. For example, Dorsal transcription regulator in *Drosophila* is an activator, but in specific *cis*-regulatory regions it associates with two DNA binding proteins, Cut and Dead ringer (Dri), and is converted into a repressor through the recruitment of co-repressor Groucho.[36-38] The well known key participants of various cellular functions RB[39] and p53 [40] each contain potentials for interactions with tens of partners conferring new properties on the complexes formed and taking part in a multitude of cellular functions. The list of such multifunctional proteins is constantly growing. It is believed that over 2,000 transcription factors take part in cell-specific gene regulation.[41] The combinatorial use (Fig. 1) of these *trans*-regulators is an inexhaustible source of unique combinations ensuring the correct expression of each of the genomic genes. The vertebrate immune system is one more example of a biological system capable of generating a great repertoire of specific responses by using combinations of a few hundred different genes.[16]

To conclude the discussion of the interrelation between the complexity and *number* of genes I would like to mention after Venter et al[14] a speculation made by Haldane in 1937 that if the number of genes were too large, each zygote would carry too many new deleterious mutations thus making the population not able to maintain itself. In 1967 H. Muller calculated that the mammalian genome would contain a maximum of about 30,000 genes. Muller's estimate for *D. melanogaster* was 10,000 genes, compared to 13,000 derived from the annotated fly genome. Although these calculations may be too simplified and even not quite correct, they suggest that increasing complexity does not necessarily mean proportionally increasing gene number. However, it should inevitably lead to an increase in functional load at least on regulatory genes, as predicted by some evolution theories (reviewed in ref. 35) to explain one more paradox emerged in 1970s: multicellular organisms seem to use a highly conserved regulatory machinery for their functioning, while having different levels of complexity.

However, it should be stressed that all the speculations have important gaps. The genome may contain many additional, small genes expressed at relatively low levels which escape the detection by modern techniques.[9]

Genes Common and Different among Various Species

In the previous paragraph I discussed the differences in the total gene number for various species. The next question is whether different species have mainly the same sets of genes just amplified in some of them, but not in the others. The scientific evidence accumulated

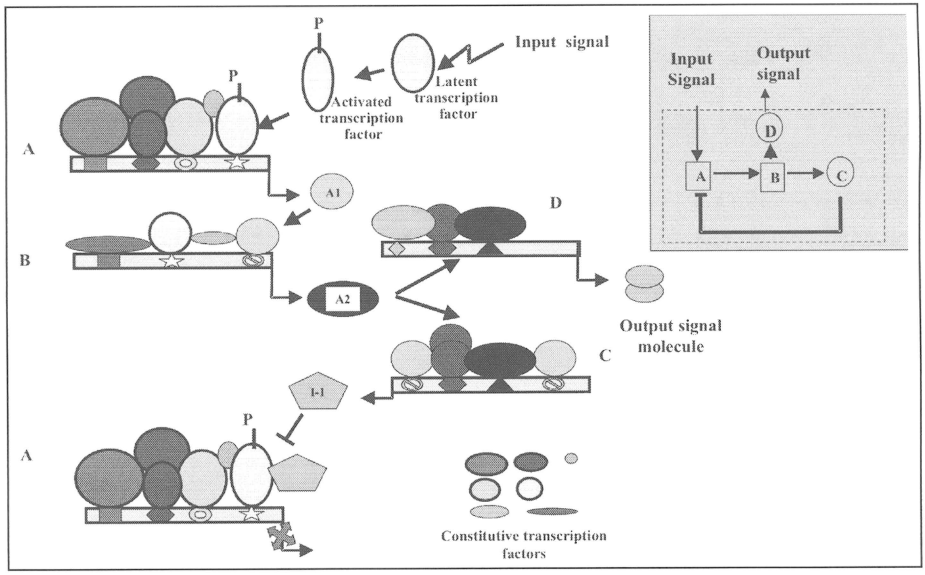

Figure 1. Combinatorial interactions of factors with their targets within cis--regulatory elements. A schematized functional module involves: (i) element A receiving an external signal and switching on synthesis of a transcriptional activator A1; (ii) element B accepting the A1 activator and switching on the next activator, A2; (iii) element C which responds to A2 activator and activates synthesis of I-1 Inhibitor which in turn inhibits the expression of element A at some step; (iv) element D which, in response to A2, directs synthesis of an output signal molecule connecting this functional module with another one. The constitutive transcription factors available in the cells participating in the regulatory events are shown as circles or ovals. The gene regulatory network representing this functional module as a formalized scheme is shown in grey box in the right upper corner.

