


Current Opinion in Genetics & Development


Volume 15, Issue 6, Pages 569-666 (December 2005)

1.  **Editorial board • EDITORIAL BOARD**
Page i

2.  **Contents • CONTENTS LIST**
Page iii


3.  **The next issue of this journal • CONTENTS LIST**
Page iv

Genomes and evolution. Edited by Stephen J O'Brien and Claire M Fraser


4.  **Genomes and evolution: The power of comparative genomics • EDITORIAL**
Pages 569-571
Stephen J O'Brien and Claire M Fraser

5.  **For better or worse: genomic consequences of intracellular mutualism and parasitism • REVIEW ARTICLE**
Pages 572-583
Jennifer J Wernegreen

6.  **Common themes in the genome strategies of pathogens • REVIEW ARTICLE**
Pages 584-588
Jeffrey G Lawrence

7.  **The microbial pan-genome • REVIEW ARTICLE**
Pages 589-594
Duccio Medini, Claudio Donati, Hervé Tettelin, Vega Massignani and Rino Rappuoli

8.  **Ancestral state reconstructions for genomes • REVIEW ARTICLE**
Pages 595-600
Christos A Ouzounis

9.  **Causes and effects of nuclear genome reduction • REVIEW ARTICLE**
Pages 601-608
Patrick J Keeling and Claudio H Slamovits

10. 

Comparative genomics of malaria parasites • REVIEW ARTICLE

Pages 609-613

Neil Hall and Jane Carlton

11. 

Hemiascomycetous yeasts at the forefront of comparative genomics • REVIEW ARTICLE

Pages 614-620

Bernard Dujon

12. 

Transposable elements, gene creation and genome rearrangement in flowering plants • REVIEW ARTICLE

Pages 621-627

Jeffrey L Bennetzen

13. 

Conserved sequences and the evolution of gene regulatory signals • REVIEW ARTICLE

Pages 628-633

Mark D Adams

14. 

Comparative genomics as a tool in the understanding of eukaryotic transcriptional regulation • REVIEW ARTICLE

Pages 634-639

Julie E Baggs, Kevin R Hayes and John B Hogenesch

15. 

Short, local duplications in eukaryotic genomes • REVIEW ARTICLE

Pages 640-644

Elizabeth E Thomas

16. 

Chromosomal sex-determining regions in animals, plants and fungi • REVIEW ARTICLE

Pages 645-651

James A Fraser and Joseph Heitman

17. 

Genomic inferences from Afrotheria and the evolution of elephants • REVIEW ARTICLE

Pages 652-659

Alfred L Roca and Stephen J O'Brien

Commentary

18.



Sex chromosomes and sex determination in reptiles: Commentary •

ARTICLE

Pages 660-665

William S Modi and David Crews

Copyright © 2006 Elsevier Ltd. All rights reserved

Editorial Enquiries

Current Opinion in Genetics & Development,
Elsevier London,
84 Theobald's Road,
London,
WC1X 8RR, UK
Tel: +44 (0)20 7611 4400
Fax: +44 (0)20 7611 4401
e-mail: COGenetdev@elsevier.com

Subscription Enquiries

e-mail: ct.subs@qss-uk.com

In-house Editor Christopher Walsh
Cross-title Editor Jenny Gillion
Editorial Coordinator Nikos Panayiotou
Publishing Manager O Claire Moulton
Illustrators The Studio
Journal Manager Rolf van der Sanden

Current Opinion in Genetics & Development
ISSN 0959-437X

is published bimonthly by

Elsevier Ltd
The Boulevard
Langford Lane
Kidlington
Oxford, OX5 1GB, UK

The **Aims** of the journal can be found at
www.elsevier.com/locate/gde

Printed by The Manson Group Ltd, St Albans, UK.

Current Opinion in Genetics & Development
is indexed and/or abstracted by
BIOSIS/Zoological Record
Biotechnology Citation Index
CAB Abstracts International
Chemical Abstracts
Current Contents (Life Science)
EMBASE
Index Medicus
Index Veterinarius
Life Sciences Collection
Medline/MEDLARS Online
Nucleic Acids Abstracts
Reference Update
Science Citation Index
SciSearch/Science Citation Index Expanded

Editorial Board

Robin Allshire UK
Andrea Ballabio Italy
David Baltimore USA
J Michael Bishop USA
Elizabeth H Blackburn USA
Piet Borst Netherlands
Marian Carlson USA
Sarah CR Elgin USA
Gerald R Fink USA
Richard A Firtel USA
Uta Francke USA
John C Gerhart USA
Peter N Goodfellow UK
Corey S Goodman USA
John Gurdon UK
Chris Higgins UK
Tony Hunter USA
Thomas Jessell USA
Mary-Claire King USA
Eric S Lander USA
Ron Laskey UK
Chris Leaver UK
Tom Maniatis USA
Anne McLaren UK
Barbara J Meyer USA
Christiane Nüsslein-Volhard Germany
Leena Peltonen USA
Mark Ptashne USA
Elizabeth J Robertson USA
Lucy Shapiro USA
Allan C Spradling USA
Robert Tijan USA
David Weatherall UK
Robert A Weinberg USA
Jean Weissenbach France
Mitsuhiro Yanagida Japan

2005 Contents

The subject of genetics and development is divided into six major sections, each of which is reviewed once a year.

February

Oncogenes and cell proliferation
Edited by Michelle D Garrett and Sibylle Mitnacht

April

Chromosomes and expression mechanisms
Edited by Barbara J Meyer and Jonathan Widom

June

Genetics of disease
Edited by Veronica van Heyningen and David FitzPatrick

August

Pattern formation and developmental mechanisms
Edited by William McGinnis and Cheryl Tickle

October

Differentiation and gene regulation
Edited by Andrew J Bannister and Tony Kouzarides

December

Genomes and evolution
Edited by Stephen J O'Brien and Claire M Fraser

Reviews

Genomes and evolution

Edited by Stephen J O'Brien and Claire M Fraser

- 569 Stephen J O'Brien and Claire M Fraser**
Editorial overview: The power of comparative genomics
- 572 Jennifer J Wernegreen**
For better or worse: genomic consequences of intracellular mutualism and parasitism
- 584 Jeffrey G Lawrence**
Common themes in the genome strategies of pathogens
- 589 Duccio Medini, Claudio Donati, Hervé Tettelin, Vega Masignani and Rino Rappuoli**
The microbial pan-genome
- 595 Christos A Ouzounis**
Ancestral state reconstructions for genomes
- 601 Patrick J Keeling and Claudio H Slamovits**
Causes and effects of nuclear genome reduction
- 609 Neil Hall and Jane Carlton**
Comparative genomics of malaria parasites
- 614 Bernard Dujon**
Hemiascomycetous yeasts at the forefront of comparative genomics
- 621 Jeffrey L Bennetzen**
Transposable elements, gene creation and genome rearrangement in flowering plants

- 628 Mark D Adams**
Conserved sequences and the evolution of gene regulatory signals
- 634 Julie E Baggs, Kevin R Hayes and John B Hogenesch**
Comparative genomics as a tool in the understanding of eukaryotic transcriptional regulation
- 640 Elizabeth E Thomas**
Short, local duplications in eukaryotic genomes
- 645 James A Fraser and Joseph Heitman**
Chromosomal sex-determining regions in animals, plants and fungi
- 652 Alfred L Roca and Stephen J O'Brien**
Genomic inferences from Afrotheria and the evolution of elephants

Commentary

- 660 William S Modi and David Crews**
Sex chromosomes and sex determination in reptiles

The cover

Swazi, a female African savannah elephant, is the matriarch of a herd imported from Africa to the San Diego Wild Animal Park. Part of a herd that was relocated to Swaziland from Kruger National Park in South Africa, where successful reproduction required reduction of the herd, she and six other savannah elephants were imported by the Zoological Society of San Diego. Mandatory blood-testing prior to importation provided sufficient DNA for the Broad Institute at the Massachusetts Institute of Technology to complete a 2x shotgun genome sequence to help annotate the human genome. In this issue of *Current Opinion in Genetics & Development*, Alfred L Roca and Stephen J O'Brien discuss the 'Genomic inferences from Afrotheria and the evolution of elephants'. Photograph courtesy of the Zoological Society of San Diego. With thanks to Oliver Ryder and Christina Simmons.

The next issue of this journal

Oncogenes and cell proliferation

Edited by Allan Balmain and Denise Montell

Will contain reviews by

Brian Druker and Thomas O'Hare

Targeted therapies in cancer treatment: overcoming Gleevec resistance

Carlos Arteaga

Inhibition of TGF β signaling in cancer therapy

Lara S Collier and David A Largaespada

Transforming science: cancer gene identification

Ari Melnick

Peptide interference as a therapeutic approach

Suzanne Eaton

Synthesis and trafficking of Hedgehogs and Wnts

Marcos Vidal and Ross L Cagan

Drosophila models for cancer research

Brad Ozanne

Fos and cell invasion or *Caenorhabditis elegans* models for cancer research

Felix H Brembeck and Walter Birchmeier

Balancing cell adhesion and Wnt signaling: the key role of β -catenin

Alea A Mills

p63: oncogene or tumour suppressor

Qi-Wen Fan and William A Weiss

Chemical genetics approaches to the development of cancer therapeutics

Scott Hammond

The oncogenic and tumour suppressor functions of microRNAs

Jeffrey Rosen

The role of stem cells in cancer cell proliferation

George Thomas

mTOR and cancer: reason for dancing at the crossroads?

Mina Bissell

MMPs and genetic instability



ELSEVIER

Genomes and evolution The power of comparative genomics

Editorial overview

Stephen J O'Brien and Claire M Fraser

Current Opinion in Genetics & Development 2005,
15:569–571

0959-437X/\$ – see front matter
© 2005 Elsevier Ltd. All rights reserved.

DOI 10.1016/j.gde.2005.10.001

Stephen J O'Brien

Laboratory of Genomic Diversity, National Cancer Institute, Frederick, MD 21702, USA
e-mail: obrien@ncifcrf.gov

Stephen J O'Brien is Chief of the Laboratory of Genomic Diversity at the National Cancer Institute in Frederick, Maryland, USA. He researches human genes that regulate response to infectious diseases, particularly AIDS, hepatitis and cancer. He also studies the evolution of mammalian genome organization, though studies of comparative genomics of the domestic cat, their wild felid relatives, and their infectious disease agents.

Claire M Fraser

The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA
e-mail: cmfraser@tigr.org

Claire M Fraser is the President and Director of The Institute for Genomic Research. Her research interests include the application of genomics-based approaches to the study of microbial diversity and evolution, and the development of novel antimicrobials and vaccines.

Introduction

The past ten years has seen remarkable progress in the sequencing and analysis of genomes, with more than 300 sequences now available from a broad representation of species across the phylogenetic tree (www.genomesonline.org). Approximately 80% of these genome sequences are from bacterial species; however, there is also a large number of archaeal and eukaryotic sequences included in this vast collection of data. While tremendous biological insights on any given organism can be derived from analysis of a single genome sequence, comparative analysis of multiple genomes reveals substantially more information on the biology and evolution of species. Moreover, this information provides a new starting point for biological investigations using technologies that enable genome-wide approaches to the study of gene and protein function. This issue of *Current Opinion in Genetics & Development* provides an overview of how comparative analyses have revealed new insights on gene, genome and species evolution in a variety of prokaryotes and eukaryotes.

Evolution of prokaryotic genomes

One of the insights to come from bacterial sequencing efforts is the fact that the bacterial genome is a dynamic entity shaped by multiple forces that include genome reduction, genome rearrangement, gene duplication, and gene loss, and gene acquisition by lateral gene transfer. An interesting bacterial group that has been targeted for genome analysis represents species that are no longer free-living and, because of genome reduction, are now dependent on their hosts for survival. These include obligate intracellular pathogens, endosymbionts, and mutualistic species. The review by [Wernegreen](#) summarizes the current thinking on reductive evolution and points out that the initial view of this process, in which minimal organisms were experiencing continual gene loss and little if any gene acquisition or recombination, should be re-examined on the basis of recent data for the Rickettsiales and Chlamydiales.

In contrast to intracellular bacteria, free-living bacterial pathogens emerge and adapt by continual changes in genome structure and gene content. The review by [Lawrence](#) describes the impact of gene loss and gene gain in the evolution of bacterial and unicellular eukaryotic pathogens. Early views that these processes represented the random acquisition of novel virulence genes and the loss of non-essential genes are probably too simplistic and need to be revised.

One of the important insights from comparative microbial genomics is that lateral gene transfer plays a more significant role in the evolution of species

diversity than was initially appreciated. This realization has been derived, largely, from the analysis of multiple isolates of the same species, which has revealed that, in some instances, members of the same bacterial species can differ in gene content by as much as 25–30%. This leads to the question of whether or not it is possible to fully identify all of the genes associated with a given species. In their review, **Medini et al.** define the microbial pan-genome as the sum of the core and non-essential genes for any given species. On the basis of large-scale analysis of multiple isolates of the human pathogen *Streptococcus agalactiae*, they demonstrate that sequencing of each new genome reveals novel genes. Mathematical modeling of these data suggests that the *S. agalactiae* pan-genome might be infinite in size.

The dynamic nature of the microbial genome has raised the question as to whether or not it is possible to analyze the history of entire genomes and construct meaningful phylogenetic trees. **Ouzounis** summarizes a number of approaches that are being used to tackle this problem and provides a framework for reconstructing the ancestral state of microbial genomes.

Unicellular eukaryotic genomes

When compared with bacterial genomes, most eukaryotic genomes are large and gene-poor. However, the smallest eukaryotic-genomes — those of the microsporidia and nucleomorphs — represent exceptions to this rule. **Keeling and Slamovits** discuss both the causes and the effects of nuclear genome reduction in eukaryotes, and the unique features of such hyper-compacted genomes.

Our understanding of the biology of malaria parasites has been considerably enhanced by the genome sequencing and functional analysis of four species of *Plasmodium* parasites. **Hall and Carlton** review the properties of these genomes and emphasize that one of the major differences among *Plasmodium* species is in the repertoire of genes encoding proteins involved in the interaction with the host immune system. Moreover, they present evidence that suggests that *Plasmodium* is an ancient parasite and that *Plasmodium vivax*, a human malaria parasite, might have arisen through mutations that resulted in a change in host specificity.

The hemiascomycetous yeasts, which include the model organism *Saccharomyces cerevisiae*, cover a very large evolutionary range. Large-scale comparative analysis of this class of fungi has been completed, and **Dujon** summarizes how these data have revealed ways in which gene and genome duplication, chromosomal rearrangements, and gene loss have shaped the evolution of these species.

Comparative genomics of eukaryotes

Despite their large genome sizes, significant progress has been made in the comparative genomics of eukaryotes in

the past several years. The availability of multiple plant and mammalian genome sequences has begun to reveal the extraordinary complexities of these species, but at the same time several recurrent themes have also emerged.

Most of the large variation observed in the size of plant nuclear genomes can be attributed to the differential expansion or retention of retrotransposons. **Bennetzen** describes the role of transposable elements both in the genome rearrangement of flowering plants and in the creation of novel genes — in particular, the contribution of transposon capture and exon-shuffling to the appearance of new genes.

For the vertebrates, including the 4500–5000 mammalian species that walk the earth today, comparative genomics is driven by the desire to annotate the newly completed human genome — approximately 2.8 billion base pairs in length. Numerous uncertainties surrounding gene identification, genome organization, replication, regulation, noncoding sequence, the sea of repeats, and footprints of evolutionary adaptation have stimulated the whole genome sequencing of a score of mammals. Species approved and funded for high-coverage (greater than 7x) whole genome sequencing include human, mouse, rat, chimpanzee, macaque, dog, cattle, marsupial opossum, and monotreme platypus. The recent molecular phylogeny of placental mammals [1] has provided a framework to select species for lower-coverage (2x) genome sequencing on the basis that their genome sequences would capture the depth of genome diversity created by the mammalian radiations. To date, 13 additional species have been selected for 2x whole genome sequence assessment by The National Institutes of Health–National Genome Research Institute: elephant, hyrax, tenrec, armadillo, tree shrew, bushbaby, rabbit, guinea pig, squirrel, common shrew, hedgehog, horseshoe bat, and domestic cat. The genome sequences of all these species will be online by March 2006.

Among the many results anticipated from the whole genome assessment of vertebrate genomes is the characterization of non-genic regulatory elements revealed by short stretches of DNA sequences that remain highly conserved among genomes of species from diverse mammalian orders (e.g. in comparisons of either human versus mouse genomes or dog versus platypus genomes). In his review, **Adams** highlights the importance of conserved sequence segments and the approaches used to identify and validate them as regulatory elements. **Baggs et al.** present an authoritative review of the application of comparative inference to elucidate the transcriptional regulation of genes and non-genic regions.

More than half of the human genome and of most mammal genomes are composed of repetitive elements of different size, abundance and degrees of polymorphism. Most repeats derive by invasion from other species

but can have important consequences for hereditary disease, phenotypes and also genomic utility (e.g. for gene mapping, forensics or population genetic structure assessment). **Thomas** discusses the dispersal of short repeats (minisatellites, microsatellites and doublets) in mammals, with emphasis on the disease causation in man, dogs, voles and other fascinating examples.

Genome evolution comprises a mix of co-evolving autosomes, mitochondrial plasmids, and sex chromosome (X and Y in mammals). **Fraser and Heitman** tell the story of the birth of the mammalian Y chromosome some 300 million years ago and trace the tortuous lineage that led to today's 50–60Mb Y chromosome, a hall of mirrors with massive palindromes engulfing a few score functional genes that are operative in man, mouse and many — but not all — living mammals. The authors reach back to fish, plants and fungi to review what we can discern from comparative inference of this sex chromosome's hapless evolutionary adventures. In their commentary, **Modi and Crews** assess the mysterious sex chromosomes and sex-determining mechanisms discovered in reptilian species. They raise a cogent case for genome sequencing of selected reptile species (i.e. green anole lizard, American alligator, garter snake, and turtle). The comparative analysis of sex-determining mechanisms in higher vertebrates — including the still poorly understood temperature-dependent sex determination in many reptile species — plus the knowledge that reptiles have dominated one of the three major geological eras of our planet since the dawn of life (trilobite invertebrates and

mammals dominate the other two) are reasons enough to consider their genome sequencing important.

The review by **Roca and O'Brien** deals with African elephants — in phylogenetic terms, the most basal species of placental mammals, and taxa with much to reveal, biologically, from genomic assessment. Applications of genomic technologies recently unearthed a cryptic African elephant species and illustrated the disjunctive evolutionary transmission-kinetics of nuclear, mitochondrial, X and Y chromosomes in newly sympatric species. Elephants are fascinating to humankind and are studied intensely for behavioral and conservational concerns. The authors argue that elephants, with their sophisticated intelligence and large brains, might also pose a plausible neuroscience subject among mammals.

The promises of comparative genomics are many and go way beyond the scope of this volume. Nonetheless, some of the highlights of the new enquiries can be gleaned from these reviews. The time for testing and validating uncounted biological hypotheses in a genomic context has come; and the advances revealed will fundamentally change our understanding of the past, the present and the future.

References

1. Murphy WJ, Eizirik E, O'Brien SJ, Madsen O, Scally M, Douady CJ, Teeling E, Ryder OA, Stanhope MJ, de Jong WW, Springer MS: **Resolution of the early placental mammal radiation using Bayesian phylogenetics**. *Science* 2001, **294**:2348-2351.

For better or worse: genomic consequences of intracellular mutualism and parasitism

Jennifer J Wernegreen

Bacteria that replicate within eukaryotic host cells include a variety of pathogenic and mutualistic species. Early genome data for these intracellular associates suggested they experience continual gene loss, little if any gene acquisition, and minimal recombination in small, isolated populations. This view of reductive evolution is itself evolving as new genome sequences clarify mechanisms and outcomes of diverse intracellular associations. Recently sequenced genomes have confirmed a trajectory of gene loss and exceptional genome stability in long-term, nutritional mutualists and certain pathogens. However, new genome data for the Rickettsiales and Chlamydiales indicate more repeated DNA, a greater abundance of mobile DNA elements, and more labile genome dynamics than previously suspected for ancient intracellular lineages. Surprising discoveries of conjugation machinery in the parasite *Rickettsia felis* and the amoebae symbiont *Parachlamydia* sp. suggest that DNA transfer might play key roles in some intracellular taxa.

Addresses

Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, 7 MBL Street, Woods Hole, MA 02543, USA

Corresponding author: Wernegreen, Jennifer J (jwernegreen@mbl.edu)

Current Opinion in Genetics & Development 2005, 15:572–583

This review comes from a themed issue on
Genomes and evolution
Edited by Stephen J O'Brien and Clarie M Fraser

Available online 17th October 2005

0959-437X/\$ – see front matter
© 2005 Elsevier Ltd. All rights reserved.

DOI 10.1016/j.gde.2005.09.013

Introduction

In bacterial evolution, the transition from a free-living existence to a close relationship with eukaryotic cells represents a frequent theme. Certain bacterial symbionts have taken such associations to the extreme by completely abandoning any semblance of a free-living phase and replicating solely within the domain of a host cell. Throughout the history of life, these obligately intracellular bacteria have acted as major evolutionary catalysts, being involved in the origin of organelles and the diversification of eukaryotes. Present-day intracellular associations include a range of parasites, mutualists and commensal symbionts that play important roles in the

ecology and physiology of their hosts (see Glossary for many of the terms mentioned in the Introduction) [1].

Owing to their medical and ecological importance, intracellular bacteria have been targets of numerous genome sequencing projects that have provided insights into the consequences of this specialized lifestyle (Table 1; Box 1). We have learned that, typically, these species have drastically reduced genomes that encode a streamlined metabolism, show rapid DNA sequence evolution and strong nucleotide compositional biases, and exhibit lower levels of genome flux (i.e. gene acquisition from foreign sources, and intragenomic changes such as inversions and translocations). The integration of population genetic processes with knowledge of bacterial physiology and ecology has helped to clarify mechanisms that might explain these common features.

In particular, current views of 'reductive evolution' emphasize that fundamental evolutionary processes — natural selection, mutation and genetic drift — might affect intracellular species differently than they do free-living ones [2,3]. For instance, genome streamlining might reflect relaxed purifying selection on metabolic functions that are dispensable in a resource-rich intracellular niche. In addition, strong effects of nucleotide mutations in intracellular bacteria might elevate rates of gene disruption, followed by erosion owing to a deletion bias in bacteria [4]. Many intracellular endosymbionts show few if any signs of gene acquisition. This is thought to reflect their generally low levels of repeats and mobile DNA, reduced recombination functions, and limited opportunities for DNA exchange among sequestered species [5]. Moreover, reduction of effective population sizes (N_e) owing to bottlenecks upon transmission [6] is expected to increase the rates of fixation of slightly deleterious mutations [7]. Lack of gene exchange would exacerbate this effect by preventing the recovery of beneficial alleles or entire gene regions that are lost [3,8].

This reductive evolution model offers a valuable framework to explain commonalities among intracellular bacteria, to identify informative exceptions, and to generate predictions that can be tested with new sequence data. The abundance of excellent reviews on this topic illustrates the utility of this conceptual framework in assimilating a wealth of new genome information and in guiding development of the field [9–15]. In this review, I discuss insights from recent — between 2004 and July 2005 — genome analyses of obligately intracellular bacteria that replicate solely within a host cell. These data

Glossary

Commensal symbiont: A symbiont that benefits from an association without conferring a serious disadvantage or advantage to the host.

Genetic drift: This describes the changes in the frequencies of alleles or genotypes as a result of chance alone. This stochastic effect plays an especially important role in small populations, in which drift can accelerate the fixation of slightly deleterious mutations [7].

Genome flux: A broad term describing changes in gene content or order owing to gene acquisition by horizontal transfer from foreign donors or recombination among related strains or species. This also includes intragenomic changes within a given genome, such as inversions, duplications, translocations and deletions.

Mobile DNA: Elements such as phage DNA, transposons, conjugative plasmids, insertion sequences and other DNA segments that move among or within genomes, typically without the need of extensive DNA sequence matches for homologous recombination. Often considered as selfish DNA that propagates at the expense of hosts and depends on occasional horizontal transmission for its maintenance.

Mutualist: A symbiont that provides a benefit to the host and, in turn, benefits from the association.

Obligately intracellular: An organism that replicates exclusively within a host cell.

Parasite: A symbiont that propagates by causing some degree of harm to the host.

Reductive evolution: A conceptual framework that considers the evolutionary and molecular mechanisms that drive genome streamlining in most intracellular bacteria. Current views suggest that gene loss reflects relaxed selection on dispensable traits, elevated mutation pressure, and even the loss of beneficial functions mutations as a result of genetic drift in small bacterial populations. Furthermore, reduced recombination documented in some intracellular associates might prevent the recovery of lost alleles or gene regions.

Symbiont: Any species that lives in close association with another. Broadly speaking, symbiosis includes obligate and facultative relationships that are parasitic, mutualistic or commensal. Among symbionts, endosymbionts are those that live within the tissues or cells of their hosts for part or all of their life cycles. Endosymbionts that can replicate within host cells are termed intracellular. Of these intracellular associates, certain highly specialized ones have lost the ability to replicate outside of host cells and are obligately intracellular — the focus of this review.

Type III secretion: An assemblage of ~20 proteins that spans the cell membrane, transports proteins out of the cell and mediates the delivery of specific proteins that suppress defenses or otherwise facilitate cell invasion.

Type IV secretion: Derived repeatedly from conjugation systems [83**], this is a secretion pathway that exports distinct DNA or protein substrates that cause various physiological changes in host cells during infection.

enable us to explore expectations of reductive evolution models — namely, that intracellular bacterial genomes are (i) severely reduced, (ii) specialized to their particular host association, and (iii) show patterns of genome dynamics that differ from those of free-living species.

Genome size reduction

The influx of genome data supports the general trend that strictly intracellular bacteria have very reduced genomes, typically in the range of 1 Mb or less (Figure 1). At 416 kb, the tiny genome of *Buchnera* BCc, associated with the cedar aphid *Cinara cedri*, is the smallest known for bacteria (A Latorre, unpublished). As expected from their small genomes, metabolisms of intracellular bacteria are

Box 1 Diverse lifestyles and host effects of intracellular bacteria.

Bacteria ‘make their living’ within host cells through a variety of strategies. At one end of the spectrum, intracellular parasites represent unwelcome invaders that spread at their hosts’ expense and offer models to understand how bacteria exploit cellular functions. Fully sequenced representatives (Table 1) include *Mycobacterium leprae* and *Coxiella burnetii*, which have adopted an obligately intracellular lifestyle quite recently, and older intracellular lineages such as *Phytoplasma*, a plant parasite and close relative of the epicellular *Mycoplasmas*, in addition to parasites within the families Rickettsiaceae and Anaplasmataceae. In addition to vertebrate pathogens, Anaplasmataceae includes the invertebrate endosymbiont *Wolbachia*, typically a parasite in insects that hijacks host reproduction to increase the production of infected females. Given that other members of *Wolbachia* are mutualists involved in development and oogenesis in nearly all filarial nematodes, this genus shows a natural lifestyle variation that facilitates comparisons between mutualists and parasites.

The exclusive intracellularity across known Rickettsiaceae and Anaplasmataceae implies that they adopted this lifestyle at least 400 mya, a conservative estimate of divergence between the two families [27]. The even more ancient Chlamydiales acquired an intracellular lifestyle ~700 mya [28**], and today include parasitic members of the Chlamydiaceae that cause respiratory, ocular and genital infections, in addition to the recently sequenced *Parachlamydia* sp., a mutualistic symbiont of free-living amoebae and occasional opportunistic pathogen of humans. Although they share an obligately intracellular lifestyle, the wide diversity in tissue tropism, host ranges, life-history nuances, and phenotypic effects of intracellular parasites remain poorly understood.

Infections don’t always turn out poorly for the host. At the other end of the symbiotic spectrum, mutualists provide benefits that increase the fitness of their hosts. In addition to the nematode and amoebae hosts mentioned above, insects as a group frequently associate with beneficial intracellular bacteria. These symbionts occur within specialized host cells, undergo maternal transmission to offspring and have co-evolved with hosts in stable associations that date back tens to hundreds of millions of years [69]. For example, estimated divergence times of host insects imply that the *Blochmannia*–ant and *Wigglesworthia*–tsetse associations originated at least 30 mya, and the even older *Buchnera*–aphid symbiosis was established ~150–200 mya. Such endosymbionts provide essential nutrients to about 10–15% of insects, most of which feed on nutritionally unbalanced diets (e.g. plant sap or blood). By enabling their hosts to exploit otherwise inadequate food sources and habitats, the acquisition of these mutualists can be viewed as a key innovation in the evolution of their hosts [70,71]. These ‘primary’ endosymbionts often co-exist with *Wolbachia* and facultative (‘secondary’) endosymbionts [72,73**]. Mutualistic symbionts and reproductive parasites might offer new tools for the modification, suppression, containment or eradication of arthropod populations of medical and/or agricultural importance (e.g. [74–76,86**]).

simpler than those of free-living or facultatively intracellular species with larger genomes. This reduction often involves massive deletions early in the transition to intracellularity (see Update). For example, reconstructions of early *Buchnera* evolution indicate deletions of large DNA segments, some of which encoded 20 open reading frames (ORFs) or more, in addition to exceptionally high levels of gene rearrangements compared with other γ -Proteobacteria [16,17*]. Sequence data from facultatively or recently intracellular species provide a window into these

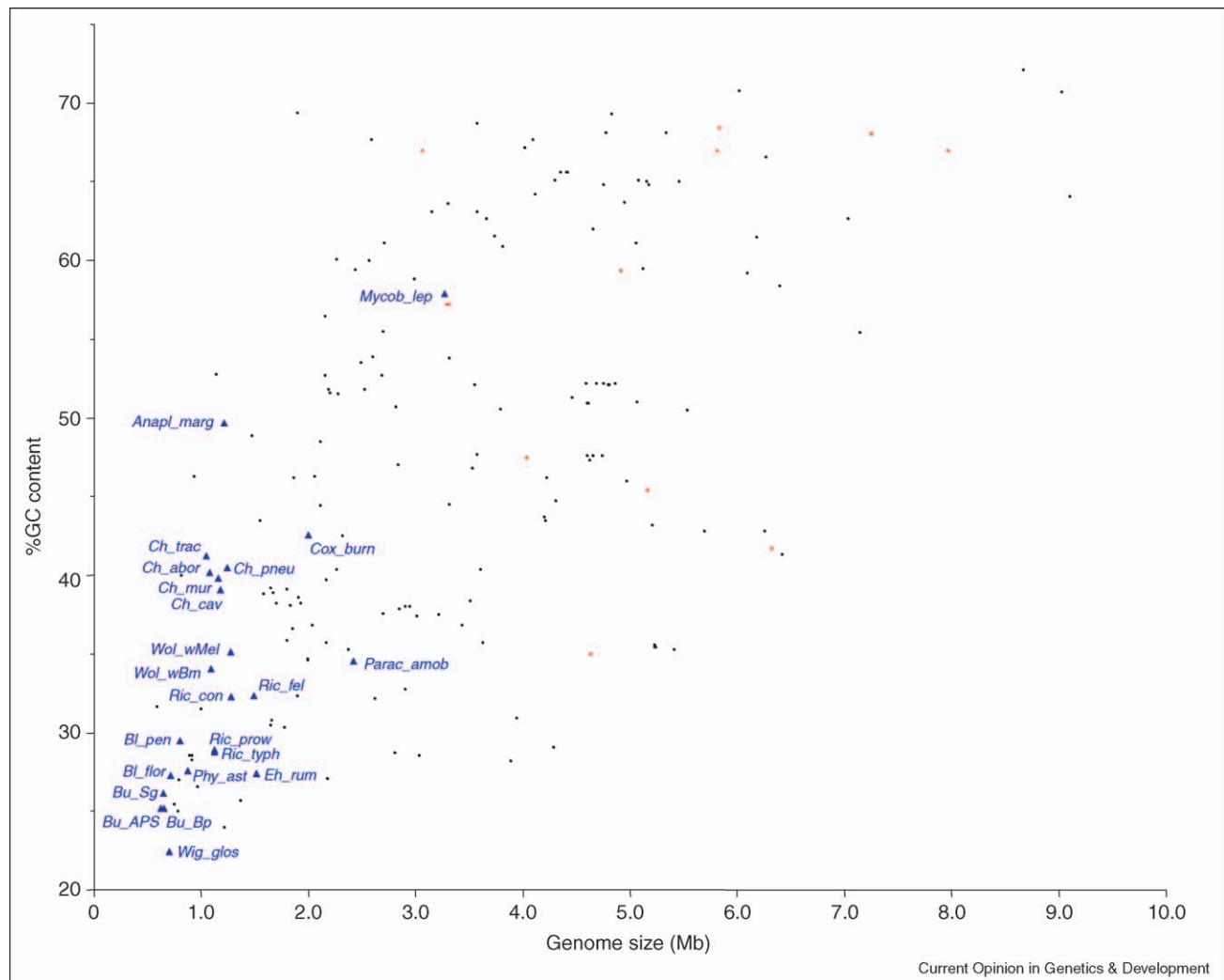
Table 1

Obligately intracellular bacteria with full genome sequence data (as of July 2005) or for which genome projects are in progress.

	Genome size (Mb)	Genome released	Host	Host effects
γ-Proteobacteria				
Enterobacteriales				
* <i>Buchnera aphidicola</i> APS	0.66	2000	Aphid, <i>Acyrtosiphon pisum</i>	Nutritional mutualist
* <i>Buchnera aphidicola</i> Sg	0.64	2002	Aphid, <i>Schizaphis graminum</i>	Nutritional mutualist
* <i>Buchnera aphidicola</i> Bp	0.62	2003	Aphid, <i>Baizongia pistaciae</i>	Nutritional mutualist
* <i>Buchnera aphidicola</i> BCc	0.42 ^a	In progress, University of Valencia	Aphid, <i>Cinara cedri</i>	Nutritional mutualist
* <i>Wigglesworthia glossinidia</i>	0.7	2002	Tsetse fly, <i>Glossina brevipalpis</i>	Nutritional mutualist
* <i>Blochmannia floridanus</i>	0.71	2003	Ant, <i>Camponotus floridanus</i>	Nutritional mutualist
* <i>Blochmannia pennsylvanicus</i>	0.79	2005	Ant, <i>Camponotus pennsylvanicus</i>	Nutritional mutualist
* <i>Baumannia cicadellincola</i>	0.69 ^b	In progress, University of Arizona and TIGR	Sharp shooter, <i>Homalodisca coagulata</i>	Likely nutritional mutualist
Legionellales				
<i>Coxiella burnetii</i>	2.03	2003	Reptiles, birds, and mammals	Q fever
α-Proteobacteria				
Rickettsiales				
Rickettsiaceae				
<i>Rickettsia conorii</i>	1.27	2000	Mammals, through insect vectors	Rocky Mountain spotted fever
<i>Rickettsia prowazekii</i>	1.11	1998	Mammals, through insect vectors	Typhus
<i>Rickettsia typhi</i>	1.11	2003	Mammals, through insect vectors	Murine typhus
<i>Rickettsia felis</i>	1.46	2005	Mammals, through insect vectors	Spotted fever
Anaplasmataceae				
<i>Anaplasma marginale</i>	1.2	2004	Mammals, through insect vectors	Bovine anaplasmosis, human granulocytic ehrlichiosis
<i>Ehrlichia ruminantium</i> (2 strains)	1.5–1.52	2005	Wild ruminants, through tick host	Heartworm disease
<i>Wolbachia</i> wMel	1.27	2004	Fruit fly, <i>Drosophila melanogaster</i>	Cytoplasmic incompatibility
* <i>Wolbachia</i> wBm	1.08	2005	Filarial nematode, <i>Brugia malayi</i>	Worm development and fertility
<i>Wolbachia</i> wAna		2005 (95% of genome recovered from Trace Archive, [62])	Fruit fly, <i>Drosophila ananassae</i>	Cytoplasmic incompatibility
<i>Wolbachia</i> wRi		2005 (75–80% of genome recovered from Trace Archive [62])	Fruit fly, <i>Drosophila simulans</i>	Cytoplasmic incompatibility
<i>Wolbachia</i> wUni		In progress, EUWOL (European <i>Wolbachia</i> Consortium)	Parasitoid wasp, <i>Muscidifurax uniraptor</i>	Induction of parthenogenesis
<i>Wolbachia</i> wVul	1.6–1.7	In progress, EUWOL	Isopod, <i>Armadillidium vulgare</i>	Induction of feminization
<i>Wolbachia</i> wRi	1.5–1.6	In progress, EUWOL	Fruit fly, <i>Drosophila simulans</i>	Cytoplasmic incompatibility
<i>Wolbachia</i> – <i>Culex</i>	~1.5	In progress (http://www.sanger.ac.uk/Projects/W_pipientis/)	Mosquito, <i>Culex quinquefasciatus</i>	Cytoplasmic incompatibility
* <i>Wolbachia</i> – <i>Onchocerca</i>	~1.1	In progress (http://www.sanger.ac.uk/Projects/Wolbachia/)	Nematode, <i>Onchocerca volvulus</i>	Worm development and fertility
Mollicutes				
Acholeplasmatales				
<i>Phytoplasma asteris</i>	0.86	2003	Plants, through insect vector	Stunted plant growth and other symptoms
Actinobacteria				
Actinomycetales				
<i>Mycobacterium leprae</i>	3.27	2001	Humans and other vertebrates	Leprosy
Chlamydiae group				
Chlamydiales				
* <i>Parachlamydia</i> sp.	2.41	2004	Free-living amoebae	
Chlamydiaceae				
<i>Chlamydia muridarum</i>	1.08	2000	Rodents	Mouse lung or genital tract infections
<i>Chlamydia trachomatis</i>	1.04	1998	Human and other mammals	Chronic genital and ocular infections
<i>Chlamydophila abortus</i>	1.14	2005	Ruminants and swine	Ruminant abortion
<i>Chlamydophila caviae</i>	1.18	2003	Guinea pig	Guinea pig inclusion conjunctivitis (GPIC)
<i>Chlamydophila pneumoniae</i> (4 strains)	1.23	1999–2003	Human and other mammals	Pneumonia, bronchitis and pharyngitis

The list of genomes for which sequencing is in progress is intended to illustrate the rapid growth of this data and is not exhaustive. * Mutualistic association. ^a Updated from Gil *et al.* [84]. ^b Updated from Moran *et al.* [85]. Initials after endosymbiont strains often refer to the invertebrate host species from which the bacteria were isolated.

