MOLECULAR MEDICAL PARASITOLOGY

Edited by

J. Joseph Marr • Timothy W. Nilsen

Richard W. Komuniecki

Molecular Medical Parasitology

This Page Intentionally Left Blank

Molecular Medical Parasitology

Edited by

J. Joseph Marr, M.D.
President, BioMed., Estes Park, Colorado 80517, USA

Timothy W. Nilsen, Ph.D.
Case Western Reserve University School of Medicine,
Cleveland, Ohio 43606, USA

Richard W. Komuniecki, Ph.D.
Department of Biology, University of Toledo,
Toledo, Ohio 43606, USA

ACADEMIC PRESS
An imprint of Elsevier Science
Amsterdam • Boston • London • New York • Oxford • Paris
San Diego • San Francisco • Singapore • Sydney • Tokyo

This book is printed on acid-free paper.

Academic Press
An Imprint of Elsevier Science
84 Theobald's Road, London WC1X 8RR, UK
http://www.academicpress.com

Academic Press
An Imprint of Elsevier Science
525 B Street, Suite 1900, San Diego, California 92101-4495, USA
http://www.academicpress.com

ISBN 0–12–473346–8

Library of Congress Catalog Number: 2002111022

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

Cover figure: Adapted from *The International Journal for Parasitology*, Vol. 31, Issue 12 (October 2001), 1343–1353, Figure 1A, courtesy of Dr K. A. Joiner and The Australian Society for Parasitology Inc.

Typeset by Charon Tec Pvt. Ltd, Chennai, India
Printed and bound in Great Britain by MPG Books, Bodmin, Cornwall
02 03 04 05 06 07 MP 9 8 7 6 5 4 3 2 1

Contents

Colour plates appear between pages 252 and 253

List of contributors

Mark Blaxter, Ph.D.
Institute of Cell, Animal and Population Biology
Darwin Building
King’s Buildings
University of Edinburgh
West Mains Road
Edinburgh EH9 3JT
UK

Stephen M. Beverley, Ph.D.
Professor and Chairman
760B McDonnell Science Building
Box 8230
Department of Molecular Microbiology
Washington University School of Medicine
660 South Euclid Avenue
St. Louis, MO 63110-1093
USA

Nicola S. Carter, Ph.D.
Assistant Professor
Department of Biochemistry and Molecular Biology
Oregon Health and Sciences University
3181 S. W. Sam Jackson Park Road
Portland, OR 97239
USA

Juan José Cazzulo, Ph.D.
Instituto de Investigaciones Biotecnológicas
Universidad Nacional de General San Martín
Av. General Paz y Albarellos
INTI, Edificio 24, Casilla de Correo 30
(1650) San Martín, Provincia de Buenos Aires
Argentina

Debopam Chakrabarti, Ph.D.
Associate Professor
Department of Molecular Biology &
Microbiology, University of Central Florida
12722 Research Parkway
Orlando, Fl 32826
USA

Christine Clayton, B.A., Ph.D.
Zentrum für Molekulare Biologie
Im Neuenheimer Feld 282
D-69120 Heidelberg
Germany

Michael J. Crawford, Ph.D.
Department of Biology
University of Pennsylvania
415 South University Avenue
Philadelphia, PA 19104-6018
USA

George A.M. Cross, Ph.D.
Andre and Bella Meyer Professor
Laboratory of Molecular Parasitology
(Box 185)
The Rockefeller University
1230 York Avenue
New York, NY 10021-6399
USA

Tim Day, Ph.D.
Assistant Professor
Biomedical Sciences
College of Veterinary Medicine
Christensen Drive
Iowa State University
Ames, IA 50011
USA

Martin J. Fraunholz, Ph.D.
Department of Biology
University of Pennsylvania
415 South University Avenue
Philadelphia, PA 19104-6018
USA

Timothy G. Geary, Ph.D.
Discovery Research
Pharmacia Corp.
301 Henrietta St.
Mailstop 7923-25-423
Kalamazoo, MI 49006
USA

Jonatha M. Gott, Ph.D.
Center for RNA Molecular Biology
Case Western Reserve University
School of Medicine
10900 Euclid Avenue
Cleveland, OH 44106-4960
USA

Arthur Günzl, Ph.D.
Associate Professor of Genetics and
Developmental Biology
Center for Microbial Pathogenesis
University of Connecticut Health Center
262 Farmington Avenue
Farmington, CT 06030-8130
USA

Erik L. Hewlett, M.D.
Professor of Medicine and Pharmacology
University of Virginia School of Medicine
Box 800419 School of Medicine
University of Virginia
Charlottesville, VA 22908
USA

John M. Kelly, Ph.D.
Pathogen Molecular Biology and
Biochemistry Unit
Department of Infectious and Tropical
Diseases
London School of Hygiene and Tropical
Medicine
Keppel Street
London WC1E 7HT
UK

Richard W. Komuniecki, Ph.D.
Distinguished University Professor of
Biological Sciences
Department of Biological Sciences
University of Toledo
2801 West Bancroft St.
Toledo, OH 43606
USA

J. Joseph Marr, M.D.
180 Centennial Dr.
Estes Park, CO 80517
USA

Richard. J. Martin, Ph.D.
Professor and Chair
Biomedical Sciences
College of Veterinary Medicine
Christensen Drive
Iowa State University
Ames, IA 50011-1250
USA

