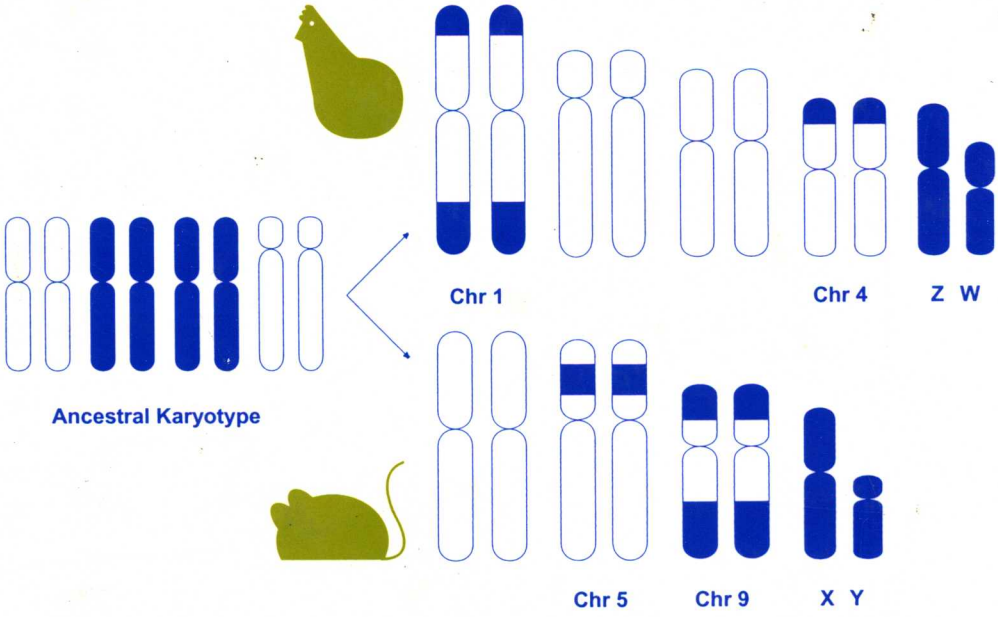
Comparative Genomics



Edited By

Melody Clark

Kluwer Academic Publishers

COMPARATIVE GENOMICS

edited by

Melody S. Clark

HGMP Resource Centre
United Kingdom

KLUWER ACADEMIC PUBLISHERS
Boston / Dordrecht / London

Distributors for North, Central and South America:
Kluwer Academic Publishers
101 Philip Drive
Assinippi Park
Norwell, Massachusetts 02061 USA
Telephone (781) 871-6600
Fax (781) 681-9045
E-Mail <kluwer@wkap.com>

Distributors for all other countries:
Kluwer Academic Publishers Group
Distribution Centre
Post Office Box 322
3300 AH Dordrecht, THE NETHERLANDS
Telephone 31 78 6392 392
Fax 31 78 6546 474
E-Mail <services@wkap.nl>


Electronic Services <http://www.wkap.nl>

Library of Congress Cataloging-in-Publication Data

Comparative genomics / edited by Melody S. Clark.
p. cm.
Includes index.
ISBN 0-412-83080-9 (alk. paper)

QH447 . C65 2000
572.8'6—dc21

00-058730

Printed on acid-free paper.

Printed in the United States of America

COMPARATIVE GENOMICS

Contributors

Clark, Melody S. Fugu Genomics HGMP Resource Centre, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SB.
email : mclark@hgmp.mrc.ac.uk.

Elgar, Greg. Fugu Genomics HGMP Resource Centre, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SB.
email: gelgar@hgmp.mrc.ac.uk

Fröniecke, Lutz. National Institute of Health, National Cancer Institute, Basic Science Laboratory, Frederick, MD 21702. USA.
email: froenickel@ncifcrf.gov

Hertzog, Paul. Centre for Functional Genomics and Human Disease. Institute of Reproduction and Development, Monash University. 27-31 Wright Street, Clayton, Vic. 3168. Australia
email: paul.hertzog@med.monash.edu.au

Jäckle, Herbert. Institut für biophysikalische Chemie, Abteilung Molekulare Entwicklungsbiologie, Am Fassberg 11, D-37077 Göttingen, Germany.
email: hjaeckl@gwdg.de

Jeffrey, William R. Department of Biology, University of Maryland, College Park, MD 20742-4415. USA.
email: wj33@umail.umd.edu