Figure 1



Genome size and %GC content of bacterial chromosome sequences, illustrating the small genome size and AT-richness of obligate intracellular associates. Intended as an update to similar published figures (e.g. [77]), this graph includes genomes that were publicly available as of July 2005. Genomes of multiple, closely related strains are presented with a single point. Blue triangles represent obligately intracellular species. Red points represent those species that possess two or more chromosomes (the mark reflects values for single chromosomes).

early stages of genome turbulence, when the proliferation of insertion sequences (ISs) and other mobile DNA elements might catalyze instability [9]. In addition, the larger genomes, numerous pseudogenes, and/or dispersed repeats in recent associates (e.g. *Coxiella burnetii*, *Mycobacterium leprae* and the *Sitophilus oryzae* [weevil] primary endosymbiont [SOPE]) also support an initial instability [18,19,20**].

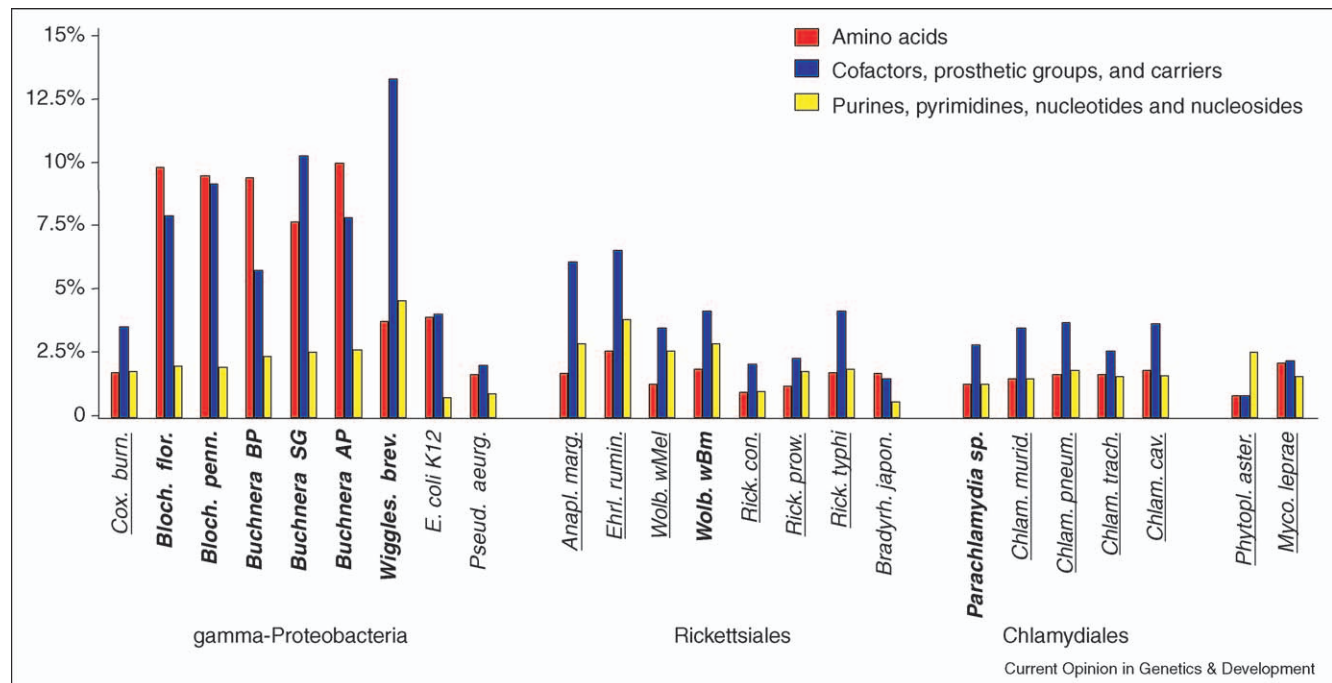
Genome size variation within endosymbiont groups indicates that streamlining has continued in the context of intracellular associations but in a much more gradual fashion. For example, in contrast to large early deletions, subsequent gene loss in *Buchnera* has tended to occur through gene disruption and gradual erosion, with inacti-

vated genes requiring ~40–60 million years to erode completely [21•]. This continued genome shrinkage leads to further metabolic loss. For instance, reduction in *Buchnera* BCc is caused by the loss of protein-coding genes, in comparison with the numbers in other *Buchnera* species, and not by the shortening of ORFs or intergenic regions [22**].

Specialization to the intracellular niche

What are the metabolic implications of severe genome reduction? Consistent with the prediction that many deleted genes were dispensable in a host cell, small intracellular genomes tend to lose genes for metabolic diversity but retain those encoding transcription, translation and other basic processes that are important regard-

Figure 2



Percentage of genes encoding particular biosynthetic functions. Y axis indicates proportion of ORFs involved in biosynthesis of (i) amino acids (red), (ii) cofactors, prosthetic groups, and carriers (blue), and (iii) purines, pyrimidines, nucleotides and nucleosides (yellow). Species in bold are obligately intracellular mutualists; those underlined are obligately intracellular pathogens. *Escherichia coli*, *Pseudomonas aeruginosa* and *Bradyrhizobium japonicum* retain a free-living phase and are included for comparison. Full names of other bacteria are listed in Table 1. Values for *E. coli* and the γ -proteobacterial nutritional mutualists were based on re-analysis of genome sequences [37]. Data for other species was downloaded from the Comprehensive Microbial Resource at The Institute for Genomic Research (TIGR, <http://www.tigr.org/>) [78], for a more consistent comparison across genomes, but might differ from original genome papers. Readers interested in particular taxa are encouraged to refer to the original genome-publications. In the rare cases that TIGR counted largely overlapping, putative ORFs as two separate genes, these were counted as a single ORF for the purposes of this figure.

less of ecological niche [2]. The preferential retention of informational genes also holds within endosymbiont groups. Analysis of partial genome regions indicates that, compared with its *Buchnera* relatives, *Buchnera BCc* has undergone a more extensive loss of metabolic than informational functions [22••].

The nutrient trade balance

Although all parasites and mutualists rely on their host for certain nutrients, they differ in the degree of their dependency. As expected from their roles in supplementing the diet of the host, primary, γ -proteobacterial mutualists of insects retain a wide spectrum of biosynthetic genes to fulfill symbiont functions, devoting a higher fraction of their genomes to biosynthesis than do free-living bacteria or pathogens (Figure 2) [11,15]. By contrast, intracellular parasites apparently rely on their eukaryotic host cell for many amino acids, cofactors, nucleotides and other compounds.

Though more subtle, the same contrast holds among species of the Rickettsiales and Chlamydiales orders, in which mutualists encode a wider array of biosynthetic

functions than do parasites. The relatively large (2.41 Mb) genome of *Parachlamydia* complicates direct comparison of genome proportions with its ~1–1.2 Mb parasitic relatives, but its retention of twice as many amino acid and cofactor biosynthetic genes suggests it imposes fewer metabolic demands on its host cell. Within Rickettsiales, the mutualistic *Wolbachia* wBm, unlike *Rickettsia*, retains the ability to synthesize riboflavin and other coenzymes [23••]. Biosynthesis of riboflavin and heme might be key functions of the symbiont, because, to date, neither pathway has been detected in the *Brugia malayi* genome [24]. *Wolbachia* wBm also retains complete pathways for biosynthesis of purines and pyrimidines, and might supplement the nematode's nucleotide pools during oogenesis and embryogenesis [23••]. Although the parasitic *Wolbachia* wMel shares many of these same functions, the nematode mutualist devotes a larger fraction of its smaller genome to these potentially host-beneficial traits.

Shared infection strategies of mutualists and pathogens

One of the surprises from molecular studies of host-associated bacteria has been the discovery of 'virulence'

genes in mutualists [25,26]. As expected, obligately intracellular parasites typically encode numerous mechanisms to infect various tissue and cell types and to evade an ever-adapting host immune system. These mechanisms include Type III secretion (in Chlamydiales), Type IV secretion (*Coxiella* and Rickettsiales) [see Glossary], and paralogous families of polymorphic surface proteins (e.g. in *R. felis*, *Anaplasma* and *Ehrlichia*) [27].

Certain obligately intracellular mutualists also possess such so-called pathogenicity genes. The discovery of Type III secretion in *Parachlamydia* highlights parallels with related parasites and implies that the ancestor of this group could infect cells [28**]. In a similar manner to certain insect endosymbionts [29], this amoebae associate might deliver specific proteins into host cells to facilitate invasion. This exciting discovery pushes the origin of infectious chlamydia back to ~700 million years ago (mya), arguably the oldest intracellular group that today spans diverse host associations and effects [30,31*]. In a similar fashion to its pathogenic relatives, *Parachlamydia* imports ATP from the host cytosol using an ATP/ADP translocase, an ‘energy-parasite’ transport system unique to Rickettsiales, Chlamydiales, and plant plastids [32*]. *Parachlamydia* apparently acquired Type IV secretion through horizontal gene transfer, making it the only chlamydia to possess this pathway [28**].

Wolbachia spp. might also use common infection strategies across diverse interaction types. As with *Wolbachia* wMel and other parasitic α -Proteobacteria, *Wolbachia* wBm possesses Type IV secretion in its stable mutualistic association with nematodes [23**]. The shared presence of ankyrin repeats in both *Wolbachia* genomes — although there are fewer in *Wolbachia* wBm — might mediate attachment of endosymbionts to the cytoskeleton, modulation of host gene expression, or other activities essential to their intracellular lifestyle. The roles of these and other *Wolbachia* genes in shaping its varied host interactions are promising new areas of research [33,34] (see Update).

Even the long-term insect mutualists possess genes once considered to be in the purview of parasitism. The urease gene cluster, which encodes significant virulence factors in some bacterial and fungal pathogens [35], is retained in *Blochmannia*, apparently enabling this ant symbiont to hydrolyze insect host waste (urea) to ammonia used in amino acid biosynthesis [36,37]. The overexpression of GroEL in certain insect mutualists [38] also occurs in some pathogens, in which this and other heat-shock proteins are major antigens and candidates for vaccine development [39–41]. The functions of GroEL in mutualists remain uncertain, but it might help the folding of endosymbiont proteins that have experienced deleterious amino acid changes [42], mediate viral transmission through certain insect hosts [43], or counter irreversible

oxidative modifications during stationary phase [44], in which endosymbionts probably spend much of their life cycle. Other parallels with pathogens include the retention of outer membrane proteins and flagellar genes, the latter of which might be involved in secretion [45]. These parallels suggest that ‘pathogenicity’ functions might have evolved in the context of beneficial interactions and today play generally important roles for host-associated bacteria.

Deleterious deletions

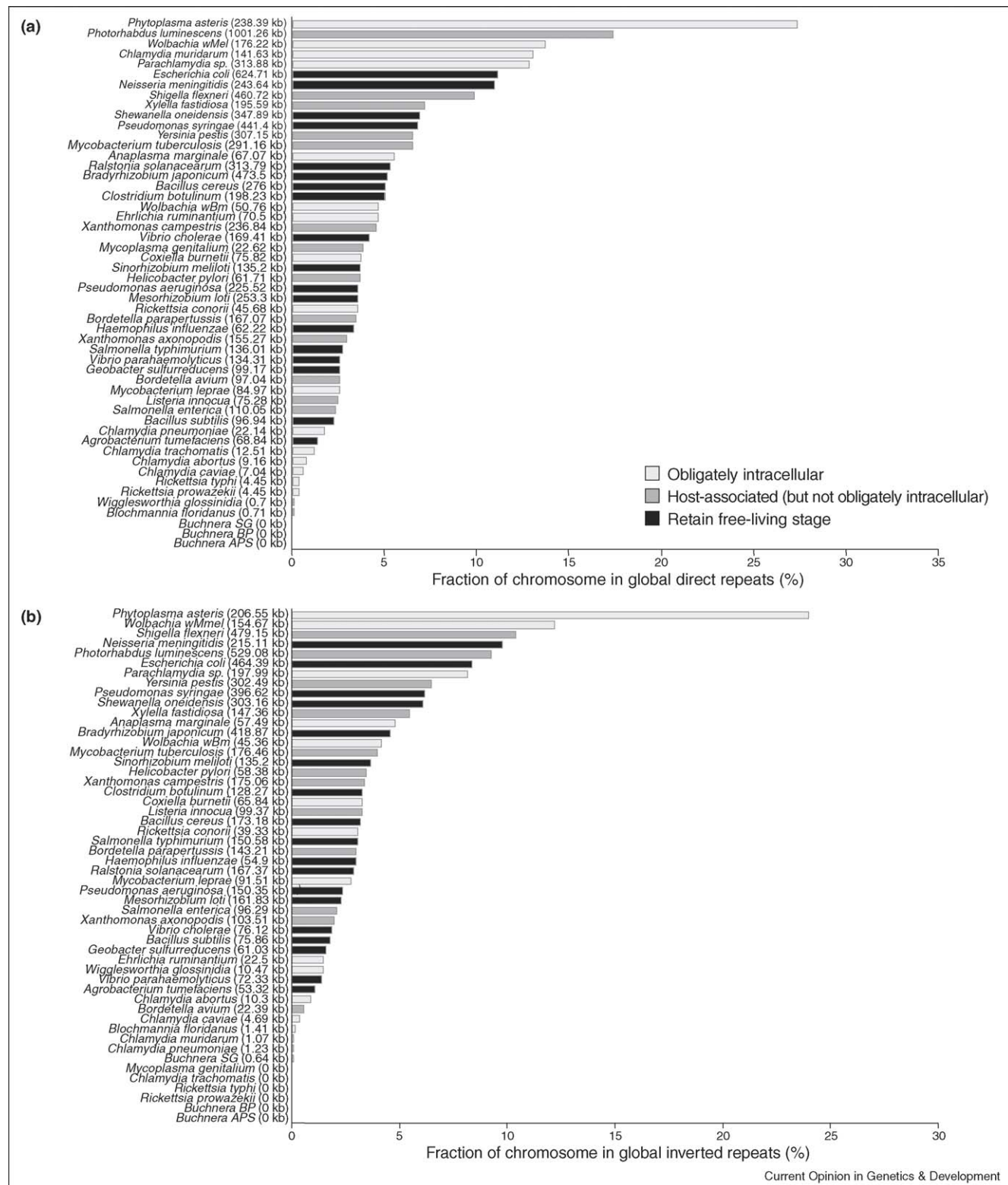
In contrast to the above examples of adaptation to an intracellular lifestyle, it is more difficult to infer cases of deleterious deletions that are fixed by genetic drift, another component of the reductive evolution model. Candidates for harmful deletions include the loss of several DNA repair functions, a convergent pattern across intracellular pathogens and mutualists. Notably, in *Buchnera* many DNA repair loci were lost in large, early deletions that included numerous genes of varied functions, a pattern that is difficult to reconcile with adaptive fine-tuning of gene content [17*]. The loss of repair functions might contribute to the drop in %GC content associated with genome reduction (Figure 1). Namely, AT bias might reflect greater exposure of an underlying GC→AT mutational pressure in small genomes that lack many DNA repair functions (but see the study by Rocha and Danchin [46] for a metabolic hypothesis for AT bias).

Genome dynamics within a host cell

As noted above, initial genome turbulence of intracellularly is thought to be followed by genetic stability associated with the consumption of recombination genes and repeated DNA in large deletions, reduced opportunities for gene exchange in a sequestered environment, and extinction of mobile DNA species that require horizontal transmission for their maintenance.

Is recent genome data consistent with genetic stability of long-term intracellular associates? In many cases, the answer is a resounding yes. Although horizontal gene transfer can be important in the evolution of new host associations [47], strictly intracellular associates often show little evidence of laterally acquired genes. In *Buchnera*, this stability extends to intragenomic dynamics, with no gene acquisition, inversions or translocations throughout 50–70 million years of evolution within aphids [48], and near-perfect conservation of gene order since the establishment of this association 150–200 mya [49] (bio-synthetic plasmids have mediated the few exceptions to genome stability in this group [50,51*]). Within the ant mutualist *Blochmannia*, lineages that diverged ~20 mya exhibit a similar pattern of chromosome stasis [37]. Such stability also characterizes certain long-term intracellular pathogens, for which rare cases of horizontal transfer (e.g. a potentially acquired 12 kb sequence in *Rickettsia typhi* [52]) appear to be the exceptions, synteny between

Figure 3



Fraction of bacterial chromosomes devoted to (a) global direct repeats and (b) global inverted repeats. Although certain obligately intracellular bacteria have few if any repeats, others have repeat densities that match or exceed those of facultatively intracellular or free-living species. Values represent only long repeats (≥ 100 nucleotides long, with at least an 80% match among copies), a category that mediates large-scale inversions, duplications and deletions. Values were obtained from the CBS (Centre for Biological Sequence Analysis) Genome Atlas Database

species implies few intragenomic rearrangements [48,53], and repeated DNA is often scarce (Figure 3) [5].

However, recent data indicate that DNA transfer might play a key role in shaping other intracellular associates. The *R. felis* genome revealed the first conjugative plasmid discovered in intracellular bacteria, and experimental analysis demonstrated conjugative pili and mating [54**]. *R. felis* also stands out among *Rickettsia* in having abundant transposases, a >six-fold higher density of repeats, paralogous gene families encoding surface proteins, and evidence for frequent inversions and translocations that disrupt synteny. Of the ORFs lacking in other *Rickettsia* spp., 91 have closest matches outside of the α -Proteobacteria, implying that a sizable fraction of this genome might be horizontally acquired. Notably, in the transmission of *R. felis* through flea vectors, co-infection with *Bartonella henselae*, *Bartonella quintana* or *Wolbachia* might enable opportunities for gene transfer [54**].

Genetic machinery for DNA transfer also occurs in *Parachlamydia* spp., each of which possesses a genomic island encoding all *tra* genes essential for F-like conjugative DNA transfer, the first evidence for a putative conjugative system in the Chlamydiales [28**,55*]. Conjugation might occur within the amoebae host containing numerous bacteria tightly packed in vacuoles [56]. Although base composition analysis revealed few signs of recent lateral gene acquisition, the apparent acquisitions of F-like conjugative DNA transfer and, as noted above, of Type IV secretion might be important in shaping host interactions [28**].

In addition to conjugation machinery, certain intracellular bacteria possess more mobile DNA, such as bacteriophage and transposable elements, than previously suspected [57] (see Update). Although missing from nutritional mutualists, such elements persist in long-term intracellular species such as *Parachlamydia* spp., *R. felis*, *Phytoplasma asteris*, *Wolbachia* wMel and, to a lesser extent, *Wolbachia* wBm [23**,54**,58**,59,60]. These and other intracellular bacteria also have surprisingly abundant repeated DNA sequences, devoting a relatively large fraction of their chromosomes to global direct and inverted repeats that can mediate large-scale intragenomic rearrangements (Figure 3). The growing list of examples of disrupted synteny implies frequent inversions and translocations [23**,28**,58**,61*].

Other examples of genome flux come from re-analysis of available genomes, often within a phylogenetic context.

Detailed phylogenetic analysis showed fifteen cases of gene transfer into the Chlamydiaceae from plant, fungal, archaeobacterial and bacterial donors [62]. Recombination among closer relatives was demonstrated by Gomes and colleagues [63*], who showed that transfer had occurred between the rodent-associated *Chlamydia muridarum* and *Chlamydia caviae*, and demonstrated frequent recombination among *Chlamydia trachomatis* strains at genes encoding polymorphic membrane proteins (pmps), a Chlamydiaceae-specific family of proteins, members of which are expressed on the cell surface. The same study discovered the first IS element in this group. Because subtle variations in gene content can lead to important differences in host ranges and phenotypic effects, just a handful of genes might influence pathogenic signatures [64]. Thus, even occasional gene transfer might be biologically significant.

At a more local genomic level, recombination among tandem repeats or among paralogous gene copies enables certain parasites to respond to the adaptive immune response of vertebrate hosts. Many intracellular parasites commit a high percentage of their tiny genomes to paralogous families of polymorphic surface molecules, suggesting that host immunity is among the 'highest priority' of the challenges they face [27] (see Update). By recombining various pseudogenes into a single expression site, *Anaplasma* spp. have generated sequential diversity of membrane proteins, thereby achieving a persistent infection [65*]. As crucial players in the generation of surface-coat antigenic variation, such pseudogenes are quite distinct from those reflecting genome erosion in Rickettsiales, *M. leprae*, *C. burnetii* and other parasites. In *Ehrlichia*, continual duplications among numerous tandemly repeated genes counter genome reduction by creating new loci that are important in immune evasion [66*]. In short, despite the exceptional stability of some intracellular associates, it is increasingly clear that intracellular lifestyle does not necessarily constrain genome flux.

Conclusions

The rapidly growing database of bacterial genomes has enhanced our understanding of processes that shape the outcomes of intimate bacterial–eukaryotic relationships. Models of reductive evolution offer a valuable framework to understand forces shaping the metabolic capabilities of obligately intracellular bacteria and their potential for genetic change. Recent genome data have upheld many expectations from these models, including severe genome-size reduction and metabolic specialization of intra-

(Figure 3 Legend Continued) [79,80] and were based on calculation methods described elsewhere [81,82]. Because searches were performed across entire chromosomes, values represent global, rather than just local, repeats. The location of the repeats along the chromosome can be visualized in the Genome Atlas or Repeat Atlas databases (<http://www.cbs.dtu.dk/services/GenomeAtlas/>). The Y-axis lists the bacterial species name and, in parentheses, the length of DNA involved in repeats. The *Rickettsia felis* genome, not yet released when this figure was developed, has an exceptionally high fraction of repeated DNA [54**].

cellular lineages. As predicted, many intracellular genomes are exceptionally stable, showing little evidence of gene acquisition by lateral transfer and few if any intragenomic changes that disrupt synteny among related strains.

Recent data have also revealed some exciting surprises that illustrate diverse modes of genome evolution within host cells. Discoveries of Type III and Type IV secretion in mutualists highlight parallels among infection strategies and add to the growing evidence that ‘virulence’ genes can play crucial roles in beneficial interactions. In addition, genomes are teaching us that intracellular associates can experience various forms of genome flux, ranging from gene acquisition from phylogenetically distant donors, to intragenomic lability with frequent inversions and rearrangements, to specific recombination-mechanisms that generate antigenic diversity. This rather surprising component of reductive evolution has prompted new research into the impact of gene transfer and recombination in these species, including the roles of mobile elements that manage to persist in certain anciently intracellular groups.

Completion of additional genomes will provide a richer context to assess mechanisms of evolution and to make predictions about functions that mediate host associations. Testing these predictions will depend on the development of new experimental approaches to clarify processes involved in various stages of bacterial infection, persistence, and transmission to new hosts. Promising experimental methods often build upon genome data, such as the use of microarrays to assess gene expression [67**] and applications of an ever-growing knowledge of natural mobile DNA sequences to develop tools for the genetic manipulation of uncultivable bacteria [68].

Update

Since the July submission of this review, several important papers in endosymbiont genomics have been published (owing to space limitations, only a few have been added below). These include an exploration by Nilsson *et al.* [87**] of deletion rates and patterns in experimental cultures of *Salmonella enterica*. The authors detected very large deletions, similar in magnitude to those considered important in the early evolution of endosymbiont genomes.

Full sequences of extrachromosomal DNA (ecDNA) of *Sodalis glossinidius*, including three plasmids and a bacteriophage of two *S. glossinidius* isolates, revealed transposases, conjugation functions and evidence for recent gene acquisition [88**]. These findings illustrate the importance of gene exchange in the evolution of some intracellular associates.

In addition, several recent studies have explored the evolution of α -proteobacterial endosymbionts [89**–

91**]. These articles include an overview of genome plasticity in mutualistic and pathogenic α -Proteobacteria [89**], an investigation of the roles of ankyrin domain genes in shaping distinct reproductive alterations caused by *Wolbachia* [90**], and a population genetic study showing a recent global replacement of *Wolbachia* throughout *Drosophila melanogaster* lab stocks and field populations [91**].

Acknowledgements

I thank Richard R Lawler and Kostas Bourtzis for helpful comments on the manuscript, David W Ussery for sharing DNA repeat calculations and helpful advice, and many investigators for providing information about genome projects in progress. My apologies to colleagues whose work I could not highlight, owing to space constraints. I gratefully acknowledge support from the National Institutes of Health (R01 GM62626-01), National Science Foundation (DEB 0089455) and the NASA Astrobiology Institute (NNA04CC04A).

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- of special interest
- of outstanding interest

1. Moran NA: **The ubiquitous and varied role of infection in the lives of animals and plants.** *Am Nat* 2002, **60**:S1–S8.
2. Andersson SG, Kurland CG: **Reductive evolution of resident genomes.** *Trends Microbiol* 1998, **6**:263–268.
3. Moran NA: **Accelerated evolution and Muller's ratchet in endosymbiotic bacteria.** *Proc Natl Acad Sci USA* 1996, **93**:2873–2878.
4. Mira A, Ochman H, Moran NA: **Deletional bias and the evolution of bacterial genomes.** *Trends Genet* 2001, **17**:589–596.
5. Frank AC, Amiri H, Andersson SG: **Genome deterioration: loss of repeated sequences and accumulation of junk DNA.** *Genetica* 2002, **115**:1–12.
6. Mira A, Moran NA: **Estimating population size and transmission bottlenecks in maternally transmitted endosymbiotic bacteria.** *Microb Ecol* 2002, **44**:137–143.
7. Ohta T: **Slightly deleterious mutant substitutions in evolution.** *Nature* 1973, **246**:96–98.
8. Muller J: **The relation of recombination to mutational advance.** *Mutat Res* 1964, **106**:2–9.
9. Moran NA, Plague GR: **Genomic changes following host restriction in bacteria.** *Curr Opin Genet Dev* 2004, **14**:627–633.
10. Subtil A, Dautry-Varsat A: **Chlamydia: five years A.G. (after genome).** *Curr Opin Microbiol* 2004, **7**:85–92.
11. Gil R, Latorre A, Moya A: **Bacterial endosymbionts of insects: insights from comparative genomics.** *Environ Microbiol* 2004, **6**:1109–1122.
12. Klasson L, Andersson SG: **Evolution of minimal-gene-sets in host-dependent bacteria.** *Trends Microbiol* 2004, **12**:37–43.
13. Renesto P, Ogata H, Audic S, Claverie JM, Raoult D: **Some lessons from Rickettsia genomics.** *FEMS Microbiol Rev* 2005, **29**:99–117.
14. Batut J, Andersson SG, O'Callaghan D: **The evolution of chronic infection strategies in the alpha-proteobacteria.** *Nat Rev Microbiol* 2004, **2**:933–945.
15. Zientz E, Dandekar T, Gross R: **Metabolic interdependence of obligate intracellular bacteria and their insect hosts.** *Microbiol Mol Biol Rev* 2004, **68**:745–770.

16. Moran NA, Mira A: **The process of genome shrinkage in the obligate symbiont *Buchnera aphidicola***. *Genome Biol* 2001, **2**: RESEARCH0054.
17. Belda E, Moya A, Silva FJ: **Genome rearrangement distances and gene order phylogeny in γ -Proteobacteria**. *Mol Biol Evol* 2005, **22**:1456-1467.
- By comparing 30 γ -proteobacterial genomes, this study quantified the number of inversions necessary to explain current gene order. *Buchnera* and *Wigglesworthia* have experienced relatively high levels of gene rearrangements. Given the conservation of gene order within *Buchnera*, this implies a very high rearrangement rate early in the history of this mutualist.
18. Seshadri R, Paulsen IT, Eisen JA, Read TD, Nelson KE, Nelson WC, Ward NL, Tettelin H, Davidsen TM, Beanan MJ *et al.*: **Complete genome sequence of the Q-fever pathogen *Coxiella burnetii***. *Proc Natl Acad Sci USA* 2003, **100**:5455-5460.
19. Cole ST, Eiglmeier K, Parkhill J, James KD, Thomson NR, Wheeler PR, Honore N, Garnier T, Churcher C, Harris D *et al.*: **Massive gene decay in the leprosy bacillus**. *Nature* 2001, **409**:1007-1011.
20. Dale C, Jones T, Pontes M: **Degenerative evolution and functional diversification of type-III secretion systems in the insect endosymbiont *Sodalis glossinidius***. *Mol Biol Evol* 2005, **22**:758-766.
- This study characterizes two Type III secretion systems in a maternally transmitted endosymbiont of tsetse flies. The authors found that both systems are closely related to extant pathogenicity islands but have undergone functional diversification in this mutualism. This study demonstrates the specific expression of these systems during cellular invasion and the subsequent proliferation of the endosymbiont.
21. Gomez-Valero L, Latorre A, Silva FJ: **The evolutionary fate of nonfunctional DNA in the bacterial endosymbiont *Buchnera aphidicola***. *Mol Biol Evol* 2004, **21**:2172-2181.
- This analysis traces the history of gene erosion during the divergence among *Buchnera* lineages. The finding that inactivated genes require 40–60 million years to disintegrate completely suggests that continued genome reduction within intracellular groups is a gradual process.
22. Perez-Brocail V, Latorre A, Gil R, Moya A: **Comparative analysis of two genomic regions among four strains of *Buchnera aphidicola*, primary endosymbiont of aphids**. *Gene* 2005, **345**:73-80.
- This preliminary analysis of two sequence regions of *Buchnera* BCc, which has the smallest known bacterial genome, showed a significant reduction in the number of genes encoding metabolic functions. This full genome sequence for this strain is in progress and will shed light on functions retained in this tiny genome.
23. Foster J, Ganatra M, Kamal I, Ware J, Makarova K, Ivanova N, Bhattacharyya A, Kapatral V, Kumar S, Posfai J *et al.*: **The *Wolbachia* genome of *Brugia malayi*: endosymbiont evolution within a human pathogenic nematode**. *PLoS Biol* 2005, **3**:e121.
- This genome sequence revealed that mutualistic *Wolbachia* retain numerous metabolic pathways that might function in its beneficial association with nematodes, such as synthesis of riboflavin, heme and nucleotides. Although this species shares many metabolic traits with *Wolbachia* wMel, the mutualist lacks phage DNA and has lower levels of repeated DNA.
24. Ghedin E, Wang S, Foster JM, Slatko BE: **First sequenced genome of a parasitic nematode**. *Trends Parasitol* 2004, **20**:151-153.
25. Hacker J, Hentschel U, Dobrindt U: **Prokaryotic chromosomes and disease**. *Science* 2003, **301**:790-793.
26. Goebel W, Gross R: **Intracellular survival strategies of mutualistic and parasitic prokaryotes**. *Trends Microbiol* 2001, **9**:267-273.
27. Palmer GH: **The highest priority: what microbial genomes are telling us about immunity**. *Vet Immunol Immunopathol* 2002, **85**:1-8.
28. Horn M, Collingro A, Schmitz-Esser S, Beier CL, Purkhold U, Fartmann B, Brandt P, Nyakatura GJ, Droege M, Frishman D *et al.*: **Illuminating the evolutionary history of chlamydiae**. *Science* 2004, **304**:728-730.
- The genome sequence of *Parachlamydia* sp., a symbiont of free-living amoebae, has many virulence factors found in parasitic chlamydia, including a Type III secretion system. This finding suggests that, 700 mya, the last common ancestor of this group had already adopted an intracellular lifestyle. The authors propose that ancient chlamydia might have originated mechanisms to exploit eukaryotic cells.
29. Dale C, Plague GR, Wang B, Ochman H, Moran NA: **Type III secretion systems and the evolution of mutualistic endosymbiosis**. *Proc Natl Acad Sci U S A* 2002, **99**:12397-12402.
30. Corsaro D, Venditti D: **Emerging chlamydial infections**. *Crit Rev Microbiol* 2004, **30**:75-106.
31. Everett K, Thao M, Horn M, Dyszynski G, Baumann P: **Novel chlamydiae in whiteflies and scale insects: endosymbionts '*Candidatus Fritschea bemisiae*' strain Falk and '*Candidatus Fritschea eriococci*' strain Elm**. *Int J Syst Evol Microbiol* 2005, **55**:1581-1587.
- Combining genetic data and microscopy, this study characterizes chlamydial endosymbionts in whiteflies and scale insects. A detailed phylogenetic analysis provides an excellent overview of the genetic diversity and lifestyle variation in this bacterial group.
32. Schmitz-Esser S, Linka N, Collingro A, Beier CL, Neuhaus HE, Wagner M, Horn M: **ATP/ADP translocases: a common feature of obligate intracellular amoebal symbionts related to Chlamydiae and Rickettsiae**. *J Bacteriol* 2004, **186**:683-691.
- This study screened bacterial endosymbionts of free-living amoebae and paramoecia for the presence of ATP/ADP translocase genes, which are necessary for the transport of ATP across a membrane in exchange for ADP. The discovery of these genes solely in the Rickettsiales and the Chlamydiales, including *Parachlamydia*, is consistent with current views about their distribution. Phylogenetic analysis suggests these translocases originated in the Chlamydiales and were horizontally transferred to Rickettsiales and plants.
33. McGraw EA, O'Neill SL: ***Wolbachia pipientis*: intracellular infection and pathogenesis in *Drosophila***. *Curr Opin Microbiol* 2004, **7**:67-70.
34. Brownlie JC, O'Neill SL: ***Wolbachia* genomes: insights into an intracellular lifestyle**. *Curr Biol* 2005, **15**:R507-R509.
35. Burne RA, Chen YY: **Bacterial ureases in infectious diseases**. *Microbes Infect* 2000, **2**:533-542.
36. Gil R, Silva FJ, Zientz E, Delmotte F, Gonzalez-Candelas F, Latorre A, Rausell C, Kamerbeek J, Gadau J, Holldobler B *et al.*: **The genome sequence of *Blochmannia floridanus*: comparative analysis of reduced genomes**. *Proc Natl Acad Sci USA* 2003, **100**:9388-9393.
37. Degnan PH, Lazarus AB, Wernegreen JJ: **Genome sequence of *Blochmannia pennsylvanicus* indicates parallel evolutionary trends among bacterial mutualists of insects**. *Genome Res* 2005, **15**:1023-1033.
38. Fares MA, Moya A, Barrio E: **GroEL and the maintenance of bacterial endosymbiosis**. *Trends Genet* 2004, **20**:413-416.
39. Bencina D, Slavec B, Narat M: **Antibody response to GroEL varies in patients with acute *Mycoplasma pneumoniae* infection**. *FEMS Immunol Med Microbiol* 2005, **43**:399-406.
40. Renesto P, Azza S, Dolla A, Fourquet P, Vestris G, Gorvel JP, Raoult D: **Proteome analysis of *Rickettsia conorii* by two-dimensional gel electrophoresis coupled with mass spectrometry**. *FEMS Microbiol Lett* 2005, **245**:231-238.
41. Hechard C, Grepinet O, Rodolakis A: **Molecular cloning of the *Chlamydomydia abortus* groEL gene and evaluation of its protective efficacy in a murine model by genetic vaccination**. *J Med Microbiol* 2004, **53**:861-868.
42. Fares MA, Ruiz-Gonzalez MX, Moya A, Elena SF, Barrio E: **Endosymbiotic bacteria: GroEL buffers against deleterious mutations**. *Nature* 2002, **417**:398.
43. Morin S, Ghanim M, Zeidan M, Czosnek H, Verbeek M, van den Heuvel JF: **A GroEL homologue from endosymbiotic bacteria of the whitefly *Bemisia tabaci* is implicated in the circulative transmission of tomato yellow leaf curl virus**. *Virology* 1999, **256**:75-84.
44. Fredriksson A, Ballesteros M, Dukan S, Nystrom T: **Defense against protein carbonylation by DnaK/DnaJ and proteases of the heat shock regulon**. *J Bacteriol* 2005, **187**:4207-4213.