Paul A.M. Michels, Ph.D.
Research Unit for Tropical Diseases
Christian de Duve Institute of Cellular Pathology
Catholic University of Louvain
Avenue Hippocrate 74–75
B-1200 Brussels
Belgium

Geoffrey Ian McFadden, Ph.D.
ARC Professorial Fellow
Plant Cell Biology Research Centre
School of Botany
University of Melbourne
Victoria 3010
Australia

Miklós Müller, M.D.
Laboratory of Biochemical Parasitology
The Rockefeller University
1230 York Avenue
New York, NY 10021-6399
USA

Timothy W. Nilsen, Ph.D.
Professor and Director
Center for RNA Molecular Biology
Case Western Reserve University
School of Medicine
10900 Euclid Avenue
Cleveland, OH 44106-4960
USA

Fred R. Opperdoes, Ph.D.
Research Unit for Tropical Diseases and Laboratory of Biochemistry
Christian de Duve Institute of Cellular Pathology
Catholic University of Louvain
Avenue Hippocrate 74–75
B-1200 Brussels
Belgium

Marc Ouellette, Ph.D.
MRC Scientist
Burroughs Wellcome Fund Scholar
Centre de Recherche en Infectiologie
CHUQ, pavillon CHUL
2705 Boul. Laurier
Québec, QuéG1V 4G2
Canada

Richard D. Pearson, M.D.
Professor of Medicine and Pathology
Division of Infectious Diseases and International Health, Box 801378
Departments of Internal Medicine and Pathology
University of Virginia
School of Medicine
University of Virginia Health System
Charlottesville, VA 22908
USA

William A. Petri, Jr., M.D., Ph.D.
Professor of Medicine, Microbiology, and Pathology
University of Virginia Health System
MR4 Building, Room 2115
P.O. Box 801340
Charlottesville, VA 22908-1340
USA

Jenny Purcell, Ph.D.
Department of Preclinical
Veterinary Sciences R. (D.) S. V. S., Summerhall
University of Edinburgh
Edinburgh EH91QH
UK

Nicolle Rager, B.A.
Senior Research Assistant
Department of Biochemistry and Molecular
Biology
Oregon Health and Sciences University
3181 S. W. Sam Jackson Park Road
Portland, OR 97239
USA

Alan P. Robertson, Ph.D.
Adjunct Assistant Professor
Biomedical Sciences
College of Veterinary Medicine
Christensen Drive
Iowa State University
Ames, IA 50011
USA

David S. Roos, Ph.D.
Merriam Professor of Biology
Director, Genomics Institute
University of Pennsylvania
415 South University Avenue
Philadelphia, PA 19104-6018
USA

Larry Ruben, Ph.D.
Professor and Chairman
Department of Biological Sciences
Southern Methodist University
6501 Airline
Dallas, TX 75275
USA

David P. Thompson, Ph.D.
Discovery Research
Pharmacia Corp.
301 Henrietta St.
Mailstop 7923-25-410
Kalamazoo, MI 49007
USA

Aloysius G.M. Tielens, Ph.D.
Department Biochemistry and Cell Biology
Faculty of Veterinary Medicine
Utrecht University
P.O. Box 80176
3508 TD Utrecht
The Netherlands

Salvatore J. Turco, Ph.D.
Anthony S. Turco Professor of Biochemistry
Department of Biochemistry
University of Kentucky Medical Center
Lexington, KY 40536
USA

Buddy Ullman, Ph.D.
Department of Biochemistry and
Molecular Biology
Oregon Health and Sciences University
3181 S. W. Sam Jackson Park Road
Portland, OR 97239
USA

Steve Ward
Walter Myers Professor of Parasitology
Liverpool School of Tropical Medicine
Pembroke Place
Liverpool L35 5QA
UK

Preface

Parasitology was born as the tropical stepchild of medicine but has become a well recognized scientific and medical discipline in its own right in our increasingly globally conscious world. It began as a descriptive medical curiosity but the remarkable adaptive mechanisms evinced by these astoundingly versatile organisms have stimulated significant research. Many advances in basic science have come from the study of this increasingly fascinating, phylogenetically diverse group of organisms. Parasitology, in the past decade, has undergone another consequential metamorphosis. The entry of molecular biology with its elucidation of the genetics, genomics, and proteomics of these organisms has provided increasingly sophisticated explanations of their capacities to persist under intense ecological and physiological pressures.

Molecular Medical Parasitology had its inception in an earlier volume entitled *Biochemistry and Molecular Biology of Parasites*. This earlier work has been subsumed in the present text. *Molecular Medical Parasitology* presents parasitology in the context of current molecular biology, biochemistry and cell biology. Throughout the text, emphasis has been placed on the commonality of biochemical and cellular biological processes among these varied organisms. In some discussions, traditional taxonomy, which grouped certain organisms according to similarities in morphology or disease processes, has not been adhered to rigorously. This has been done judiciously in order to emphasize the universality of biochemical and molecular biological mechanisms. Wherever appropriate, information from one chapter has been cross-referenced to another in order to strengthen the important molecular relationships among groups.

The first section, entitled Molecular Biology, opens with a chapter on genomics that is the stage on which the next five chapters play. These chapters include RNA editing and processing, transcription, and post-transcriptional events and describe the interplay of molecular biology and physiology that is manifest in such specific topics as antigenic variation of African trypanosomes and the genetics of virulence.