Kola, Ismail. 7245-24-110, Pharmacia and Upjohn, 301 Henrietta Street, Kalamazoo, MI 49007. USA
email: Ismail.kola@am.pnu.com

Lazner, Francesca. Centre for Functional Genomics and Human Disease. Institute of Reproduction and Development, Monash University. 27-31 Wright Street, Clayton, Vic. 3168. Australia
email: Francesca.cristiano@med.monash.edu.au

Lipkin, Ehud. Department of Genetics, The Hebrew University of Jerusalem, 91904 Jerusalem. Israel.
email: lipkin@vms.huji.ac.il

Marshall Graves, Jennifer. A. Department of Genetics and Evolution, La Troube University, Melbourne, Victoria 3083, Australia.
email: genjmg@plasmid.gen.latrobe.edu.au

Schäfer, Ulrich. Institut für biophysikalische Chemie, Abteilung Molekulare Entwicklungsbiologie, Am Fassberg 11, D-37077 Göttingen, Germany.
email: uschaef@gwdg.de

Shetty, Swathi. Department of Genetics and Evolution, La Troube University, Melbourne, Victoria 3083. Australia.
email: swathi@gen.latroube.edu.au

Soller, Morris. Department of Genetics, The Hebrew University of Jerusalem, 91904 Jerusalem. Israel.
email: soller@vms.huji.ac.il

Stanyon, Roscoe. National Institute of Health, National Cancer Institute, Basic Science Laboratory, Frederick, MD 21702. USA.
email: stanyonr@ncifcrf.gov

Wienberg, Johannes. National Institute of Health, National Cancer Institute, Basic Science Laboratory, Frederick, MD 21702. USA.
email: wienbergj@ncifcrf.gov

Wilson, Trevor. Centre for Functional Genomics and Human Disease. Institute of Reproduction and Development. Monash University. 27-31 Wright Street, Clayton, Vic. 3168. Australia.
email: Trevor.Wilson@med.monash.edu.au

Contents

Contributors.....vii

1. **Comparative genomics: An introduction: sequencing projects and model organisms.**
Melody S. Clark.....1

2. **Drosophila melanogaster: A genetic tool**
Ulrich Schäfer and Herbert Jäckle.....23

3. **Tunicates: Models for chordate evolution and development at low genomic complexity.**
William R. Jeffrey.....43

4. **Fugu rubripes: A fish model genome**
Melody S. Clark and Greg Elgar.....71

5. **The mouse and the genomic era**
Trevor J. Wilson, Francesca Lazner, Ismail Kola and Paul J. Hertzog.....97

6. **Quantitative Trait Loci in domestic animals – Complex inheritance patterns.**
Ehud Lipkin and Morris Soller.....123

7. **Comparative genomics of vertebrates and the evolution of sex chromosomes.**
Jennifer A. Marshall Graves and Swathi Shetty.....153

8. **Insights into mammalian genome organization evolution by molecular cytogenetics.**
Johannes Wienberg, Lutz Fröniecke and Roscoe Stanyon.....207

Index.....245

1 COMPARATIVE GENOMICS: AN INTRODUCTION: SEQUENCING PROJECTS AND MODEL ORGANISMS

Melody S. Clark, Fugu Genomics, HGMP Resource Centre, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SB, UK.

INTRODUCTION: WHAT IS COMPARATIVE GENOMICS?

At its most literal the term means comparing genomes. This immediately brings to mind DNA and protein sequences and inevitably comparison with the human genome. However, Comparative genomics is more than that. It applies to the comparison of any organism at a variety of levels: DNA or protein sequences, mapping positions and maps, function and evolution. The aim is to decipher how genes function and provide an understanding of the link between genotype and phenotype. Often this is with particular reference to a set of heritable characters or disease, as these are clearly more attractive funding possibilities (even more so when human studies enter into the experimental equation). With livestock, such as cattle, sheep, pigs, fish etc, which are of great economic importance to any country, there are clear commercial requirements to being able to understand the inheritance patterns of advantageous characters and also disease. However, any commercial applications are underpinned by a vast array of academic or "basic" research.