45. Akman L, Yamashita A, Watanabe H, Oshima K, Shiba T, Hattori M, Aksoy S: **Genome sequence of the endocellular obligate symbiont of tsetse flies, *Wigglesworthia glossinidia***. *Nat Genet* 2002, **32**:402-407.
46. Rocha EP, Danchin A: **Base composition bias might result from competition for metabolic resources**. *Trends Genet* 2002, **18**:291-294.
47. Ochman H, Moran NA: **Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis**. *Science* 2001, **292**:1096-1099.
48. Tamas I, Klasson L, Canback B, Naslund AK, Eriksson AS, Wernegreen JJ, Sandstrom JP, Moran NA, Andersson SG: **50 million years of genomic stasis in endosymbiotic bacteria**. *Science* 2002, **296**:2376-2379.
49. Van Ham RC, Kamerbeek J, Palacios C, Rausell C, Abascal F, Bastolla U, Fernandez JM, Jimenez L, Postigo M, Silva FJ *et al.*: **Reductive genome evolution in *Buchnera aphidicola***. *Proc Natl Acad Sci USA* 2003, **100**:581-586.
50. Van Ham RC, Gonzalez-Candelas F, Silva FJ, Sabater B, Moya A, Latorre A: **Postsymbiotic plasmid acquisition and evolution of the repA1-replicon in *Buchnera aphidicola***. *Proc Natl Acad Sci USA* 2000, **97**:10855-10860.
51. Sabater-Munoz B, van Ham RC, Moya A, Silva FJ, Latorre A: **Evolution of the leucine gene cluster in *Buchnera aphidicola*: insights from chromosomal versions of the cluster**. *J Bacteriol* 2004, **186**:2646-2654.
- This study describes an exceptional case of gene rearrangements in *Buchnera*. The authors discovered that variable position of a leucine biosynthesis gene cluster (*leuABCD*) reflects independent chromosomal insertions from ancestral plasmid-encoded genes.
52. McLeod MP, Qin X, Karpathy SE, Gioia J, Highlander SK, Fox GE, McNeill TZ, Jiang H, Muzny D, Jacob LS *et al.*: **Complete genome sequence of *Rickettsia typhi* and comparison with sequences of other rickettsiae**. *J Bacteriol* 2004, **186**:5842-5855.
53. Mira A, Klasson L, Andersson SG: **Microbial genome evolution: sources of variability**. *Curr Opin Microbiol* 2002, **5**:506-512.
54. Ogata H, Renesto P, Audic S, Robert C, Blanc G, Fournier PE, Parinello H, Claverie JM, Raoult D: **The genome sequence of *Rickettsia felis* identifies the first putative conjugative plasmid in an obligate intracellular parasite**. *PLoS Biol* 2005, **3**:e248.
- The genome of this flea-associated pathogen is substantially different from those of its *Rickettsia* relatives. In addition to the exciting discovery of a conjugative plasmid, this study revealed numerous transposases and repeated DNA in this genome. The authors also show evidence for two types of pili, hemolytic activity, β -lactamase activity and actin-polymerization-driven intracellular motility.
55. Greub G, Collyn F, Guy L, Roten CA: **A genomic island present along the bacterial chromosome of the *Parachlamydiaceae* UWE25, an obligate amoebal endosymbiont, encodes a potentially functional F-like conjugative DNA transfer system**. *BMC Microbiol* 2004, **4**:48.
- This re-analysis of the *Parachlamydia* sp. genome sequence identified a genomic island that has atypical GC-content and is flanked by inverted repeats. This island encodes an F-like conjugative DNA transfer mechanism involved in sex pilus retraction and mating pair stabilization, and might provide a genetic tool for chlamydia research.
56. Horn M, Wagner M: **Bacterial endosymbionts of free-living amoebae**. *J Eukaryot Microbiol* 2004, **51**:509-514.
57. Bordenstein SR, Reznikoff WS: **Mobile DNA in obligate intracellular bacteria**. *Nat Rev Microbiol* 2005, **3**:688-699.
58. Wu M, Sun LV, Vamathevan J, Riegler M, Deboy R, Brownlie JC, McGraw EA, Martin W, Esser C, Ahmadijad N *et al.*: **Phylogenomics of the reproductive parasite *Wolbachia pipientis* wMel: A streamlined genome overrun by mobile genetic elements**. *PLoS Biol* 2004, **2**:E69.
- This genome, the first published for *Wolbachia*, has exceptionally high levels of repetitive DNA and mobile elements that might mediate genome flux. Initially considered highly unusual for reduced genomes, similar patterns have since been discovered in other long-term intracellular α -Proteobacteria such as *Rickettsia felis*, *Ehrlichia ruminantium*, *Anaplasma marginale* and, to a lesser extent, the mutualistic *Wolbachia* wBm.
59. Oshima K, Kakizawa S, Nishigawa H, Jung HY, Wei W, Suzuki S, Arashida R, Nakata D, Miyata S, Ugaki M *et al.*: **Reductive evolution suggested from the complete genome sequence of a plant-pathogenic phytoplasma**. *Nat Genet* 2004, **36**:27-29.
60. Lee IM, Zhao Y, Bottner KD: **Novel insertion sequence-like elements in phytoplasma strains of the aster yellows group are putative new members of the IS3 family**. *FEMS Microbiol Lett* 2005, **242**:353-360.
61. Salzberg SL, Hotopp JC, Delcher AL, Pop M, Smith DR, Eisen MB, Nelson WC: **Serendipitous discovery of *Wolbachia* genomes in multiple *Drosophila* species**. *Genome Biol* 2005, **6**:R23.
- The authors discovered sequences for three new *Wolbachia* species in the Trace Archive (a publicly available repository for unanalyzed sequence data; <http://www.ncbi.nlm.nih.gov/Traces>). The bacteria are associated with *Drosophila ananassae*, *Drosophila simulans* and *Drosophila mojavensis*, and available sequences encompass 95%, 75–80% and 6–7% of the *Wolbachia* genomes, respectively. Comparisons with published *Wolbachia* genomes confirmed high levels of genome flux in this group, as numerous rearrangements, insertions and hundreds of novel genes were detected.
62. Ortutay C, Gaspari Z, Toth G, Jager E, Vida G, Orosz L, Vellai T: **Speciation in *Chlamydia*: genomewide phylogenetic analyses identified a reliable set of acquired genes**. *J Mol Evol* 2003, **57**:672-680.
63. Gomes JP, Bruno WJ, Borrego MJ, Dean D: **Recombination in the genome of *Chlamydia trachomatis* involving the polymorphic membrane protein C gene relative to ompA and evidence for horizontal gene transfer**. *J Bacteriol* 2004, **186**:4295-4306.
- This phylogenetic analysis of numerous *C. trachomatis* strains demonstrated clear evidence for frequent, whole-gene recombination among polymorphic membrane proteins that are expressed on the cell surface. The authors suggest that genome plasticity might produce adaptive changes in tissue tropism and pathogenesis.
64. Read TD, Myers GS, Brunham RC, Nelson WC, Paulsen IT, Heidelberg J, Holtzapple E, Khouri H, Federova NB, Carty HA *et al.*: **Genome sequence of *Chlamydia caviae* (*Chlamydia psittaci* GPIC): examining the role of niche-specific genes in the evolution of the Chlamydiaceae**. *Nucleic Acids Res* 2003, **31**:2134-2147.
65. Brayton KA, Kappmeyer LS, Herndon DR, Dark MJ, Tibbals DL, Palmer GH, McGuire TC, Knowles DP. Jr: **Complete genome sequencing of *Anaplasma marginale* reveals that the surface is skewed to two superfamilies of outer membrane proteins**. *Proc Natl Acad Sci USA* 2005, **102**:844-849.
- The genome of this tick-borne mammalian pathogen revealed two gene families that underlie its rapid generation of antigenic variation. Pseudogenes in this species do not reflect genome erosion; rather, they play important roles in generating variation in surface coat proteins.
66. Collins NE, Liebenberg J, de Villiers EP, Brayton KA, Louw E, Pretorius A, Faber FE, van Heerden H, Josemans A, van Kleef M *et al.*: **The genome of the heartwater agent *Ehrlichia ruminantium* contains multiple tandem repeats of actively variable copy number**. *Proc Natl Acad Sci USA* 2005, **102**:838-843.
- In contrast to the stable genomes of some intracellular bacteria, this pathogen has numerous tandemly repeated and duplicated sequences that mediate frequent translocations, inversions and duplications that have given rise to new genes.
67. Moran NA, Dunbar HE, Wilcox JL: **Regulation of transcription in a reduced bacterial genome: nutrient-provisioning genes of the obligate symbiont *Buchnera aphidicola***. *J Bacteriol* 2005, **187**:4229-4237.
- Using full-genome microarrays, this study showed that transcription in *Buchnera* changes little in response to host dietary changes, suggesting that loss of regulation-genes might constrain the ability to respond to environmental changes.
68. Qin A, Tucker AM, Hines A, Wood DO: **Transposon mutagenesis of the obligate intracellular pathogen *Rickettsia prowazekii***. *Appl Environ Microbiol* 2004, **70**:2816-2822.

69. Buchner P: *Endosymbiosis of Animals with Plant Microorganisms*. New York: Interscience Publishers, Inc; 1965.
70. Baumann P, Moran N, Baumann L: **Bacteriocyte-associated endosymbionts of insects**. In *The Prokaryotes, A Handbook on the Biology of Bacteria; Ecophysiology, Isolation, Identification, Applications*. Edited by Dworkin M. Springer-Verlag; Release 3.1, 2000.
71. Moran N, Telang A: **Bacteriocyte-associated symbionts of insects**. *Bioscience* 1998, **48**:295-304.
72. Aksoy S, Rio RV: **Interactions among multiple genomes: tsetse, its symbionts and trypanosomes**. *Insect Biochem Mol Biol* 2005, **35**:691-698.
73. Moran NA, Russell JA, Koga R, Fukatsu T: **Evolutionary relationships of three new species of enterobacteriaceae living as symbionts of aphids and other insects**. *Appl Environ Microbiol* 2005, **71**:3302-3310.
- This combination of phylogenetic analysis and electron microscopy describes three new species of secondary endosymbionts that coexist with *Buchnera* within aphid hosts.
74. Brownstein JS, Hett E, O'Neill SL: **The potential of virulent *Wolbachia* to modulate disease transmission by insects**. *J Invertebr Pathol* 2003, **84**:24-29.
75. Rio RV, Hu Y, Aksoy S: **Strategies of the home-team: symbioses exploited for vector-borne disease control**. *Trends Microbiol* 2004, **12**:325-336.
76. Sinkins SP, Godfray HC: **Use of *Wolbachia* to drive nuclear transgenes through insect populations**. *Proc Biol Sci* 2004, **271**:1421-1426.
77. Moran NA: **Microbial minimalism: genome reduction in bacterial pathogens**. *Cell* 2002, **108**:583-586.
78. Peterson JD, Umayam LA, Dickinson T, Hickey EK, White O: **The Comprehensive Microbial Resource**. *Nucleic Acids Res* 2001, **29**:123-125.
79. Hallin PF, Ussery DW: **CBS Genome Atlas Database: a dynamic storage for bioinformatic results and sequence data**. *Bioinformatics* 2004, **20**:3682-3686.
80. Ussery DW, Binnewies TT, Gouveia-Oliveira R, Jarmer H, Hallin PF: **Genome update: DNA repeats in bacterial genomes**. *Microbiology* 2004, **150**:3519-3521.
81. Skovgaard M, Jensen LJ, Friis C, Stærfeldt H-H, Worning P, Brunak S, Ussery DW: **The atlas visualisation of genome-wide information**. *Methods Microbiol* 2002, **33**:49-63.
82. Jensen LJ, Friis C, Ussery DW: **Three views of microbial genomes**. *Res Microbiol* 1999, **150**:773-777.
83. Frank AC, Alsmark CM, Thollesson M, Andersson SG: **Functional divergence and horizontal transfer of type IV secretion systems**. *Mol Biol Evol* 2005, **22**:1325-1336.
- Using a phylogenetic approach, the authors illustrate the importance of horizontal transfer in the evolution and functional divergence of type IV secretion pathways. Repeatedly, transfer events have corresponded to shifts from the ancestral state of conjugation to secretion of effector molecules.
84. Gil R, Sabater-Munoz B, Latorre A, Silva FJ, Moya A: **Extreme genome reduction in *Buchnera* spp.: toward the minimal genome needed for symbiotic life**. *Proc Natl Acad Sci USA* 2002, **99**:4454-4458.
85. Moran NA, Dale C, Dunbar H, Smith WA, Ochman H: **Intracellular symbionts of sharpshooters (Insecta: Hemiptera: Cicadellinae) form a distinct clade with a small genome**. *Environ Microbiol* 2003, **5**:116-126.
86. Zabalou S, Riegler M, Theodorakopoulou M, Stauffer C, Savakis C, ●● Bourtzis K: ***Wolbachia*-induced cytoplasmic incompatibility as a means for insect pest population control**. *Proc Nat Acad Sci USA* 2004, **101**:15042-15045.
- This study illustrates the potential use of *Wolbachia*-induced reproductive incompatibilities to suppress insect host populations. Using embryonic cytoplasm transfer, the authors infected Mediterranean fruit fly (medfly) hosts with *Wolbachia* from a related fruit fly species. This new infection was successfully established and shown to cause complete cytoplasmic incompatibility in the medfly host. Moreover, the release of infected males in cage experiments caused the suppression of lab populations.
87. Nilsson AI, Koskiniemi S, Eriksson S, Kugelberg E, Hinton JC, ●● Andersson DI: **Bacterial genome size reduction by experimental evolution**. *Proc Natl Acad Sci USA* 2005, **102**:12112-12116.
- Taking an experimental approach, the authors examined rates, sizes and chromosomal locations of gene deletions during serial passage of *Salmonella enterica* cultures. Deletion rates were particularly high (~2.5 bp/chromosome/generation) in a strain with disrupted mismatch repair functions (*mutS*⁻). The large sizes (up to 202 kb) of many deletions corroborate the view that massive deletion events impacted the early evolution of certain intracellular bacteria.
88. Darby AC, Lagnel J, Matthew CZ, Bourtzis K, Maudlin I, ●● Welburn SC: **Extrachromosomal DNA of the symbiont *Sodalis glossinidius***. *J Bacteriol* 2005, **187**:5003-5007.
- Three plasmids and a bacteriophage were sequenced from two strains of *Sodalis glossinidius*, a secondary endosymbiont of tsetse flies. Annotation of plasmid genes revealed two putative symbiotic islands showing signatures of recent horizontal acquisition. In addition, the presence of pilus genes suggests that conjugation is important in the evolution of this symbiont. These sequence data illustrate the important role of gene transfer in some intracellular bacteria, and further support the idea that parasites and mutualists, such as *Sodalis*, often use similar genes to mediate host interactions.
89. Sallstrom B, Andersson SG: **Genome reduction in the α-Proteobacteria**. *Curr Opin Microbiol* 2005, **8**:579-585.
- This review highlights evidence for genome plasticity, including computational estimates of gene losses, acquisitions and duplications, among medically, agriculturally and environmentally significant α-Proteobacteria. The authors emphasize the importance of recombination in generating antigenic variability within pathogen populations, and note common patterns of genome reduction in intracellular species. This review also notes examples of genome reduction in 'real-time' during the course of a pathogen's infection of a single host individual and during lab cultivation of a plant symbiont.
90. Iturbe-Ormaetxe I, Burke GR, Riegler M, O'Neill SL: ●● **Distribution, expression, and motif variability of ankyrin domain genes in *Wolbachia pipientis***. *J Bacteriol* 2005, **187**:5136-5145.
- The authors address the molecular mechanisms shaping the diversity of reproductive alterations that *Wolbachia* causes in hosts, with a focus on the potential roles of ankyrin (ANK) domain genes. These domains, known to mediate protein-protein interactions, are abundant in the *Wolbachia* genome. Comparisons of ANK domain genes with different effects across *Wolbachia* species revealed significant variation that might affect the functions of encoded proteins.
91. Riegler M, Sidhu M, Miller WJ, O'Neill SL: **Evidence for a global *Wolbachia* replacement in *Drosophila melanogaster***. *Curr Biol* 2005, **15**:1428-1433.
- Using polymorphic molecular markers to distinguish closely related *Wolbachia* strains, the authors sampled *Drosophila melanogaster* hosts from stock collections and field populations. They discovered that one *Wolbachia* strain has replaced all others in this fruit fly species during the past century. It is difficult to explain this global replacement by cytoplasmic incompatibility, but it might be linked to the recent invasion by P elements in this host and an establishment of a maternally transmitted repressive P cytotype.



ELSEVIER

Common themes in the genome strategies of pathogens

Jeffrey G Lawrence

Genomes of pathogenic bacteria evolve by large-scale changes in gene inventory. The continual acquisition of genomic islands, which refines their metabolic arsenal, is offset by gene loss. Far from this being a passive deletion of genes no longer useful to pathogens, the removal of genes encoding problematic metabolic process and immunogenic surface antigens might be strongly beneficial. Genomes of virulent eukaryotes show the footprint of similar genomic alterations, including acquisition of genes by lateral transfer, and genome degradation in obligate pathogens. These common features suggest that unicellular pathogens share common strategies for adaptation.

Addresses

Department of Biological Sciences, University of Pittsburgh,
Pittsburgh, PA 15260, USA

Corresponding author: Lawrence, Jeffrey G (jlawrenc+@pitt.edu)

Current Opinion in Genetics & Development 2005, 15:584–588

This review comes from a themed issue on
Genomes and evolution
Edited by Stephen J O'Brien and Claire M Fraser

Available online 26th September 2005

0959-437X/\$ – see front matter

© 2005 Elsevier Ltd. All rights reserved.

DOI 10.1016/j.gde.2005.09.007

Introduction

Gene inventories provide a window into the physiological capabilities of bacteria. Unlike multicellular eukaryotes, which adapt by altering the expression of common sets of genes to enable morphological distinctiveness, bacteria living in different environments have markedly different sets of genes that provide biochemical distinctiveness. Comparative genomics opens the doorway not only for uncovering the genetic differences between strains — as was demonstrated strikingly in the comparison of pathogenic and non-pathogenic strains of *Escherichia coli* that share only 40% of their common gene-pool [1] — but also for understanding the pathways along which bacteria evolve. Pathogens are characterized by very clear differences from their non-pathogenic relatives: they have evolved the ability to cause disease in one or more hosts. Therefore, the genomic changes seen in these organisms can often be interpreted using a framework in which potential causes of genomic changes can be more precisely defined.

As with all genomes, those of bacterial pathogens evolve by three major processes: modification of existing genes;

loss of genes no longer under selection for function; and gain of genes that confer a benefit in their current ecological niche. There are notable examples in which the increased pathogenicity of a strain has resulted from the modification of genes that are also found in less virulent relatives, such as alteration of the *Salmonella pmrD* gene to become regulated by the PhoPQ two-component regulatory system [2], or mutations leading to increased expression of the *Bordetella pertussis ptxA* toxin genes [3]. But the fruits of definitive analyses of differential gene-expression are hard won, and little is understood beyond exemplar cases. Rather, initial progress in comparative genomics lies in interpretation of changes in genome inventory.

Here, I review advances in understanding the impact of gene gain and gene loss in the evolution of pathogens. Recent work has led to a more sophisticated interpretation of both processes, beyond the arbitrary accumulation of genes that enable virulence, and the sloughing of genes not contributing to the pathogenic lifestyle. In addition, analyses of the genomes of pathogenic eukaryotic are showing that they are influenced by many of the same evolutionary processes that shape the genomes of bacterial pathogens.

Two avenues of gene gain

Gene gain by lateral gene-transfer has long been a hallmark of pathogen evolution, first in the acquisition of antibiotic resistance genes [4] and then in the acquisition of virulence determinants as pathogenicity islands [5]. First described in animal pathogens, genomic islands are now well documented in plant pathogens, including *Erwinia carotovora* [6] and *Leifsonia xyli* [7], providing these bacteria with the molecular toolbox necessary to exploit this new ecological niche. Yet gene acquisition is no longer viewed simply as a road taken to achieve pathogenicity; rather, ongoing gene transfer may impart new virulence determinants on pathogen genomes. For example, the genome of *Salmonella enterica* serovar Choleraesuis contains two pathogenicity islands not found in the genomes of serovars Typhimurium, Typhi, Enteritidis or Paratyphi [8], and antibiotic resistance genes continue to be introduced into *Staphylococcus aureus* genomes [9,10], enabling it to adapt to a changing environment. Genes need not be introduced from other bacteria; species of *Xanthomonas* might have acquired PNP genes — encoding plant natriuretic peptides — from their host plant genomes, potentially enabling molecular mimicry [11], and *Entamoeba*-infecting *Legionella* have numerous eukaryotic-like proteins that are predicted to interact with host cell proteins [12].

Horizontal exchange plays another role in pathogen evolution: genes acquired by lateral transfer are shared among strains of a bacterial species by homologous recombination, enabling pathogens to synergize their gene arsenals as they adapt to new or existing hosts. For instance, frequent recombination leads to the emergence of well defined groups — strains of *Neisseria meningitidis*, *Neisseria lactamica* and *Neisseria gonorrhoeae* form discrete clades [13] — that share virulence loci and show distinct patterns of infectivity. Here, the balance between recombination and selection enables rapid exploration of 'protein space' by bringing combinations of new alleles together [14]. Recombination in *Borrelia* also helps maintain diversity at virulence loci, including the *ospC* cell-surface locus [15], has diversified otherwise clonal strains of *Vibrio parahaemolyticus* at antigenic loci [16] and has led to the emergence of a newly virulent strain of *Vibrio vulnificus* by combining strains from otherwise separate populations [17]. Transfer of antibiotic resistance gene-bearing plasmids among *E. coli* strains is evident in a non-clinical setting in Australia [18]. Thus, lateral gene-transfer serves not only to create pathogens from non-pathogens by the introduction of genomic islands but also continually rearranges genes within pathogen populations to enable ongoing adaptation to their hosts.

Two avenues of genome loss

Given that bacterial genomes are not ever-increasing in size, gene acquisition must be balanced by gene loss. During genome reduction, symbionts and pathogens experience net gene-loss as their genomes shrink relative to those of their free-living siblings and ancestors [19]. Here, gene gain by lateral transfer fails to keep pace with gene loss, and gene content dwindles. Dramatic examples, such as the pseudogene-laden *Mycobacterium leprae* genome [20], remind us that genome reduction can proceed rapidly and remove large percentages of an organism's gene complement. Given that both obligate symbionts (e.g. *Buchnera* and *Wolbachia*) and obligate pathogens (e.g. *Rickettsia* and *Mycoplasma*) with highly reduced genomes live in relatively constant environments, genome reduction was thought to reflect a lack of utility of many metabolic pathways, leading to the passive loss of genes no longer needed. Yet comparative genomics and analytical genetics have revealed two inter-dependent pathways that fuel an active process of gene loss.

First, the action of some gene products might be detrimental to the pathogenic lifestyle. This possibility was noted some time ago, when the loss of the *cadA* gene from *Shigella* was correlated with an increase in pathogenicity [21]. Recent analyses of metabolic pathways in *Chlamydomophila* genomes again show that the loss of metabolic pathways — including loss of tryptophan metabolism and purine nucleotide-cycling in *Chlamydomophila abortus*, and loss of biotin biosynthesis in *Chlamydomophila caviae* — is

correlated with the onset of virulence [22]. In addition, there appears to have been selection for the loss of metabolic pathways in *Mycobacterium tuberculosis*, primarily those pathways involved in the synthesis of antigens [23]. Therefore, as seen with pseudogene formation in *Yersinia pestis* [24] and *Salmonella enterica* serovar Choleraesuis [8], genes can be lost because their absence is beneficial, not neutral. This possibility increases the relevance of recombination within pathogen populations, in which pseudogenes can be propagated among strains with as much benefit as newly acquired pathogenicity islands.

Second, genes encoding cell-surface determinants appear to be preferentially lost from genomes of pathogens. This phenomenon was noted more than a decade ago in *Shigella*, in which loss of the *ompT* gene was correlated with increased pathogenicity [25], and increasing numbers of examples suggest that it is a common strategy to enable pathogens to avoid immune detection. Analysis of genome reduction in *Bordetella*, the causative agent of whooping cough, showed that genes encoding cell-surface determinants were over-represented among those lost in *B. pertussis* but retained in its less virulent, sibling species *Bordetella brochiseptica* [3]. There is also extensive loss of genes encoding cell-surface determinants among strains of *Mycobacterium tuberculosis* [23]; for example, many strains lack the antigen-encoding *plcA* and *plcD* genes in addition to the *lppA*, *lppB*, *lppC*, *lpqH*, *lpqS* and *lprP* genes, which probably encode lipoproteins. Parallel loss of the same genes in independent lineages suggests that gene loss was under selection [23]. Similarly, strains of *Francisella tularensis* have lost genes encoding outer membrane proteins and pilins [26], and comparison of strains of *Pseudomonas aeruginosa* isolated from natural and clinical environments showed a loss of flagellar loci in clinical isolates [27], again suggesting a preferential loss of surface-antigens in pathogenic strains. As with the loss of metabolic pathways, these steps in genome reduction are not neutral but confer a benefit in enabling better escape from immune surveillance.

Although some genomes serve as exemplars of certain processes — gene acquisition by enterohemorrhagic *E. coli* O157, or gene loss by *Mycobacterium leprae* — the processes of gene gain and gene loss do not operate independently. For example, both gene transfer and gene loss have played roles in the turnover of genes encoding effectors for Type III secretion systems in *Pseudomonas syringae* [28]. A mixture of lateral gene transfer and genome degradation is also evident in the evolution of the *Treponema denticola* [29] and *Corynebacterium diphtheria* [30] genomes: that is, genomes are not segregated into those experiencing gene transfer and those experiencing gene loss; the processes are inter-related, and both contribute to the ongoing evolution of pathogenicity. Only in strictly intracellular obligate pathogens and symbionts

does the rate of lateral transfer become so low as to make genome reduction the dominant evolutionary force.

Similar strategies seen in eukaryotic pathogens

Lateral gene-transfer is not thought to be widespread among multi-cellular eukaryotes; this might reflect the lack of opportunity in these organisms — genes must be introduced into the germ line and become expressed in the appropriate tissues — or lack of utility — adaptation to new environments rarely requires the deployment of novel biochemistries conferred by acquired genes. But protozoa do not share these traits, and one might predict that single-celled eukaryotes could evolve by gene gain as much as do Bacteria and Archaea. Lateral gene-transfer is not uncommon among unicellular eukaryotes [31], including pathogenic and parasitic lineages [32]. Examples such as the acquisition of degradative loci by ruminant fungi [33] show that eukaryotes can adapt to new ecological niches by lateral gene-transfer from bacterial donors. Analyses of the genome sequences of several eukaryotic pathogens has shown not only that lateral transfer has played a role in their evolution but that other features of prokaryotic genome evolution were evident as well.

Multiple lines of evidence suggest that pathogenicity loci have been acquired by the pea pathogen *Nectria haematococca* [34]. The PEP (pea pathogenicity) cluster contains four genes required for maximum virulence; this clustering is reminiscent of bacterial genomic islands. The genes have been characterized by atypical nucleotide composition, atypical codon-usage bias, features often correlated with recently acquired genes in bacteria genomes. A more recent phylogenetic analysis [34] shows that these genes are absent from closely related fungal taxa, again consistent with their acquisition by lateral gene-transfer. Transfer is implicated in the history of a bacterial-like catalase gene into the microsporidian *Nosema locustae* [35], the citrate synthase gene in the ciliate *Tetrahymena* [36], and two tRNA synthase genes into the ancestor of diplomonad algae [37]. Similarly, the *Entamoeba histolytica* genome shows evidence of numerous lateral gene-transfer events [38]. A detailed analysis of the *isc* locus — encoding iron-sulfur proteins — shows that these genes cluster very strongly with bacterial homologues, especially in the ϵ -Proteobacteria *Campylobacter* and *Helicobacter* [39^{*}]; these data strongly support recent transfer of these loci into the *Entamoeba* genome, as opposed to their movement from the mitochondrial genome, as is seen for large numbers of genes in the yeast genome [40]. Similar large-scale gene transfer is proposed for the *Trypanosoma brucei* genome, purportedly increasing its metabolic repertoire [41^{*}]. In addition, there are ~1000 non-expressed variable surface-glycoprotein genes in *T. brucei*; although some are fully functional, most appear to be non-functional [41^{*}]. Intra-genic recombination has been proposed as a mechanism

that maintains these non-functional copies as a sort of genomic reservoir, a strategy also seen in some bacterial pathogens.

Demonstrating further similarities to bacterial pathogens, eukaryotic pathogens accumulate transposons and pseudogenes on the path towards genome reduction. Approximately 50% of the *Trypanosoma cruzi* genome comprises repetitive elements [42], and 5% of the *Cryptococcus neoformans* genome comprises mobile genetic element [43]; in addition, *Cryptococcus* genes are highly intron-rich, unlike genes of related fungi [43]. Full-scale genome reduction has been observed both for the apicomplexan *Theileria parva* [44] and for the microsporidians *Encephalitozoon cuniculi* and *Antonospora locustae* [45]. Therefore, we may conclude that gene transfer from unrelated taxa, gene shuffling by recombination, and selective gene loss have played similar roles in the evolution of both prokaryotic and eukaryotic pathogens.

Beyond pathogenicity

Legionella pneumophila is the causative agent of Legionnaire's disease. Yet it has become clear that *Legionella* is not an obligate human pathogen; its natural prey are fresh-water amoebae [46]. Protozoan hosts might be used by many human pathogens [47], including *Salmonella* and *Pseudomonas* [48]; the same gene products that enable invasion of prey amoeba, survival in and escape from feeding vacuoles, and replication within the host cytoplasm can be deployed to enable infection of human cells. When considering disease as one part of pathogen life history, two spectres are raised: (i) that human pathogens might simply be bacteria using their metabolic repertoire on an unfortunately chosen host; and (ii) virulence loci might be genes the primary functions of which have little to do with pathogenicity. Although the first scenario can often be dismissed for many pathogens that are not found outside of their human hosts, the second scenario is less easily discounted.

Genes found in pathogenicity islands often contribute to the pathogenic lifestyle in clear and well-explained ways. For example, Type III secretion systems deliver effector molecules to the cytoplasm of host cells, adhesions provide a mechanism for cell attachment, and siderophores enable metal ion uptake in ion-poor host environments. Yet the benefits of some virulence loci are less clear. Recent insight into bacteria–protozoa interactions raises the possibility that virulence loci increase the fitness of pathogens by enabling them to interact more favourably with the protozoa they encounter, thus indirectly enabling more efficient invasion of host tissues. For example, Wildschutte *et al.* [49^{*}] showed that variability at the *Salmonella rfb* locus, which encodes the O-antigen biosynthetic machinery, might enable *Salmonella* to escape particular protozoan predators. This scenario is satisfying because, unlike the systemic pathogens *Neisseria* and *Haemophilus*,

Salmonella do not alter their O-antigens upon infection, so the great diversity of antigenic types is not readily explained by the conventional argument that antigenic variation enables escape of immune surveillance. In another intriguing study, some *Salmonella* captured by the ciliate *Tetrahymena* were released in vesicles that promoted their survival in the face of subsequent ciliate grazing [50]. It is not clear that *Salmonella* promoted their release from the *Tetrahymena* predator or that expression of virulence loci enabled long-term survival in this environment. Yet both studies highlight the importance of examining the role of microbe–microbe interactions in assessing the importance of virulence loci, because predation and competition play critical roles in the survivorship of virtually all organisms.

Conclusions

Both prokaryotic and eukaryotic pathogens can adapt rapidly by gene acquisition and by gene loss. The benefits of both processes span the evolutionary life-span of pathogens, enabling continual adaptation to their hosts. Unlike the scenario in multicellular eukaryotes, extensive lateral gene-transfer is seen in pathogenic eukaryotes, suggesting that major forces in genome evolution are shared among unicellular organisms.

Acknowledgements

Research in the Lawrence laboratory is supported by grants from the National Science Foundation and the National Institutes of Health.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- of special interest
 - of outstanding interest
1. Welch RA, Burland V, Plunkett G III, Redford P, Roesch P, Rasko D, Buckles EL, Liou SR, Boutin A, Hackett J *et al.*: **Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli***. *Proc Natl Acad Sci USA* 2002, **99**:17020-17024.
 2. Winfield MD, Groisman EA: **Phenotypic differences between *Salmonella* and *Escherichia coli* resulting from the disparate regulation of homologous genes**. *Proc Natl Acad Sci USA* 2004, **101**:17162-17167.
The authors report that the dramatic differences between the *Escherichia coli* and *Salmonella enteric pmrD* homologues have resulted in the ability of the iron-responsive PmrAB system to respond to low magnesium concentrations through the PhoPQ-regulated PmrD.
 3. Parkhill J, Sebahia M, Preston A, Murphy LD, Thomson N, Harris DE, Holden MT, Churcher CM, Bentley SD, Mungall KL *et al.*: **Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica***. *Nat Genet* 2003, **35**:32-40.
 4. Davies J: **Origins and evolution of antibiotic resistance**. *Microbiologia* 1996, **12**:9-16.
 5. Hacker J, Kaper JB: **Pathogenicity islands and the evolution of microbes**. *Annu Rev Microbiol* 2000, **54**:641-679.
 6. Yap MN, Barak JD, Charkowski AO: **Genomic diversity of *Erwinia carotovora* subsp. *carotovora* and its correlation with virulence**. *Appl Environ Microbiol* 2004, **70**:3013-3023.
 7. Monteiro-Vitorello CB, Camargo LE, Van Sluys MA, Kitajima JP, Truffi D, do Amaral AM, Harakava R, de Oliveira JC, Wood D, de Oliveira MC *et al.*: **The genome sequence of the gram-positive sugarcane pathogen *Leifsonia xyli* subsp. *xyli***. *Mol Plant Microbe Interact* 2004, **17**:827-836.
 8. Chiu CH, Tang P, Chu C, Hu S, Bao Q, Yu J, Chou YY, Wang HS, Lee YS: **The genome sequence of *Salmonella enterica* serovar *Choleraesuis*, a highly invasive and resistant zoonotic pathogen**. *Nucleic Acids Res* 2005, **33**:1690-1698.
 9. Gill SR, Fouts DE, Archer GL, Mongodin EF, Deboy RT, Ravel J, Paulsen IT, Kolonay JF, Brinkac L, Beanan M *et al.*: **Insights on evolution of virulence and resistance from the complete genome analysis of an early methicillin-resistant *Staphylococcus aureus* strain and a biofilm-producing methicillin-resistant *Staphylococcus epidermidis* strain**. *J Bacteriol* 2005, **187**:2426-2438.
 10. Holden MT, Feil EJ, Lindsay JA, Peacock SJ, Day NP, Enright MC, Foster TJ, Moore CE, Hurst L, Atkin R *et al.*: **Complete genomes of two clinical *Staphylococcus aureus* strains: evidence for the rapid evolution of virulence and drug resistance**. *Proc Natl Acad Sci USA* 2004, **101**:9786-9791.
 11. Nembaware V, Seoighe C, Sayed M, Gehring C: **A plant natriuretic peptide-like gene in the bacterial pathogen *Xanthomonas axonopodis* may induce hyper-hydration in the plant host: a hypothesis of molecular mimicry**. *BMC Evol Biol* 2004, **4**:10.
 12. Cazalet C, Rusniok C, Bruggemann H, Zidane N, Magnier A, Ma L, Tichit M, Jarraud S, Bouchier C, Vandenesch F *et al.*: **Evidence in the *Legionella pneumophila* genome for exploitation of host cell functions and high genome plasticity**. *Nat Genet* 2004, **36**:1165-1173.
 13. Hanage WP, Fraser C, Spratt BG: **Fuzzy species among recombinogenic bacteria**. *BMC Biol* 2005, **3**:6.
 14. Andrews TD, Gojbori T: **Strong positive selection and recombination drive the antigenic variation of the PIIe protein of the human pathogen *Neisseria meningitidis***. *Genetics* 2004, **166**:25-32.
 15. Qiu WG, Schutzer SE, Bruno JF, Attie O, Xu Y, Dunn JJ, Fraser CM, Casjens SR, Luft BJ: **Genetic exchange and plasmid transfers in *Borrelia burgdorferi* sensu stricto revealed by three-way genome comparisons and multilocus sequence typing**. *Proc Natl Acad Sci USA* 2004, **101**:14150-14155.
Draft sequences of two *Borrelia* genomes were compared with the available complete genome to identify regions under positive selection for variability. Data from further clinical isolates establish recombination as a dominant force in *Borrelia* population genetics, with previous conclusions of clonality attributed to sampling bias.
 16. Chowdhury NR, Stine OC, Morris JG, Nair GB: **Assessment of evolution of pandemic *Vibrio parahaemolyticus* by multilocus sequence typing**. *J Clin Microbiol* 2004, **42**:1280-1282.
 17. Bisharat N, Cohen DI, Harding RM, Falush D, Crook DW, Peto T, Maiden MC: **Hybrid *Vibrio vulnificus***. *Emerg Infect Dis* 2005, **11**:30-35.
 18. Sherley M, Gordon DM, Collignon PJ: **Evolution of multi-resistance plasmids in Australian clinical isolates of *Escherichia coli***. *Microbiology* 2004, **150**:1539-1546.
 19. Andersson JO, Andersson SG: **Insights into the evolutionary process of genome degradation**. *Curr Opin Genet Dev* 1999, **9**:664-671.
 20. Cole ST, Eiglmeier K, Parkhill J, James KD, Thomson NR, Wheeler PR, Honore N, Garnier T, Churcher C, Harris D *et al.*: **Massive gene decay in the leprosy bacillus**. *Nature* 2001, **409**:1007-1011.
 21. Day WA Jr, Fernandez RE, Maurelli AT: **Pathoadaptive mutations that enhance virulence: genetic organization of the *cadA* regions of *Shigella* spp.** *Infect Immun* 2001, **69**:7471-7480.
 22. Thomson NR, Yeats C, Bell K, Holden MT, Bentley SD, Livingstone M, Cerdeno-Tarraga AM, Harris B, Doggett J, Ormond D *et al.*: **The *Chlamydophila abortus* genome sequence reveals an array of variable proteins that contribute to interspecies variation**. *Genome Res* 2005, **15**:629-640.
Comparison of the *Chlamydophila abortus* genome with those of congeners shows no evidence for lateral transfer in these intracellular patho-

gens, but plasticity is evident in pseudogenes affecting various metabolic pathways and production of membrane proteins.