The second section encompasses the biochemistry and cell biology of the protozoa and then the helminths. Energy metabolism, probably the most thoroughly studied aspect of the biochemistry of these organisms, is presented first in each part. In the sub-section on protozoa, chapters on amino acid and nucleic acid metabolism are followed by specific topics of special interest including surface antigens, intracellular signaling, and intracellular organelles, each with a special emphasis on the commonalities and notable differences in the genomics of the organisms involved. In the sub-section on helminths, the chapter on energy metabolism is followed by an important chapter on neurotransmitters and their receptors. These are critical to the parasite in maintaining its niche in the host and, from a medical perspective, are major therapeutic targets. This section concludes with a chapter on the structure and function of helminth surfaces with emphasis the anatomy and physiology of these critical interfaces that protect the parasites from most host defenses.

Throughout the volume, the authors and editors have emphasized the actual or potential medical importance of major biochemical or molecular biological advances. These considerations are expanded in the third section. The first chapter is on drug resistance, which, in fact, is a medical manifestation of molecular biology and biochemistry bringing about alterations in the cell biology as a result of environmental pressures. It has become a significant medical problem in recent years. The chapter on therapy discusses the implications of the basic science presented in the earlier sections as well as specifics of treatment.

This is the first parasitology text that integrates current molecular biology, biochemistry, and cell biology with the control of these heterogeneous organisms. The authors are among the best in their respective fields and the knowledgeable scientist will recognize their contributions. They have written clearly, comprehensively, and well. Presentations by these seasoned investigators should be of interest to the experienced scientist, the graduate student, and the physician.

We must list first among the acknowledgements, our authors. Much credit, however, must go to Ms. Claire Minto, an extraordinary editor, who has been an exceptional resource in the preparation of this book.

J. Joseph Marr, M.D.
Richard W. Komuniecki, Ph.D.
Timothy W. Nilsen, Ph.D.

SECTION I

MOLECULAR BIOLOGY

This Page Intentionally Left Blank

CHAPTER

1

Parasite genomics

Mark Blaxter
Institute of Cell, Animal and Population Biology,
University of Edinburgh, Edinburgh, UK

INTRODUCTION

Genomics, like parasitology, is a research field that thrives on the intersection between different disciplines. Parasitologists study a phylogenetically disparate assemblage of organisms chosen from global diversity on the basis of their trophic relationships to other 'host' organisms, and use the tools and paradigms of biochemistry, molecular biology, physiology and behaviour (amongst others) to illuminate the biology of these important taxa. Genomics uses data arising from karyotypic analysis, genetic and physical mapping of traits and anonymous markers, DNA sequencing and bioinformatic prediction of function-structure relationships. The meld of parasitology and genomics is thus necessarily and productively hybrid.

Genomics research in parasitology can be divided, pragmatically, into two sectors. One is a drive to generate resources: clone banks, sequence, annotated genes, functional genomics platforms. The other is a hypothesis-driven search for pattern and process in the structure, expression and evolution of genomes: how does the organism self-assemble given this set of genotypic data? These two sectors overlap, as resource generation necessarily underlies the testing of hypotheses of genome-wide function. While the methodologies used to analyse the genomes of protozoan, nematode and platyhelminth genomes may differ because of the ways the genomes of these organisms are organized, the aims of programs on individual species are in general the same:

1. The determination of the complete sequence of the chromosomal (and plastid) genome of the organism
2. The identification of the coding genes (both protein and RNA) on the sequence (also termed 'gene discovery')
3. The prediction of function of each of the genes, and the prediction of function of operator/promoter/control regions in the non-coding DNA
4. The integration of functional, sequence and architectural information into biological models of the structure of the chromosomes

and of the interaction between the expressed parts of the genome

5. Investigating natural variation in the genome in the context of the host, population structure, drug treatment and other selective forces.

Along this difficult path additional goals can be found, such as the identification of candidate sequences, genes, or gene products that may be of utility in diagnosis, surveillance, drug targeting or vaccine component development.

Genomics and genome sequencing is still a young field. The first genomes sequenced were those of parasites: viruses infecting bacteria (phiX174 and lambda phage are landmarks). Progress to whole genome sequencing of self-reproducing organisms had as steppingstones the determination of the complete genome sequence of the human mitochondrion (again relevant to parasitology as mitochondria arise from an ancient symbiotic event). The first genome sequence determined for a self-reproducing organism was that of *Helicobacter pylori*, an important human-pathogenic bacterium. In the field of bacterial genomics, the focus has remained on pathogenic species, and most of the over 100 sequenced genomes are from human pathogens. For parasitology, these genomes give insight into the differences at the level of the genome between free-living bacteria (such as *Escherichia coli*) and endoparasitic bacteria (such as the Chlamydiae and Rickettsiales). Importantly, it is now technically feasible to sequence the genomes of eukaryotes with large genomes (>20 Mb) and thus several parasite genome projects are underway. As with the sequencing of the nematode *Caenorhabditis elegans*, the fly *Drosophila melanogaster* and the human genomes, this will in turn bring about a revolution in the way parasite biology research is done.