When embarking on a research project, it is not always possible to decide categorically which organism to study and which set of genes or heritable characteristics within that organism. Not all organisms are amenable to experimentation, humans being the classic example! This is where "Model Organisms" enter into the subject. The term is self explanatory and an increasing number of different species are being used as tools in our attempts to understand how genes function and the interplay of complex factors such as control sequences, immediate gene environment, the importance of non-coding elements (repeat sequences, retroelements etc.) and the macroenvironment surrounding the organism itself. For example; transgenics can be performed in mouse; mutation studies in yeast, *C. elegans, Drosophila*, zebrafish; analysis of quantitative traits in livestock, identification of evolutionary conserved control elements in *Fugu* and global

comparisons of genome rearrangements in any number of species; the list is endless. As the worldwide sequencing capacity increases and high throughput functional assays are developed, the comparative approach will prove increasingly important, in terms of both sequence comparison and the use of biological models of function.

This chapter is intended as an introduction to the subject of comparative genomics. The aim is to give a brief overview of the subject, concentrating on some areas, such as the genome sequencing projects and the varied utility of model organisms, which can be used to help decipher gene function and evolution. The range of the subject matter and approaches in following chapters in this book is diverse, reflecting the wide variety of ways in which this subject is tackled. Comparative genomics will not only tell us much about how human genes function, but also the genotype-phenotype link in many other organisms and the process of evolution.

So why are Comparative Genomics and model organisms so important, when, by the time this book is published, the completion of the human draft sequence will have been announced, with total sequence available (no gaps) by 2003?

THE HUMAN SEQUENCE AND COMPARATIVE GENOMICS IN HISTORICAL CONTEXT

Today's society is very "instant" and scientific breakthroughs are often represented in the media as though they happened overnight, when, in actual fact, it is probably fair to say that this is never the case. The same is true of the human genome sequence. Whilst the sequence data itself is the result of a worldwide collaborative effort involving highly specialised laboratories, the scientific understanding, which enabled this to happen, has its basis over a hundred years ago and in many different scientific disciplines.

It is difficult to determine an exact start point, but perhaps the original definition of cell theory in 1830 by Theodor Schwann and Matthius Scleiden (physiologist and botanist respectively) is appropriate. Thirty-nine years later, a chemist, Friedrich Miescher analysed cell extracts and demonstrated that they contained protein and an unusual phosphorus-containing compound, which he called nuclein (or nucleic acid, as this is now known to be). The first person to have described chromosomes is said to have been Flemming (1843-1905) whilst working on salamander and Mendel, the acknowledged "Father of Genetics" was a plant breeder in his spare time as an Abbot. The early 1900's saw an explosion in the field of genetics, heralded by the re-discovery and confirmation of Mendel's work. Much of this confirmatory work was carried out on plants, grasshoppers and sea urchins, organisms that were readily available and amenable to manipulation (premier characteristics of model organisms). So, although the concept of the "model organism" and comparative genomics was still along way off, the application of these particular aspects of genetics was very much in evidence from the beginning. The grandfather of all model organisms: *Drosophila*, makes a major appearance in 1911, with the publication of the first linkage map containing five genes. T.H. Morgan

could not have possibly realised that his work provided the basis for the *Drosophila* genome sequencing project which was completed eighty-nine years later.

Surprisingly, the term "Genome" is not a modern one, having been developed in 1920 by Winkler. It could not have had exactly the same meaning as ascribed to it today, due to the fact that it was not discovered until 1944 that DNA coded for the genetic material of the cell. Up until then it was firmly believed that the genetic material must be protein, as protein is chemically complex; nucleic acid is simple; genes are complex therefore genes must be made from protein! It is incredible, the speed that genetics has moved at in the latter part of the 20th Century. A strange quirk in this is that the landmark discovery of the double helix structure of DNA by Watson and Crick in 1953 actually pre-empted the definitive proof that normal humans have a chromosome complement of 46 (finally determined in 1956 by Tjio and Levan, after years of debate ranging from numbers of 16 to 40). 1977 saw the cloning of the first human gene whilst twenty-two years later; in 1999 the first human chromosome was completely sequenced (Dunham et al, 1999).

It is probably fair to say that genetics has become an increasingly specialised science over the past fifty years and this is particularly true of molecular biology. However, now that so much sequence data is available, the emphasis will shift to determining function, which will require a far more multidisciplinary approach and a wider appreciation of "biology". Model organisms and comparative genomics can contribute significantly, as exemplified by the wide range of approaches described in each of these book chapters.