over the past quarter of the century suggests that many essential mechanisms underlying similar functions of various organisms, from bacteria and yeasts to *Drosophila* and man, are highly conservative. In a review devoted to the problem of the molecular mechanisms underlying the evolution of greater biological complexity, Duboule and Wilkins[35] refer to the famous seminal article "Evolution and tinkering" by F. Jacob:[42] "What distinguishes a butterfly from a lion,... or a worm from a whale is much less a difference in chemical constituents than in the organization and distribution of these constituents". The authors raise a very important problem of to what extent phenotype diversity depends on new inventions of evolution and to what extent on the process of tinkering—reiterative usage of the same genes in different contexts. They left the final answer to this important question to the next 5-10 years, "...as comparative genomics broadens our knowledge about gene functions in different organisms". Five years have passed since the review was published in 1998, a great amount of structural information has been accumulated, and, though functional data has lagged far behind, one can try to look at the problem with this newly acquired knowledge. The most straightforward approach would be, of course, to compare total ensembles of proteins (proteomes) specified by various genomes.

Certainly, such an analysis will be incomplete due to difficulties in gene finding and functional classification of the proteins deduced from genomic sequences. Furthermore, as far as functional predictions are based on similarity to protein sequences with known functions, only

basic biochemical functions can be assigned (rather than higher order cellular processes).[a] Besides, in such a way only about 60% of all deduced proteins can be assigned to certain broad functional categories, like "cytoskeletal structural protein", "ion channel", "transcription factor", or "cell adhesion".[14,15] In addition, a very serious problem is the identification of the "orthologs" for each of the human genes in other organisms under comparison. Orthologs, by definition, are the genes that appeared in the organisms by descent from a common ancestor. To analyze the evolutionary events, it is critical to separate orthologs from paralogs, that is the genes formed in a given species by duplication events in more than one copy (e.g., see refs. 43, 44). As a criterion, both human genome sequencing groups[14,15] used the highest sequence identity levels between pairs of cognate proteins in organisms under comparison. Since it is not a perfect criterion[b] of orthology, some of the derived evolutionary conclusions are probably not quite correct. However, despite all the limitations of such analyses, they allow one to draw some general conclusions and to get a deeper insight into the functional commonalties and diversity among eukaryotes.

Below I will confine myself to the human genome analysis of Venter et al,[14] but that of the Public consortium[15] is rather similar. In total, 2758 strict human-fly and 2031 human-worm orthologs were identified. 1523 were common to both sets, and they were defined as an evolutionarily conserved set of human proteins. Not surprisingly, the most basic cellular functions such as basic metabolism, transcription, translation, and DNA replication remained conservative since the divergence of single-celled yeast and bacteria. 60% of predicted human proteins display some sequence similarity to proteins from other species with sequenced genomes. About 40% of the human proteins show similarity with fruitfly or worm proteins. And 61% of fruitfly proteins, 43% of worm proteins and 46% of yeast proteins have sequence similarity to predicted human proteins.

The recently sequenced mouse genome demonstrated that approximately 99% of mouse genes had homologs in the human genome.[9] It is still unclear whether the remaining 1% predicted mouse genes do have no human homologs or this is just due to the unsequenced part of the human genome.

Very informative is the comparison of not the whole proteins but their discrete structural units, called domains (modules). The number of protein domains was estimated to be about 1,500-2,000, that by combining in different fashions can form tens of thousands of modular proteins. It appeared that only a relatively small proportion of various known protein domains have been invented in the vertebrate lineage, and that most domains trace at least as far back as a common animal ancestor: of 1,262 investigated domain and protein families only 94 (7%) representing 24 domain families and 70 protein families were 'vertebrate-specific'. The 94 vertebrate-specific families include defence and immunity proteins and proteins that function in the nervous system all of them contributing to important physiological differences between vertebrates and other eukaryotes. They emerged recently in evolution and/or were subject to rapid divergence.[15]

---

[a] It is important to differentiate what type of functions is under question. The term "function" is uncertain in itself. Just biochemical function like "protein kinase", "protein phosphatase" does not tell us much of the real functional role of an individual protein in the cell or in the organism: the same biochemical function can be used at different time during development, in different cellular compartments and in different signal transduction pathways (for review see ref. 45).

[b] See comment in ref. 46. Here a sober look at the problem of orthology identification is presented: this is an extremely difficult problem that can not be solved simply based on the highest sequence homology between two genes in two or more species. Therefore, all conclusions made using this criterion should be taken with caution.