23. Tsolaki AG, Hirsh AE, DeRiemer K, Enciso JA, Wong MZ, Hannan M, Goguet de la Salmoniere YO, Aman K, Kato-Maeda M, Small PM: **Functional and evolutionary genomics of *Mycobacterium tuberculosis*: insights from genomic deletions in 100 strains.** *Proc Natl Acad Sci USA* 2004, **101**:4865-4870.
Microarray analysis of hundreds of clinical isolates of *Mycobacterium tuberculosis* shows widespread gene-loss affecting all classes of genes, including those involved basic metabolic functions in addition to those encoding membrane proteins, suggesting that gene loss might be more prevalent in this organism than was suspect.
24. Tong Z, Zhou D, Song Y, Zhang L, Pei D, Han Y, Pang X, Li M, Cui B, Wang J *et al.*: **Pseudogene accumulation might promote the adaptive microevolution of *Yersinia pestis*.** *J Med Microbiol* 2005, **54**:259-268.
25. Nakata N, Tobe T, Fukuda I, Suzuki T, Komatsu K, Yoshikawa M, Sasakawa C: **The absence of a surface protease, OmpT, determines the intercellular spreading ability of *Shigella*: the relationship between the *ompT* and *kcpA* loci.** *Mol Microbiol* 1993, **9**:459-468.
26. Samrakandi MM, Zhang C, Zhang M, Nietfeldt J, Kim J, Iwen PC, Olson ME, Fey PD, Duhamel GE, Hinrichs SH *et al.*: **Genome diversity among regional populations of *Francisella tularensis* subspecies *tularensis* and *Francisella tularensis* subspecies *holartctica* isolated from the US.** *FEMS Microbiol Lett* 2004, **237**:9-17.
27. Finnan S, Morrissey JP, O'Gara F, Boyd EF: **Genome diversity of *Pseudomonas aeruginosa* isolates from cystic fibrosis patients and the hospital environment.** *J Clin Microbiol* 2004, **42**:5783-5792.
28. Rohmer L, Guttman DS, Dangel JL: **Diverse evolutionary mechanisms shape the type III effector virulence factor repertoire in the plant pathogen *Pseudomonas syringae*.** *Genetics* 2004, **167**:1341-1360.
29. Seshadri R, Myers GS, Tettelin H, Eisen JA, Heidelberg JF, Dodson RJ, Davidsen TM, DeBoy RT, Fouts DE, Haft DH *et al.*: **Comparison of the genome of the oral pathogen *Treponema denticola* with other spirochete genomes.** *Proc Natl Acad Sci USA* 2004, **101**:5646-5651.
30. Nishio Y, Nakamura Y, Usuda Y, Sugimoto S, Matsui K, Kawarabayasi Y, Kikuchi H, Gojobori T, Ikeo K: **Evolutionary process of amino acid biosynthesis in *Corynebacterium* at the whole genome level.** *Mol Biol Evol* 2004, **21**:1683-1691.
31. Andersson JO: **Lateral gene transfer in eukaryotes.** *Cell Mol Life Sci* 2005, **62**:1182-1197.
32. Richards TA, Hirt RP, Williams BA, Embley TM: **Horizontal gene transfer and the evolution of parasitic protozoa.** *Protist* 2003, **154**:17-32.
33. Garcia-Vallve S, Romeu A, Palau J: **Horizontal gene transfer of glycosyl hydrolases of the rumen fungi.** *Mol Biol Evol* 2000, **17**:352-361.
34. Temporini ED, VanEtten HD: **An analysis of the phylogenetic distribution of the pea pathogenicity genes of *Nectria haematococca* MPVI supports the hypothesis of their origin by horizontal transfer and uncovers a potentially new pathogen of garden pea: *Neocosmospora boniensis*.** *Curr Genet* 2004, **46**:29-36.
35. Fast NM, Law JS, Williams BA, Keeling PJ: **Bacterial catalase in the microsporidian *Nosema locustae*: implications for microsporidian metabolism and genome evolution.** *Eukaryot Cell* 2003, **2**:1069-1075.
36. Mukai A, Endoh H: **Presence of a bacterial-like citrate synthase gene in *Tetrahymena thermophila*: recent lateral gene transfers (LGT) or multiple gene losses subsequent to a single ancient LGT?** *J Mol Evol* 2004, **58**:540-549.
37. Andersson JO, Sarchfield SW, Roger AJ: **Gene transfers from nanoarchaeota to an ancestor of diplomonads and parabasalids.** *Mol Biol Evol* 2005, **22**:85-90.
38. Loftus B, Anderson I, Davies R, Alsmark UC, Samuelson J, Amedeo P, Roncaglia P, Berriman M, Hirt RP, Mann BJ *et al.*: **The genome of the protist parasite *Entamoeba histolytica*.** *Nature* 2005, **433**:865-868.
39. van der Giezen M, Cox S, Tovar J: **The iron-sulfur cluster assembly genes *iscS* and *iscU* of *Entamoeba histolytica* were acquired by horizontal gene transfer.** *BMC Evol Biol* 2004, **4**:7.
Phylogenies of two iron-sulfur assembly proteins in the amitochondrial protist *Entamoeba histolytica* each cluster the *Entamoeba* protein with the ϵ -proteobacterial homologues, supporting both a bacterial origin for these proteins and one distinct from homologues in mitochondria-bearing eukaryotes.
40. Esser C, Ahmadinejad N, Wiegand C, Rotte C, Sebastiani F, Gelius-Dietrich G, Henze K, Kretschmann E, Richly E, Leister D *et al.*: **A genome phylogeny for mitochondria among alpha-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes.** *Mol Biol Evol* 2004, **21**:1643-1660.
41. Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, Bartholomeu DC, Lennard NJ, Caler E, Hamlin NE, Haas B *et al.*: **The genome of the African trypanosome *Trypanosoma brucei*.** *Science* 2005, **309**:416-422.
The genome of *Trypanosoma brucei* uncovers not only a wealth of biology but evidence for the acquisition of dozens of genes from bacteria, including those involved in the electron transport chain and glycoprotein attachment, offering potential drug targets.
42. El-Sayed NM, Myler PJ, Bartholomeu DC, Nilsson D, Aggarwal G, Tran AN, Ghedin E, Worthey EA, Delcher AL, Blandin G *et al.*: **The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease.** *Science* 2005, **309**:409-415.
43. Loftus BJ, Fung E, Roncaglia P, Rowley D, Amedeo P, Bruno D, Vamathevan J, Miranda M, Anderson JJ, Fraser JA *et al.*: **The genome of the basidiomycetous yeast and human pathogen *Cryptococcus neoformans*.** *Science* 2005, **307**:1321-1324.
44. Gardner MJ, Bishop R, Shah T, de Villiers EP, Carlton JM, Hall N, Ren Q, Paulsen IT, Pain A, Berriman M *et al.*: **Genome sequence of *Theileria parva*, a bovine pathogen that transforms lymphocytes.** *Science* 2005, **309**:134-137.
45. Slamovits CH, Fast NM, Law JS, Keeling PJ: **Genome compaction and stability in microsporidian intracellular parasites.** *Curr Biol* 2004, **14**:891-896.
46. Swanson MS, Hammer BK: ***Legionella pneumophila* pathogenesis: a fateful journey from amoebae to macrophages.** *Annu Rev Microbiol* 2000, **54**:567-613.
47. Harb OS, Gao LY, Abu Kwaik Y: **From protozoa to mammalian cells: a new paradigm in the life cycle of intracellular bacterial pathogens.** *Environ Microbiol* 2000, **2**:251-265.
48. Skriwan C, Fajardo M, Hagele S, Horn M, Wagner M, Michel R, Krohne G, Schleicher M, Hacker J, Steinert M: **Various bacterial pathogens and symbionts infect the amoeba *Dictyostelium discoideum*.** *Int J Med Microbiol* 2002, **291**:615-624.
49. Wildschutte H, Wolfe DM, Tamewitz A, Lawrence JG: **Protozoan predation, diversifying selection and the evolution of antigenic diversity in *Salmonella*.** *Proc Natl Acad Sci USA* 2004, **101**:10644-10649.
Salmonella bearing different O-antigen carbohydrates are able to escape predation by different sets of amoeboid predators, suggesting that variability at this virulence locus increases *Salmonella* fitness by enabling it to escape surveillance by predators, although not necessarily the immune system.
50. Brandl MT, Rosenthal BM, Haxo AF, Berk SG: **Enhanced survival of *Salmonella enterica* in vesicles released by a soilborne *Tetrahymena* species.** *Appl Environ Microbiol* 2005, **71**:1562-1569.



ELSEVIER

The microbial pan-genome

Duccio Medini¹, Claudio Donati¹, Hervé Tettelin², Vega Masiagnani¹ and Rino Rappuoli¹

A decade after the beginning of the genomic era, the question of how genomics can describe a bacterial species has not been fully addressed. Experimental data have shown that in some species new genes are discovered even after sequencing the genomes of several strains. Mathematical modeling predicts that new genes will be discovered even after sequencing hundreds of genomes per species. Therefore, a bacterial species can be described by its pan-genome, which is composed of a 'core genome' containing genes present in all strains, and a 'dispensable genome' containing genes present in two or more strains and genes unique to single strains. Given that the number of unique genes is vast, the pan-genome of a bacterial species might be orders of magnitude larger than any single genome.

Addresses

¹ Immunobiological Research Institute of Siena (IRIS), Chiron Vaccines, via Fiorentina 1, 53100 Siena, Italy

² Department of Microbial Genomics, The Institute for Genomic Research (TIGR), 9712 Medical Center Drive, Rockville, MD 20850, USA

Corresponding author: Rappuoli, Rino (rino_rappuoli@chiron.com)

Current Opinion in Genetics & Development 2005, 15:589–594

This review comes from a themed issue on Genomes and evolution
Edited by Stephen J O'Brien and Claire M Fraser

Available online 26th September 2005

0959-437X/\$ – see front matter

© 2005 Elsevier Ltd. All rights reserved.

DOI 10.1016/j.gde.2005.09.006

Introduction to the pan-genome

Ten years after the first sequence of a free-living organism was revealed, public databases contain 239 complete bacterial genomes. However, as shown in Table 1, in 83% and 8% of the cases, only one or two genomes per bacterial species have been sequenced, respectively. In a recent work [1•], eight genomes representative of the serogroup (see Glossary) diversity among group B *Streptococcus* (GBS) strains were analyzed to answer the question of how many genomes are needed to fully describe a bacterial species. Each GBS strain was found to contain an average of 1806 genes that are present in every strain (core genome [see Glossary]), plus 439 genes that are absent in one or more strains (dispensable genome [see Glossary]).

The dispensable genes are also divided into genes present in two or more but not all strains (18% of the genome)

and genes unique to each strain (1.5% of the genome). Mathematical modeling based on the eight genomes showed that unique genes will continue to emerge even after hundreds or thousands of genomes are sequenced [1•]. Hence, core and dispensable genes represent the essence and the diversity of the species, respectively.

The surprising conclusion from the study is that, in theory, the bacterial species will never be fully described, because new genes will be added to the genome of the species with each new genomic sequence. Therefore, the best approximation to describe a species is to use the concept of the pan-genome ('pan' — 'παν' in Greek — means 'whole' [see Glossary]), which is made up of the sum of core and dispensable genomes (Figure 1). In the case of GBS, presently, the pan-genome contains 2713 genes, of which 1806 belong to the core genome, and 907 belong to the dispensable genome. The GBS pan-genome is predicted to grow by an average of 33 new genes every time a new strain is sequenced (Figure 1). Similar analysis [1•] carried out on five strains of *Streptococcus pyogenes* revealed a similar genomic diversity, indicating an asymptotic value of 27 specific genes for each new genome added, leading, again, to an 'open' pan-genome. A different behavior was observed in the study of eight independent *Bacillus anthracis* isolates. In this case, the number of specific genes added to the pan-genome was found to rapidly converge to zero after the addition of only a fourth genome [1•]. Hence, the *B. anthracis* species has a 'closed' pan genome, and four genome sequences are sufficient to completely characterize this species.

In this review, we discuss how the concept of the pan-genome might fit with the available data and consider which experiments need to be done to address the questions raised by this concept.

A large microbial gene pool driving evolution

Much indirect evidence had already hinted at the concept of the pan-genome, even before it was properly defined by mathematical quantification [1•]. Several studies of subtractive hybridization and comparative genome hybridization (CGH) using multiple isolates of the same species had shown that bacterial species such as *Helicobacter pylori*, *Staphylococcus aureus* and *Escherichia coli* display an extensive genetic diversity, with an average of 20–35% of genes being specific for a single strain [2–4].

The presence of so many strain-specific genes in each of these species suggests that — as in the case of GBS — they could also display an open pan-genome. This raises

Glossary

Core genome: The pool of genes shared by all the strains of the same bacterial species.

Dispensable genome: The pool of genes present in some — but not all — strains of the same bacterial species.

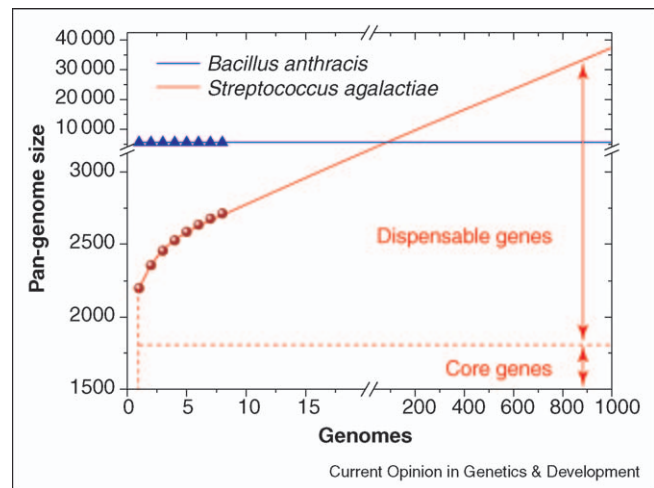
Lateral gene transfer: Mechanism by which an individual of one species transfers genetic material (i.e. DNA) to an individual of a different species.

Pan-genome: The global gene repertoire of a bacterial species: core genome + dispensable genome.

Serogroup: Group of related bacterial strains characterized by the same composition of the capsular polysaccharide.

the question of whether the microbial world contains enough genes to fit the prediction of such a vast gene pool generated by the pan-genome.

For instance, it has been shown that a single environmental sample of DNA from marine water encodes more than 1.2 million previously unknown genes from 1800 predicted genomic species [5^{••}]. Similar results have also been obtained for a totally different ecosystem, the human gastrointestinal tract. In this case, almost 400 different bacterial phylotypes were identified, of which 244 were novel [6[•]]. Even more impressive is the recent estimate of 10^7 distinct bacterial species in a 10 gram soil sample containing a total of approximately 10^{10} cells [7^{••}], a species diversity two orders of magnitude larger than previous estimates [8], showing that further quantitative and rational explorations of microbial ecology are strongly needed [9[•]]. Furthermore, a great heterogeneity was also identified when looking at a single species within a well-defined natural bacterial population: 16S RNA sequencing and pulse-field gel electrophoresis (PFGE) analysis were performed to study the diversity associated with the species *Vibrio splendidus* within coastal bacterioplankton, revealing in this same species the presence of as many as

Figure 1

The set of genes pertaining to a species, or species pan-genome, depends on the number of available genome sequences. In this figure, the size of *S. agalactiae* (red dots) and *B. anthracis* (blue triangles) pan-genomes are shown as a function of the number of sequenced strains. The curves represent a mathematical extrapolation of the data to a large number of strains. The size of a species pan-genome can grow with the number of sequenced strains, or quickly saturate to a limiting value. The *S. agalactiae* pan-genome is 'open'; the *B. anthracis* one is 'closed'. After sequencing a large number of strains, the number of dispensable genes in an open pan-genome is orders of magnitude larger than the size of the core genome, forcing us to reconsider the definition of a bacterial species.

1287 distinct genotypes, most of which are differentiated by the insertion or deletion of large genomic elements [10^{••}]. Although new genes can originate through duplication of existing sequences, followed by diversification, the most common way to acquire new functions is by the

Table 1**Number of genomes sequenced in different bacterial species.**

Species with sequenced genome(s)	Number of species (% of the total)	Number of genomes sequences per species
<i>Streptococcus agalactiae</i> , <i>Bacillus anthracis</i> , <i>Burkholderia mallei</i>	3 (1.2%)	8
<i>Burkholderia pseudomallei</i>	1 (0.4%)	7
<i>Staphylococcus aureus</i> , <i>Streptococcus pyogenes</i>	2 (0.8%)	6
<i>Salmonella enterica</i> , <i>Escherichia coli</i> , <i>Bacillus cereus</i> , <i>Chlamydomydia pneumoniae</i> , <i>Haemophilus influenzae</i> , <i>Listeria monocytogenes</i> , <i>Xylella fastidiosa</i>	7 (2.8%)	5
<i>Prochlorococcus marinus</i> , <i>Buchnera aphidicola</i> , <i>Burkholderia cenocepacia</i> , <i>Ehrlichia ruminantium</i> , <i>Legionella pneumophila</i> , <i>Pseudomonas syringae</i> , <i>Streptococcus thermophilus</i> , <i>Yersinia pestis</i>	8 (3.2%)	3
<i>Streptococcus pneumoniae</i> , <i>Mycobacterium tuberculosis</i> , <i>Neisseria meningitidis</i> , <i>Bacillus licheniformis</i> , <i>Bifidobacterium longum</i> , <i>Campylobacter jejuni</i> , <i>Chlorobium phaeobacteroides</i> , <i>Corynebacterium glutamicum</i> , <i>Haemophilus somnus</i> , <i>Helicobacter pylori</i> , <i>Lactococcus lactis</i> , <i>Leptospira interrogans</i> , <i>Mycoplasma genitalium</i> , <i>Pseudomonas aeruginosa</i> , <i>Shigella flexneri</i> , <i>Staphylococcus epidermidis</i> , <i>Synechococcus elongates</i> , <i>Thermus thermophilus</i> , <i>Tropherymaa whipplei</i> , <i>Vibrio vulnificus</i> , <i>Xanthomonas campestris</i>	21 (8.3%)	2
Various species	211 (83.3%)	1

Bacterial species for which multiple sequenced strains are available represent a small fraction of the species for which only one strain has been sequenced to date. In this table, we report the number of strains for these species, and the percentage that they represent over the total number of sequenced bacterial species.

transfer of genetic material from unrelated organisms. The importance of the mechanisms of lateral gene transfer (see Glossary) in evolutionary processes has been hotly debated in recent years [11–16,17*,18], but it is now generally accepted that it represents an evolutionary ‘fast route’, which enables an organism to quickly adapt to a changing environment.

Genes from this large pool are continuously exchanged within and between bacterial species by three main processes: (i) by transformation, when genetic material can be taken up from the environment; (ii) by transduction, when the DNA is delivered by a virus; and (iii) by conjugation, when DNA is directly exchanged between cells. Transformation and conjugation require that the source and target organisms live in close contact, and bacteriophages might enable bacterial species populating different environments to exchange genetic material, which often contains genes that are crucially important for pathogenesis [19*]. Considering that the global population of phages has been estimated to be in the range of 10^{31} and that they are responsible for an average of 10^{23} infections per second [20], it is easy to conclude that the global pool of genes

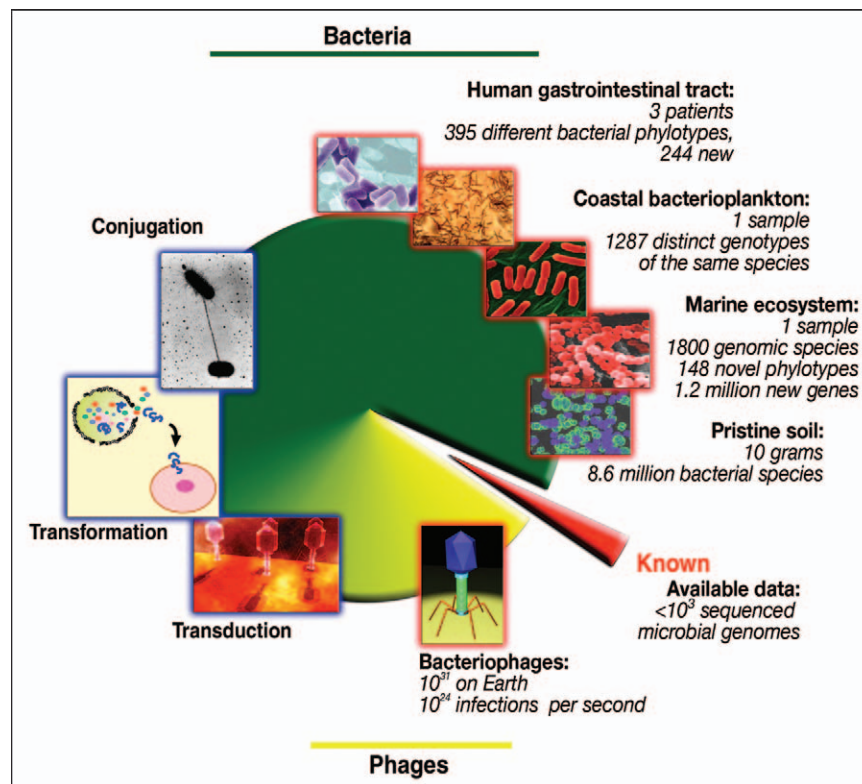
present in the microbial world is likely to exceed by several orders of magnitude any estimate that has been made to date, and that the presence of billions of genes is no longer unexpected (Figure 2).

The surprisingly large gene pool described here suggests that, during evolution, the vast majority of novel functions were probably generated in the microbial world and not in large animals, such as humans, which have only 25 000–35 000 genes. The consequence of this would be that microbes and large animals might have totally different roles in evolution. In fact, under this theory, microbes would generate new genes and functional modules, whereas large animals would evolve by first taking up modules generated by microbes and then by rearranging them in many different ways within the genome itself and by alternative splicing of the mRNAs.

Core and dispensable genes

In general, the core genome includes all genes responsible for the basic aspects of the biology of a species and its major phenotypic traits. By contrast, dispensable genes contribute to the species diversity and might encode

Figure 2

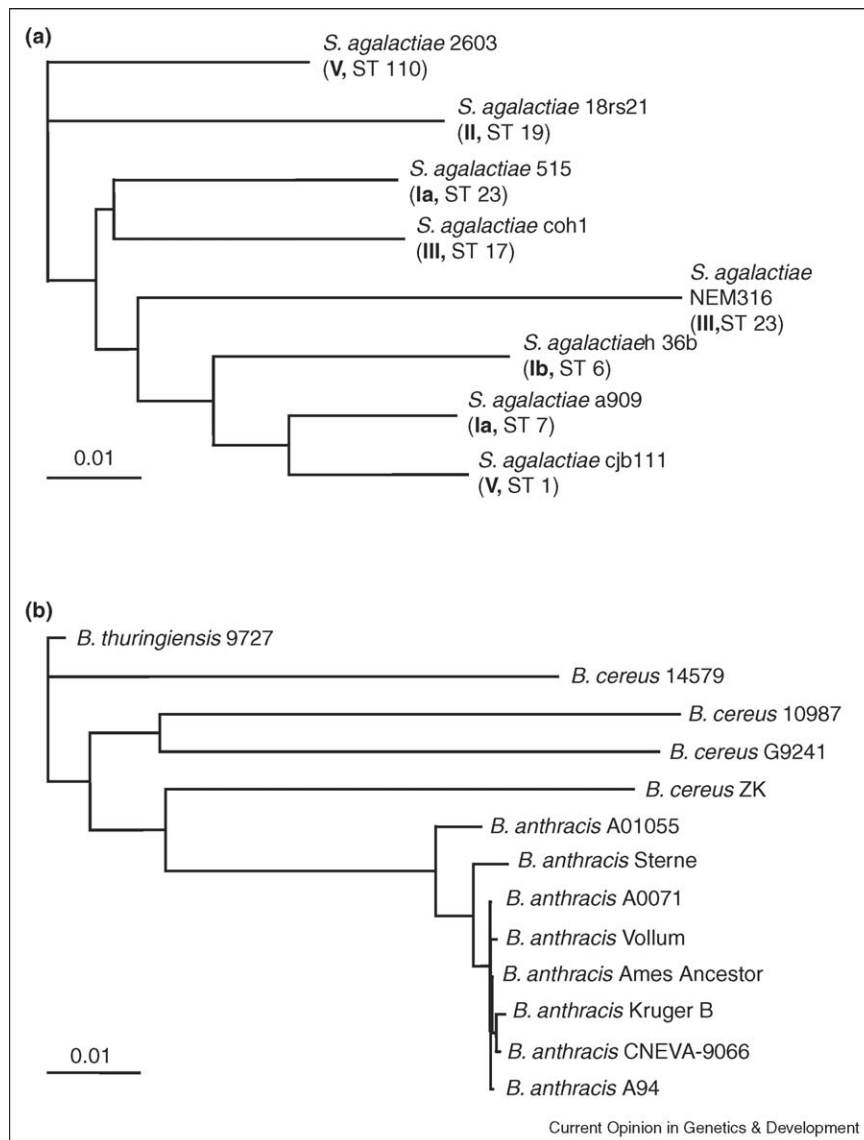


Diversity of the microbial universe. Recent experimental findings have shown that the genetic diversity within bacterial populations is much higher than expected [5**,6*,10**,20], raising the question of what is the origin of this genetic diversity. Bacteria can acquire genes from the environment by conjugation, transformation and phage infection (transduction), and experimental studies on environmental samples have shown that the amount of genetic material available in single ecosystems is large enough to constitute a virtually infinite reservoir of new genes. Estimates of the contribution of mechanisms of lateral gene transfer to the innovation rate show that genes acquired through this route give an essential contribution to species diversification [18,19*,20,21]. To date, of all this diversity, less than 1000 microbial genomes have been completely sequenced.

supplementary biochemical pathways and functions that are not essential for bacterial growth but which confer selective advantages, such as adaptation to different niches, antibiotic resistance, or colonization of a new host. Such genes are generally clustered on large genomic islands that are typically flanked by short repeated DNA sequences and are characterized by an abnormal G + C content. Investigation and functional annotation of dispensable genes reveals that hypothetical, phage- and

transposon-related genes account for the vast majority of findings, whereas in a typical genome this type of gene represents much smaller percentages [21]. The fact that these genes are mostly associated with a limited number of strains indicates a weak positive selection for these functions and shows that mobile elements contribute poorly to the overall fitness and differentiation of the species, although sometimes they can carry important genes [19^{*},22^{**}]. Given that these genes are not necessary for

Figure 3



Dendrograms of the eight *Streptococcus agalactiae* (a) and thirteen *B. cereus* group (b) genomes. The fraction of genes of one strain that is not shared with other strains was used to define a distance matrix. The matrix was then used to build a dendrogram with the neighbor-joining method, as implemented in the NEIGHBOR program of the PHYLIP suite (<http://evolution.genetics.washington.edu/phylip.html>). Tree branches are proportional to the fraction of the gene content not shared between the different isolates, and the ruler shows the length corresponding to 1% difference in gene content.

For each *S. agalactiae* strain, serogroup (bold, roman letters) and sequence type (ST) are reported in brackets. The *S. agalactiae* and *B. cereus* group genomes are available for downloading at <http://www.ncbi.nlm.nih.gov>. From the figure, it is evident that the distance between two *S. agalactiae* strains is comparable to the distance between *B. anthracis* stains and other *B. cereus* group species, making the definition of *B. anthracis* as an autonomous species questionable.

survival or maintenance of the species, they can also be deleted from the genome; however, in pathogenic species, this loss is often accompanied by a parallel reduction in virulence. For example, a spontaneous loss of the genes coding for fimbriae, hair-like projections thought to have an important role in colonization, has been observed in successive passages of *in vitro* cultures of *Haemophilus influenzae* and *E. coli*. Similarly, GBS was recently found to encode a pilus-like structure, which is not ubiquitous in all strains, and the presence or absence of which could be related to either gene acquisition or loss [23].

Serotypes and sequence types do not correlate with genomic diversity

Classical methods to catalogue bacterial species are based on knowledge convenient phenotypic traits. The most popular is the agglutination of bacterial cells by specific antisera against the capsular polysaccharide surrounding many pathogens. For a variety of encapsulated bacteria, this method has been widely used for epidemiology studies and vaccine design, assuming that all strains belonging to the same serogroup are similar. More recently, techniques such as multilocus enzyme electrophoresis (MLEE) and multilocus sequence typing (MLST), which are based on the detection of variability associated with housekeeping genes, were applied to several bacterial species and led to the classification of strains into 'clonal complexes' and sequence types, respectively.

However, comparison of the whole genome sequences of GBS strains has shown that the genomic diversity does not segregate with serotypes or MLST sequence-types (Figure 3a). In fact, the analysis revealed that, often, isolates belonging to different serogroups are more closely related than are isolates of the same serogroup, and that strains of the same sequence type can be genetically very distant (Figure 3a). The reason for the absence of correlation between serotypes and genetic diversity is likely to reside in the fact that capsular specificity genes are present in the dispensable genome, which is exchanged freely between strains with different genetic background. By contrast, the genes used to determine the MLST type belong to the core genome, and they do not pick up similarities present in the dispensable genome, which often are linked to pathogenic features.

Challenging the concept of species

Species can have an open or a closed pan-genome. An open pan-genome is typical of those species that colonize multiple environments and have multiple ways of exchanging genetic material. *Streptococci*, *Meningococci*, *H. pylori*, *Salmonellae* and *E. coli* have these properties and are likely to have an open pan-genome. By contrast, other species such as *B. anthracis*, *Mycobacterium tuberculosis* and *Chlamydia trachomatis*, which are known to be more conserved, live in isolated niches with limited access to the global microbial gene pool. Such species,

with a low capacity to acquire foreign genes, have a closed pan-genome. An extreme example is represented by *Buchnera aphidicola*, an endosymbiont of aphids, the genome of which has undergone no chromosome rearrangements, duplications or horizontal gene transfer in the past 50 million years, thus demonstrating the most extreme genome stability observed to date [24].

A closer look at the structures of the genetic trees of open pan-genomic species (such as GBS; see Figure 3a) and closed pan-genomic species (such as *B. anthracis*) shows that the latter species resembles a clone of a *B. cereus* species rather than being a true independent species (Figure 3b). *B. anthracis* is, in fact, genetically very closely related to other members of the *B. cereus* group (*B. cereus* and *Bacillus thuringiensis* species), and the main feature that distinguishes these organisms is the acquisition of two virulence plasmids, one of which codes for anthrax toxin [25]. Although this feature is extremely important in justifying the classification of *B. anthracis* as an independent species, genetically, this is just a phenotypic trait encoded by the dispensable genome of the *B. cereus* group. This example shows that the criteria used to define microbial species might be inconsistent with the genetic information. In the future, we will need to consider how to handle these inconsistencies.

Conclusions and practical implications

The need to sequence multiple genomes from each species to better understand the diversity of bacterial species is not just a theoretical exercise. Recently, it has been shown that the design of a universal vaccine against GBS was only possible using dispensable genes [26*]. In addition, sequencing of multiple genomes was instrumental in discovering the presence of the pilus in GBS, group A *Streptococcus*, and *Pneumococcus*, an essential virulence factor that had been missed by all conventional technologies for a whole century [23]. It is very likely that the study of the bacterial pan-genome will continue to surprise us with fascinating discoveries that cannot be predicted with the conventional methods used to date in microbiology.

Acknowledgements

The authors would like to thank Michael Cieslewicz, Antonello Covacci and John Telford for their contribution to the pan-genome concept. They are also grateful to the IRIS Bioinformatic group, to the TIGR information technology and database server groups led by Vadim Sapiro and Michael Heaney, respectively, and to Giorgio Corsi for artwork.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- of special interest
 - of outstanding interest
1. Tettelin H, Massignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones A, Durkin AS et al.: **Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial 'pan-genome'**. *Proc Natl Acad Sci USA* 2005, in press.

This study reports the first quantification of the diversity of a single prokaryotic species on the basis of genomic sequences of multiple strains. The authors introduce the pan-genome concept: the gene pool pertaining to a single species, which can be orders of magnitude larger than the genome of any single isolate.

2. Fitzgerald JR, Sturdevant DE, Mackie SM, Gill SR, Musser JM: **Evolutionary genomics of *Staphylococcus aureus*: insights into the origin of methicillin-resistant strains and the toxic shock syndrome epidemic.** *Proc Natl Acad Sci USA* 2001, **98**:8821-8826.
3. Dorrell N, Mangan JA, Laing KG, Hinds J, Linton D, Al-Ghusein H, Barrell BG, Parkhill J, Stoker NG, Karlyshev AV *et al.*: **Whole genome comparison of *Campylobacter jejuni* human isolates using a low-cost microarray reveals extensive genetic diversity.** *Genome Res* 2001, **11**:1706-1715.
4. Fukiya S, Mizoguchi H, Tobe T, Mori H: **Extensive genomic diversity in pathogenic *Escherichia coli* and *Shigella* strains revealed by comparative genomic hybridization microarray.** *J Bacteriol* 2004, **186**:3911-3921.
5. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W *et al.*: **Environmental genome shotgun sequencing of the Sargasso sea.** *Science* 2004, **304**:66-74.
 The first work in which the 'whole-genome shotgun sequencing technique' was applied to an environmental sample. It greatly improves our understanding of how complex and variable is the world of uncultured bacteria.
6. Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M, Gill SR, Nelson KE, Relman DA: **Diversity of the human intestinal microbial flora.** *Science* 2005, **308**:1635-1638.
 A study showing that intestinal flora contains a much higher number of species than expected, with relevant variations both between different sites and between different subjects.
7. Gans J, Wolinsky M, Dunbar J: **Computational improvements reveal great bacterial diversity and high metal toxicity in soil.** *Science* 2005, **309**:1387-1390.
 This excellent study introduces innovative computational methods for measuring species abundance in environmental samples, leading to an increase of more than two orders of magnitude in current estimates of species diversity.
8. Sandaa R, Torsvik V, Enger Ø, Daae FL, Castberg T, Hahn D: **Analysis of bacterial communities in heavy metal-contaminated soils at different levels of resolution.** *FEMS Microbiol Ecol* 1999, **30**:237-251.
9. Curtis TP, Sloan WT: **Exploring microbial diversity – a vast • below.** *Science* 2005, **309**:1332-1333.
 Illuminating perspective on the new frontiers opened in microbial ecology by the use of advanced mathematical and computational approaches to the quantification of prokaryotic diversity.
10. Thompson JR, Pacocha S, Pharino C, Klepac-Ceraj V, Hunt DE, Benoit J, Sarma-Rupavtarm R, Distel DL, Polz MF: **Genotypic diversity within a natural coastal bacterioplankton population.** *Science* 2005, **307**:1311-1313.
 This study presents illuminating data on the genetic diversity of a bacterial population, clearly showing that macroscopic differences in genetic content are present also within seemingly homogeneous populations from a single ecosystem.
11. Kurland CG, Canback B, Berg OG: **Horizontal gene transfer: a critical view.** *Proc Natl Acad Sci USA* 2003, **100**:9658-9662.
12. Lawrence JG, Hendrickson H: **Lateral gene transfer: when will adolescence end?** *Mol Microbiol* 2003, **50**:739-749.
13. Koonin EV: **Horizontal gene transfer: the path to maturity.** *Mol Microbiol* 2003, **50**:725-727.
14. Ochman H, Lerat E, Daubin V: **Examining bacterial species under the specter of gene transfer and exchange.** *Proc Natl Acad Sci USA* 2005, **102**(Suppl 1):6595-6599.
15. Simonson AB, Servin JA, Skophammer RG, Herbold CW, Rivera MC, Lake JA: **Decoding the genomic tree of life.** *Proc Natl Acad Sci USA* 2005, **102**(Suppl 1):6608-6613.
16. Hooper SD, Berg OG: **Duplication is more common among laterally transferred genes than among indigenous genes.** *Genome Biol* 2003, **4**:R48.
17. Dobrindt U, Hochhut B, Hentschel U, Hacker J: **Genomic islands • in pathogenic and environmental microorganisms.** *Nat Rev Microbiol* 2004, **2**:414-424.
 Comprehensive review presenting the results on lateral gene transfer obtained for pathogenicity islands in pathogenic bacteria. It shows that lateral gene transfer is a universal mechanism of evolution.
18. Kazazian HH Jr: **Mobile elements: drivers of genome evolution.** *Science* 2004, **303**:1626-1632.
19. Brussow H, Canchaya C, Hardt WD: **Phages and the evolution • of bacterial pathogens: from genomic rearrangements to lysogenic conversion.** *Microbiol Mol Biol Rev* 2004, **68**:560-602.
 A clear review highlighting the importance of phages as a source of novel genes in bacterial evolution.
20. Hendrix RW: **Bacteriophage genomics.** *Curr Opin Microbiol* 2003, **6**:506-511.
21. Daubin V, Ochman H: **Bacterial genomes as new gene homes: the genealogy of ORFans in *E. coli*.** *Genome Res* 2004, **14**:1036-1042.
22. Feil EJ: **Small change: keeping pace with microevolution. •• Nat Rev Microbiol** 2004, **2**:483-495.
 Outstanding review, detailing the latest experimental results for characterizing species variability, with a special focus on genome sequencing, microarray and MLST data. The consequences on the concept of species and on evolutionary mechanisms in bacteria are thoroughly discussed.
23. Lauer P, Rinaudo CD, Soriani M, Margarit I, Maione D, Rosini R, Taddei AR, Mora M, Rappuoli R, Grandi G *et al.*: **Genome analysis reveals pili in Group B *Streptococcus*.** *Science* 2005, **309**:105.
24. Tamas I, Klasson L, Canback B, Naslund AK, Eriksson AS, Wernegreen JJ, Sandstrom JP, Moran NA, Andersson SG: **50 million years of genomic stasis in endosymbiotic bacteria.** *Science* 2002, **296**:2376-2379.
25. Rasko DA, Altherr MR, Han CS, Ravel J: **Genomics of the *Bacillus cereus* group of organisms.** *FEMS Microbiol Rev* 2005, **29**:303-329.
26. Maione D, Margarit I, Rinaudo CD, Massignani V, Mora M, Scarselli M, Tettelin H, Brettoni C, Iacobini ET, Rosini R *et al.*: **Identification of a universal group B *Streptococcus* vaccine by multiple genome screen.** *Science* 2005, **309**:148-150.
 A clear demonstration of the practical importance of sequencing multiple strains of a single pathogen for the formulation of an effective vaccine.



ELSEVIER

Ancestral state reconstructions for genomes

Christos A Ouzounis^{1,2}

The recent expansion of phylogenetic analysis from the traditional field of molecular evolution, analyzing histories of genes, to the nascent field of 'genomic evolution', analyzing histories of entire genomes, enables the construction of trees based on genome information, the quantification of the key processes that shape genome content and, ultimately, plausible parsimony reconstructions of ancestral genomes. Thus, when genomes are considered as phylogenetic characters, it is possible to reconstruct not only the history of species but also the ancestral states in terms of genome structure or function. In the future, we might be able to accurately reconstruct — or retrodict — a chain of events that led to the emergence of a specific genome sequence and, ultimately, to synthesize ancestral genomes at will, creating a 'Jurassic database' of genomes.

Addresses

¹ Computational Genomics Group, The European Bioinformatics Institute, EMBL Cambridge Outstation, Cambridge, CB10 1SD, UK

² Computational Genomics Unit, National Center for Research and Technology, GR-57001 Thessaloniki, Greece

Corresponding author: Ouzounis, Christos A (ouzounis@ebi.ac.uk)

Current Opinion in Genetics & Development 2005, **15**:595–600

This review comes from a themed issue on
Genomes and evolution
Edited by Stephen J O'Brien and Clare M Fraser

Available online 7th October 2005

0959-437X/\$ – see front matter
© 2005 Elsevier Ltd. All rights reserved.