THE HUMAN GENOME SEQUENCE

The sequencing of the human genome is a fantastic scientific feat and represents the culmination of years of work by hundreds of laboratories round the world. In the final stages, the race to complete the human genome between the private company Celera and the publicly funded bodies lead by the NIH (National Institute of Health) in the US and the Wellcome Trust in the UK has generated enormous amounts of publicity and put genetics into the spotlight. Whilst great claims (usually by the media) are made about what the availability of the human sequence will mean to the average person and the scare stories of insurance implications abound, for science, the human sequence will be a tremendous resource. It is the first vertebrate total genomic sequence available; it is publicly accessible and will provide a reference genome for comparative studies. It is just the start, other genomes will follow, of different organisms and different ethnic human groups, telling us much about the importance of gene order and content between species and polymorphism and its implications both within and between species. Evolution and population genetics, areas of biology, which became slightly unfashionable for a while, are back in the limelight.

DOES SEQUENCE EQUAL FUNCTION?

"unknown" genes

In the popular press I have seen the human genome sequence referred to as similar to trying to read either thirty-two volumes of the Encyclopaedia Britannica or the bible without any paragraphs, headings or punctuation present. A somewhat difficult task! This is not entirely true. Gene prediction programmes have been developed which are organism-specific and can identify putative exons and/or genes with high efficiencies (Claverie, 1997; Burset and Guigo, 1996). A quick glance at the databases will reveal that many of the *C. elegans* genes are annotated by cosmid ID and are therefore "putative genes". Some of these genes exhibit sequence similarity to other characterised genes in the databases and therefore can be ascribed a "putative" function or assigned to a gene family. This is a start, but after this how much remains unknown? (see Table 1)

Table 1. Genome sizes of some completed genomes with predicted number of genes (ORFs: Open Reading Frames) and percentage of genes with no known match in the databases.

ORGANISM	GENOME SIZE (Mb)	PREDICTED ORFS	UNKNOWN GENES %	
Mycoplasma genitalium	0.58	470	20	Fraser et al, 1995
Haemophilus influenzae	1.83	1,743	40	Fleischmann et al, 1995
Escherichia coli	4.63	4,288	38	Blattner et al, 1997
Saccharomyes cerevisiae	12.1	6,034	25	Botstein et al, 1997
Caenorhabditis elegans	97	19,099	24	C. elegans sequencing consortium, 1998
Drosophila melanogaster	120*	13,600	23	Adams et al, 2000

*This refers to the sequenced euchromatic part of the genome and does not include the additional 60Mb of heterochromatic DNA present.

This table simplifies the situation. It certainly appears on first glance that there are more "unknown" genes in the prokaryotes *E.coli* and *H.influenzae* compared to yeast, *C. elegans* and *Drosophila*. However, it should be noted that these figures were taken from when the genomes were first published and therefore

the functional assignments date back to that time when the databases were considerably smaller and functional assays were only just starting.

Closer examination of the eukaryotic data reveals a more detailed picture. When the yeast genome sequence was published 60% of its genes had no experimentally determined function. However, of these, the majority showed some sequence similarity or motif suggesting possible functions, leaving approximately 25% with no clue whatever (Botstein et al, 1997), hence the 25% entered into the table. With *C. elegans* 42% of the predicted genes had cross-phylum matches, most of which had putative functional information. A further 34% matched only other nematode sequences (C.elegans sequencing consortium, 1998) i.e. probably *C. elegans* or *C. briggsae* cDNAs, few of which had been functionally characterised, so a more appropriate figure for "unknown" genes would be 58%. As regards *Drosophila*, 23% of predicted genes had no known database match and a further 27% were only matched against ESTs (Adams et al, 2000), many of which are not well annotated. So again, a revised figure of 50% is probably more accurate when discussing known functional data. The situation of the human sequence will be similar to that of the other eukaryotes and much work will be required to ascribe function to putative genes.

Alternative splicing

Identifying genes is only the first part of the long path towards determining function. Computer programmes can predict genes and confirmation is usually either via database searching against EST databases or screening cDNA libraries. An EST sequence match confirms that a "putative" gene is "real". These EST sequences usually only represent incomplete single pass sequencing of a cDNA clone. Obviously, further confirmation of structure can be obtained by sequencing the whole clone, but here another factor enters into the equation: that of alternative splicing.