THE SIZE OF THE PROBLEM

Bacterial genomes are relatively small (0.6 to 15 Mb) compared to those of eukaryotes (10 Mb to >10 000 Mb). Parasitic eukaryote genomes range from ~9 Mb (*Theileria*) to 5000 Mb (*Ascaris*) and above (Table 1.1). The number of genes encoded by a genome is roughly proportional to its size, but is modified by the presence of intronic DNA and of junk, or non-coding repetitive DNA. For example, while the genome of the nematode *Caenorhabditis elegans* is 100 Mb, and contains 20 000 protein coding genes, the human genome is 3000 Mb (30-fold larger) but encodes only 30 000–40 000 genes. The average gene density in *C. elegans* is thus about one gene per 5 kb, while in humans it is one gene per >70 kb. Protozoa have relatively small genomes that are often rich in non-coding repeats, and are likely to have in the region of 6000 to 15 000 protein coding genes. Parasitic nematodes have genomes of a similar size to *C. elegans* in the main, but several species have much larger DNA contents per haploid genome. In *Ascaris* and related taxa, the genome is both highly repetitive and much larger than that of *C. elegans*. Overall, parasitic nematodes are likely to have similar gene complements to *C. elegans* (20 000). The genomes of platyhelminths are much less well known, but *Schistosoma* species have genomes of ~270 Mb that are rich in repetitive sequences. Again the gene count is likely to be in the 20 000 range. Arthropod parasites have larger genomes than the model arthropod, *Drosophila melanogaster*, which has 15 000 genes in a 160 Mb genome. For example, *Anopheles* has a genome of 280 Mb but is expected to have a gene count similar to *D. melanogaster*.

The multitude and phylogenetic diversity of parasites means that genomic approaches to parasite biology and control need to be carefully

TABLE 1.1 Parasite genomes: genome sizes, karyotype, gene number and genome project status of selected par

Species	Genome size	Karyotype (2n)	Genome survey sequences in dbGSS	Genome sequencing status	Methods used
Nematode parasites					
Brugia malayi	100 Mb	12	18 000	Selected genome segments	
Ascaris suum	5000 Mb	48		–	
Haemonchus contortus	100 Mb	12		–	
Platyhelminth parasites					
Schistosoma mansoni	270 Mb	14	42 000		Physical map based
Apicomplexan parasites					
Plasmodium falciparum	35 Mb	14	(also 18 000 GSS from other plasmodial species)	Completed	Whole genome and chromosome-b shotgun
Theileria annulata	10 Mb	4		Genome sequencing initiated	
Trypanosomatid parasites					
Trypanosoma cruzi	35 Mb Haploid	35	21 000	Near completion	Chromosome-b and physical map based
Trypanosoma brucei	35 Mb	11	90 000	Near completion	Chromosome-b shotgun and physical map based
Leishmania major	33.6 Mb	36	15 000	Near completion	Chromosome-b shotgun and physical map based
Other protozoan parasites					
Entamoeba histolytica	<20 Mb	14	80 000	Near completion	Whole genome shotgun
Giardia intestinalis	12 Mb	5		Near completion	Whole genome shotgun
Vectors of parasites and arthropod parasites					
Anopheles gambiae	280 Mb	3	60 000	Near completion	Whole genome shotgun

tailored to each target organism. The World Health Organisation in collaboration with the national funding agencies of both endemic and developed countries have therefore sponsored genome projects on target organisms representing the major human and animal parasitic diseases. Each project has used tools based on the peculiarities of their system and the knowledge/skill base present in the interested community. The parasite genome projects are models of north–south, endemic–developed cooperation, and, in this open spirit, most of the data produced is freely available through the internet to interested researchers.

GENERATING GENOMICS DATA

Genomics uses data from many sources. The parasite genome projects use layers of related data types to build first physical and genetic maps of the target genomes, followed by finer detail sequence and expression maps, ultimately yielding an annotated genome. Most of the projects are still in the midst of the data generation part of the process (see http://www.ebi.ac.uk/parasites/parasite-genome.html for the latest news on the various parasite genome projects), and no simple summary will adequately cover all the projects. The field is also changing extremely rapidly, and a summary given today may be rendered obsolete with tomorrow's database release.

Genetic maps

Genetic maps are available for many parasitic organisms. The maps are built by examining the genotype of recombinant cross progeny of marked parents. The markers can be phenotypic (eye color, resistance to filarial nematode infection) or anonymous genetic markers (microsatellite sequence tagged sites or restriction fragment length polymorphisms for example). The result is a linkage map showing the association of the markers and their relative order. This map is of utility in verifying the correctness of related genome maps made at the physical (DNA) level, as markers placed adjacent by genetics should also be adjacent in any physical map. Genetic mapping is necessarily restricted to organisms that reproduce sexually, and operationally is further restricted by considerations of practicality (is it possible to carry out controlled crosses and score progeny in the laboratory?).

To overcome this need for sex, a method for genetic mapping without sexual recombination has been developed, called HAPPY mapping. HAPPY mapping is based on the observation that in a population of large DNA fragments generated by random shearing of a complete genome, the chance that two sequence tagged markers will be on the same individual molecule is proportional to their separation on the genome. This mapping procedure uses PCR-based genotyping assays to screen sub-haploid quantities of sheared DNA for association between markers, and the association is then used to build a 'genetic' map as one would with real genetic data. The benefit of the HAPPY map is that the markers are cloned and sequenced at the outset, allowing rapid progression to complete physical mapping (see below).

Karyotyping

Chromosomes are the units of genome organization. Mapping of genes or other molecular markers to physical chromosomes is a useful and often central step in genomics. At a gross level, chromosomes can be separated by morphology (for example the filarial nematode sex determining X and Y chromosomes) and by differential banding staining with intercalary dyes. In the protozoa, the chromosomes are often too small to be resolved usefully by microscopy, but are within the range that is resolved by pulsed field gel electrophoresis (PFGE). PFGE karyotypes are available for all of the major parasitic protozoan species, and comparative karyotyping of strains and related species has yielded valuable information on conservation of linkage, and patterns of genome evolution. Fluorescence *in situ* hybridization (FISH) involves the 'painting' of a chromosomal copy of a gene with a fluor-labelled probe in a preparation of metaphase cells. It is useful in confirming linkage of cloned markers, and in joining otherwise unlinked segments of a physical map. For chromosomes separable by PFGE, Southern hybridization can be used to similar effect.