DOI 10.1016/j.gde.2005.09.011

Introduction

Evolution — from genes to genomes

In recent years, we have witnessed an unprecedented increase of genome information, marked by the milestones of completed genome projects for 100 species in 2003 [1], 200 species in 2005 [2], and growing. It was thus unavoidable that, at some point, the evolutionary analysis of individual genes would expand to encompass genome-wide characters [3]. Such distinguishing genome traits include amino acid composition [4], genomic structure [5], gene fusions [6], gene clusters [7], and global patterns of genome divergence (e.g. [8*]).

Apart from conveying important structural and functional detail about the evolution of genomes, these characters — markedly divergent across species — have also been exploited for the construction of so-called 'genome trees'.

These can be defined as dendrograms based on the clustering and tree representation of genome relationships, according to various measures of genome divergence, distance metrics and approaches for tree construction. The fact that we are able to construct genome trees — reviewed recently [9*] — is important by itself, but this issue forms only a part of the main theme of this review.

Instead, we summarize here recent progress in the reconstruction of ancestral genome states — using genome and other trees — the quantification of forces that shape genome content, advances that challenge the notion of a tree of life, future directions and open problems. It is crucial to realize that although, from a methodological viewpoint, the reconstruction of ancestral states for genomes does not strictly require genome trees, the explicit use of such trees offers a host of superior solutions. This is because genome trees provide the opportunity to take into account the key processes shaping genomes, primarily gene loss and horizontal gene transfer (HGT). Using gene trees, it is not usually feasible to obtain evidence on whether HGT has or has not taken place, what is its relative frequency, or which genes are absent and why, for a particular species. By considering genome trees, these effects are attenuated, and the reliability of detected species-relationships increases, by quantifying gene presence/absence patterns.

The structure of this review is as follows: first, the general patterns of genome structure that serve as phylogenetic characters are discussed; second, the treatment of evolutionary forces by a handful of methods is explained; third, the reconstruction of ancestral states is presented, taking into account these factors; and, finally, an outline for certain open problems and future directions is provided.

Genome patterns as phylogenetic characters

The fundamental premise of molecular evolution is that individual characters under consideration differ substantially across species and, thus, can convey evolutionary evidence for the reconstruction of species relationships [9*]. Moreover, it has been pointed out many times that in traditional work, in which gene or protein sequences are considered, the dendrograms reflect, mostly, the history of molecules and not necessarily that of species — unless specific conditions are met [10]. This important issue is certainly valid for gene-based trees and possibly for the emerging field of genome trees as well.

The first genome trees appeared six years ago, simultaneously produced by several groups [11–13]. Despite

differences in approach, the common theme in these analyses was the consideration of the presence/absence of genes or global measures of similarity as the principal elements in the construction of a distance matrix and tree (Figure 1a).

All detected homologies were counted in two methods [11,13], with slightly unsatisfactory results: the generated trees are less sharp than they are when using orthology [12], and possibly somewhat inaccurate. It is indicative

that the homology approach has not been used subsequently.

Other advances, taking into account orthologous clusters [14] or protein folds [15[•]], have been made. For a justification of using protein folds, and an insightful discussion of eliminating paralogy, see the study by Yang *et al.* [15[•]].

An important issue realized early was the need for genome size normalization [12]: larger genomes have a

Figure 1

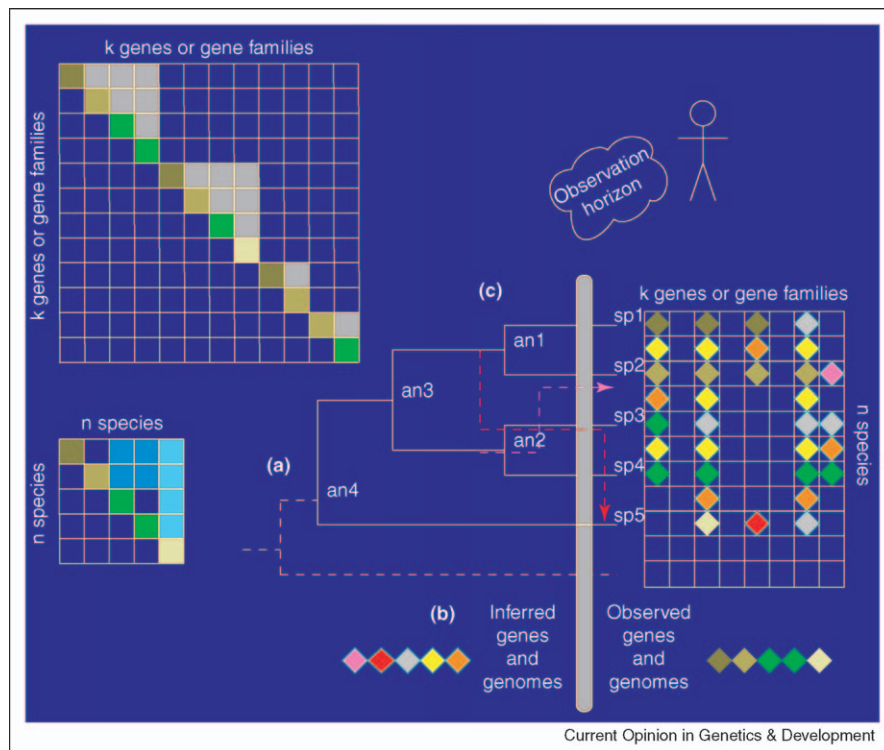


Illustration of methods discussed in the three main sections of the text. **(a)** Genome trees; **(b)** Ancestral state inference; **(c)** The net of life.

The initial data are typically a list of k genes or gene families and their allocation across n species (right side, or 'observed genomes'). For simplicity, only four gene families are displayed in odd-numbered 1/3/5/7 columns, and even-numbered columns 2/4/6 are left empty; column 8 is used for an alternative scenario of column 7. In this example, species are shown (sp1–sp5) in five rows, and are color-coded in dark-brown, light-brown, dark-green, light-green and beige colors. Present, native genes are shown as diamonds in the corresponding colors only ('observed genes'). In detail, sp2 and sp4 contain four native genes each, sp1 contains three native genes, and sp3 and sp5 contain one native gene each. In all, 13 'observed' genes are distributed across five species and four families. (a) After sequence comparison, a k - k matrix is produced, listing the genes by column from the k - n matrix, so that gene (protein) families are easily seen (the pairwise matches, and an arbitrary criterion of similarity, are displayed as grey boxes). Using various procedures (see 'Genome patterns as phylogenetic characters'), an n - n matrix is created that represents species relationships (shown in three hues of blue, and directly reflecting the similarities shown in the species tree), from which a genome tree is finally generated. Note that in this example tree four ancestral nodes are present (an1–an4). An outgroup is shown as a dotted line. (b) Algorithms that assess the likelihood that a gene was acquired or lost, based on the observed family distribution patterns, examine each family and infer a parsimonious scenario of its evolutionary history (see 'Quantifying the forces that shape gene content'). Inferred genes are shown as cyan-lined diamonds ('inferred genes'): orange is the last ancestor of each family, yellow depicts all other ancestors, light-grey represents a lost gene (or gene family), and the two purple/red genes correspond to cases of HGT; for simplicity, these were not included in the k - k matrix. The inferred genomes for the four ancestral states can thus be compiled ('inferred genomes'): an4 has two genes, and sp5 loses one gene and acquires another one (red gene, in family 3). An easy case to resolve is family 1, one gene (or gene family) per species for sp1 to sp4, an1 and an2 contain the gene and an3 is the last ancestor. The most complex case is family 4, in which two alternative scenarios can be imagined (shown in columns 7 and 8). Either family 4 was present in an4 and was subsequently lost in sp1, sp3 and sp5, and was conserved only in sp2 and sp4 or it was first invented in an2 (or an1), lost once in sp3 (sp1) and horizontally transferred from sp4 to sp2 (or vice versa; cases in parentheses not shown in the figure). (c) After these phases, it is possible to observe which families are likely to have been horizontally transferred, thus creating a 'net of life', overlaid on the tree of life (see 'A net of histories for extant and extinct genomes'). These examples, shown as dotted arrows, correspond to the two possible cases of HGT (purple and red).

chance to share more genes with similarly large genomes, irrespective of their relative phylogenetic distance. Conversely, smaller genomes are penalized when raw counts of shared genes are considered, even if they might be closely related to a species with a large genome. Therefore, normalization schemes were introduced, limiting the number of genes that can be shared across two genomes by dividing with the number of genes in the smallest genome: this is the maximum ‘share’ value that can be attained [8*].

This is, in fact, reminiscent of a similar concept at the protein level, called ‘conservation score’, which represents a robust measure of similarity [16].

Average DNA or protein sequence similarity was further proposed as an alternative measure for genome tree construction. The first such report [17] independently introduced the concept of normalized score, which was named ‘self-matching score’, for average sequence similarity.

This was the first case in which the large-scale clustering of all Proteobacteria resulted in a monophyletic grouping [17]. This particular example typically serves as a casual benchmark for several of these methods, owing to the complexity of genome relationships, independent genome reduction events (massive gene loss) and extensive species representation in this group of bacteria. Another method based on genomic DNA comparison was subsequently reported [18], with similarly impressive results. Other, less scalable methods have also been investigated, namely substitution rates across proteins [19], and bidirectional best hits [14]. It remains to be seen whether the application of the latter methods to hundreds of genomes is possible.

Finally, a composite measure for genome conservation, for which all hits were considered, has been presented recently [8*]. A significant advantage of this measure is that it takes into account both gene presence and sequence similarity. The robust clustering of 153 species by the assembly of a genome similarity ‘heat map’ suggests that most of these approaches are resistant against the effects of HGT. In fact, they challenge the notion that it is not possible to build a species tree from genome sequences, owing to the effects of HGT [20,21], because HGT is limited in the context of entire genome sequences [22]. We encounter issues of quantifying HGT and the reconstruction of ancestral genomes in the following sections of this review.

Taking these issues into account, the importance of genome trees in this framework is not so much about their ability to discern evolutionary relationships across species as an alternative to gene-based trees as it is about the reassurance that genome trees capture relationships

across taxa, without suffering from the symptoms of gene trees (for example, and most importantly, HGT and gene loss). In fact, it initially came as a surprise that genome trees are robust to these effects, and yet at the same time capture global genome similarities [8*,9*,22]. Quoting a critical review on HGT, it is stated that: “Other discontinuous genome events such as gene loss, gene duplication, and the segregation of paralogs as well as the generation of orphan sequences provide much more frequent challenges to genome phylogeny than does HGT” [23].

Quantifying the forces that shape gene content

Building genome or species trees is, therefore, crucial to the exploration of evolutionary relationships. Suppose, however, that a tree was provided by an independent method that didn’t rely on genome evolution and, thus, didn’t suffer from certain kinds of artefacts. Then, would it be possible to overlay our gene or genome information on such a tree and explore various aspects of molecular evolution? Although this is indeed the ideal scenario, unfortunately no such tree exists, with the possible exception of the vertebrate tree, and our reliance on the vertebrate fossil record. Indeed, this ‘chicken-and-egg’ problem — the challenge of both delineating species relationships and, at the same time, quantifying the relative contributions of evolutionary forces — has mired progress to date.

However, several groups have assessed recently the relative contribution of HGT events in species phylogenies [24–26]. To achieve this, all these studies employed various trees, including those (e.g. rRNA trees and curated taxonomy trees) that do not rely on genome comparisons. This strategy reflects a classification consensus from multiple sources, in addition to the use of genome trees, and is thus robust to ‘discontinuous genome events’ (see above).

Remarkably, most of the protein family distributions for extant species can be explained by patterns of vertical evolution, namely gene genesis and loss [24,25], regardless of the tree used, thus strengthening the notion that genome trees (in both cases, gene content trees) are robust to HGT (Figure 1b).

At least two groups have arrived independently at the same idea of using parsimony and an optimal HGT penalty value that effectively represents the expected relative loss versus HGT frequency [24,25]. The first study [24] demonstrated that gene fusion or duplication is virtually independent of HGT, in the case of Archaea and Proteobacteria, for a total of 17 genomes. The HGT penalty was explored as a parameter, and alternative scenarios for the relative contributions of gene gain (genesis and HGT) versus loss were investigated.

In a subsequent study [25], encompassing 51 microbial species, two constraints were exploited to obtain an optimal HGT penalty value: first, that the expected relative frequency of HGT versus loss should correspond to the observed ratio of these events; and, second, that the average genome size in prokaryotes remains constant [25]. Intriguingly, both these constraints result in highly similar value ranges, between 2 and 3, suggesting that HGT is probably three times less frequent than gene loss. In both cases, only protein families were considered, thus eliminating the issue of paralogy across genomes.

Together, these studies demonstrate that the majority of protein family distributions can be explained most parsimoniously by gene loss rather than by HGT. It remains to be seen whether more sophisticated methods will be able to infer taxa-specific parameters for gene gain and loss, what is the role of paralogy and differential gene loss within gene families, and what the relative contributions of these factors will be in the case of eukaryotic genomes (not examined in these studies). Most of these issues will be resolved when genome data availability becomes sufficient for these types of analyses.

A net of histories for extant and extinct genomes

One aim of the above mentioned studies was clearly the relative quantification of gene gain (genesis and HGT) versus gene loss. Yet implicit in these methods is the notion that all internal nodes of a species dendrogram, be it a curated taxonomy or a genome tree, contain a number of protein families that can be inferred on the basis of both the modeling of these evolutionary processes and the ensuing dendrogram.

The parsimony analysis of extant genomes can, therefore, result in the reconstruction of the internal nodes of any tree at any level, given the correct parameters and a parsimony algorithm. A side-product of one of these studies was the description of an empirical procedure developed for this analysis, called GeneTrace [27]. The reconstruction of ancestral gene content could reveal not only ancestral patterns of genome structure [3,5] but also, most importantly, genome function. For instance, the characterization of the last universal common ancestor, also known as LUCA [28], has been achieved by taking into account gene loss processes [26,29]. Previously, this had been accomplished only by pairwise genome comparisons [28] and was thus deemed as an underestimate, with the expansion of available genome sequences [29].

Armed with reasonable estimates for the major forces that shape genome content and robust trees, it is possible to further explore the structure of the microbial, and ultimately the entire, evolutionary space. It has long been suspected that the structure of the 'tree of life' resembles a net [30]. Various frameworks have been proposed to

capture structural aspects of the tree, including mathematical formalisms [31], statistical measures [32], probabilistic procedures [33], modeling and simulation (not considering HGT) [34], and software components that assess incongruence against 'standard' trees [35]. It remains to be seen whether any of the above methods are scalable and applicable for genome-wide studies and hundreds of species.

These attempts were always made amid doomsday scenarios that kept suggesting that a tree of life is impossible to derive, owing to HGT [20,21,36]. Two examples of such criticisms have appeared [20,36], one against an analysis of the proteobacterial tree [37] and the other against a novel method called 'conditional reconstruction' [38], respectively.

However, a recent study provides for the first time a visual in addition to a quantitative representation of the forces that shape gene content. This was achieved by using the GeneTrace algorithm for the explicit reconstruction of ancestral states (internal nodes), the overlaying of vertical and horizontal transfer events (or gene flows), and the independent adoption of the term 'vine' to capture horizontal gene flows along the 'trunk' of the tree of life [39]. This study currently represents one of the most sophisticated approaches to overlay all available information within a single framework: 165 species; approximately 200 000 protein families across two major databases; three genome tree construction methods, including genome conservation [8]; and network analysis to examine the topology of the resulting network (Figure 1c). The conclusions are unequivocal: over 600 000 vertical transfers; over 80 000 gene loss events; and less than 40 000 HGT cases. In this case, one could name this process as 'tree de-construction', because the analysis reveals the inner workings of the processes that shape genome content, structure and function, by inferring the gene complement of internal (i.e. ancestral) nodes of the 'net of life' [39]. Thus, the long suspected case of a network or ring of life [40,41] has been quantified for the first time. In the future, it will be crucial to overcome certain shortcomings of this approach (e.g. the demarcation of HGT directionality [39]).

Future perspectives: real or virtual ancestral genomes

The reconstruction of ancestral states for genomes is a key step towards our understanding of genome structure and function. This endeavor is expected to expand, by connecting both to more established and mature frameworks for the estimation of ancestral states for individual sequences [42,43] and to functional genomics data, such as the estimation of ancestral states for gene expression [44].

The most important challenge is the creation of family-specific evolutionary change frequencies (c.f. residue-

specific, in the case of individual proteins). In the same spirit that nucleotides or amino acids are counted within multiple alignments and their propensity is assessed with regard to different states (including other residues and gaps), gene families could be counted within species, and their propensity to duplicate, be lost or horizontally transferred can be quantified. Then, we will be able to achieve a continuum of methods between traditional molecular evolutionary studies based on precise models and the novel genomic evolutionary approaches based on gene content. This rich environment will provide an unprecedented level of detail for genome evolution, because it will enable the inference of ancestral states (inner nodes) of the tree on the basis of extant species (terminal nodes).

Conclusions

It is conceivable that, one day, entire genomes will be synthesized or engineered, reminiscent of the spirit applied to the 'resurrection' of entire proteins and their subsequent biochemical characterization [45^{••}], to test hypotheses of genome evolution or recreate extinct ancestral species. From a bioinformatics perspective, it will be interesting to see how these new types of extinct genetic evidence might be incorporated in novel databases, and to what extent they will become regularly used, in the same manner that extant sequences and genomes are.

Acknowledgements

I wish to express my gratitude to all the members of the Computational Genomics Group for criticisms, discussions, and a stimulating collegial atmosphere that is ending soon. Particular thanks to Victor Kunin, who pioneered this work. Also, many thanks to our collaborators, especially Ben Blencowe (University of Toronto), Anton J Enright (Sanger Institute, UK), Paul Janssen (Belgian Nuclear Research Center), Peter D Karp (SRI International, CA), David Kreil (University of Vienna), Isidore Rigoutsos (IBM Research, NY), Chris Sander (Memorial Sloan-Kettering Cancer Center, NY) and others, too numerous to list here.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- of special interest
- of outstanding interest

1. Janssen P, Audit B, Cases I, Darzentas N, Goldovsky L, Kunin V, Lopez-Bigas N, Peregrin-Alvarez JM, Pereira-Leal JB, Tsoka S *et al.*: **Beyond 100 genomes.** *Genome Biol* 2003, **4**:402.
2. Janssen P, Goldovsky L, Kunin V, Darzentas N, Ouzounis CA: **The challenge of annotating 200 genomes with 4 million publications.** *EMBO Rep* 2005, **6**:397-399.
3. Moret BM, Warnow T: **Advances in phylogeny reconstruction from gene order and content data.** *Methods Enzymol* 2005, **395**:673-700.
4. Kreil DP, Ouzounis CA: **Identification of thermophilic species by the amino acid compositions deduced from their genomes.** *Nucleic Acids Res* 2001, **29**:1608-1615.
5. Sankoff D, Blanchette M: **Phylogenetic invariants for genome rearrangements.** *J Comput Biol* 1999, **6**:431-445.
6. Enright AJ, Ouzounis CA: **Functional associations of proteins in entire genomes via exhaustive detection of gene fusion.** *Genome Biol* 2001, **2**:r0034.0031-r0034.0037.
7. Huynen MA, Bork P: **Measuring genome evolution.** *Proc Natl Acad Sci USA* 1998, **95**:5849-5856.
8. Kunin V, Ahren D, Goldovsky L, Janssen P, Ouzounis CA: **Measuring genome conservation across taxa: divided strains and united kingdoms.** *Nucleic Acids Res* 2005, **33**:616-621.
An empirical study comparing the performance of gene content, average sequence similarity and a new composite measure called 'genome conservation', in their ability to reconstruct species trees. The surprising contrast function provided by genome conservation (see Figure 2 of their study) results in very reliable trees and appears to be extremely resilient to noise arising from HGT and loss.
9. Snel B, Huynen MA, Dutilh BE: **Genome trees and the nature of genome evolution.** *Annu Rev Microbiol* 2005, **59**:191-209.
This is an insightful review by a group that pioneered the use of gene content trees in evolutionary studies. It contains a detailed comparison of the merits and shortcomings of different methods, in addition to suggestions for future improvements.
10. Nichols R: **Gene trees and species trees are not the same.** *Trends Ecol Evol* 2001, **16**:358-364.
11. Tekaiia F, Lazcano A, Dujon B: **The genomic tree as revealed from whole proteome comparisons.** *Genome Res* 1999, **9**:550-557.
12. Snel B, Bork P, Huynen MA: **Genome phylogeny based on gene content.** *Nat Genet* 1999, **21**:108-110.
13. Fitz-Gibbon ST, House CH: **Whole genome-based phylogenetic analysis of free-living microorganisms.** *Nucleic Acids Res* 1999, **27**:4218-4222.
14. Wolf YI, Rogozin IB, Grishin NV, Tatusov RL, Koonin EV: **Genome trees constructed using five different approaches suggest new major bacterial clades.** *BMC Evol Biol* 2001, **1**:8.
15. Yang S, Doolittle RF, Bourne PE: **Phylogeny determined by protein domain content.** *Proc Natl Acad Sci USA* 2005, **102**:373-378.
This is an impressive study of genome comparison, using the presence or absence of protein domains. In this challenging mode, it is reported that protein domain presence — free from gene duplication, domain shuffling, and spurious detections owing to the multi-domain or paralogy problems — provides a robust measure of genome similarity and reflects the classification of taxa into monophyletic groups.
16. Lopez-Bigas N, Ouzounis CA: **Genome-wide identification of genes likely to be involved in human genetic disease.** *Nucleic Acids Res* 2004, **32**:3108-3114.
17. Clarke GD, Beiko RG, Ragan MA, Charlebois RL: **Inferring genome trees by using a filter to eliminate phylogenetically discordant sequences and a distance matrix based on mean normalized BLASTP scores.** *J Bacteriol* 2002, **184**:2072-2080.
18. Henz SR, Huson DH, Auch AF, Nieselt-Struwe K, Schuster SC: **Whole-genome prokaryotic phylogeny.** *Bioinformatics* 2005, **21**:2329-2335.
19. Grishin NV, Wolf YI, Koonin EV: **From complete genomes to measures of substitution rate variability within and between proteins.** *Genome Res* 2000, **10**:991-1000.
20. Baptiste E, Boucher Y, Leigh J, Doolittle WF: **Phylogenetic reconstruction and lateral gene transfer.** *Trends Microbiol* 2004, **12**:406-411.
21. Baptiste E, Susko E, Leigh J, MacLeod D, Charlebois RL, Doolittle WF: **Do orthologous gene phylogenies really support treethinking?** *BMC Evol Biol* 2005, **5**:33.
22. Kurland CG: **What tangled web: barriers to rampant horizontal gene transfer.** *Bioessays* 2005, **27**:741-747.
23. Kurland CG, Canback B, Berg OG: **Horizontal gene transfer: a critical view.** *Proc Natl Acad Sci USA* 2003, **100**:9658-9662.
24. Snel B, Bork P, Huynen MA: **Genomes in flux: the evolution of archaeal and proteobacterial gene content.** *Genome Res* 2002, **12**:17-25.

25. Kunin V, Ouzounis CA: **The balance of driving forces during genome evolution in prokaryotes.** *Genome Res* 2003, **13**:1589-1594.
26. Mirkin BG, Fenner TI, Galperin MY, Koonin EV: **Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes.** *BMC Evol Biol* 2003, **3**:2.
27. Kunin V, Ouzounis CA: **GeneTRACE-reconstruction of gene content of ancestral species.** *Bioinformatics* 2003, **19**:1412-1416.
28. Kyrpides N, Overbeek R, Ouzounis CA: **Universal protein families and the functional content of the last universal common ancestor.** *J Mol Evol* 1999, **49**:413-423.
29. Ouzounis CA, Kunin V, Darzentas N, Goldovsky L: **A minimal estimate for the gene content of the last universal common ancestor – exobiology from a terrestrial perspective.** *Res Microbiol* 2005, in press.
30. Hilario E, Gogarten JP: **Horizontal transfer of ATPase genes – the tree of life becomes a net of life.** *Biosystems* 1993, **31**:111-119.
31. Makarenkov V, Legendre P: **From a phylogenetic tree to a reticulated network.** *J Comput Biol* 2004, **11**:195-212.
32. Kim J, Salisbury BA: **A tree obscured by vines: horizontal gene transfer and the median tree method of estimating species phylogeny.** *Pac Symp Biocomput* 2001:571-582.
33. Suchard MA: **Stochastic models for horizontal gene transfer: taking a random walk through tree space.** *Genetics* 2005, **170**:419-431.
34. Huson DH, Steel M: **Phylogenetic trees based on gene content.** *Bioinformatics* 2004, **20**:2044-2049.
35. Trooskens G, De Beule D, Decouttere F, Van Crielinge W: **Phylogenetic trees: visualizing, customizing and detecting incongruence.** *Bioinformatics* 2005. DOI:10.1093/bioinformatics/bti590.
36. Baptiste E, Walsh DA: **Does the 'Ring of Life' ring true?** *Trends Microbiol* 2005, **13**:256-261.
37. Lerat E, Daubin V, Moran NA: **From gene trees to organismal phylogeny in prokaryotes: the case of the γ -Proteobacteria.** *PLoS Biol* 2003, **1**:E19.
38. Lake JA, Rivera MC: **Deriving the genomic tree of life in the presence of horizontal gene transfer: conditioned reconstruction.** *Mol Biol Evol* 2004, **21**:681-690.
39. Kunin V, Goldovsky L, Darzentas N, Ouzounis CA: **The net of life: reconstructing the microbial phylogenetic network.** *Genome Res* 2005, **15**:954-959.
- This analysis overlays a genome tree for 165 microbial species with horizontal transfer events, thus creating a net of life in which both vertical and horizontal gene flow is depicted. This work unravels certain species groups with a high propensity to exchange genes across the tree and attempts a rough characterization of the network of life in terms of graph topology, with implications for genome evolution.
40. Philippe H, Douady CJ: **Horizontal gene transfer and phylogenetics.** *Curr Opin Microbiol* 2003, **6**:498-505.
41. Simonson AB, Servin JA, Skophammer RG, Herbold CW, Rivera MC, Lake JA: **Decoding the genomic tree of life.** *Proc Natl Acad Sci USA* 2005, **102**(Suppl 1):6608-6613.
42. Pagel M, Meade A, Barker D: **Bayesian estimation of ancestral character states on phylogenies.** *Syst Biol* 2004, **53**:673-684.
43. Cai W, Pei J, Grishin NV: **Reconstruction of ancestral protein sequences and its applications.** *BMC Evol Biol* 2004, **4**:33.
44. Rossmes R, Eidhammer I, Liberles DA: **Phylogenetic reconstruction of ancestral character states for gene expression and mRNA splicing data.** *BMC Bioinformatics* 2005, **6**:127.
45. Gaucher EA, Thomson JM, Burgan MF, Benner SA: **Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins.** *Nature* 2003, **425**:285-288.
- This seminal study demonstrates the synthesis of an ancestral protein and its characterization in the laboratory with the aim to infer the environmental conditions in which ancient bacteria thrived. Despite the fact that it is not directly related to the reconstruction of ancestral states for genomes, it certainly points the way.

Elsevier joins major health information initiative

Elsevier has joined with scientific publishers and leading voluntary health organizations to create patientINFORM, a groundbreaking initiative to help patients and caregivers close a crucial information gap. patientINFORM is a free online service dedicated to disseminating medical research and is scheduled to launch in 2005.

Elsevier will provide the voluntary health organizations with increased online access to our peer-reviewed biomedical journals immediately upon publication, together with content from back issues. The voluntary health organizations will integrate the information into materials for patients and link to the full text of selected research articles on their websites.

patientINFORM has been created to allow patients seeking the latest information about treatment options online access to the most up-to-date, reliable research available for specific diseases.

'Not only will patientINFORM connect patients and their caregivers with the latest research, it will help them to put it into context. By making it easier to understand research findings, patientINFORM will empower patients to have a more productive dialogue with their physicians and make well-informed decisions about care', said Harmon Eyre, M.D., national chief medical officer of the American Cancer Society.

For more information, visit www.patientinform.org



ELSEVIER

Causes and effects of nuclear genome reduction

Patrick J Keeling and Claudio H Slamovits

Eukaryotic nuclear genomes are generally considered to be large and gene-sparse, but extreme reduction has taken place several times, resulting in small genomes with a high gene-density. This process involves losing genes, compacting those that remain, or often both. Recently sequenced nuclear genomes include several that have converged to similar gene-densities by many means: variation in numbers and lengths of genes, intergenic regions and introns all contribute, but not equally in any given genome. Genomes of microsporidia and nucleomorphs have taken compaction much further, and in these hyper-compacted genomes there is evidence that some basic processes such as gene expression might be affected by genome form. In these genomes, normally weak forces might become more significant drivers of genome evolution.

Addresses

Canadian Institute for Advanced Research, Botany Department, University of British Columbia, 3529-6270 University Boulevard, Vancouver, BC, V6T 1Z4, Canada

Corresponding author: Keeling, Patrick J (pkeeling@interchange.ubc.ca)

Current Opinion in Genetics & Development 2005, **15**:601-608

This review comes from a themed issue on
Genomes and evolution
Edited by Stephen J O'Brien and Claire M Fraser

Available online 26th September 2005

0959-437X/\$ – see front matter
© 2005 Elsevier Ltd. All rights reserved.

DOI 10.1016/j.gde.2005.09.003

Introduction

At the bottom of the rabbit hole, Alice found a bottle labeled, “Drink Me”. When she did, Alice shrank to a perfectly functioning, ten-inch miniature of herself. In reality, shrinking can be more difficult than simply drinking a potion, because the component parts of many systems are not themselves shrinkable, and so the system fails to function properly. In the world of eukaryotic nuclear genomes this is probably true, despite the fact that they vary in size by factors of hundreds of thousands (Figure 1), much more than all of Alice’s many transformations combined.

Variations in genome size have been a persistent puzzle, mostly because genome size does not correspond to organismal complexity, often referred to as the ‘C-value paradox’. Many other characteristics have been tied to genome size, including metabolic rates, body size, effec-

tive population size, and cell size or nucleus size [1–3,4*], the latter being the characteristic that most uniformly correlates with genome size. As with all complex characteristics, however, genome size is probably controlled by a variety of factors of varying importance in different organisms. These correlations are probably important across a broad spectrum of eukaryotes, but do not explain everything. Moreover, some genomes seem to depart from any otherwise stable trends, and these are often at the extremes of genome size [4*].

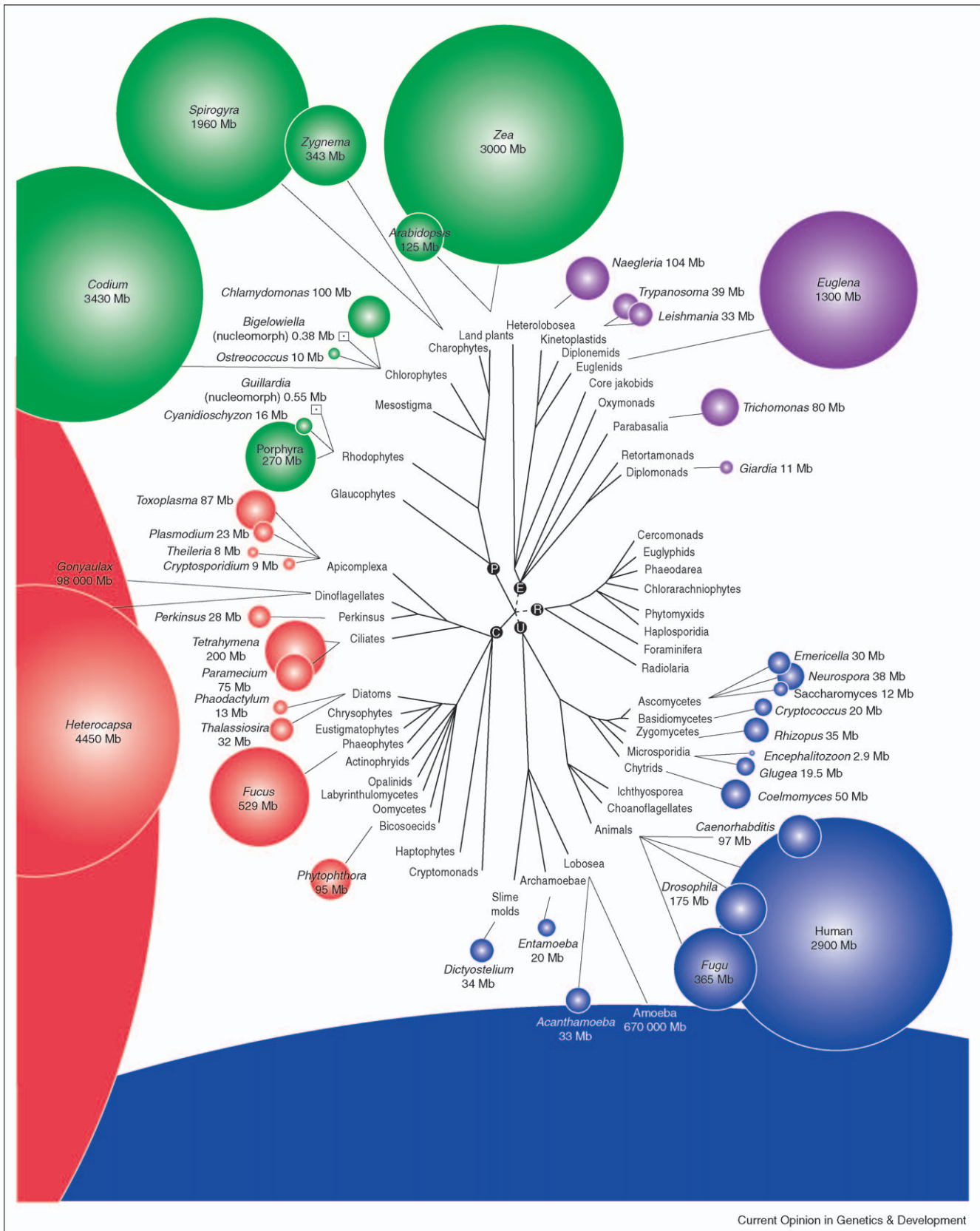
Here, we review the smallest eukaryotic genomes, illustrate some of the recent findings on how compaction impacts not only genome form but also function, and, along the way, point out some interesting characteristics of cells that have enabled compaction to take place. Some of the more familiar ‘compact’ genomes, such as that of yeast, are in reality only mildly compacted. Mildly compacted genomes (with gene densities of approximately 2 kb/gene) are not particularly uncommon and probably follow most of the same rules as larger genomes. A few genomes, however, are hyper-compacted (with densities closer to 1 kb/gene) and it is emerging that, at some point, such genomes might depart from some rules that govern other genomes — a point when compaction begins to affect basic processes such as replication, expression or recombination. At this point, processes and pressures that are common to all genomes but are generally relatively insignificant might become more important in genome evolution.

Different ways to shrink a genome: reduction versus compaction

Before examining a small genome, it is helpful to consider a simplistic division of ways that a genome can shrink: it can lose genes (elimination) or it can pack genes into a smaller space (compaction) [5]. Gene elimination results in a genome with a reduced coding capacity and a simplified proteome, whereas compaction by itself does not change the coding capacity but results in a higher gene density. It is relatively easy to explain gene elimination in many instances because the organisms are endosymbionts or intracellular parasites. These genomes have shed a large number of genes because they rely on a host cell for nutrients and biosynthesis of small molecules. Why compaction takes place is less obvious, and its effects on the genome are also more subtle, although not insignificant.

Returning to Alice, depending on what she ate (e.g. eating mushrooms versus eating cake), the effect on the way she shrank was different: sometimes she simply got smaller

Figure 1



and sometimes parts of her shrank more than others. Examining different groups of eukaryotes with compacted genomes we also see different combinations of elimination versus compaction.

We are fortunate now to have complete genome sequences for several eukaryotes with mildly compacted genomes, and also to have more than one closely related species in a couple of cases (Table 1). Focusing on protists, the best resource of complete genome sequences is in the Apicomplexa, a diverse phylum of intracellular parasites. In addition to *Plasmodium* [6,7], there are now complete genomes for two species of *Cryptosporidium* [8*,9*] and two species of *Theileria* [10*,11*], all of which have mildly compacted genomes with densities in the order of twice that seen in *Plasmodium*. Each has achieved this compaction by different means. All have eliminated genes to about the same extent, but *Theileria* has reduced its intergenic spaces more than *Cryptosporidium*, whereas *Cryptosporidium* has reduced its intron content considerably more than *Theileria* (Table 1). Comparing the genomes of congeneric relatives has already revealed a high degree of conservation in gene order in *Theileria*, and many features relating to the expansion of species-specific gene families in both genera [8*–11*]. In other groups, even just considering the simplistic characteristics of coding capacity, gene density and intron numbers, we see similar variations in modes of reduction: *Dictyostelium* has retained many genes and introns, but reduced its intergenic regions [12], whereas *Entamoeba* has reduced all three characteristics [13], and *Cyanidioschyzon* has reduced gene number and lost nearly all of its introns [14]. Interestingly, even though many of these genomes have reduced by eliminating genes, there is also evidence for acquiring new genes by horizontal gene transfer in several parasites with reduced genomes [13,15–20].

One of the more interesting compacted genomes is that of the ciliate *Paramecium* [21**]. This genome is relatively large and contains an estimated 30 000 genes. It has clearly eliminated a small amount, and yet it has compacted this large repertoire of genes significantly. It has retained many introns but shrunk them near to the lower limit known, and, most interestingly, has drastically reduced intergenic lengths (Table 1). This genome shows, perhaps better than does any other eukaryote to date, that elimination and compaction do not necessarily go together. We do not know why, but one possible factor is the separation of germ and somatic nuclei. In the development of somatic chromosomes from germ-line

chromosomes, a substantial amount of sequence is discarded: nearly all non-genic, including most or all of the transposable elements in the genome. The somatic genome, therefore, has a higher gene density: perhaps this separation of selective pressures shields the somatic genome from negative selection on some otherwise deleterious effects of compaction.