The current data on the *Drosophila* sequence predicts 13,601 genes, which is considerably less than the 19,099 predicted for *C. elegans*. However, current cDNA data indicates that although there are only 13,601 genes, these encode at least 14,113 transcripts through alternative splicing and the number of transcripts is considered a substantial underestimate (Adams et al, 2000). It is not just *Drosophila* where this phenomenon occurs. For example; the WT1 gene, which is involved in mammalian genitourinary development, encodes sixteen different protein isoforms in human (Hastie, 1994). So far, it has been estimated that over 30% of human genes are affected by alternative splicing (Hanke et al, 1999; Mironov et al, 1999), this is further complicated by the possibilities for post-translational modification, for which figures are not yet available (Bork, 2000). The PTHrP gene (which is described more fully in the chapter on *Fugu*) generates three isoforms in human (Yasuda et al, 1989a; Mangin et al, 1989), but only one in other mammals and *Fugu* (Mangin et al, 1990; Yasuda et al, 1989b; Thiede and Routledge 1990, Power et al, 2000). The role isoform generation plays in function and evolution, is only just beginning to be explored. This is partly as a consequence of the worldwide increased sequencing

capacity and the popularity of EST sequencing projects. The question of post-translational modification will increasingly come to the fore with protein functional studies.

So the answer to the question "Does sequence equal function?" is clearly "No". There are still many gaps in our knowledge with regard to gene function, even with the sequence of complete genomes and anecdotal evidence from others (Bork et al, 1998). Our current ability to assign function relies heavily on database annotation, and computer prediction programmes. This is particularly true with regard to routine annotation of mass sequencing data. Bork (2000) estimates feature annotation of sequences to be 70% accurate. The main problem is that the gap between the amount of sequence data available and experimental characterisation of proteins is widening. Sequence data can only reveal a certain amount; more effort is required on protein characterisation and experimentation.

THE NON-CODING PORTION OF THE GENOME

One of the great advantages of the human genome sequence is that it will include all the non-coding sequences. The majority of sequence data in the databases for most organisms is in the form of cDNAs, the expressed part of the genome. Whilst the coding sequence is important, the control elements are found in the non-coding portion of the genome. Gene prediction programmes are well advanced, but programmes to decipher control and promoter regions and 5' and 3' untranslated regions (UTRs) are still in their infancy (Fickett and Hatzigeorgiou, 1997).

Subtracting the amount of DNA required for gene control from the genome still leaves the majority with no ascribed function and to date, very little is known. Although repeat elements have been endlessly classified, the numbers of them which have been determined as having functional implications are few. Unstable triplet repeats are associated with several genetic diseases such as Huntington's disease and Myotonic muscular dystrophy (Caskey et al, 1992). Many repeat elements are known to be of retroviral origin. Some of these retroelements have been implicated in genome evolution and genome plasticity (Pickeral et al, 2000). They have probably been most intensively studied in connection with the Major Histocompatibility Complex (MHC) region where it is thought that via their ability to cause gene rearrangements, they have played a significant role in its the evolution (Abdulla et al, 1996; Kulski et al, 1997; Dawkins et al, 1999).

The C-value paradox

Many students labour under the misapprehension that the more complex the organism, the higher the DNA content of the nucleus. Whilst this holds for the current sequencing projects, there are still many more organisms out there with massively divergent (and massive) genomes (see Table 2). Why does a particular species of lily have fifteen times more DNA than a human? What is the significance

of all this "extra" DNA? This is termed "the C-value paradox". Only now are we approaching the stage where we may be able to start answering this conundrum.

Table 2. DNA content and haploid chromosome number in a variety of eukaryotes. Adapted from Clark and Wall (1996).

SPECIES	COMMON NAME	1C NUCLEAR DNA CONTENT (pg)	n
Fritillaria davisii	Lily species	98.4	12
Protopterus	Lungfish	50	19
Avena sativa	Oat	21.5	21
Triticum aestivum	Bread wheat	18.1	21
Allium cepa	Onion	16.8	8
Homo sapiens	Human	3.7	23
Mus musculus	Mouse	2.5	10
Drosophila	Fruit fly	0.1	4
Arabidopsis thaliana	Mouse ear cress	0.07	5
Saccharomyces cerevisiae	Yeast	0.026	15

The availability of several complete reference genomes will allow research to expand into this previously neglected field of non-coding (or "junk") DNA. It provides the potential to answer the question of what does the rest of the DNA really do?