For many organisms, including the nematodes, the chromosomes are too large (>10 Mb) to be separated by PFGE and too small to be useful for FISH and banding studies. It is possible to separate these chromosomes using a fluorescence-activated cell sorting instrument, though this technique has not been used yet in parasite genomics.

Physical maps

It is often useful to have a genomic copy of a gene of interest cloned. Large-insert genomic DNA clones can be constructed in a number of different vector–host systems. These range from lambda bacteriophage (maximal insert capacity ~21 kb of foreign DNA), through cosmids (~35 kb), bacterial artificial chromosomes (BAC, ~200 kb) and yeast artificial chromosomes (YAC, ~3000 kb). Each vector–host combination also differs in copy number within the host cell: in general vectors maintained at low copy number tend to be more stable against recombination, rearrangement and deletion. Yeast host cells are often more tolerant of skewed base-composition insert DNA, such as that from *Plasmodium*, and of repeat-rich insert DNA.

The inserts of large-insert clone libraries can be compared to each other and the overlap data used to build a map of the cloned genome, a physical map. Overlap between clones can be predicted in two ways. One is derived from restriction enzyme fingerprinting of each clone. A fingerprint is the pattern of bands observed when the clone is cut with one or two enzymes. Clones containing DNA from the same genomic region will share more fingerprint bands than would be expected to occur by chance, and can be overlapped on the basis of shared fragments. The other method of building a physical map is by sequence-tagged site mapping, where the library is screened with probes by hybridization or clones are identified using STS-based PCR. The two methods (fingerprinting and STS mapping) can be, and usually are, combined in the production of a map. Maps have been produced or are in production for many parasites. FISH hybridization to spread chromosomal segments can also be used to build maps, and for smaller genomes it is possible to construct restriction fragment-based maps using stretched chromosomes cut *in situ* on the slide.

Physical map construction is compromised by the sheer volume of data that must be produced and analysed, and the known sorts of confounding errors that can occur. In fingerprinting, there is (usually known) error in band size estimation, and two bands can be scored as matching by size despite being different in sequence. The method is very sensitive to the number of shared bands required to score a real overlap, as too high a score requirement will result in failure to link overlapping clones, whereas too low a score will result in multiple, incompatible overlaps being accepted. In STS mapping, errors can arise from the presence of repetitive sequence (either duplicated genes or non-coding repeats) that will wrongly join two distinct genomic regions. The number of clones required to map a genome depends on the genome size, the mean size of the inserts (and the distribution around the mean), and the representativeness of the library. Some genomic regions clone poorly (if at all), and often multiple cloning systems must be used to obtain closure.

The experience of the *C. elegans* project is of relevance here, as it was one of the first to build a 'complete' physical map. A map built from 17 000 cosmid fingerprints yielded ~3000 contigs of overlapping clones. When a larger-insert YAC (yeast artificial chromosome) library was added to the map by hybridization, the number of contigs dropped to ~600. Much additional work, involving constructing and screening libraries in multiple vector–host systems, was necessary to achieve the final 98% coverage. In *P. falciparum*, YACs have also been used to construct a physical map, but in this case the process was facilitated by using hybridization to PFGE blots, and hybridization of PFGE-separated chromosomes to YAC libraries, to assign cloned YACs to chromosomes. Similarly, in *S. mansoni*, FISH is being used to assign clone contigs to the chromosomes.

Genome sequencing

Expressed sequence tags

The genome sequence of a parasitic organism can be obtained in a number of convergent ways. The choice of experimental route is dependent on the resources available, the genomic biology of the organism and the needs of the researchers. For some parasites, genomics effort has focused on gene discovery, and rapid and cost effective methods have been used to obtain sequence tags on many of the genes of the organisms. The coding portion of a genome (the portion that is transcribed as RNA, and is translated to give protein) is typically less than 50% for eukaryotic organisms. For eukaryotes with introns this proportion drops further still. A method that sampled and sequenced only the expressed portions of a genome would thus be an efficient gene discovery tool.

The expressed sequence tag (EST) strategy is one such method. To generate ESTs, a cDNA library, representative of the genes expressed in a particular stage, sex or tissue of the parasite, is sampled at random. From each random clone, a single-pass sequence is generated. This sequence serves both to tag the transcript from which it derived, and also offers sequence data that can be used to perform informatic analyses to identify the function of the encoded protein.

For some parasitic genome projects, ESTs are the main or only mode of genomics data production, while in others they play a minor role: the balance is based on the needs and opportunities available for each target species. The dbEST division of the public databases contains over 10 million ESTs, from over 390 organisms (Table 1.2). Of these, only ~2% (200 000) are from parasitic organisms and their vector hosts, but parasites make up ~15% of the different species represented. The overrepresentation by species is due to the generally smaller size of the parasitic datasets than those from humans and model organisms. EST acquisition is relatively cheap, and is a 'low tech' genomics option for laboratories and communities without funding and infrastructure for larger programs. The yield, in terms of 'interesting new genes' per unit of effort, is very high, and EST projects can substitute for more 'hypothesis-driven' gene cloning efforts where the aim is to define the transcriptional features of a particular stage or tissue of the parasite of interest.