Extreme compaction: microsporidia and nucleomorphs

Even the most gene-dense of the genomes described above do not fall below 2 kb/gene overall, suggesting that beyond this density functional difficulties might arise. It is easy to imagine what kinds of problem might be faced: for instance, regulatory regions interfering with one another, or transcriptional fronts colliding [22]. However, in three different cases, nuclear gene densities have hyper-compacted: in microsporidia and the nucleomorphs of cryptomonads and chlorarachniophytes.

Microsporidia are a group of diverse obligate intracellular parasites now known to be closely related to fungi. This has been a contentious issue because they were formerly thought to be early-diverging eukaryotes, until conflicting phylogenetic data suggested that they were fungi [23]. The reason behind this incongruence has now been shown in a genome-wide phylogenetic analysis, which demonstrated a strong correlation between how highly divergent a gene was and its tendency to show microsporidia branching early in eukaryotes, a classic phylogenetic artefact [24*]. Microsporidia are highly adapted to their parasitic way of life: they have a highly advanced and sophisticated infection mechanism (Figure 2), but they are also degenerate in many ways. Known microsporidian genomes range in length from 19.5 Mb to a mere 2.3 Mb [25–28], and the complete 2.9 Mb sequence of the *Encephalitozoon cuniculi* genome has been determined and shown to be a model of both elimination and compaction [29,30]. The genome contains just fewer than 2000 protein-coding genes, a great proportion of which are involved in replication and gene expression. Many biosynthetic pathways have been lost — particularly those involving the biosynthesis of small molecules such as nucleotides or amino acids — which, not surprisingly, suggests a heavy reliance on the host cell for nutrients. In terms of compaction, the average intergenic size is only 129 bp; by comparison, this is approximately a quarter of the size of that found in the mildly compacted genome of *Saccharomyces*. Furthermore, there are few short repeats, only one repeated block of the genome, and no evidence

(Figure 1 Legend) The tree of eukaryotes, showing some variations in genome size. The tree is a composite based on a variety of available data according to the study by Keeling *et al.* [48]. There are five hypothetical supergroups, indicated by black circles at their bases: Excavates (E), Unikonts (U), Plants (P), Chromalveolates (C) and Rhizaria (R). Genome sizes are derived either from complete sequences or from estimates based on methods such as CHEF (contour-clamped homogeneous electric-field) gel electrophoresis or quantifying DNA content. Some estimates — in particular, those of larger genomes — are probably erroneous (for example, as a result of polyploidy) but it is nevertheless clear that the range of genome sizes in eukaryotes is vast. Genome sizes are taken from too many sources to list them all, but useful compilations can be found in the studies by Lynch and Conery [2] and Kapraun [3], and at www.cbs.dtu.dk/databases/DOGS.

Table 1

Some characteristics of model genomes compared with compact genomes*

Organism	Genome size (Mb)	Number of genes	Gene density (kb/gene)	Mean intergenic distance (bp)	% Intron-containing †	Mean intron size (bp)
<i>Homo sapiens</i>	2851	22 287	127.90	~10 ⁴	85	3365
<i>Arabidopsis thaliana</i>	125	25 498	4.90	2900	79	170
<i>Saccharomyces cerevisiae</i>	12.50	5770	2.09	500	5	287
Apicomplexa						
<i>Plasmodium falciparum</i>	12.03	5268	4.34	1694	54	179
<i>Theileria parva</i>	8.31	4035	2.20	405	74	94
<i>Theileria annulata</i>	8.35	3792	2.05	369	71	69
<i>Cryptosporidium parvum</i>	9.11	3952	2.30	566	5	‡
<i>Cryptosporidium hominis</i>	9.16	3994	2.29	716	5–20	‡
Red Algae						
<i>Cyanidioschyzon merolae</i>	16.52	5331	3.10	‡	0.5	‡
Amoebozoa						
<i>Dictyostelium discoideum</i>	33.82	12 500	2.72	‡	69	146
<i>Entamoeba histolytica</i>	23.75	9938	2.39	‡	25	‡
Ciliate						
<i>Paramecium tetraurelia</i> ∞	~100	~30 000	2.14	202	84	25
Microsporidia						
<i>Encephalitozoon cuniculi</i>	2.51	1997	1.25 [§]	129	0.6	‡
Nucleomorphs						
<i>Guillardia theta</i>	0.551	486	1.13 [§]	70	1.8	46
<i>Bigelowiella natans</i>	0.373	~308	1.21 [§]	113	80	19

* Values for several of these characteristics vary among published reports, in part because they change as annotations improve, and in part because they are measured in different ways. Accordingly, the values should be taken as an approximation that demonstrates the general trends in the genome. In general, the values reported here are taken from the most current source or from the report describing the genome

‡ Value not available.

§ Overall gene density is based on genome size divided by protein-coding gene number. However in the smallest genomes, essential non-coding regions (e.g. telomeres) and small RNA genes (e.g. tRNAs) have a disproportionate affect on density values compared with their affect on large genomes. The effective gene density in the chromosomal cores can be much higher (e.g. 20% higher in *E. cuniculi*), so this value is useful for comparison, but a more interesting value with respect to the effects of compaction is the mean intergenic length.

† The number of genes containing introns is most commonly reported, but in terms of compaction this value is only partly informative since the number of introns in each of these genes is also relevant, or the total number of genes across the genome.

∞ Based on partial genome.

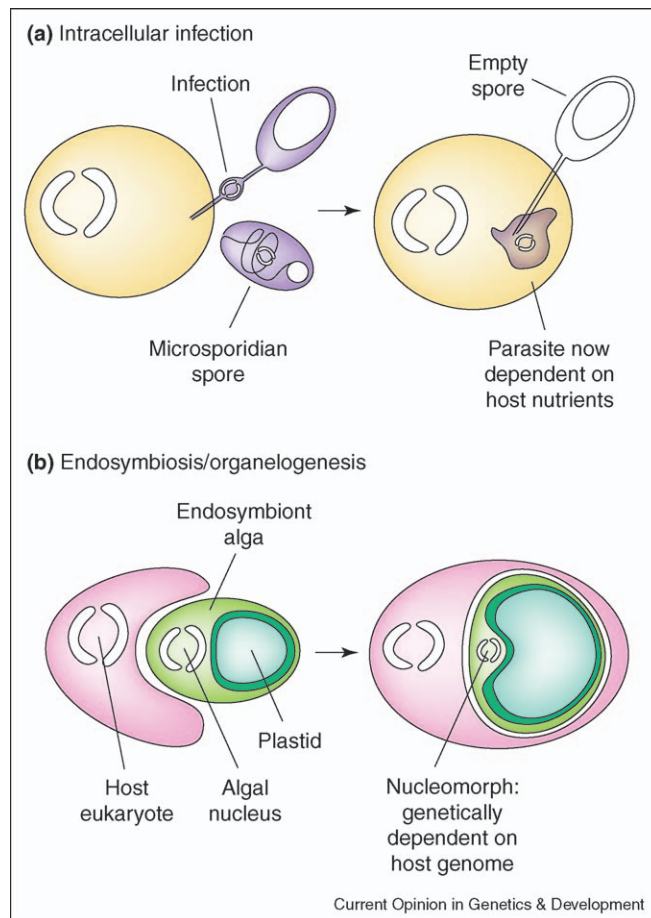
of transposons. There are only 13 introns, located within 12 genes, known in the entire genome. Overall, these characteristics result in a gene density of 1.25 bp/gene [30].

Nucleomorphs are relict nuclei of endosymbiotic algae found in cryptomonads and chlorarachniophytes (Figure 2) [31,32]; other groups have endosymbiotic algae that retain nuclei [33], but virtually nothing is known about their genome form or content. The apparent function of nucleomorph genomes is to supply a few proteins to the plastid with which they are associated. In most secondary plastids, these protein genes have all moved to the host nucleus, rendering the endosymbiont genome obsolete [34,35]. In cryptomonads and chlorarachniophytes, however, the endosymbionts have retained a small number of essential genes — encoding plastid proteins and, perhaps, also proteins involving in targeting — that cannot be lost until these genes also move to the nucleus. Cryptomonad and chlorarachniophyte nucleomorphs evolved independently from red and green algae,

respectively, but they share many features of overall structure in common [31]. Their genomes are all composed of three linear chromosomes with gene-dense cores and rRNA operons as subtelomeric repeats. They are by far the smallest nuclear genomes known: the genomes of cryptomonads range from 450 to 710 kb [36], and those of chlorarachniophytes range from 373 to only 455 kb [31,37].

Complete genome sequences are known from a representative of both groups: the cryptomonad *Guillardia theta* [38] and the chlorarachniophyte *Bigelowiella natans* ([39] and PR Gilson, V Su, CH Slamovits, ME Reith, PJ Keeling and GI McFadden, unpublished). These genomes are similar to microsporidia in terms of compaction, but have eliminated far more genes because of their even greater dependence on their host. The potential for reduction in nucleomorphs is higher than that in microsporidia for an intriguing reason: not only are they dependent upon their hosts for energy and small molecules but, because secondary algae have transferred many genes to

Figure 2



Intracellular life and small genomes. **(a)** Spores are the only stage of microsporidian obligate intracellular parasites that survive outside another cell. To the left, two spores (purple) are shown: one is in its dormant state (below); and one (above) is infecting a host (yellow). Infection takes place using a projectile tube that injects the parasite directly into the host cytoplasm, as shown to the right. Because all growth and development occurs inside this host cell, the parasite become deeply dependent upon the host for energy and nutrients, enabling them to lose a great number of genes, and their genomes to shrink (although this does not explain why they compact). **(b)** Nucleomorphs arose when an alga (green) was eaten by another eukaryote (pink), as shown on the left. Usually, this meal would be digested, but on several occasions the alga was retained and the two cells integrated in a process called secondary endosymbiosis, forming a new algal lineage. Sometimes secondary endosymbionts lose their nucleus altogether, but in two instances the nucleus of the endosymbiont was not lost and, instead, degenerated substantially to what we now call a nucleomorph (shown on right). The endosymbiont is dependent on the host for energy and nutrients, and so the degeneration of nucleomorph genomes is similar in many ways to that seen in microsporidian genomes. In addition, however, many nucleomorph genes appear to have moved to the host nucleus, and their products are targeted back to the endosymbiont, enabling nucleomorphs to degenerate even further than a parasite might. Overall, endosymbionts and intracellular parasites and the relationship of each to their host differ in many ways, but there are also many similarities and parallels.

the host and developed a mechanism to target the protein products back to the plastid [34], they have also become dependent on their host for most of their proteins. Accordingly, nucleomorphs might encode fewer proteins than are needed for the most basic functions. Indeed, the complete *G. theta* nucleomorph genome is missing genes for several essential proteins, such as DNA polymerases [38]. It has long been speculated that the genes for such proteins have moved to the host nucleus, but, until recently, no such gene was found. Now, however, *G. theta* nuclear genes for several putatively endosymbiont-tar-

geted proteins have been found [40••]. These genes are one of the 'holy grails' of nucleomorph research because they hold many keys to important questions about protein-trafficking. Indeed, expressing these genes as green fluorescent protein (GFP)-fusions in a genetically tractable diatom has already proposed an intriguing mechanism for traversing the long-mysterious third membrane of complex plastids: it is suggested that a plastid outer-membrane complex is duplicated and present in both membranes, although the two complexes are subtly different [40••].

Genome-wide effects of compaction

Aside from altering the form of the genome, one of the first features of a functional nature to be noted in the hyper-compacted genomes of both microsporidia and nucleomorphs was that the proteins encoded in them tend to be smaller than their homologues in related genomes [30,38]. In *E. cuniculi* it was hypothesized that these smaller proteins have not arisen directly as a result of compaction, however, because a shrinking proteome could result in simpler interaction networks, which would, in turn, facilitate the loss of protein domains responsible for these interactions [30].

Perhaps a stronger link to compaction is seen in conservation of the overall gene order of the genome in microsporidia. Comparisons between the distantly related microsporidia *E. cuniculi* and *Antonospora locustae* revealed a relatively high number of conserved gene-pairs and showed that the intergenic regions between these conserved pairs were markedly shorter in both genomes [41^{••}]. It was hypothesized that the short intergenic regions slowed genomic rearrangements simply by reducing the number of possible breakpoints [41^{••}]. In the mildly compacted genome of yeasts, conservation of synteny has been linked to short intergenic regions, but, overall, other factors such as co-expression appear to be more important [42]. The extreme conditions in microsporidian genomes might, therefore, place new prominence on otherwise less significant forces. Currently, there is no comparative genomic data from nucleomorphs, but they offer a very interesting system to study synteny over time, because the host nuclear genome has been evolving in parallel with the nucleomorph. If one could demonstrate that two host genomes shared significantly less synteny than the nucleomorph genomes from the same species, it would prove that the nucleomorph genome was evolving more slowly than that of the host, because both host and nucleomorph genomes would have diverged at exactly the same time. On a more practical level, the conserved synteny in microsporidia is also an ideal guide for gene discovery, particularly as the divergent nature of their genes makes identifying them a challenge. Recently, this conservation was used to identify, in the absence of sequence similarity, *A. locustae* homologues of two *E. cuniculi* proteins that are crucial for infection [43[•]]. Functional studies confirm this identification and highlight the practical value of understanding the dynamics of a genome.

Although these characteristics are each intriguing aspects of reduction, they do not obviously explain the rarity of hyper-reduction in nuclear genomes. One potential explanation has emerged not from the genomes themselves but from examining gene expression within them. Expressed sequence tag (EST) projects from microsporidia and both nucleomorph genomes have revealed a high frequency of overlapping transcription in all three sys-

tems [44^{••}]. In most eukaryotes, transcripts from one gene do not typically overlap with those of adjacent genes, and if they are engineered to do so they often have disastrous effects on expression of one or both genes [22,45,46]. In microsporidia, however, about 15% of transcripts appear to initiate within the upstream gene, terminate within or beyond the downstream gene, or both [44^{••},47]. In nucleomorphs, little initiation within upstream genes was observed, but most transcripts read through into downstream genes (97% in *G. theta* and 82% in *B. natans*), sometimes encoding all or parts of as many as four genes [44^{••}]. It is possible that one of the forces containing mildly compacted genomes to about 2 kb/gene is the potential for adjacent genes to interfere with one another's expression, and that hyper-compact genomes have overcome this constraint, but this hypothesis needs a great deal of work for confirmation.

Conclusions and more questions

Small genomes tend to be the first to get sequenced, so even in these early days of comparative nuclear genomics there are already several reduced genomes for further analysis and comparison, but this is only the beginning. New genomes of potential interest to reduction and compaction are on the horizon, including those of the reduced green alga *Ostreococcus*, the ciliate *Tetrahymena*, and additional Apicomplexa and microsporidia, some with much larger genomes than those presently analysed. The number of genomes with densities around 2 kb/gene is interesting, but whether it represents some functional line that is difficult to cross needs not only more genomic data to test the robustness of the observation but also experimental data that might explain what are the constraints on compaction. Moreover, there appear to be correlations between compaction, synteny and transcriptional properties in hyper-compacted genomes, but, again, the full strength of comparative analyses have not yet been brought to bear on these questions.

Acknowledgements

We thank GI McFadden, PR Gilson and V Su for allowing us to discuss unpublished data from the collaborative *B. natans* nucleomorph genome project, F Delbac and U-G Maier for giving us access to in-press manuscripts, and B Williams for critical comments. Work on microsporidian genomics is supported by the Canadian Institutes for Health Research. PJK is a Fellow of the Canadian Institute for Advanced Research and a New Investigator of the Michael Smith Foundation for Health Research and the Canadian Institutes for Health Research.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- of special interest
 - of outstanding interest
1. Vinogradov AE: **Evolution of genome size: multilevel selection, mutation bias or dynamical chaos?** *Curr Opin Genet Dev* 2004, **14**:620-626.
 2. Lynch M, Conery JS: **The origins of genome complexity.** *Science* 2003, **302**:1401-1404.

3. Kapraun DF: **Nuclear DNA content estimates in multicellular green, red and brown algae: phylogenetic considerations.** *Ann Bot (Lond)* 2005, **95**:7-44.
4. Cavalier-Smith T: **Economy, speed and size matter: evolutionary forces driving nuclear genome miniaturization and expansion.** *Ann Bot (Lond)* 2005, **95**:147-175.
A current review that covers a range of questions surrounding eukaryotic genome size by synthesizing data from the molecular, cell and population levels.
5. Keeling PJ: **Reduction and compaction in the genome of the apicomplexan parasite, *Cryptosporidium parvum*.** *Dev Cell* 2004, **6**:614-616.
6. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S *et al.*: **Genome sequence of the human malaria parasite *Plasmodium falciparum*.** *Nature* 2002, **419**:498-511.
7. Carlton JM, Angiuoli SV, Suh BB, Kooij TW, Pertea M, Silva JC, Ermolaeva MD, Allen JE, Selengut JD, Koo HL *et al.*: **Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*.** *Nature* 2002, **419**:512-519.
8. Abrahamsen MS, Templeton TJ, Enomoto S, Abrahante JE, Zhu G, Lancto CA, Deng M, Liu C, Widmer G, Tzipori S *et al.*: **Complete genome sequence of the apicomplexan, *Cryptosporidium parvum*.** *Science* 2004, **304**:441-445.
See annotation [9*].
9. Xu P, Widmer G, Wang Y, Ozaki LS, Alves JM, Serrano MG, Puiu D, Manque P, Akiyoshi D, Mackey AJ *et al.*: **The genome of *Cryptosporidium hominis*.** *Nature* 2004, **431**:1107-1112.
This study and that by Abrahamsen *et al.* [8*] report the complete genomes of *C. parvum* and *C. hominis*, respectively, together providing a comparison between the modes of reduction in two closely related, mildly compacted genomes. In these *Cryptosporidium* species, unlike in *Theileria* [11*], there is almost no variation in gene content.
10. Pain A, Renaud H, Berriman M, Murphy L, Yeats CA, Weir W, Kerhornou A, Aslett M, Bishop R, Bouchier C *et al.*: **Genome of the host-cell transforming parasite *Theileria annulata* compared with *T. parva*.** *Science* 2005, **309**:131-133.
See annotation [11*].
11. Gardner MJ, Bishop R, Shah T, de Villiers EP, Carlton JM, Hall N, Ren Q, Paulsen IT, Pain A, Berriman M *et al.*: **Genome sequence of *Theileria parva*, a bovine pathogen that transforms lymphocytes.** *Science* 2005, **309**:134-137.
As above, this study and that by Gardner *et al.* [10*] describe two complete sequences of closely related species with mildly compacted genomes. These reports once again show different modes of reduction in addition to interesting species-specific amplifications of gene families, and a high degree of gene-order conservation.
12. Eichinger L, Pachebat JA, Glockner G, Rajandream MA, Suggang R, Berriman M, Song J, Olsen R, Szafranski K, Xu Q *et al.*: **The genome of the social amoeba *Dictyostelium discoideum*.** *Nature* 2005, **435**:43-57.
13. Loftus B, Anderson I, Davies R, Alsmark UC, Samuelson J, Amedeo P, Roncaglia P, Berriman M, Hirt RP, Mann BJ *et al.*: **The genome of the protist parasite *Entamoeba histolytica*.** *Nature* 2005, **433**:865-868.
14. Matsuzaki M, Misumi O, Shin IT, Maruyama S, Takahara M, Miyagishima SY, Mori T, Nishida K, Yagisawa F, Yoshida Y *et al.*: **Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10D.** *Nature* 2004, **428**:653-657.
15. Fast NM, Law JS, Williams BA, Keeling PJ: **Bacterial catalase in the microsporidian *Nosema locustae*: implications for microsporidian metabolism and genome evolution.** *Eukaryot Cell* 2003, **2**:1069-1075.
16. Huang J, Mullapudi N, Sicheritz-Ponten T, Kissinger JC: **A first glimpse into the pattern and scale of gene transfer in Apicomplexa.** *Int J Parasitol* 2004, **34**:265-274.
17. Huang J, Mullapudi N, Lancto CA, Scott M, Abrahamsen MS, Kissinger JC: **Phylogenomic evidence supports past endosymbiosis, intracellular and horizontal gene transfer in *Cryptosporidium parvum*.** *Genome Biol* 2004, **5**:R88.
18. Richards TA, Hirt RP, Williams BA, Embley TM: **Horizontal gene transfer and the evolution of parasitic protozoa.** *Protist* 2003, **154**:17-32.
19. Striepen B, Pruijssers AJ, Huang J, Li C, Gubbels MJ, Umejiego NN, Hedstrom L, Kissinger JC: **Gene transfer in the evolution of parasite nucleotide biosynthesis.** *Proc Natl Acad Sci USA* 2004, **101**:3154-3159.
20. Striepen B, White MW, Li C, Guerini MN, Malik SB, Logsdon JM Jr, Liu C, Abrahamsen MS: **Genetic complementation in apicomplexan parasites.** *Proc Natl Acad Sci USA* 2002, **99**:6304-6309.
21. Zagulski M, Nowak JK, Le Mouel A, Nowacki M, Migdalski A, Gromadka R, Noel B, Blanc I, Dessen P, Wincker P *et al.*: **High coding density on the largest *Paramecium tetraurelia* somatic chromosome.** *Curr Biol* 2004, **14**:1397-1404.
Paramecium has a large genome with many genes but, despite this, the somatic chromosome described here is highly compacted with very short intergenic regions. This shows that compaction can take place without reducing the proteomic complexity of the organism. In the case of ciliates, this might be related to the generation of a somatic nucleus from a germ line nucleus, which includes the elimination of a great deal of non-coding DNA from the genome.
22. Shearwin KE, Callen BP, Egan JB: **Transcriptional interference – a crash course.** *Trends Genet* 2005, **21**:339-345.
23. Keeling PJ, Fast NM: **Microsporidia: biology and evolution of highly reduced intracellular parasites.** *Annu Rev Microbiol* 2002, **56**:93-116.
24. Thomarat F, Vivarès CP, Gouy M: **Phylogenetic analysis of the complete genome sequence of *Encephalitozoon cuniculi* supports the fungal origin of microsporidia and reveals a high frequency of fast-evolving genes.** *J Mol Evol* 2004, **59**:780-791.
The debate over the phylogenetic position of microsporidia has endured for many years because different genes often yield different trees. Here, a large proportion of the protein-coding genes from the *E. cuniculi* genome were analysed and a strong correlation was found between the level of divergence and the phylogeny, which should put this debate to rest. Divergent genes tend to show microsporidia branching early whereas more-conserved genes show them allied with fungi.
25. Streett DA: **Analysis of *Nosema locustae* (Microsporidia: Nosematidae) chromosomal DNA with pulsed-field gel electrophoresis.** *J Invert Pathol* 1994, **63**:301-303.
26. Biderre C, Pagès M, Méténier G, David D, Bata J, Prensier G, Vivarès CP: **On small genomes in eukaryotic organisms: molecular karyotypes of two microsporidian species (Protozoa) parasites of vertebrates.** *C R Acad Sci III* 1994, **317**:399-404.
27. Biderre C, Pagès M, Méténier G, Canning EU, Vivarès CP: **Evidence for the smallest nuclear genome (2.9 Mb) in the microsporidian *Encephalitozoon cuniculi*.** *Mol Biochem Parasitol* 1995, **74**:229-231.
28. Peyretailade E, Biderre C, Peyret P, Duffieux F, Méténier G, Gouy M, Michot B, Vivarès CP: **Microsporidian *Encephalitozoon cuniculi*, a unicellular eukaryote with an unusual chromosomal dispersion of ribosomal genes and a LSU rRNA reduced to the universal core.** *Nucleic Acids Res* 1998, **26**:3513-3520.
29. Peyret P, Katinka MD, Duprat S, Duffieux F, Barbe V, Barbazanges M, Weissenbach J, Saurin W, Vivarès CP: **Sequence and analysis of chromosome I of the amitochondriate intracellular parasite *Encephalitozoon cuniculi* (Microspora).** *Genome Res* 2001, **11**:198-207.
30. Katinka MD, Duprat S, Cornillot E, Méténier G, Thomarat F, Prensier G, Barbe V, Peyretailade E, Brottier P, Wincker P *et al.*: **Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*.** *Nature* 2001, **414**:450-453.
31. Gilson PR, McFadden GI: **Jam packed genomes—a preliminary, comparative analysis of nucleomorphs.** *Genetica* 2002, **115**:13-28.
32. Cavalier-Smith T: **Nucleomorphs: enslaved algal nuclei.** *Curr Opin Microbiol* 2002, **5**:612-619.
33. Dodge JD: **Observations on the fine structure of the eyespot and associated organelles in the dinoflagellate *Glenodinium foliaceum*.** *J Cell Sci* 1969, **5**:479-493.

34. McFadden GI: **Primary and secondary endosymbiosis and the origin of plastids.** *J Phycol* 2001, **37**:951-959.
35. Keeling PJ: **The diversity and evolutionary history of plastids and their hosts.** *Am J Bot* 2004, **91**:1481-1493.
36. Rensing SA, Goddemeier M, Hofmann CJ, Maier UG: **The presence of a nucleomorph *hsp70* gene is a common feature of Cryptophyta and Chlorarachniophyta.** *Curr Genet* 1994, **26**:451-455.
37. Gilson PR, McFadden GI: **Molecular, morphological and phylogenetic characterization of six chlorarachniophyte strains.** *Phycol Res* 1999, **47**:7-19.
38. Douglas S, Zauner S, Fraunholz M, Beaton M, Penny S, Deng LT, Wu X, Reith M, Cavalier-Smith T, Maier UG: **The highly reduced genome of an enslaved algal nucleus.** *Nature* 2001, **410**:1091-1096.
39. Gilson PR, McFadden GI: **The miniaturized nuclear genome of a eukaryotic endosymbiont contains genes that overlap, genes that are cotranscribed, and the smallest known spliceosomal introns.** *Proc Natl Acad Sci USA* 1996, **93**:7737-7742.
40. Gould SB, Sommer MS, Hadfi K, Zauner S, Kroth PG, Maier U-G: **Protein targeting into the complex plastid of cryptophytes.** *J Mol Evol*, in press.
- It has long been known that cryptomonad and chlorarachniophyte algae must have nuclear-encoded genes for proteins that are targeted to the cytosol of the eukaryotic endosymbiont. These proteins are crucially important to our understanding of targeting to secondary plastids: because they only travel part way to the plastid, they can be seen as a 'natural dissection' of the targeting pathway. This report identifies and characterises the first such genes and analyses their localization in a diatom genetic system.
41. Slamovits CH, Fast NM, Law JS, Keeling PJ: **Genome compaction and stability in microsporidian intracellular parasites.** *Curr Biol* 2004, **14**:891-896.
- Comparing the frequency of conserved gene pairs in two distantly related microsporidia, the authors revealed a high level of conservation, and this conservation was more likely in cases in which intergenic regions were short. The study concluded that the compaction of the genome probably slowed down the rate of naturally occurring reorganizations by limiting the number of non-deleterious breakpoints.
42. Hurst LD, Williams EJ, Pal C: **Natural selection promotes the conservation of linkage of co-expressed genes.** *Trends Genet* 2002, **18**:604-606.
43. Polonais V, Prensier G, Méténier G, Vivarès CP, Delbac F: **Microsporidian polar tube proteins: highly divergent but closely linked genes encode PTP1 and PTP2 in members of the evolutionarily distant *Antonospora* and *Encephalitozoon* groups.** *Fungal Genet Biol* 2005, **43**:491-803.
- Functionally annotating genes that are poorly conserved at the primary sequence level is a huge challenge, even in small genomes. Here, the authors demonstrated that synteny can be used to identify homologues of important genes when no detectable sequence conservation exists. Functional studies were also used to confirm the identity of the genes, suggesting that synteny will be an important tool for translating functional annotation from one microsporidian to another.
44. Williams BAP, Slamovits CH, Patron NJ, Fast NM, Keeling PJ: **A high frequency of overlapping gene expression in compacted eukaryotic genomes.** *Proc Natl Acad Sci USA* 2005, **102**:10936-10941.
- Expressed sequence tag (EST) data from a microsporidian and two nucleomorphs showed that cDNAs in all three systems frequently include sequence of more than one gene. Such overlapping transcription can be a problem in other nuclear genomes, and this transcriptional interference is one possible barrier to further reduction in many mildly compacted genomes. In these hyper-compacted genomes, however, overlapping transcription might be tolerated at very high levels: nearly all *G. theta* nucleomorph transcripts extend into downstream genes.
45. Prescott EM, Proudfoot NJ, Furger A, Dye MJ, Greger IH: **Transcriptional collision between convergent genes in budding yeast. Integrating mRNA processing with transcription Poly(A) signals control both transcriptional termination and initiation between the tandem *GAL10* and *GAL7* genes of *Saccharomyces cerevisiae*.** *Proc Natl Acad Sci USA* 2002, **99**:8796-8801.
46. Springer C, Valerius O, Strittmatter A, Braus GH: **The adjacent yeast genes *ARO4* and *HIS7* carry no intergenic region.** *J Biol Chem* 1997, **272**:26318-26324.
47. Slamovits CH, Keeling PJ: **Class II photolyase in a microsporidian intracellular parasite.** *J Mol Biol* 2004, **341**:713-721.
48. Keeling PJ, Burger G, Durnford DG, Lang BF, Lee RW, Pearlman RE, Roger AJ, Gray MW: **The tree of eukaryotes.** *Trends Ecol Evol* 2005, in press.

Elsevier.com – Dynamic New Site Links Scientists to New Research & Thinking

Elsevier.com has had a makeover, inside and out. Designed for scientists' information needs, the new site, launched in January, is powered by the latest technology with customer-focused navigation and an intuitive architecture for an improved user experience and greater productivity.

Elsevier.com's easy-to-use navigational tools and structure connect scientists with vital information – all from one entry point. Users can perform rapid and precise searches with our advanced search functionality, using the FAST technology of Scirus.com, the free science search engine. For example, users can define their searches by any number of criteria to pinpoint information and resources. Search by a specific author or editor, book publication date, subject area – life sciences, health sciences, physical sciences and social sciences – or by product type. Elsevier's portfolio includes more than 1800 Elsevier journals, 2200 new books per year, and a range of innovative electronic products. In addition, tailored content for authors, editors and librarians provides up-to-the-minute news, updates on functionality and new products, e-alerts and services, as well as relevant events.

Elsevier is proud to be a partner with the scientific and medical community. Find out more about who we are in the About section: our mission and values and how we support the STM community worldwide through partnerships with libraries and other publishers, and grant awards from The Elsevier Foundation.

As a world-leading publisher of scientific, technical and health information, Elsevier is dedicated to linking researchers and professionals to the best thinking in their fields. We offer the widest and deepest coverage in a range of media types to enhance cross-pollination of information, breakthroughs in research and discovery, and the sharing and preservation of knowledge. Visit us at Elsevier.com.

Elsevier. Building Insights. Breaking Boundaries.



ELSEVIER

Comparative genomics of malaria parasites

Neil Hall and Jane Carlton

In the past few years, the area of comparative genomics of malaria parasites has begun to come of age, with the completion of genome sequencing projects of four *Plasmodium* species, and several functional genomics studies. A picture is emerging of a parasite genome that is highly adapted to its mammalian and vector hosts, and which uses post-transcriptional gene-silencing as one method for the control of gene expression. The genome is compartmentalized into a core of conserved housekeeping genes, sandwiched between subtelomerically located genes encoding surface antigens. Species-specific gene families shape the preference of the parasite for host cells, in addition to determining interactions with the host immune-system. Recent research has led to the description of a motif that is conserved across *Plasmodium* species and which plays a central role in protein export into the host cell.

Addresses

The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20851, USA

Corresponding authors: Hall, Neil (nhall@tigr.org); Carlton, Jane (carlton@tigr.org)

Current Opinion in Genetics & Development 2005, 15:609–613

This review comes from a themed issue on
Genomes and evolution
Edited by Stephen J O'Brien and Claire M Fraser

Available online 22nd September 2005

0959-437X/\$ – see front matter

© 2005 Elsevier Ltd. All rights reserved.

DOI 10.1016/j.gde.2005.09.001

Introduction

Malaria infects 300–600 million people and causes more than one million deaths annually in tropical and subtropical parts of the world, making it one of the most important diseases affecting mankind [1]. The disease is caused by species of the genus *Plasmodium*, intracellular protozoan parasites that are transmitted from host to host by mosquito vectors. There are four species of *Plasmodium* that infect humans, including *Plasmodium falciparum*, which is the most virulent species, and *Plasmodium vivax*, which is the most prevalent. As *Plasmodium* parasites are host-restricted, making studies of the human-infective species difficult, there has been considerable interest in model malaria parasites that can be adapted to growth in laboratory rodents: these model parasites include *Plasmodium berghei*, *Plasmodium chabaudi* and *Plasmodium yoelii*, and species, such as

Plasmodium knowlesi and *Plasmodium cynomolgi*, that infect non-human primates. In addition, the avian malaria parasite *Plasmodium gallinaceum* is used as a model for the study of the mosquito stages.

Genome sequencing projects of four *Plasmodium* species have now been published: the complete genome sequence of *P. falciparum* and the rodent malaria parasite *P. y. yoelii* in 2002 [2,3], and two further rodent malaria species in 2005 [4•]. Several other *Plasmodium* species are currently being sequenced (see [5•] for review). Examination of data from all of these organisms shows that the *Plasmodium* genomes are haploid, have a standard size of approximately 22–26 Mb and are distributed among 14 linear chromosomes with a size range of 0.5–3.0 Mb. Genome composition varies among species, from the extremely (A + T)-rich genome of *P. falciparum* and the rodent malaria species (~80%) to the more (G + C)-rich *P. vivax* and *P. cynomolgi* genomes (~70%), which, in addition, have an isochore structure with regions of high (G + C) content interspersed between regions of high (A + T) content [6]. Each *Plasmodium* species appears to have 5000–6000 predicted genes per genome [2,3], ~60% of which are orthologous among the species. Many of the genes unique to each species are located within subtelomeric regions, and many code for immunodominant antigens. The difference in gene number between species is caused by (i) differential gene expansion in distinct lineages; and (ii) in some species, the presence of a large, variant gene family, the *Plasmodium* interspersed repeats (PIRs) family, members of which are predicted to be involved in antigenic variation. Finally, studies involving mapping of conserved genes to separations of *Plasmodium* chromosomes [7–9], and the generation of a whole genome alignment map among the rodent malaria species and *P. falciparum* [3] have shown that gene location and order, and even exon–intron boundaries and the fine-scale organization of genes, are preserved over large regions across *Plasmodium* species. The degree of conservation of synteny is greatest when comparing genomes of more closely related species.

In this article, we review the most recent studies of comparative genomics emerging in the wake of the completion of four *Plasmodium* genome sequencing projects. These studies are in their infancy but promise to provide a wealth of data greater than that provided by analysis of the individual genomes alone.

Comparative analysis of *Plasmodium* antigens, and antigenic variation

It is now well established that the major differences in gene content among *Plasmodium* species occur between

genes involved in interaction with the host immune system. Many of these genes are located in the subtelomeric regions of the parasite genome, and sequencing of these regions in several species has identified such antigen families, some of which are conserved between species (for example, the *P. vivax* [10] and rodent malaria species [3,11••] PIR families, which are clearly related and might have shared a common ancestor with the *rifin* genes of *P. falciparum* [11••]). *P. falciparum* contains other gene families that encode proteins involved in antigenic variation and evasion of immune responses (the *var*, *rifin*, *stevor* gene families, with 60, 140 and 25 copies each, respectively; reviewed in [12]). In *P. knowlesi*, the *SICA-var* (*Schizont-infected cell agglutination variant antigen*) gene family has also been described [13]. This gene encodes a protein that is expressed on the surface of infected erythrocytes and is implicated in antigenic variation in this species. There is little homology between the *SICA-var* and the *var* genes, despite their common function. Recently, a comparative study has identified the *P. falciparum* *SURFIN* gene family [14•], which forms a clade with the gene encoding the *P. vivax* transmembrane protein PvSTP1 (*P. vivax* subtelomeric protein 1) [10]; together, these proteins contain features of several exported and surface-expressed proteins from human, rodent and monkey malaria species. These studies suggest that species-specific evolution of antigen genes, most probably in response to pressure from differing host immune systems, has led to the current diverse repertoire of malaria antigens found in different species.

More recent whole genome analysis studies have demonstrated that proteins which are exported to the red blood cell surface contain a motif (termed the Pexel motif) that appears to be conserved across several *Plasmodium* species [15•,16]. A search of all the *P. falciparum* proteins identified many containing the Pexel motif, indicating that many as yet uncharacterized subtelomeric proteins are exported to the infected red blood cell surface. Although this motif is present in *var*, *rifin* and *stevor* genes and in members of several hypervariable gene families of *P. yoelii*, *P. vivax* and *P. gallinaceum*, it is not present in any members of the *pir* gene family [15•], which might suggest that proteins of this family are not exported to the red blood cell surface and might, indeed, have a different role than originally inferred.

Comparative gene expression and regulation

Until recently, very little was known regarding the regulation of gene expression in *Plasmodium*. Genes are monocistronically transcribed, implying the presence of regulatory sequence elements flanking coding regions: and the few promoters that have been identified appear to conform to the standard eukaryotic promoter structure [i.e. a basal promoter regulated by upstream enhancer elements (reviewed in [17]). Few of the identified DNA

elements direct the transcription of *Plasmodium* genes, although one promising candidate, the G-box, was recently identified upstream of several *P. falciparum* heat shock protein genes, and was found to be conserved across several *Plasmodium* species [18]. Since publication of the *P. falciparum* genome sequence, several transcriptome [19,20,21•,22,23] and proteome [3,24,25] studies of various life-cycle stages have been completed. These sequences have enabled further insight into gene regulation in *Plasmodium*; for example, microarray studies have shown that steady-state RNA levels for many transcripts change throughout the parasite life-cycle, indicating transcriptional regulation at the level of RNA synthesis and/or stability. Bozdech *et al.* [22] generated microarray data for the *P. falciparum* asexual stages and suggested that a small number of transcription factors with overlapping binding site specificities could account for the mechanical character of transcriptional control. These studies have also been cross-referenced to proteomic studies of the asexual, sexual and mosquito stages of *P. falciparum* [24,25], subsequently revealing that large proportions of the genome encode proteins that are used in multiple stages of the life-cycle. This has led to theories that a complex, multi-layer regulatory network is employed by the parasite for gene expression, a different mode of regulation than that observed in other eukaryotes [22].