GENOME SEQUENCING PROJECTS

The human genome sequencing project is, quite rightly, currently enjoying the spotlight of media attention; it is an amazing achievement. However, the technology which enabled this was developed on less complex genomes. It appears at first glance that genome sequencing is becoming more routine: Genomes On Line Database (GOLD 1.0) (Kyrpides, 1999) (http://igweb.integratedgenomics.com/GOLD/) lists all completed and ongoing genome projects. As at 09/03/00, there were 25 complete genomes in the databases with 106 prokaryotic and 31 eukaryotic ongoing. This explosion in genome sequencing, particularly of eukaryotes is a reflection of the success of whole genome shotgun sequencing, which was first reported for *Haemophilus influenzae* (Fleischmann et al, 1995) and now is being tried on more complex organisms (Adams et al, 2000).

Prokaryotic sequencing projects

The first microbial genome to be sequenced was the 5386bp of bacteriophage ΦX174 (Sanger et al, 1978). Amazingly, this was only one year after Sanger's paper on dideoxy-sequencing methodology was published. At this point in time, only the relatively small viral genomes presented the opportunity of sequencing with the technology available at the time (manual radioactive sequencing). It was not until 1995 that the first bacterial genome, the 1.83Mb *Haemophilus influenzae* strain Rd was completed (Fleischmann et al, 1995) using the technique of whole genome shotgun sequencing. This is very effective for small genomes which are compact, gene rich, do not contain introns and large amounts of repeat DNA and therefore can be assembled without the requirement for a detailed genetic map or a complex scaffold provided by large insert libraries.

The advent of such a mass of sequence data has changed the approach and focus of research on microorganisms. The sequencing of *Mycoplasma genitalium* (the smallest genome recorded (580kb) so far of any free living organism) has enabled the definition of the minimum gene set required for a self-replicating cell (Fraser et al, 1995). The relative ease of sequencing these "small" genomes has meant that comparative studies are in advance of eukaryotes (Perrière et al, 2000). In-depth knowledge of these organisms has many commercial and medical applications. The evaluation of the process of prokayote evolution and phylogenetic relationships can be used as a tool to determine the spectrum of a drug target (Allsop, 1998). Comparative studies can shed light on the molecular mechanisms of pathogenesis.: identify the functions of individual genes and determine how genes interact to form complex traits such as virulence (Field et al, 1999). One of the aims of the prokaryotic sequencing projects is to compare the gene set of an infectious strain with an attenuated lab strain to examine factors for virulence and host specificity (Saunders and Moxon, 1998). With the availability of the human genome, it is possible to evaluate the pathogen within the genetic context of the host (Field et al, 2000). Understanding of host-microbe interactions is also important for diseases of livestock, which have huge economic implications. Commercially, the spin off of this increased knowledge should be more precise drug targeting and new vaccine development (Allsop, 1998)

Biochemically and genetically, *E.coli* and many other microorganisms have been studied for over 50 years. There are essential biochemical pathways common to all organisms and much of the understanding of these was carried out on bacterial genomes. Whilst there are many differences between genes and gene structure of prokaryotes and eukaryotes, comparative analysis even between human and *E.coli* can still provide information on gene function. Although the number of orthologous genes between vertebrates and *E.coli* is low, individual protein domains are conserved. This is in line with the theory the explosion in genes associated with the metazoan radiation and the construction of multidomain extracellular and cell surface proteins, essential requirements for the evolution of multicellular organisms, was facilitated by exon (or domain) shuffling (Patthy, 1999). This domain conservation has allowed insights into poorly characterised vertebrate genes. Many positionally cloned genes encode large multidomain proteins, some of which contain putative

enyzmatic domains of unknown function. Motif detection and structural modelling using bacterial genes (Mushegian et al, 1997) has revealed putative functional sites that previously escaped detection with standard approaches. Three domains with homology to a nuclease, a 3'-5' proof-reading exonuclease and a helicase were identified in Werner Syndrome (a disease associated with features of premature aging) indicating that the protein may be involved in DNA repair and processing (Mushegian et al, 1997). This provides an entry point into dissecting the exact molecular nature of the human disease.

These relatively simple organisms have much to contribute to our understanding of genetics and evolution. The genome sequencing projects and the subsequent analyses promise much in the field of health care and preventive medicine.