The diversity of genes represented in an EST dataset reflects not only the size of the dataset,

TABLE 1.2 EST datasets from parasitic organisms (December 2001)

Species	Number of ESTs	Expected number of genes per genome
Nematode parasites	104 222	20 000
Brugia malayi	22 439	
Onchocerca volvulus	14 922	
Strongyloides stercoralis	11 392	
Ascaris suum	7 410	
Ancylostoma caninum	7 259	
Strongyloides ratti	6 562	
Meloidogyne javanica	5 600	
Haemonchus contortus	4 843	
Parastrongyloides trichosuri	4 541	
Heterodera glycines	4 327	
Trichinella spiralis	4 238	
Meloidogyne arenaria	3 334	
Trichuris muris	1 388	
Globodera pallida	1 246	
Ancylostoma ceylanicum	1 110	
Necator americanus	961	
Globodera rostochiensis	894	
Toxocara canis	519	
Ostertagia ostertagi	450	
Teladorsagia circumcincta	315	
Litomosoides sigmodontis	198	
Wuchereria bancrofti	131	
(seven additional species with less than100 ESTs each)		
Platyhelminth parasites	19 709	20 000
Schistosoma mansoni	16 813	
Schistosoma japonicum and *S. haematobium*	2 097	
Echinococcus granulosus	799	
Apicomplexan parasites	39 138	7 000
Plasmodium yoelii yoelii	12 471	
Eimeria tenella	11 438	
Plasmodium falciparum	6 769	
Plasmodium berghei	5 345	
Plasmodium yoelii	3 091	
Theileria parva and *T. annulata*	24	
Trypanosomatid parasites	17 479	10 000
Trypanosoma cruzi	10 133	
Trypanosoma brucei, T. b. brucei and *T. b. rhodesiense*	5 133	
Leishmania major, L. infantum and *L. mexicana*	2 213	
Other protozoan parasites	1 070	6 000+
Entamoeba histolytica	463	
Acanthamoeba healyi	377	
Entamoeba dispar	139	
Giardia intestinalis	91	
Vectors of parasites and arthropod parasites	9 587	15 000+
Anopheles gambiae	6 037	
Aedes aegypti	1 518	
Biomphalaria glabrata	1 426	
Sarcoptes scabiei	396	
Boophilus microplus	143	
Culex pipiens pallens and *Anopheles stephensi*	64	

but also the representativeness of the library, and the faithfulness of the cDNA cloning procedure. In general, the abundance of ESTs corresponding to one gene transcript will reflect the steady state mRNA levels of the transcript in the organism, but smaller mRNAs are reverse transcribed and cloned more easily than larger ones, and thus some bias can arise. To gain access to low-expression-level transcripts, a large number of ESTs must be sequenced from a given library.

To improve the efficiency of new gene discovery in EST libraries, subtraction or normalization procedures can be carried out. Normalization aims to make the levels of each transcript in the library approximately equivalent by selecting against highly expressed genes. Subtraction aims to eliminate from the library sequences that derive from transcripts also present in another stage or tissue. Normalization and subtraction can be carried out previous to the cloning stage, or on libraries by hybridization-elimination of unwanted clones.

The EST strategy brings with it problems. One is that the EST dataset will only be as good as the library from which it derives. Another is that it becomes increasingly more difficult to identify new genes as sequences are accumulated from a species: the yield of new genes per sequence can drop to less than one per ten ESTs quite rapidly. If libraries are not available from all life-cycle stages, it will not be possible to sample all the genes of the organism, as many will have close stage-specific regulation. Finding rarely expressed genes will be a stochastic process. For example, even in 'mature' genome projects, EST analysis typically yields only ~50% of the genes later discovered by genome sequencing. Normalization and subtraction can improve this, but in general the returns for effort fall off rapidly. The bioinformatic analysis of ESTs is discussed below.

Genome survey sequencing

ESTs can only sample a gene when it is expressed in the tissue or stage from which the cDNA library is made. In addition, many genes may be expressed at such low levels (for example in one neuron of a metazoan parasite), or in very particular environmental circumstances (for example during the process of entering a new host cell) that it is very unlikely that they will be identified by ESTs. While genes are represented in cDNA libraries in proportion to their level of expression, in genomic DNA libraries they are present in proportion to their representation in the genome. For relatively unbiased gene discovery the genomic equivalent of ESTs, genome survey sequences (GSS), are often useful.

GSS surveys of eukaryotic genomes yield interagenic, non-coding, intronic and coding sequences. This feature makes them ideal for establishing an overview of the patterns of sequence present in a genome. For organisms with few or no introns, and thus relatively gene-dense genomes, GSS surveys can be as efficient as EST surveys for gene discovery. In *Trypanosoma brucei* for example, coding sequence makes up ~50% of the genome, and thus GSS surveying yields many open reading frames. For metazoan genomes, which typically contain less than 25% coding sequence, GSS are less efficient at gene discovery, but are still a valuable adjunct to EST-based analyses.

In addition to coding genes, GSS can help define repeat sequences useful for diagnostic or population genetics programs, and also reveal features such as transposons and retrotransposons. As outlined below, the GSS concept can be taken further to provide shotgun sequencing for most of a genome. GSS are usually determined from large insert libraries in cosmid or BAC vectors, using primers that yield reads of each end of the cloned insert (and thus called, for example, 'BAC end sequences').