More recently, a search for transcription-associated proteins within the complete *P. falciparum* genome found that the parasite contained far fewer than expected in comparison with those of other eukaryotes [26], leading to the conclusion that *Plasmodium* protein levels might be primarily determined by post-transcriptional mechanisms. A global analysis of transcript and protein levels in *P. falciparum* also led to a similar conclusion [21•]. As a step towards identifying motifs involved in such gene regulation, a recent study took advantage of available *P. berghei* microarray studies and proteomic data and identified a motif in the 3'UTR (untranslated region) of genes that are upregulated in gametocyte stages but whose protein products appear after transmission from the vertebrate to the invertebrate host [4••]. This enabled identification of the motif as a putative control-element involved in the translational repression of transcripts produced during the sexual stages. Downstream regions of *P. falciparum* orthologs of these genes did not contain the same motif, indicating that the regulation of gene expression could be a major contributor to *Plasmodium* host specificity and parasite diversification.

Finally, the first proteomic analysis of separated male and female gametocytes in *P. berghei* has shown that expression of sex-specific proteins is probably controlled by the 5'UTR and not by the 3'UTR nor through post-translational processes [27•]. It remains to be seen how applicable this finding is to the control of gene expression in other *Plasmodium* species.

Comparative evolutionary studies

Comparative genomic methods have been essential in attempting to understand the evolutionary history of malaria parasites in addition to the selective pressures acting upon them (for a recent review, see [28^{*}]). For example, several studies using a limited number of genes have proposed a 'Malaria's Eve' hypothesis, which suggests that, on the basis of the sparse genetic diversity exhibited by the parasite, extant *P. falciparum* originated from a population bottleneck around 6000–10 000 years ago [29,30]. By contrast, Mu *et al.* generated a SNP (single nucleotide polymorphism) density map of *P. falciparum* chromosome 3 from five geographically distinct isolates, and based on the level of divergence, dated the origin of *P. falciparum* to 100 000–180 000 years ago [31], significantly older than Malaria's Eve. Another comparative genomics study, which somewhat reconciled these two extreme views, analyzed the genome sequence of the 6 kb mitochondrial genome from 100 *P. falciparum* isolates world-wide, and provided compelling evidence of an ancient origin for the species (50 000–100 000 years old) but with a recent expansion in the African malaria parasite population [32]. Such large-scale mitochondrial sequencing studies have been repeated using *P. vivax*, and this species, too, was found to be an ancient parasite [33^{**}] that most probably arose by a host switch from macaque monkeys [34].

Understanding the diversity of human *Plasmodium* isolates is of serious consequence for control measures, because a genetically homogenous population is easier to control than a variable, heterogenous one through the rational use of drugs and/or vaccines. In addition to the studies mentioned above, several large-scale genomics approaches have been used to study *P. vivax* diversity. Feng *et al.* [35] sequenced a 100 kb region syntenic to *P. falciparum* chromosome 3 in five *P. vivax* isolates and compared the perceived evolutionary histories of the orthologs between species and within species. A highly diverse *P. vivax* genome was revealed, and orthologous genes between the species were found to evolve at different rates and with different mutation patterns. More recently, whole genome analysis of *P. falciparum* and its closest relative, the chimpanzee species, *Plasmodium reichenowi*, has shown that the highest level of divergence occurs within genic sequences at four-fold synonymous sites, followed by introns and then intergenic sequences [36^{*}]. This is similar to the pattern seen in primates and suggests that the greater level of conservation in intergenic sites might be caused by conserved regulatory sequences [36^{*}]. Comparison of evolution rates in duplicated versus non-duplicated genes in *P. falciparum* and *P. y. yoelii* has demonstrated that duplicated genes are evolving more rapidly at the nucleotide level and have accelerated rates of intron gain and loss [37]. This supports the theory that paralogous gene family expansion and diversification is playing a major role in the evolution of malaria parasites.

Finally, identification of rapidly evolving genes in *Plasmodium* species is of considerable interest because such genes might be interacting with the host immune system. A genome-wide analysis of selective constraints in the genomes of *P. berghei* and *P. chabaudi* demonstrated that putative surface-proteins that are expressed in the vertebrate host are evolving more rapidly than those expressed in the mosquito vector [4^{**}]. As the genome sequences of more *Plasmodium* species are completed, this comparative analysis will become more powerful and promises to become a useful method for identifying host-interacting proteins.

Conclusion

Notwithstanding the recent tremendous advances in understanding *Plasmodium* biology that have been facilitated by comparative genomics of *Plasmodium* parasites, there remains much more that can be done. To all intents, the whole genome comparisons completed to date have been with only two species of *Plasmodium*, because the three rodent malaria genomes sequenced are too closely related to provide more than a single reference point. With the completion of the genome sequence of a second human species (*P. vivax*) and a second model species genome (*P. knowlesi*), expected in Spring 2006, novel analyses such as determining 'original synteny' and dissecting the evolutionary pathway of gene expression regulation might be possible. The landscape of comparative genomics of malaria parasites is set to change significantly over the next few years.

Acknowledgements

The authors would like to acknowledge the Burroughs Wellcome Fund, National Institute of Allergy and Infectious Diseases, US Department of Defense and the Wellcome Trust for their support of the *Plasmodium* genome sequencing initiatives.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- of special interest
- of outstanding interest

1. Snow RW, Guerra CA, Noor AM, Myint HY, Hay SI: **The global distribution of clinical episodes of *Plasmodium falciparum* malaria.** *Nature* 2005, **434**:214-217.
2. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S *et al.*: **Genome sequence of the human malaria parasite *Plasmodium falciparum*.** *Nature* 2002, **419**:498-511.
3. Carlton JM, Angiuoli SV, Suh BB, Kooij TW, Perteau M, Silva JC, Ermolaeva MD, Allen JE, Selengut JD, Koo HL *et al.*: **Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*.** *Nature* 2002, **419**:512-519.
4. Hall N, Karras M, Raine JD, Carlton JM, Kooij TW, Berriman M, Florens L, Janssen CS, Pain A, Christophides GK *et al.*: **A comprehensive survey of the *Plasmodium* life cycle by genomic, transcriptomic, and proteomic analyses.** *Science* 2005, **307**:82-86.

This study includes comparative genome sequence analysis across four species, plus a microarray and proteomic study of *P. berghei*. Integration of these datasets enabled the authors to study evolution rates of genes

expressed in different stages, and to identify a sequence involved in post-transcriptional repression.

5. Carlton J, Silva J, Hall N: **The genome of model malaria parasites, and comparative genomics.** *Curr Issues Mol Biol* 2005, **7**:23-37.

This is a particularly useful review describing the structure of *Plasmodium* genomes in detail for readers without a genomic background. Although it predates the publication of the *P. berghei* and *P. chabaudi* genomes it provides broad comparisons across the most studied species, and detailed comparisons of *P. falciparum* with *P. yoelii*

6. McCutchan TF, Dame JB, Miller LH, Barnwell J: **Evolutionary relatedness of *Plasmodium* species as determined by the structure of DNA.** *Science* 1984, **225**:808-811.
7. Janse CJ, Carlton JM, Walliker D, Waters AP: **Conserved location of genes on polymorphic chromosomes of four species of malaria parasites.** *Mol Biochem Parasitol* 1994, **68**:285-296.
8. Carlton JM, Galinski MR, Barnwell JW, Dame JB: **Karyotype and synteny among the chromosomes of all four species of human malaria parasite.** *Mol Biochem Parasitol* 1999, **101**:23-32.
9. Carlton JM, Vinkenoog R, Waters AP, Walliker D: **Gene synteny in species of *Plasmodium*.** *Mol Biochem Parasitol* 1998, **93**:285-294.
10. del Portillo HA, Fernandez-Becerra C, Bowman S, Oliver K, Preuss M, Sanchez CP, Schneider NK, Villalobos JM, Rajandream MA, Harris D *et al.*: **A superfamily of variant genes encoded in the subtelomeric region of *Plasmodium vivax*.** *Nature* 2001, **410**:839-842.
11. Janssen CS, Phillips RS, Turner CM, Barrett MP: ***Plasmodium* interspersed repeats: the major multigene superfamily of malaria parasites.** *Nucleic Acids Res* 2004, **32**:5712-5720.

This is an excellent bioinformatics study of the *pir* gene family of *Plasmodium* species. It describes the structure and diversity of the genes and, most importantly, suggests that the *rifin* family is in fact a member of the *pir* superfamily, which is found in all other sequenced species.

12. Rasti N, Wahlgren M, Chen Q: **Molecular aspects of malaria pathogenesis.** *FEMS Immunol Med Microbiol* 2004, **41**:9-26.
13. al-Khedery B, Barnwell JW, Galinski MR: **Antigenic variation in malaria: a 3' genomic alteration associated with the expression of a *P. knowlesi* variant antigen.** *Mol Cell* 1999, **3**:131-141.
14. Winter G, Kawai S, Haeggstrom M, Kaneko O, von Euler A, Kawazu S, Palm D, Fernandez V, Wahlgren M: **SURFIN is a polymorphic antigen expressed on *Plasmodium falciparum* merozoites and infected erythrocytes.** *J Exp Med* 2005, **201**:1853-1863.
- This study identifies a new family of proteins in *P. falciparum*, members of which are directed to the surface of the red blood cell. Comparison to other species suggests several conserved motifs with other gene families such as *vir* and *SICAvar*.
15. Knuepfer E, Rug M, Klonis N, Tilley L, Cowman AF: **Trafficking of the major virulence factor to the surface of transfected *P. falciparum*-infected erythrocytes.** *Blood* 2005, **105**:4078-4087.
- This study identifies and validates a motif involved in targeting *Plasmodium* proteins to the surface of the red blood cell. The authors demonstrate conservation of this motif across several *Plasmodium* species.
16. Hiller NL, Bhattacharjee S, van Ooij C, Liolios K, Harrison T, Lopez-Estrano C, Haldar K: **A host-targeting signal in virulence proteins reveals a secretome in malarial infection.** *Science* 2004, **306**:1934-1937.
17. Horrocks P, Decherig K, Lanzer M: **Control of gene expression in *Plasmodium falciparum*.** *Mol Biochem Parasitol* 1998, **95**:171-181.
18. Militello KT, Dodge M, Bethke L, Wirth DF: **Identification of regulatory elements in the *Plasmodium falciparum* genome.** *Mol Biochem Parasitol* 2004, **134**:75-88.
19. Le Roch KG, Zhou Y, Blair PL, Grainger M, Moch JK, Haynes JD, De La Vega P, Holder AA, Batalov S, Carucci DJ *et al.*: **Discovery of gene function by expression profiling of the malaria parasite life cycle.** *Science* 2003, **301**:1503-1508.

20. Le Roch KG, Zhou Y, Batalov S, Winzeler EA: **Monitoring the chromosome 2 intraerythrocytic transcriptome of *Plasmodium falciparum* using oligonucleotide arrays.** *Am J Trop Med Hyg* 2002, **67**:233-243.
21. Le Roch KG, Johnson JR, Florens L, Zhou Y, Santrosyan A, Grainger M, Yan SF, Williamson KC, Holder AA, Carucci DJ *et al.*: **Global analysis of transcript and protein levels across the *Plasmodium falciparum* life cycle.** *Genome Res* 2004, **14**:2308-2318.
- The authors compare mRNA and protein abundance from a number of studies to quantify the role of post-transcriptional regulation in parasite development. Much of the abundance data between mRNA and proteins are in agreement, suggesting that mRNA regulation is the predominant form of expression regulation. The authors did find that the gametocyte transcriptome was most similar to the gamete proteome, which suggested that post-transcriptional regulation was also important.
22. Bozdech Z, Zhu J, Joachimiak MP, Cohen FE, Pulliam B, DeRisi JL: **Expression profiling of the schizont and trophozoite stages of *Plasmodium falciparum* with a long-oligonucleotide microarray.** *Genome Biol* 2003, **4**:R9.
23. Bozdech Z, Llinas M, Pulliam BL, Wong ED, Zhu J, DeRisi JL: **The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*.** *PLoS Biol* 2003, **1**:E5.
24. Lasonder E, Ishihama Y, Andersen JS, Vermunt AM, Pain A, Sauerwein RW, Eling WM, Hall N, Waters AP, Stunnenberg HG *et al.*: **Analysis of the *Plasmodium falciparum* proteome by high-accuracy mass spectrometry.** *Nature* 2002, **419**:537-542.
25. Florens L, Washburn MP, Raine JD, Anthony RM, Grainger M, Haynes JD, Moch JK, Muster N, Sacci JB, Tabb DL *et al.*: **A proteomic view of the *Plasmodium falciparum* life cycle.** *Nature* 2002, **419**:520-526.
26. Coulson RM, Hall N, Ouzounis CA: **Comparative genomics of transcriptional control in the human malaria parasite *Plasmodium falciparum*.** *Genome Res* 2004, **14**:1548-1554.
27. Khan SM, Franke-Fayard B, Mair GR, Lasonder E, Janse CJ, Mann M, Waters AP: **Proteome analysis of separated male and female gametocytes reveals novel sex-specific *Plasmodium* biology.** *Cell* 2005, **121**:675-687.
- This proteomic study of male and female gametocytes in *P. berghei* identified male- and female-specific proteins. Regulatory sequences are validated using green fluorescent protein (GFP) tagging, and gene disruptions are carried out to validate putative signal-transduction proteins, enabling the identification of new signal-transduction mechanisms.
28. Hartl DL: **The origin of malaria: mixed messages from genetic diversity.** *Nat Rev Microbiol* 2004, **2**:15-22.
- This review gives an excellent overview of the differing data either supporting or refuting the 'Malaria's Eve' hypothesis.
29. Volkman SK, Barry AE, Lyons EJ, Nielsen KM, Thomas SM, Choi M, Thakore SS, Day KP, Wirth DF, Hartl DL: **Recent origin of *Plasmodium falciparum* from a single progenitor.** *Science* 2001, **293**:482-484.
30. Rich SM, Licht MC, Hudson RR, Ayala FJ: **Malaria's Eve: evidence of a recent population bottleneck throughout the world populations of *Plasmodium falciparum*.** *Proc Natl Acad Sci USA* 1998, **95**:4425-4430.
31. Mu J, Duan J, Makova KD, Joy DA, Huynh CQ, Branch OH, Li WH, Su XZ: **Chromosome-wide SNPs reveal an ancient origin for *Plasmodium falciparum*.** *Nature* 2002, **418**:323-326.
32. Joy DA, Feng X, Mu J, Furuya T, Chotivanich K, Krettli AU, Ho M, Wang A, White NJ, Suh E *et al.*: **Early origin and recent expansion of *Plasmodium falciparum*.** *Science* 2003, **300**:318-321.
33. Jongwutiwes S, Putaporntip C, Iwasaki T, Ferreira MU, Kanbara H, Hughes AL: **Mitochondrial genome sequences support ancient population expansion in *Plasmodium vivax*.** *Mol Biol Evol* 2005, **22**:1733-1739.
- This study of 106 mitochondrial genomes dates the emergence of *P. vivax* to ~30 000 years ago, suggesting that it emerged with the host species. A similar finding was reached in a study of mitochondrial genome polymorphisms in *P. falciparum* [32].

34. Mu J, Joy DA, Duan J, Huang Y, Carlton J, Walker J, Barnwell J, Beerli P, Charleston MA, Pybus OG *et al.*: **Host switch leads to emergence of *Plasmodium vivax* malaria in humans.** *Mol Biol Evol* 2005, **22**:1686-1693.
35. Feng X, Carlton JM, Joy DA, Mu J, Furuya T, Suh BB, Wang Y, Barnwell JW, Su XZ: **Single-nucleotide polymorphisms and genome diversity in *Plasmodium vivax*.** *Proc Natl Acad Sci USA* 2003, **100**:8502-8507.
36. Neafsey DE, Hartl DL, Berriman M: **Evolution of noncoding and silent coding sites in the *Plasmodium falciparum* and *P. reichenowi* genomes.** *Mol Biol Evol* 2005, **22**:1621-1626.

This study, although based on an early and fragmented assembly of *P. reichenowi*, measures the evolution rates of different silent sites in the *P. falciparum* genome, using comparative analysis. The authors demonstrate that the intergenic sites are not mutating as rapidly as intron or four-fold synonymous coding sites. They conclude that there is evidence of selection on the intergenic sites, probably as a result of regulatory sequences.

37. Castillo-Davis CI, Bedford TB, Hartl DL: **Accelerated rates of intron gain/loss and protein evolution in duplicate genes in human and mouse malaria parasites.** *Mol Biol Evol* 2004, **21**:1422-1427.

**Have you contributed to an Elsevier publication?
Did you know that you are entitled to a 30% discount on
books?**

A 30% discount is available to ALL Elsevier book and journal contributors when ordering books or stand-alone CD-ROMs directly from us.

To take advantage of your discount:

1. Choose your book(s) from www.elsevier.com or www.books.elsevier.com

2. Place your order

Americas:

TEL.: +1 800 782 4927 for US customers

TEL.: +1 800 460 3110 for Canada, South & Central America customers

FAX: +1 314 453 4898

E-MAIL: author.contributor@elsevier.com

All other countries:

TEL.: +44 1865 474 010

FAX: +44 1865 474 011

E-MAIL: directorders@elsevier.com

You'll need to provide the name of the Elsevier book or journal to which you have contributed. Shipping is FREE on pre-paid orders within the US, Canada, and the UK.

If you are faxing your order, please enclose a copy of this page.

3. Make your payment

This discount is only available on prepaid orders. Please note that this offer does not apply to multi-volume reference works or Elsevier Health Sciences products.

For more information, visit www.books.elsevier.com



ELSEVIER

Hemiascomycetous yeasts at the forefront of comparative genomics

Bernard Dujon

With more than a dozen species fully sequenced, as many as this partially sequenced, and more in development, yeasts are now used to explore the frontlines of comparative genomics of eukaryotes. Innovative procedures have been developed to compare and annotate genomes at various evolutionary distances, to identify short *cis*-acting regulatory elements, to map duplications, or to align syntenic blocks. Human and plant pathogens, in addition to yeasts that show a variety of interesting physiological properties, are included in this multidimensional comparative survey, which encompasses a very broad evolutionary range. As major steps of the evolutionary history of hemiascomycetous genomes emerge, precise questions on the general mechanisms of their evolution can be addressed, using both experimental and *in silico* methods.

Addresses

Unité de Génétique moléculaire des levures (associated with CNRS and University Pierre and Marie Curie), Institut Pasteur, 25, rue du Docteur Roux, F-75724, Paris Cedex 15, France

Corresponding author: Dujon, Bernard (bdujon@pasteur.fr)

Current Opinion in Genetics & Development 2005, **15**:614–620

This review comes from a themed issue on
Genomes and evolution
Edited by Stephen J O'Brien and Claire M Fraser

Available online 26th September 2005

0959-437X/\$ – see front matter

© 2005 Elsevier Ltd. All rights reserved.

DOI 10.1016/j.gde.2005.09.005

Introduction

Less than ten years have now passed since the first DNA sequence of a eukaryotic organism — that of the baker's yeast, *Saccharomyces cerevisiae* — was entirely unveiled [1]. This remarkable achievement quickly contributed to the emergence of functional genomics. But rare were those at this time who anticipated that, a few years later, the genome sequences of many other yeast species would also become available, promoting these unicellular fungi to the forefront of comparative genomics. Presently, the complete, near complete or partial genome sequences of more than two dozen yeast species have been reported, offering a collection of genomic information without equal among other eukaryotic groups (Figure 1). The significance of this novel situation, made possible by the progress in sequencing techniques, emerges from the fact that, despite their similar morphology and common life

styles, yeasts form a much diversified group. Furthermore, several of them, none more so than *S. cerevisiae*, are favoured organisms for genetic experiments. Most yeasts sequenced to date are members of the Hemiascomycete class, the group of fungi to which budding yeasts belong and which, from genome analysis, was recently discovered to cover an evolutionary range larger than that of the entire phylum of Chordates [2**]. Other yeasts belonging to the Archiascomycetes or the Basidiomycetes have also been sequenced but will not be discussed here, because the phylogenetic distances among those fungal groups are so considerable that it is difficult to compare genomes in any detail. By contrast, comparisons within the Hemiascomycetes can be performed at various phylogenetic distances, depending on the type of question examined.

The large-scale comparative exploration of hemiascomycetous genomes started five years ago. Thirteen yeast species, selected to sample various branches of the known phylogenetic tree, were sequenced at low coverage, and each was compared with *S. cerevisiae* [3]. The results indicated the power of rapid genome survey to identify conserved or specific genes, to examine the evolution of functional categories or to compare genetic maps in search of the mechanisms of genome evolution. But yeast comparative genomics has considerably accelerated over the past two years, with the successive publications of the complete or high-coverage sequences of a large panel of yeast species, selected on the basis of their intrinsic interest and/or for their phylogenetic position. Some species are major human pathogens; others are used in food processing. Some are able to propagate on a variety of natural substrates; others show specific niche adaptation. The novel genomic data were used to examine questions of general significance regarding eukaryotic genome evolution, but they also served to explore and develop novel methods and strategies of general applicability for comparative genomics. Using the yeast sequences, a large variety of biological questions can now be addressed by experimental and/or *in silico* analyses. This short review only focuses on a limited number of prominent results obtained during the past two years.

Comparative genomics on a short evolutionary range: gene discovery, speciation and identification of conserved regulatory sites

Several species of the *Saccharomyces sensu stricto* clade have been sequenced and compared [4,5]. Their sequence divergence is significant but they share very high map-synteny (see Glossary), interrupted only by a limited

Glossary

Allotetraploidy: The status of a cell or an organism having four full sets of chromosome complements, two derived from one diploid species, the other two from another, different, diploid species.

Aneuploidy: The status of a cell or an organism having a non-uniform number of the different chromosomes. This status can be caused, for example, by the loss of one chromosome from a complete diploid set, or by the addition of a supernumerary chromosome copy to a complete chromosome set.

Autotetraploidy: The status of a cell or an organism having four full sets of chromosome complements derived from the duplication of an originally diploid set.

Collinearity: Relates to objects (for example, a gene and its corresponding protein, or two chromosomes) having corresponding parts arranged in the same linear order.

Gene conversion: The process by which a gene sequence (acceptor) is partially replaced by a copy of another gene sequence (donor) from the same genome. In general, the donor and acceptor sequences must share a sufficient degree of sequence similarity (for example, the two alleles in a diploid, or two paralogs).

Paralogy: Homology between two non-allelic genes of the same genome, derived by duplication from a common ancestor.

Synonymous substitution: A nucleotide substitution, in a gene sequence encoding a protein, that does not result in an amino-acid change.

Synteny: The common presence of genes along a given chromosome or chromosomal segment. The notion generally also implies the order of those genes. Hence, conservation of synteny indicates the conservation of the order of homologous genes between two chromosomes or between chromosomal segments of different species.

number of chromosomal translocations and a higher number of single gene-deletions [6]. Species definition is made on the basis of the post-zygotic barrier: viable hybrids easily form and are mitotically stable — some are used for industrial fermentations — but they are generally sterile at meiosis. In artificially produced interspecific hybrids, the viability of meiotic spores can be partly restored by engineered reconstructions of map collinearity (see Glossary), but a high trend to aneuploidy (see Glossary) remains [7].

One of the immediate impacts of genome comparisons between related organisms is the significant improvement of sequence annotation. This has been true even for *S. cerevisiae*, the initial sequence interpretation of which was considerably simplified, in comparison with that of multicellular organisms, by the compactness of its genome and the paucity of introns. Nonetheless, a few dozen overlooked small genes, often with introns, were found by comparisons; the coordinates of several other open reading frames (ORFs) were corrected; and about a tenth of the initially proposed ORFs, often partially overlapping other genes, were shown to be spurious [3–5,8,9]. Current estimates predict the actual number of protein-coding genes of *S. cerevisiae* to be around 5700 (± 100), a number that includes the approximately 450 pairs of paralogs (see Glossary) that remain after the ancestral whole-genome duplication, and the approximately 1600 other genes that are members of multigene families originating from other ancestral duplications (see below). Obviously, the

improved annotation of *S. cerevisiae* can, in turn, facilitate annotation of other yeasts and other eukaryotic organisms. The novel genes need to be included in future global functional studies.

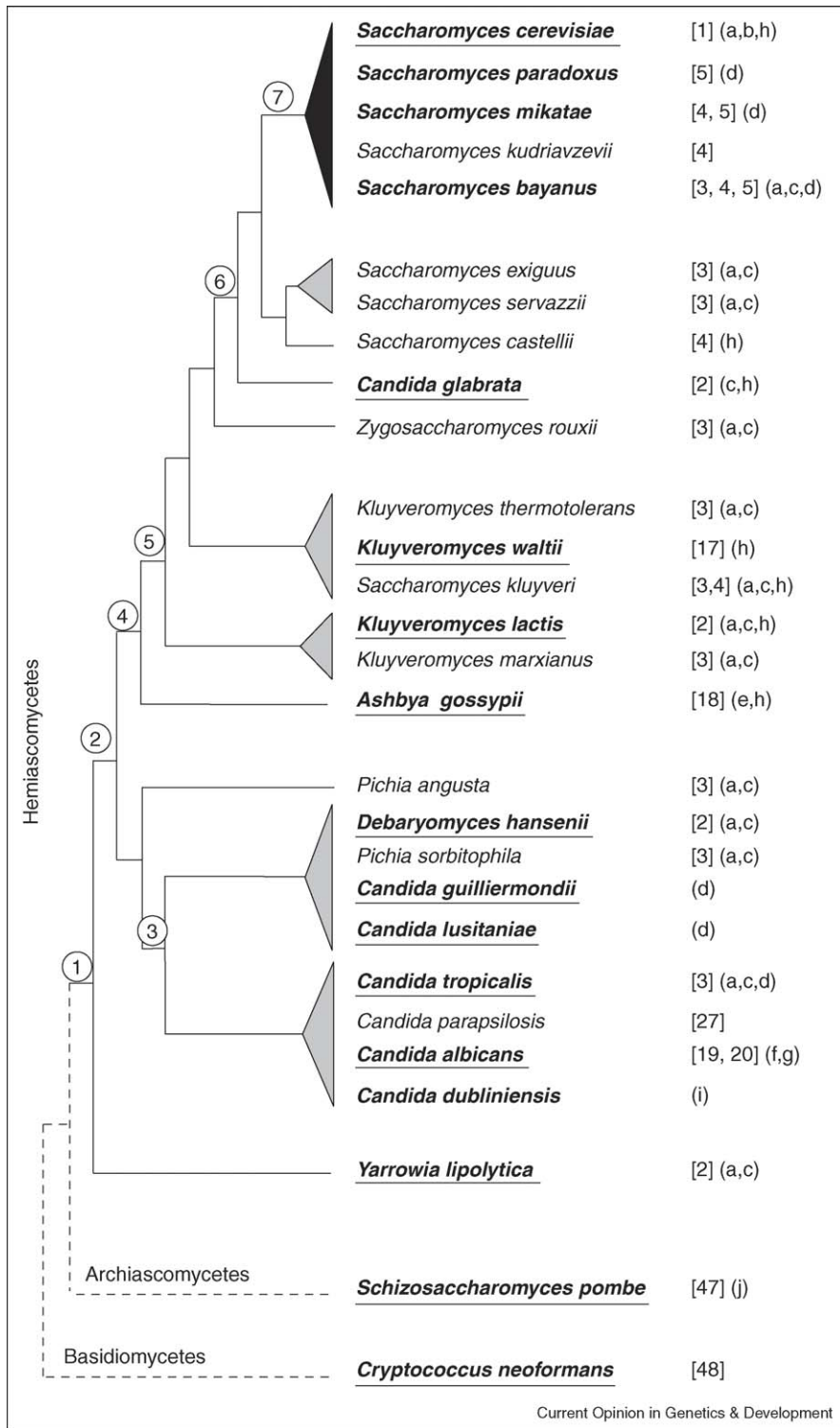
The *Saccharomyces sensu stricto* clade was used to evaluate the power of comparative genomics to identify novel *cis*-regulatory sequences, which are otherwise difficult to recognize. Sequence alignments of intergenic regions between these species, as well as with more distant yeasts, facilitated the identification of numerous novel regulatory motifs [4,5]. Elements containing these motifs can then be experimentally assayed or compared against the sequences recently determined to be bound to transcriptional regulators [10]. The recent systematic mapping of transcriptional start sites in *S. cerevisiae* will also be helpful [11]. The success of the comparative methods that have been developed [12,13], which are of general interest for investigation of many other organisms, obviously depends on the set of sequenced species available [14]. But, recently, the procedure has been successfully extended to larger evolutionary distances, such as those found in filamentous fungi [15].

The broad evolutionary range covered by Hemiascomycetes: synteny, genome content, pathway conservation and niche adaptation

Estimated to have separated from the fission yeast, *Schizosaccharomyces pombe*, between 350 and 1000 million years ago [16], Hemiascomycetes cover a broad evolutionary range. Judging by the general distributions of conserved amino-acid identities between orthologous proteins, *Candida glabrata* and *S. cerevisiae*, for example, are as distant from each other as are man and fishes [2^{**}]. And much broader distances from *S. cerevisiae* exist for other clades (for example, from *Candida albicans* or *Yarrowia lipolytica*; see Figure 1). The considerable reshuffling of genetics maps is congruent with these large evolutionary distances between clades [2^{**},17^{**},18^{**}]. When comparing species of different clades, mosaics of short conserved syntenic blocks, separated by numerous breakpoints and often containing internal inversions of a few genes, are found between all chromosomes.

Despite the evolutionary distances, there exists a large set of protein families that are common among these yeasts, and most of these families are also common to other groups of fungi or are universally conserved [2^{**},17^{**},18^{**},19,20]. Within some families, specific expansion or contraction of gene numbers occurs in the various yeasts and can be related to their known physiological properties or used to suggest novel ones. Against this common heritage, each species harbours several specific genes the function and origin of which is often unclear but probably contributes to its originality. Specificity is also obtained by the loss of certain genes that are common to other species. This phenomenon is frequent

Figure 1



The yeast species presently sequenced, and a chart of their evolutionary history. All species, except for *S. pombe* [45] and *Cryptococcus neoformans* [46] (used here as outgroups), belong to the Hemiascomycete class, the general phylogenetic topology of which is indicated [47,48]. Closely related species are defined as clades (grey triangles). The extensively studied *Saccharomyces sensu stricto* clade is shown by a black triangle. Completed or essentially completed sequences are bold and underlined; high coverage (greater than six genome equivalents) shotgun sequences are bold; others are medium- (approximately 3X) or low-coverage shotgun sequences and/or work in progress. Only publicly available sequences are indicated. References are in square brackets. URLs of specialized sites where data can be accessed:

and has happened in every yeast lineage. The same pathway can be lost independently in several evolutionary branches, as in the case of galactose utilization [21]. Reductive evolution by gene loss is particularly striking in the case of the pathogenic yeast *C. glabrata*, which has specifically lost several functional pathways that are present in related species [2**].

The evolutionary conservation of functional pathways has been studied at a large scale using yeast genome sequences: novel transporters have been identified [22]; carefully measured metabolic-fluxes can be compared between yeasts and related to their gene content, as was done for the glucose utilization pathways [23]; the conservation of genes involved in replication, recombination and repair has been systematically examined [24]; proteins involved in gene silencing have been shown to evolve rapidly [25], indicating that several independent solutions to this problem have been explored throughout evolution; the evolution of genes involved in mating-type and sexual cycle has been studied in detail in relation to the multiple, independent loss of sexuality in the various Hemiascomycete lineages [25–27]. Finally, the structure and gene content of subtelomeric regions has been compared between *Kluyveromyces lactis* and *S. cerevisiae* [28]: these regions appear as highly dynamic structures, offering a preferred location for genes involved in rapid adaptive evolution, and they contribute to a significant degree of the global genome redundancy.

A whole genome duplication in the ancestry of some Hemiascomycete yeasts

The ancient whole-genome duplication in the ancestry of *S. cerevisiae*, postulated several years ago [29] on the basis of the numerous pairs of chromosomal homologous regions, has been recently confirmed by two independent criteria. As expected from this hypothesis, the genomes of *Kluyveromyces waltii* [17**] and *Ashbya gossypii* [18**], which have not inherited this duplication, appear as a succession of segments, covering nearly their entire lengths, which show conserved synteny, simultaneously, with two distinct segments of the genome of *S. cerevisiae*. Comparison with *C. glabrata* [2**] shows an extensive coincidence of the chromosomal homologous regions of the two species, indicating that they have inherited the same ancestral

duplication event, which can be more precisely located on the phylogenetic tree (Figure 1). But the precise nature of this ancient event remains uncertain. In *S. cerevisiae*, autotetraploids (see Glossary) show a highly elevated rate of chromosomal instability and fail to arrest in glucose-limited stationary phase, resulting in low rates of survival [30]. By opposition, allotetraploids (see Glossary) appear healthy and relatively stable in mitotic growth but tend to be meiotically impaired [31]. In agreement with commonly held views on evolution, the majority of the anciently duplicated genes have been lost. Deletions appear to be random and essentially concern single genes, not segments, resulting in the observed mosaic nature of the chromosomal homologous regions [17**,18**]. In *C. glabrata*, deletions have been so numerous as to leave only approximately 2% of the postulated ancestral pairs of paralogs, compared with the approximately 8% that remain in *S. cerevisiae* [2**]. Relics of genes lost by massive accumulation of deleterious mutations are also visible in the genome [6,32]. Cases of functional specialization between the duplicated copies have been mentioned [17**], but, in general, a rapid divergence of expression after duplication seems to have occurred, causing important functional asymmetry between the copies [33]. Synonymous substitutions (see Glossary) between the remaining active pairs of paralogs are not uniform, suggesting a concerted evolution by gene conversion (see Glossary) [34*,35].

Segmental duplications, tandem gene arrays, and single gene duplication

Comparative genomics also illustrates the role of other duplication processes in the evolution of yeast genomes. Traces of a few segmental duplications were recognized in the genome of *S. cerevisiae*, taking into account the presence of gene relics [32]. Segmental duplications are also regularly observed in subtelomeric regions [28] and were recognized in the genomes of several yeast species [2**]. The spontaneous formation of large segmental duplications, in which dozens or hundreds of neighboring genes are simultaneously duplicated, was recently demonstrated experimentally using a gene dosage recovery assay in *S. cerevisiae* [36**]. These events are observed at a frequency of between approximately 10^{-9} and 10^{-10} per mitosis in haploid cell cultures, suggesting that, given

(Figure 1 Legend continued) (a) <http://mips.gsf.de/genre/proj/yeast/>; (b) <http://www.yeastgenome.org/>; (c) <http://cbl.labri.fr/Genolevures/>; (d) <http://www.broad.mit.edu/annotation/fgi/>; (e) <http://agd.unibas.ch/>; (f) <http://www-sequence.stanford.edu/group/candida/>; (g) <http://genolist.pasteur.fr/CandidaDB/>; (h) <http://wolfe.gen.tcd.ie/ygob/>; (i) http://www.sanger.ac.uk/Projects/C_dubliniensis/; (j) <http://www.genedb.org/genedb/pombe/>. Other yeast genome projects have been mentioned [49,50] but sequences remain proprietary. Genome size varies between approximately 9 Mb (in the case of *A. gossypii*) and 14 Mb (*C. albicans*) for all Hemiascomycetes, except for *Y. lipolytica*, the species on the most external branch sequenced to date, in which it reaches 20 Mb and yet has only a slightly higher gene number. Major evolutionary events at branch points are the following: (1) origin of the Hemiascomycetes (budding yeasts); (2) genome size control (range approximately 9–14 Mb); (3) deviation from universal genetic code; (4) emergence of short, tripartite centromeres and mating-type cassettes; (5) acquisition of HO endonuclease (pseudo-homothalism by mating-type switching); (6) whole-genome duplication, emergence of petite-positive yeasts; and (7) emergence of the *Saccharomyces sensu stricto* group, multiplication of sugar utilization genes. Numerous secondary events occurred individually in the branches, such as: gene loss (sometimes extensive) or inactivation (relics); loss of transposons; loss of sex; formation of segmental duplications; tandem gene formation; horizontal gene transfer (rare); transposon-mediated gene duplications (retrogenes); loss and acquisition of introns; chromosomal translocations and rearrangements; and divergence of duplicate gene regulations.

the size of natural yeast populations, they must occur very frequently over time. Intrachromosomal direct-tandem duplications are the most frequent events, but tend to be unstable at meiosis, hence limiting their possible evolutionary role. But interchromosomal duplications also occur frequently and might occasionally generate a super-numerary chromosome. The mechanism at the origin of the spontaneous segmental duplications has not yet been elucidated, but indirect evidences suggest accidental secondary firing of some replicons during S phase.

Short tandem gene arrays are also observed in all yeasts. Globally, they are more numerous in some species, in which a few, specific, larger arrays are also observed [2**]. The absence of coincidence of tandem arrays between species, and the dynamics of expansion and contraction of some large arrays within populations [37] suggest that such gene arrays are the sites of rapid adaptive evolution.

Dispersed copies of paralogous genes are also observed in all yeast genomes and are generally in higher numbers than those identifiable to all above mechanisms; however, their origin remains uncertain. The duplication of single genes at ectopic locations seems improbable. Dispersed paralogs might be the remnants of ancient segmental duplications after deletion of all other genes. But the retrotransposon-mediated duplication of partial gene copies that has been recently demonstrated in *S. cerevisiae*, using a genetic selection system [38*], offers an attractive alternative hypothesis. An important consequence of this mechanism, along with the segmental duplication mechanism, is the formation of chimeric genes at junctions. Although probably non-functional in the majority of cases, chimeric proteins with two distinct functional domains are likely to emerge over time, and several interesting examples are observed in yeast genomes [39**].