Ongoing eukaryotic sequencing projects

Of the ongoing eukaryotic sequencing projects, only two (mouse and human) are vertebrates. One of the others, *Drosophila*, has been completed during the process of compiling this book (Adams et al, 2000) and will be discussed in greater detail in chapter 2; the rest are a mixture of protozoa (*Cryptosporidium parvum*, *Giardia lamblia*, *Leishmania major* etc.), fungi (*Pneumocystis carinii, Neurospora crassa* etc.) and plants (*Arabidopsis thaliana*, rice, maize etc.) Genomes On Line Database (GOLD 1.0) (Kyrpides, 1999) (http://igweb.integratedgenomics.com/GOLD/). The reasons behind the protozoa and fungi sequencing projects are similar to the prokaryotes; understanding pathogenesis and disease control. The plants are of great economic importance.

Plant genomics

In many ways, plants too, mirror the prokaryotic sequencing projects, in that they tend to be viewed as a separate field with no overlap to the more prestigious (in some circles) vertebrate projects. However, plant genomics has much to offer and no overview of Comparative Genomics would be complete without them. The sequencing focus of plant genomics is *Arabidopsis thaliana* with its minimal genome of 120Mb. Crop plants typically have complex genomes that can be substantially larger than the human genome, the haploid content of barley, for example is 5300Mb. Several mechanisms have contributed to the expansion in genome size in some plants such as genome duplications (wheat is hexaploid) and expansion of repeat elements, in which retroelements play a large role (Bennetzen and Kellogg, 1997).

Progress in *Arabidopsis* sequencing so far is 54.8Mb completed across all the five chromosomes with 15.1Mb in the finishing stages. Detailed analysis of the complete sequence from chromosome 4 revealed that, similar to other sequenced genomes, only 60% of the genes of *Arabidopsis* have established functions (Bevan et al, 1999). So the gap between sequence generation and functional understanding holds true for plants too. One problem that has arisen with the *Arabidopsis* project is

a biological one. *Arabidopsis* is a dicotyledonous plant, and the vast majority of crops are monocotyledonous (wheat, barley, rice, maize, sorghum, oats and sugarcane etc.). Therefore direct transfer of technology using information relying on factors such as colinearity of genes is often approached via rice, which is one of the smaller monocot genomes (440Mb). In fact rice has become the subject of a major sequencing effort based in Japan (Sasaki et al, 1996) and is proposed as a good second model genome for plants.

Plants are perceived as being very different from animals, but the data from plant genome sequencing projects, like that of the prokaryotes, can contribute to our overall knowledge of gene function. The inclusion of plant genes in databases assembled of orthologous gene clusters will help identify gene function based on conserved motifs and draw in knowledge of gene function from diverse organisms. This will add a new range of plant-specific biological functions to the process of determining gene function in other organisms (Bevan and Murphy, 1999).

Plant genetics has always thrived by research on a vast array of species, the basic biology, evolution, adaptation, genome research etc. on these gives added value to those few chosen for in-depth genomic sequencing. It is clearly apparent, reviewing the current literature that all the questions being posed by animal comparative geneticists, such as uniformity of gene density, genome duplications, synteny, conserved gene order, assignment of orthology etc. are also under scrutiny by the plant geneticists. The two areas are clearly not so different and it will be interesting to see how both develop.

Indeed, in his review of plant genomics, Bennetz (1999) proposes a plan for plant comparative genomics, which animal/vertebrate geneticists would do well to consider. He suggests the initial genomic sequencing of two plants species *Arabidopsis* and rice, which would serve as reference genomes and the foundation for gene discovery and characterisation in all plants. Physical maps of a few species would be constructed and he termed these "nodal" species, chosen because they have relatively small genomes and could serve as surrogates for important and phylogenetically diverse plant families. He uses the examples of sorghum for maize and lotus for soybean. A larger number of species would be subject to medium deep (circa 50,00 clones) EST projects, as this approach is the most economical route for gene discovery and investigation into allelic diversity. In addition, these ESTs would provide the species-specific sequences needed for precise DNA chip analysis of gene expression.

Not all plant and animal species can be sequenced, but surely the most economic route is via the total genomic sequencing of a few species, with high density maps and EST projects for the others of either economic importance or those which occupy pivotal positions in evolution.

Completed eukaryotic sequencing projects: *Saccharomyces cerevisiae and Caenorhabditis elegans*

These two eukaryotic organisms both have completely sequenced genomes (yeast finished in 1996 and *C.elegans* in 1998) and were instrumental in developing the