Map-based genome sequencing

For genomes where it is valid and possible to proceed to whole genome sequencing, two main approaches are taken. One is to use the resource of a physical map of the genome, and to sequence it 'clone by clone'. The other is to perform a shotgun sequencing project on the whole genome, or PFG separated individual chromosomes. Both approaches require the ability to assemble large (usually >35 kb) genomic segments of DNA from a large number of individual reads of ~500 bp each. To robustly and credibly assemble a sequence, it is usual to first make one (or several) small insert sublibraries (inserts of ~2 kb) from the target fragment, and to sequence a number of these selected at random to give a 6–10-fold sequence coverage of the fragment. This is called shotgun sequencing. The redundancy of 6–10-fold is required both to ensure the correct sequence is determined free of errors (by verifying the nature of any particular base using independent sequence reads) and to attempt to cover the whole fragment. If the whole of the fragment were equally clonable, then ∼6-fold coverage would be required to ensure that all regions of the fragment are sampled. In practice, not all regions are equally clonable, and assembly of shotgun reads is usually followed by a 'finishing' phase where missing regions are sequenced by more directed methods (such as primer walking) and ambiguities in the sequence are clarified.

Sequencing a genome using a physical map thus involves a large number of small shotgun-assembly-finishing projects based on a set of clones that overlap minimally and cover the whole (or most of) the genome. The regions of overlap between the clones serve to add confirmation to the determined sequence and to the mapping process.

Rather than build a map first, a map-as-you-go strategy has been proposed, utilizing extensive GSS data from large insert clones. The BAC or other library is first completely end sequenced to yield one GSS every 5–10 kb of the genome on average. A BAC clone is selected (at random) and shotgun sequenced. Using the finished sequence, a new clone that minimally overlaps is selected on the basis of end sequence comparison. In this manner, the project can 'walk' from each seed clone into the flanking genome.

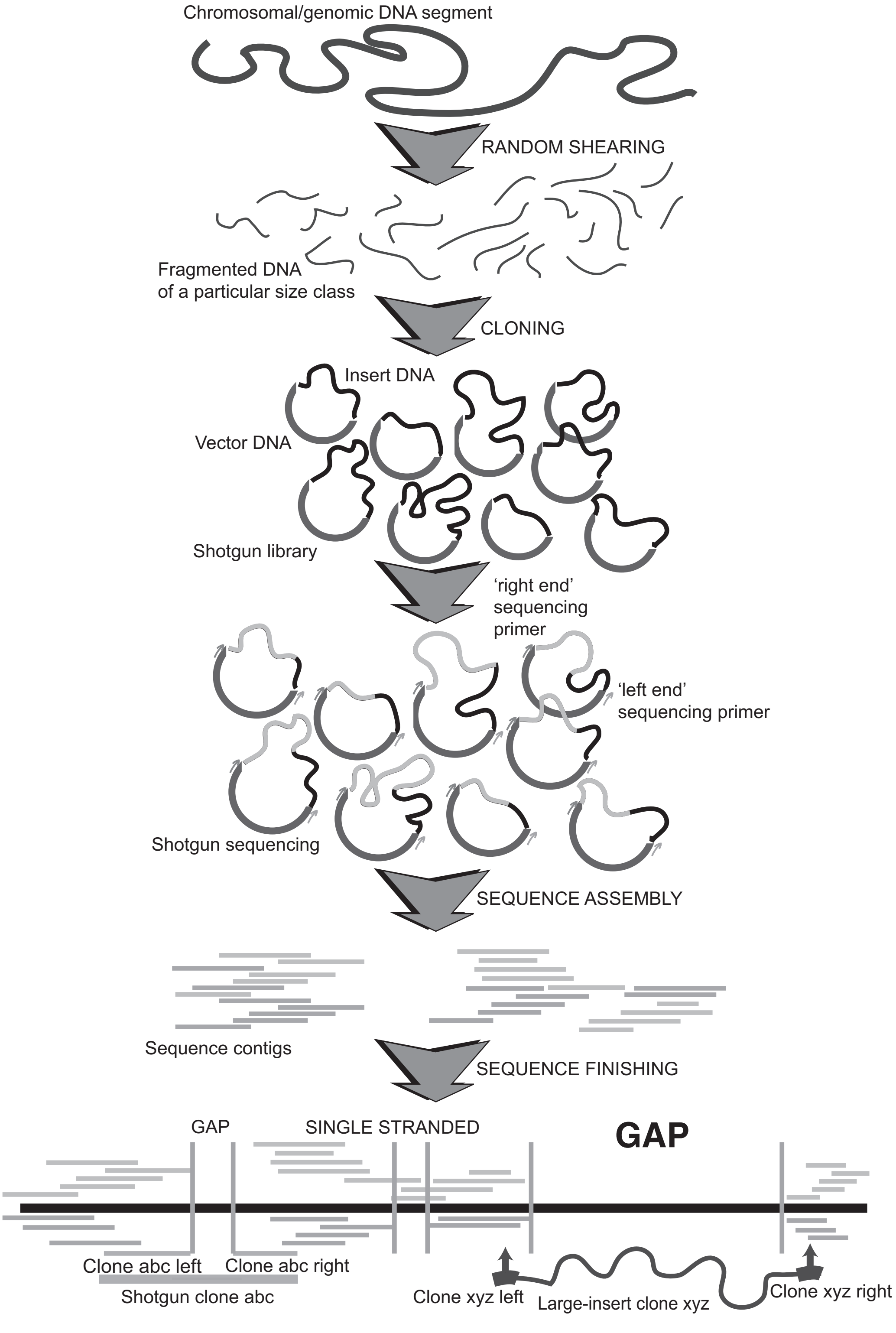
Shotgun sequencing of whole chromosomes and whole genomes

Initially, the complexity of genomic DNA, in particular the presence of local and disseminated repeats, suggested that it might not be possible to assemble and finish fragments of DNA larger than ∼100 kb. The limits were set by the efficiency of the computer algorithms used for assembly, and the overall sequencing strategy. However, with advances in sequencing, algorithms and computing power, it is now possible to assemble even the largest genomes in one go, from tens of millions of individual reads (Figure 1.1).

The success of the clone-by-clone strategy is based in part on its breaking down of the assembly problem in to a set of smaller, manageable ones. For many protozoan parasites, chromosomes can be separated by PFGE. These chromosomes, ranging from <100 kb to over 7 Mb, can also be shotgun sequenced and thus the whole-genome assembly is reduced to a set of smaller projects. Chromosome shotgun sequencing is now a mainstay of many protozoan genome sequencing projects. For organisms with larger genomes, where chromosomal separations are not possible, the success of the human (3000 Mb) and *Drosophila melanogaster* (160 Mb) whole genome shotgun assemblies suggests that this method should also be applicable to metazoan parasites.

For whole genome and whole chromosome shotgun projects, the large numbers of reads from small insert (2 kb) clones are usually supplemented with reads from shotgun libraries with larger mean insert size (10 kb and above; BAC end GSS are also used). These longer clones serve as a scaffold that is used to orientate and affirm the assembly made with the 500 bp reads from the 2 kb clones. If a sequence contig suggests that the two ends of a large insert clone are too close to each other, it is likely to be in error. Similarly, large-insert sequences can serve to link contigs derived from shotgun sequencing.

The whole genome shotgun method requires a large amount of data to be effective. For a 100 Mb genome (such as is found in many nematode parasites) a one-fold shotgun is


Chromosomal/genomic DNA segment
RANDOM SHEARING
Fragmented DNA
of a particular size class
CLONING
Insert DNA
Vector DNA
Shotgun library
'right end'
sequencing
primer
'left end'
sequencing primer
Shotgun sequencing
SEQUENCE ASSEMBLY
Sequence contigs
SEQUENCE FINISHING
GAP
SINGLE STRANDED
GAP
Clone abc left
Clone abc right
Shotgun clone abc
Clone xyz left
Large-insert clone xyz
Clone xyz right

~200 000 sequences of 500 bp. A ten-fold shotgun is thus ~2 million sequences. For each shotgun project, the reads have to be assembled with resolution of ambiguities resulting from the presence of repetitive DNA. The finishing process is necessarily more protracted, and utilizes data inherent in the scaffold of larger insert clones, and often also long-range restriction mapping and PFG southern blotting.

A genome sequence is a hypothesis

The result of a shotgun sequencing project, be it of a clone, a chromosome or a genome, is a DNA sequence that has been verified to the best ability of the sequencers. The error rate in most sequencing projects is estimated to be one miscalled base in 10 000. This is a maximal estimate of the error, and often independent resequencing surveys reveal much lower actual error rates. The final public sequence must therefore be regarded as a hypothesis, and used and interpreted with reference to the strength of the supporting evidence. In particular it is technically difficult to resolve the sequence of tandem short repeats and regions of low complexity or biased base composition. The genome sequencers will strive to resolve all conflicts in the data, but rely on the user communities to communicate to them any errors found.

During the shotgun and finishing phases of genome sequencing projects, many genome sequencing centres will release preliminary assemblies of the data. These are works in progress, and users should be aware that contiguated sequences present in one day's preliminary data release might be absent from the next due to the discovery of errors in assembly. Even the published sequence will change as errors are corrected, and care should be taken to use the latest data release in analyses.

The problems and benefits of the reference strain

Sequencing a genome, even the relatively small genomes of bacteria, is a major undertaking, and resources are unlikely to be available for multiple genome sequences of disease organisms. For model organisms, the choice of strain for full genome sequencing is often obvious: the history of genetic research will have defined an isolate or strain as being the 'wild type', and this will be the most appropriate for sequencing. Thus, for *C. elegans* the N2 strain was sequenced, and for *Arabidopsis thaliana,* the Landsberg ecotype was chosen. For parasites the choice is rarely as easy. Parasitologists are often interested in the diversity of their target organisms, and research focuses on between-strain and between-species variation in virulence and other important traits. The best studied, best 'domesticated' strains used in many laboratories may have lost key traits during adaptation to laboratory hosts, or due to inbreeding or other genetic selection. The best known strain may include in its phenotypes the

FIGURE 1.1 (See also Color Plate 1) Sequencing a genome. This figure illustrates the steps involved in determining the sequence of a genome (or genomic segment, such as a chromosome or large-insert clone). The DNA is first sheared and cloned to make a shotgun library, which is then sequenced using universal 'left' and 'right' primers to generate a shotgun sequence set for the segment. This shotgun sequence set is then assembled into contigs of sequence reads that overlap each other. These contigs are turned into a finished product through the use of linking clone data: either matching left and right reads from individual clones within the shotgun sequence dataset, or end-sequences from larger-insert clones. It is usual to finish the sequence by obtaining sequence from both strands, resolving all ambiguities in the predicted consensus, and linking all contigs.