Accidental horizontal gene transfers

Contrary to its important role in bacteria, horizontal gene transfer is numerically limited in yeast genomes, for which only a few cases (less than 0.2% of the total gene number) have been recorded [2**,40,41]. But the contribution of these rare events might become significant for niche specialization over time. When functionally identified, yeast genes originating from horizontal gene transfer almost always correspond to enzymatic functions, and, in several cases, they are duplicated in the species in which they reside, suggesting a selective advantage. A 'prokaryotic-type' gene encoding a dihydroorotate dehydrogenase in *S. cerevisiae* and other related yeasts, has been proposed to be at the origin of those yeast species able to grow in complete anaerobiosis, because the corresponding enzyme is active in the absence of oxygen, contrary to the case for the common 'eukaryotic-type' enzyme [40,41]. However, other differences between strictly aerobic yeasts and facultative anaerobes exist, in particular in the large-scale modulation of the transcriptional network [42**].

Another example of horizontal gene transfer concerns an alkylsulfatase-encoding gene of *S. cerevisiae*, which is believed to have been horizontally transferred from α -Proteobacteria, and which confers to its new yeast 'host' the ability to grow on sulfur-free minimal medium [41].

Conclusions

The multiple genome comparisons now possible among a large and rapidly increasing number of yeast species gradually reveal with ever increasing detail the evolutionary history of this diversified group of eukaryotes, at the same time as they unveil novel dimensions in our understanding of gene and genome evolution and offer multiple tools to explore them. The active evolutionary dynamics encountered, illustrated by the various modes of duplication, numerous chromosomal rearrangements, extensive gene loss, rewiring of transcriptional networks as briefly summarized above, is such that novel surprises are likely in the future exploration of novel, carefully selected yeast genomes. The formation of novel genes that, for lack of homologs, seem to have occurred in every yeast lineage remains puzzling. Several other exciting aspects that could not be addressed in this short review concern the non-coding RNA genes, introns, transposable elements, repeated DNA and protein segments. Yeasts are now a favoured case for fundamental studies on phylogenies [43] and, with *S. cerevisiae* in particular, will soon enable us to explore population genomics. At the same time, the data collected have important consequences for applications in biotechnology (with the discovery of novel enzymes or the efficient manipulation of industrial strains), in medicine and in agronomy (with the complete genetic characterization of important human and plant pathogens, and the possibility of identifying novel drug targets). If the variety of known yeast species is large, their hidden variety is probably much larger, because many more species remain to be isolated and identified from the variety of natural environments, as can be judged from recent explorations [44].

Acknowledgements

I wish to acknowledge my colleagues from the Génolevures Consortium (GDR 2354 Centre National de la Recherche Scientifique [CNRS]) in addition to the members of my own laboratory for efficient collaboration and fruitful discussions. Sequences analyzed by the Génolevures Consortium were produced by Génoscope (Jean Weissenbach and Patrick Wincker) and the Génopole Institut Pasteur-Ile de France (Christiane Bouchier). Work in my laboratory is supported by the Institut Pasteur, CNRS, Université Pierre-et-Marie-Curie and grants from Association pour la Recherche sur le Cancer and Action Coordonnée Incitative. BD is a member of Institut Universitaire de France.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- of special interest
- of outstanding interest

1. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M *et al.*: **Life with 6000 genes.** *Science* 1996, **274**:563-567.

2. Dujon B, Sherman D, Fischer G, Durrrens P, Casaregola S, Lafontaine I, De Montigny J, Marck C, Neuveglise C, Talla E *et al.*: **Genome evolution in yeasts.** *Nature* 2004, **430**:35-44.
This work presents the complete sequence and comparative analysis of four species spanning the entire class of Hemiascomycetes: *C. glabrata*, *K. lactis*, *Debaryomyces hansenii* and *Y. lipolytica*. The authors discuss the multiple mechanisms of evolution and conclude with a reconstitution of the evolutionary history of yeasts.
3. Souciet J-L, Aigle M, Artiguenave F, Blandin G, Bolotin-Fukuhara M, Bon E, Brottier P, Casaregola S, de Montigny J, Dujon B *et al.*: **Genomic exploration of the hemiascomycetous yeasts.** *FEBS Lett* 2000, **487**:3-147.
4. Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterston R, Cohen BA, Johnston M: **Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting.** *Science* 2003, **301**:71-76.
5. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES: **Sequencing and comparison of yeast species to identify genes and regulatory elements.** *Nature* 2003, **423**:241-254.
6. Fischer G, Neuveglise C, Durrrens P, Gaillardin C, Dujon B: **Evolution of gene order in the genomes of two related yeast species.** *Genome Res* 2001, **11**:2009-2019.
7. Delneri D, Colson I, Grammenoudi S, Roberts IN, Louis EJ, Oliver SG: **Engineering evolution to study speciation in yeasts.** *Nature* 2003, **422**:68-72.
8. Brachat S, Dietrich FS, Voegeli S, Zhang Z, Stuart L, Lerch A, Gates K, Gaffney T, Philippsen P: **Reinvestigation of the *Saccharomyces cerevisiae* genome annotation by comparison to the genome of a related fungus: *Ashbya gossypii*.** *Genome Biol* 2003, **4**:R45.
9. Kessler MM, Zeng Q, Hogan S, Cook R, Morales AJ, Cottarel G: **Systematic discovery of new genes in the *Saccharomyces cerevisiae* genome.** *Genome Res* 2003, **13**:264-271.
10. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne J-B, Reynolds DB, Yoo J *et al.*: **Transcriptional regulatory code of a eukaryotic genome.** *Nature* 2004, **431**:99-104.
11. Zhang Z, Dietrich FS: **Mapping of transcription start sites in *Saccharomyces cerevisiae* using 5' SAGE.** *Nucleic Acids Res* 2005, **33**:2838-2851.
12. Chiang DY, Moses AM, Kellis M, Lander ES, Eisen MB: **Phylogenetically and spatially conserved word pairs associated with gene-expression changes in yeasts.** *Genome Biol* 2003, **4**:R43.
13. Li X, Wong WH: **Sampling motifs on phylogenetic trees.** *Proc Natl Acad Sci USA* 2005, **102**:9481-9486.
14. Eddy SR: **A model of the statistical power of comparative genome sequence analysis.** *PLoS Biol* 2005, **3**:e10.
15. Gasch AP, Moses AM, Chiang DY, Fraser HB, Berardini M, Eisen MB: **Conservation and evolution of cis-regulatory systems in ascomycete fungi.** *PLoS Biol* 2004, **2**:e398.
16. Berbee ML, Taylor JW: **Systematics and evolution.** In *The Mycota VII B*. Edited by McLaughlin DJ, McLaughlin EG, Lemke PA. Berlin: Springer; 2001:229-245.
17. Kellis M, Birren BB, Lander ES: **Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*.** *Nature* 2004, **428**:617-624.
This article reports the high-coverage sequencing of *K. waltii*, and subsequent comparison of its genome with that of *S. cerevisiae*. The comparison reveals a global one-to-two map coincidence as predicted for an ancestral whole genome duplication. The authors continue with an analysis of duplicated gene-loss and of the evolution of gene pairs.
18. Dietrich FS, Voegeli S, Brachat S, Lerch A, Gates K, Steiner S, Mohr C, Pöhlmann R, Luedi P, Choi S *et al.*: **The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome.** *Science* 2004, **304**:304-307.
The complete sequence of *A. gossypii* and its comparison to *S. cerevisiae* shows a one to two relationship as predicted by a whole genome ancestral duplication. Similar conclusions to those made by Kellis *et al.* [17**].
19. Jones T, Federspiel NA, Chibana H, Dungan J, Kalman S, Magee BB, Newport G, Thorstenson YR, Agabian N, Magee PT *et al.*: **The diploid genome sequence of *Candida albicans*.** *Proc Natl Acad Sci USA* 2004, **101**:7329-7334.
20. Braun BR, van het Hoog M, d'Enfert C, Martchenko M, Dungan J, Kuo A, Inglis DO, Uhl MA, Hogues H, Berriman M *et al.*: **A human-curated annotation of the *Candida albicans* genome.** *PLoS Genetics* 2005, **1**:e1.
21. Hittinger CT, Rokas A, Carroll SB: **Parallel inactivation of multiple GAL pathway genes and ecological diversification in yeasts.** *Proc Natl Acad Sci USA* 2004, **101**:14144-14149.
22. De Hertogh B, Talla E, Tekai F, Beyne E, Sherman D, Baret PV, Dujon B, Goffeau A: **Novel transporters from Hemiascomycete yeasts.** *J Mol Microbiol Biotechnol* 2003, **6**:19-28.
23. Blank LM, Lehmebeck F, Sauer U: **Metabolic-flux and network analysis in fourteen hemiascomycetous yeasts.** *FEMS Yeast Res* 2005, **5**:545-558.
24. Richard G-F, Kerrest A, Lafontaine I, Dujon B: **Comparative genomics in Hemiascomycete yeasts: genes involved in DNA replication, repair and recombination.** *Mol Biol Evol* 2005, **22**:1011-1023.
25. Fabre E, Muller H, Therizols P, Lafontaine I, Dujon B, Fairhead C: **Comparative genomics in hemiascomycete yeasts: evolution of sex, silencing and subtelomeres.** *Mol Biol Evol* 2005, **22**:856-873.
26. Butler G, Kenny C, Fagan A, Kurischko C, Gaillardin C, Wolfe KH: **Evolution of the MAT locus and its Ho endonuclease in yeast species.** *Proc Natl Acad Sci USA* 2004, **101**:1632-1637.
27. Logue ME, Wong S, Wolfe KE, Butler G: **A genome sequence survey shows that the pathogenic yeast *Candida parapsilosis* has a defective *MTLa1* allele at its mating type locus.** *Eukaryot Cell* 2005, **4**:1009-1017.
28. Fairhead C, Dujon B: **Structure of *K. lactis* subtelomeres: duplications and gene content.** *FEM Yeast Res* 2005, in press.
29. Wolfe KH, Shields DC: **Molecular evidence for an ancient duplication of the entire yeast genome.** *Nature* 1997, **387**:708-713.
30. Andalis AA, Storchova Z, Styles C, Galitski T, Pellman D, Fink GR: **Defects arising from whole-genome duplications in *Saccharomyces cerevisiae*.** *Genetics* 2004, **167**:1109-1121.
31. Greig D, Louis EL, Borts RH, Travisano M: **Hybrid speciation in experimental populations of yeast.** *Science* 2002, **298**:1773-1775.
32. Lafontaine I, Fischer G, Talla E, Dujon B: **Gene relics in the genome of the yeast *Saccharomyces cerevisiae*.** *Gene* 2004, **335**:1-17.
33. Gu X, Zhang Z, Huang W: **Rapid evolution of expression and regulatory divergences after yeast gene duplication.** *Proc Natl Acad Sci USA* 2005, **102**:707-712.
34. Gao L-Z, Innan H: **Very low gene duplication rate in the yeast genome.** *Science* 2004, **306**:1367-1370.
Using a method distinct from the classical molecular clock, the authors show that the gene duplication rates in yeast are two orders of magnitude lower than was previously suggested.
35. Sugino RP, Innan H: **Estimating the time to the whole genome duplication and the duration of concerted evolution via gene conversion in yeast.** *Genetics* 2005, DOI:10.1534/genetics.105.043869.
36. Koszul R, Caburet S, Dujon B, Fischer G: **Eucaryotic genome evolution through the spontaneous duplication of large chromosomal segments.** *EMBO J* 2004, **23**:234-243.
This work demonstrates the spontaneous formation of segmental duplications in *S. cerevisiae* and discusses the impact of this mechanism on genome evolution.
37. Leh-Louis V, Wirth B, Potier S, Souciet JL, Despons L: **Expansion and contraction of the *DUP240* multigene family in *Saccharomyces cerevisiae* populations.** *Genetics* 2004, **167**:1611-1619.

38. Schacherer J, Tourette Y, Souciet J-L, Potier S, de Montigny J:
 • **Recovery of a function involving gene duplication by retroposition in *Saccharomyces cerevisiae*.** *Genome Res* 2004, **14**:1291-1297.
 The authors report a direct experimental demonstration of the mechanism of single-gene duplication by transposon-mediated retroposition.
39. Pereira-Leal JB, Teichmann SA: **Novel specificities emerge by stepwise duplication of functional modules.** *Genome Res* 2005, **15**:552-559.
 Using functionally characterized protein complexes from *S. cerevisiae*, the authors demonstrate that they evolved by step-wise partial duplications. They illustrate how module duplication is associated with functional specialization.
40. Gojkovic Z, Knecht W, Zameitat E, Warneboldt J, Coutelis J-B, Pynyaha Y, Neuveglise C, Moller K, Löffler M, Piskur J: **Horizontal gene transfer promoted evolution of the ability to propagate under anaerobic conditions in yeasts.** *Mol Genet Genomics* 2004, **271**:387-393.
41. Hall C, Brachat S, Dietrich FS: **Contribution of horizontal gene transfer to the evolution of *Saccharomyces cerevisiae*.** *Eukaryot Cell* 2005, **4**:1102-1115.
42. Ihmels J, Bergman S, Gerami-Nejad M, Yanai I, McClellan M, Berman J, Barkai N: **Rewiring of the yeast transcriptional network through the evolution of motif usage.** *Science* 2005, **309**:938-940.
 The authors describe the large-scale modulation of the transcription programs between two distantly related yeasts and show the fundamental differences between a strict aerobe and a facultative anaerobe. This study demonstrates that the changes in gene expression are connected with the loss of numerous *cis*-regulatory elements following the apparent whole-genome duplication event.
43. Rokas A, Williams BL, King N, Carroll SB: **Genome-scale approaches to resolving incongruence in molecular phylogenies.** *Nature* 2003, **425**:798-804.
44. Boekhout T: **Gut feeling for yeasts.** *Nature* 2005, **434**:449-450.
45. Wood V, Gwilliam R, Rajandream MA, Lyne M, Lyne R, Stewart A, Sgouros J, Peat N, Hayles J, Baker S *et al.*: **The genome sequence of *Schizosaccharomyces pombe*.** *Nature* 2002, **415**:871-880.
46. Loftus BJ, Fung E, Roncaglia P, Rowley D, Amadeo P, Bruno D, Vamathevan J, Miranda M, Anderson IJ, Fraser JA *et al.*: **The genome of the basidiomycetous yeast and human pathogen *Cryptococcus neoformans*.** *Science* 2005, **307**:1321-1324.
47. Kurtzman CP: **Phylogenetic circumscription of *Saccharomyces*, *Kluveromyces* and other members of the *Saccharomycetaceae*, and the proposal of the new genera *Lachancea*, *Nakaseomyces*, *Naumovia*, *Vanderwaltozyma* and *Zygorulaspota*.** *FEMS Yeast Res* 2003, **4**:233-245.
48. Diezmann S, Cox CJ, Schönián G, Vilgalys R, Mitchell TG: **Phylogeny and evolution of medical species of *Candida* and related taxa: a multigenic analysis.** *J Clin Microbiol* 2004, **42**:5624-5635.
49. Nakao Y, Kodama Y, Nakamura N, Ito T, Hattori M, Shiba T, Ashikari T: **Whole genome sequence of a lager brewing yeast.** In *Proceedings of the 29th European Brewery Convention Congress*: 2003 May 17-22; Dublin:524-530.
50. Ramezani-Rad M, Hollenberg CP, Lauber J, Wedler H, Griess E, Wagner C, Albermann K, Hani J, Piontek M, Dahlems U, Gellisen G: **The *Hansenula polymorpha* (strain CBS4732) genome sequencing and analysis.** *FEMS Yeast Res*. 2003, **4**:207-215.

Five things you might not know about Elsevier

1.

Elsevier is a founder member of the WHO's HINARI and AGORA initiatives, which enable the world's poorest countries to gain free access to scientific literature. More than 1000 journals, including the *Trends* and *Current Opinion* collections, will be available for free or at significantly reduced prices.

2.

The online archive of Elsevier's premier Cell Press journal collection will become freely available from January 2005. Free access to the recent archive, including *Cell*, *Neuron*, *Immunity* and *Current Biology*, will be available on both ScienceDirect and the Cell Press journal sites 12 months after articles are first published.

3.

Have you contributed to an Elsevier journal, book or series? Did you know that all our authors are entitled to a 30% discount on books and stand-alone CDs when ordered directly from us? For more information, call our sales offices:

+1 800 7 8 2 4927 (US) or +1 800 460 3110 (Canada, South & Central America)
 or +44 1865 474 010 (rest of the world)

4.

Elsevier has a long tradition of liberal copyright policies and for many years has permitted both the posting of preprints on public servers and the posting of final papers on internal servers. Now, Elsevier has extended its author posting policy to allow authors to freely post the final text version of their papers on both their personal websites and institutional repositories or websites.

5.

The Elsevier Foundation is a knowledge-centered foundation making grants and contributions throughout the world. A reflection of our culturally rich global organization, the Foundation has funded, for example, the setting up of a video library to educate for children in Philadelphia, provided storybooks to children in Cape Town, sponsored the creation of the Stanley L. Robbins Visiting Professorship at Brigham and Women's Hospital and given funding to the 3rd International Conference on Children's Health and the Environment.



ELSEVIER

Transposable elements, gene creation and genome rearrangement in flowering plants

Jeffrey L Bennetzen

Plant genome structure is largely derived from the differing specificities, abundances and activities of transposable elements. Recent studies indicate that both the amplification and the removal of transposons are rapid processes in plants, accounting for the general lack of intergenic homology between species that last shared a common ancestor more than 10 million years ago. Two newly discovered transposon varieties, *Helitrons* and Pack-MULEs, acquire and fuse fragments of plant genes, creating the raw material for the evolution of new genes and new genetic functions. Many of these recently assembled, chimeric gene-candidates are expressed, suggesting that some might escape epigenetic silencing and mutational decay, but a proven case of gene creation by any transposable element activity in plants remains to be demonstrated.

Addresses

Department of Genetics, University of Georgia, Athens, GA 40602-7223, USA

Corresponding author: Bennetzen, Jeffrey L (maize@uga.edu)

Current Opinion in Genetics & Development 2005, **15**:621-627

This review comes from a themed issue on Genomes and evolution Edited by Stephen J O'Brien and Claire M Fraser

Available online 10th October 2005

0959-437X/\$ - see front matter

© 2005 Elsevier Ltd. All rights reserved.

DOI 10.1016/j.gde.2005.09.010

Introduction

The use of DNA markers to generate comparable genetic maps [1-3] led to the suggestion that the nuclear genomes of flowering plants, the angiosperms, were highly similar in the number of genes they contained and in their colinear order on the chromosomes [4]. This apparent genomic colinearity was observed despite more than 800-fold variation in genome size across the angiosperms [5].

Subsequent investigations of local genome structure in plants, primarily by the sequencing of nuclear DNA inserts within bacterial artificial chromosome (BAC) vectors, indicated that most variation in the size of nuclear genomes was caused by the differential amplification and/or retention of retrotransposons that contain long terminal repeats (LTRs) [6,7]. In large genome species such as maize, barley or wheat, genes were commonly found in

small islands surrounded by seas of LTR-retrotransposons [6,8,9]. At this scale of comparison, orthologous genes often exhibited extensive colinearity, otherwise known as microcolinearity. However, the transposable elements and other sequences within the intergenic regions were usually completely different, even in comparisons between species such as sorghum and maize or wheat and barley that had last shared common ancestors less than 15 million years ago (mya) [10,11]. In addition, a few non-colinear genes were found in most orthologous genome segments from different plant species, including in those instances in which very closely related species such as maize and sorghum or *Arabidopsis thaliana* and *Capsella rubella* were compared [10,12]. Most of these comparative analyses were conducted in the cereals, especially maize, sorghum and rice. Even with conservative criteria for gene identification, approximately 35% of genes appeared to have moved to new locations in the ~12 million years since maize and sorghum diverged from a common ancestor [13,14]. This instability contrasts dramatically with the higher degree of conservation in gene order and content seen in mammals, for instance between mouse and human lineages over ~80 million years of independent evolutionary descent [15].

Most recently, exceptional frequencies of intraspecific gene rearrangement were described within rice and maize [16-19,20]. The conclusions that many rice genes [16,17] and >30% of maize genes [20] were non-colinear even within the species was, seemingly, largely incompatible with the observation that any colinearity could be conserved in more distant comparisons. However, further analyses indicated that most or all of the non-colinear rice genes were transposable elements that had been misannotated as genes [17,21,22], and that most of the non-colinear sequences that had been annotated as genes in maize were actually gene fragments [20]. In this review, I describe recent discoveries that explain the origin of the non-colinearities observed in the genomes of flowering plants. These observations suggest aggressive processes for gene creation in the angiosperms, which compete with persistent mechanisms for gene removal and genome shrinkage.

Adding and removing DNA sequences

As with all other eukaryotes, plants have increased the DNA content of their nuclear genomes by polyploidy, segmental duplication and transposon amplification. For reasons unknown, the rates of at least some of these processes are exceptionally high in plants. Essentially 100% of flowering plants are current polyploids or can

Glossary

Gene models: The genes and their structures predicted by 'gene-finding' computer programs. They are sometimes confirmed by cDNA sequence or peptide sequence data.

Indels: Term used to describe a difference in DNA sequence content in a region if one does not know whether it is caused by an insertion in one chromosomal segment or a deletion in the other orthologous chromosomal region.

Paleopolyploid: A polyploid that was generated in ancient times, usually many millions of year before, and that no longer shows the chromosome pairing or segregation properties of its polyploid origins, but does show genome-wide duplication of chromosomal segments, often heavily rearranged.

Solo-LTR: The single long terminal repeat that remains after unequal intra-strand recombination between the two LTRs in a single LTR-retrotransposon. The other product of this unequal recombination is a LTR-retrotransposon circle that contains only one LTR and is usually degraded or diluted by subsequent mitoses (i.e. it is lost from the genome).

be traced to one or more 'paleopolyploid' (see Glossary) events within the past 200 million years [23]. Comparative genetic maps indicate segmental duplications, some comprising whole chromosomes or chromosome arms [24], and BAC sequence analyses indicate a wealth of tandem gene families. In all angiosperms with haploid genomes larger than 2000 Mb that have been investigated to date, more than 50% of the nuclear DNA has been found to consist of LTR-retrotransposons and other repeats [25]. Given that the mean 1C angiosperm genome size is about 5600 Mb (IJ Leitch, unpublished), it can be concluded that the majority of plant DNA on the planet is composed of transposable elements. These great abundances and recent genome size contributions of transposable elements are also indicated by the fact that most are recent insertions. For instance, more than 80% of the intact LTR-retrotransposons in all analyzed angiosperms can be dated as insertions that occurred within the past five million years [25–28].

This exceptional rate of growth in the DNA content of plant genomes is in competition with a very high rate of sequence removal. Unequal homologous recombination, commonly by intra-chromatid events, can convert an intact LTR-retrotransposon into a solo-LTR (see Glossary) [6,8,21*,27,28] or can remove the chromosomal sequences between two LTR-retrotransposons of the same family [27,28]. Most DNA sequence removal in *Arabidopsis*, rice and wheat, the only species investigated at this level, appears to be associated with tiny deletions that can be ascribed to illegitimate recombination [21*,27–29]. The precise mechanism(s) of illegitimate recombination are not known in these cases, but deletion associated with the repair of double-strand breaks [30] is a likely candidate.

Although the most common size of deletions in flowering plants — at least, as measured in the rice genome — is 1–2 bp [21*], these and larger deletions can lead to a very effective removal of all classes of DNA that are not

retained by natural selection. For instance, the half-life of LTR-retrotransposon sequences in rice is less than three million years and can be associated with removal of >194 Mb of LTR-retrotransposon DNA in the past eight million years [21*,28]. Despite this, the genomes of the two rice subspecies *japonica* and *indica* have grown >2% over the past few hundred thousand years, primarily by a frequency of LTR-retrotransposon amplification that has outstripped the progressive removal of DNA by small deletions [21*].

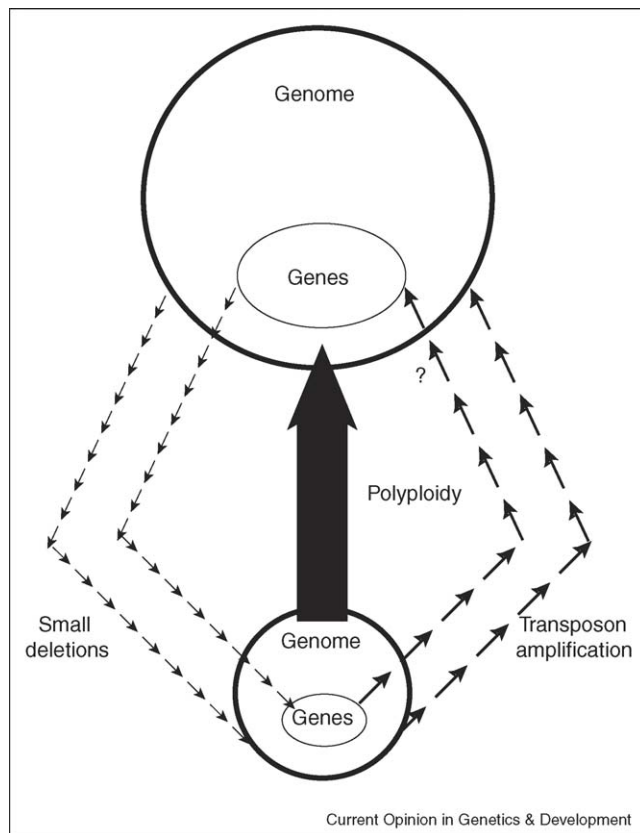
A simple explanation for the great variation in angiosperm genome size is either that lineages will differ in the frequency of genome growth, for instance, as a result of rare polyploidy and/or episodic transposon bursts, or that they might differ in the qualitative and/or quantitative properties of DNA sequence removal (Figure 1). For instance, the repair of double-strand breaks in the small genome of *A. thaliana* is accompanied by fewer insertions and a larger average size of deletions than in the larger genome of *Nicotiana tabacum* [30]. It is also possible that the activities of these competing genome-growth and -shrinkage mechanisms could be significantly influenced by dramatic changes in the internal or external environment, as indicated by the activation of transposable elements by genome stress (e.g. chromosome breakage) or the induction of sequence removal processes by the establishment of a *de novo* polyploid state [31–33].

Gene creation

Most closely related organisms share the same types of genes, although gene copy numbers and regulation can vary over short times of evolutionary divergence. Even apparently 'novel' or 'orphan' genes can often be traced to extensive primary sequence divergence from a clear ancestral gene [34,35]. Hence, most truly novel genes were probably created, perhaps from raw genomic sequence, hundreds of millions of years ago. The more recent creation of chimeric genes with novel genetic functions has been proposed by the process of 'exon shuffling', wherein fragments of genes are fused together, partly relying on the fact that chimeric introns would often be processed to yield intact exons in a final mRNA product [36].

Recent discoveries in flowering plants suggest a very high rate of gene creation by transposon capture and exon shuffling [37**,38*,39*,40**]. Early studies in maize had shown that transposons could acquire specific genic sequences and amplify them across the genome [41–43]. In the case of the *Bs1* LTR-retrotransposon of maize, a portion of a plasma membrane proton ATPase gene had been acquired — presumably at the RNA level, because all introns were missing — whereas subsequent divergence had selectively retained the transmembrane domains [41]. Although maize *Bs1* was the first LTR-retrotransposon observed, in any organism, to acquire a

Figure 1



A schematic representation of the processes that generate qualitative and quantitative variation in gene number and overall DNA content in plant nuclear genomes. Small arrows indicate minor events, and large arrows indicate large events. Segmental duplication can also be a factor that increases gene number and genome size, but on a scale less dramatic than polyploidization. The question mark indicates that it is not yet clear to what degree transposons like *Helitrons* and Pack-MULEs can actually create new genes, rather than just assemble chimeric structures that look like new genes or duplicate current genes.

gene fragment, numerous instances of the acquisition of genes by a closely related mobile DNA, the retroviruses of animals, have been described [44].

Mutator-like DNA elements (MULEs) had also been seen to acquire genic sequences in both maize and *Arabidopsis* [43,45]. However, Jiang and *et al.* [37**] found that MULE capture of gene fragments was not an interesting oddity but, rather, a major feature of rice genome structure. More than 3000 MULEs containing fragments of genes were found in rice genome sequence and were named 'Pack-MULEs'. The small genomic fragments (averaging 325 bp, with a range of 47–986 bp) had been acquired at the DNA level, with both exons and introns present, but various rearrangements were sometimes identified within the acquired sequence. These rearrangements could have occurred either during or after the acquisition process. Gene fragments from different rice

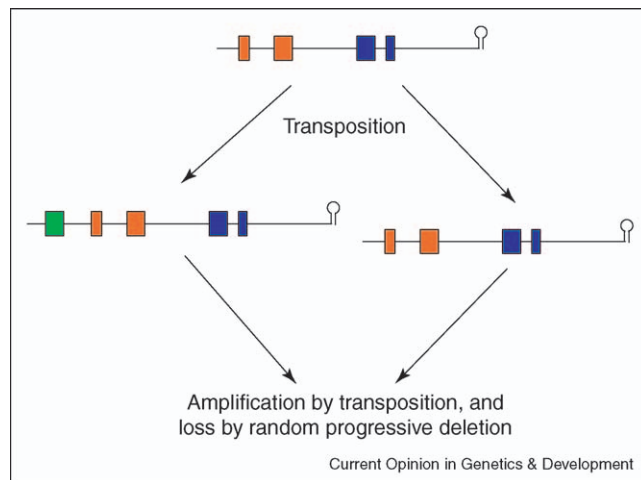
genes were found together in ~23% of Pack-MULEs, and at least 5% of Pack-MULEs were found to be expressed, as evidenced by full-length cDNAs with an identical DNA sequence match. More than 90% of these expressed Pack-MULEs appear to have been transcriptionally initiated within the element itself [37**]. Hence, by the criterion of expression at the RNA level, many of these Pack-MULEs are already new genes.

A mechanism for gene fragment acquisition by Pack-MULEs has not been proven, but ectopic gene-conversion across a nicked cruciform structure [46] has been proposed. Although MULEs are abundant components of most or all angiosperm genomes, the frequency of Pack-MULEs has not been calculated in any species other than rice [37**]. However, preliminary analysis of genomic sequence data from *Lotus japonicum*, a distant angiosperm relative of rice, indicates a similar abundance of Pack-MULEs (J Jiang and SR Wessler, unpublished), thereby suggesting that they will be major components of most or all flowering plant genomes.

Research in maize has shown the presence of numerous clusters of gene fragments, with all fragments in the same predicted transcriptional and translational orientation [47]. Lal *et al.* [48] found that one such cluster was, in fact, a *Helitron* that had recently inserted into the maize *Sh2* gene to cause an inactivational mutation. *Helitrons* are a new class of eukaryotic transposable element, initially discovered through database analyses in *A. thaliana*, rice and *Caenorhabditis elegans* [49], and now found in insects, vertebrates and fungi as well [50,51]. *Helitrons*, unlike other transposons in eukaryotes, do not have terminal repeats and do not cause duplications of target site DNA, so their detection can be quite challenging. *Helitrons* often contain open reading frames (ORFs) that are predicted to encode a protein with replication initiator and helicase activities associated with rolling-circle replication of bacterial transposons [52], plus an RPA (replication protein A)-like protein that could also provide a singled-stranded DNA binding activity needed for DNA replication. Hence, rolling-circle replication seems a likely model for *Helitron* transposition [49]. The 3' ends of *Helitrons* contain a region that could form a 12–18 bp hairpin (with a 2–4 bp unpaired loop) that is followed 5–8 bp downstream by the consensus sequence CTRR at the end of the element, whereas the 5' end contains the terminal sequence TC. These conserved components might be the only *cis* sequences required for *Helitron* transposition. *Helitron* insertions are inserted within the target sequence AT, so that the 5' end is always A/TC and the 3' end is always CTRR/T.

The mechanism of gene sequence acquisition by *Helitrons* is not known, although it appears that gene fragments are incorporated progressively at several possible locations, including the 5' end or near the 3' end

Figure 2



The structure of a parental non-autonomous *Helitron* (top) and two *Helitrons* derived from it by transposition (bottom). The boxes indicate exons, and the colors indicate different genes that donated these exons and the intervening introns. The *Helitron* on the lower left has acquired a new gene fragment. The stem-loop structure at the right (3') end of each *Helitron* is a conserved structure in these elements, but it is drawn much larger than scale.

(Figure 2) [53^{*}]. Identical copies of some *Helitrons* can be found at more than one genomic location, indicating that transposition does not require element rearrangement. No size limit for sequence acquisition is known, but most insertions are relatively small, such that larger *Helitrons* are commonly composed of small insertions from multiple genes rather than a few large genic insertions. For instance, the 17.7 kb *Helitron* insertion in maize *Sh2* contains fragments of 12 identified genes [48]. A particularly interesting aspect of new sequence acquisition is the observation that gene fragments are usually in the same predicted transcriptional and translational orientation as seen for the predicted helicase and RPA-like ORFs, even in defective elements that lack these ORFs.

As seen with Pack-MULEs, transcripts have been observed that have homology to *Helitrons*, including those with multiple gene fragment insertions [40^{**},53^{*}]. These transcripts are mostly spliced correctly to remove introns, thus fusing exons from different genes into a single transcript. Hence, in a manner similar to Pack-MULEs, *Helitrons* in maize are producing chimeric transcripts that could lead to the creation of new genes.

The rapid loss of recently created genes

Both Pack-MULEs and *Helitrons* appear to be abundant in many plant genomes, but the copy numbers of individual elements are quite low within any single genome. For instance, although there were >3000 Pack-MULEs found in rice [37^{**}], most individual elements had copy numbers of less than five. This is in stark contrast to the

more abundant plant LTR-retrotransposons that can have copy numbers exceeding 10 000 per genome [7,25]. This observation suggests that individual Pack-MULEs and *Helitrons* are quickly silenced [54] before they can amplify to high abundances within the nucleus. Given their structure and frequent expression, many of these elements could produce chimeric RNAs and fused peptides that would interact with all of those genes and gene products from which they have borrowed. Hence, it is likely that many chimeric *Helitron* or Pack-MULE RNAs will not only induce their own epigenetic regulation (e.g. silencing) but also contribute to the epigenetic regulation of the intact genes that have donated gene fragments to the element [40^{**}]. If translated, chimeric Pack-MULE or *Helitron* peptides might also alter cellular enzymology and/or regulation, for instance by serving as dominant-negative inhibitors through the poisoning of multiprotein complexes. In cases in which these chimeric RNAs or proteins affect phenotype, selection for or against their expression will become a significant factor in their retention or removal, respectively.

If the recently assembled chimeric sequences inside *Helitrons* or Pack-MULEs do not provide a trait beneficial to the host plant, point mutations and indels (see Glossary) will eventually occur within *cis*-essential components of the elements, thereby making permanent an inactivation that might have been initially epigenetic. As with all other sequences in plant genomes, Pack-MULEs and *Helitrons* will be exposed to the persistent processes of sequence removal, which are primarily associated with the small deletions created by illegitimate recombination [21^{*},27–29]. Hence, most potential chimeric genes created by Pack-MULEs or *Helitrons* will be lost within a few million years.

Those rare exceptions in which a chimeric gene survives over a long time period, and is thus shared in a conserved state by descendant genomes, will provide compelling evidence for a significant contribution of these transposons to gene creation and resultant biological diversification. The gene fragment inserted into *Bs1*, and some (>10%) of the predicted chimeric genes within Pack-MULEs exhibit the DNA sequence characteristics associated with selection for a conserved protein function [37^{**},41]; however, this selection could be for element function (e.g. more effective transposition) and not for any host biological process. Hence, convincing evidence for retained function at the sequence level in any chimeric gene would need to be manifested in Pack-MULE- or *Helitron*-derived genes that had lost their mobility (e.g. by terminal deletions). Equally convincing proof for the creation of a new gene by Pack-MULE- or *Helitron*-mediated exon-shuffling would be the identification of a mutant phenotype in the plant by a mutation (e.g. inactivation) in any element-derived chimeric gene. Neither of these forms of evidence has yet been described for any predicted new gene created by a *Helitron* or Pack-

MULE, at least partly because such statistical and experimental analyses have not been pursued to any comprehensive degree in any plant species.

Gene and genome rearrangement

The hyper-variability of the non-genic sequences that make up the majority of angiosperm genomic DNA is primarily because most of these are mobile sequences, and because plants very rapidly remove nuclear DNA that is not retained by natural selection (Figure 1). Some transposable elements also stimulate other types of genome rearrangement, including inversion, duplication or deletion of adjacent DNA, by chromosome-breaking, by aborted transposition, or by ectopic recombination between homologous transposable elements at different chromosomal locations. Hence, genome structure in any organism is, largely, the outcome of transposable element action and of the cellular processes that act on transposons.

In plants, most genes appear to retain similar or identical function despite their very different chromosomal environments in different plant species. For instance, the *adh1* gene and one adjacent gene of unknown function were moved to a new chromosome in a common ancestor of maize and sorghum [13], and thus *adh1* is found in a non-syntenic location when compared with its location in more distantly related grasses like rice or wheat. Despite this movement, plus the subsequent deletion of the adjacent mobilized gene, and the accumulation of surrounding seas of LTR-retrotransposons, the tissue specificity, induction profile and developmental timing of *adh1* gene expression appear to be unchanged. This routine observation suggests that plant genes are very well insulated from their surrounding chromosomal environments. Hence, movement of an intact gene to a new chromosomal location will often yield a functional locus, thereby permitting its retention under natural selection.

How often do transposable elements mediate these gene movements? Although gross chromosomal rearrangements are commonly traced to the action of transposable elements in maize, for instance, the more common, single-gene rearrangements observed in comparisons of different plant species [13,14] are not yet associated with any proven molecular mechanism. Reciprocal tandem duplications and/or deletions, and small inversions, are likely to be caused by unequal recombination, but it is not clear whether gene movement from chromosome to chromosome often occurs by this same process in plants. Given the random loss of unselected DNA in plants, the sequence hallmarks for a transposable element vector in gene movement would be rapidly lost, and thus only observed in very recent gene-rearrangements [11].

Conclusions

The frequent acquisition of gene fragments by Pack-MULEs and *Helitrons* suggests a possible role for these

elements in the redistribution of genes across the genome. However, in both element types, the fragments acquired are significantly smaller than most intact genes. It is possible that rare fragment-acquisitions will include a complete gene, but no such case has yet been found. Also significant, at an analytical level, is the fact that predicted chimeric gene fusion products within Pack-MULEs or *Helitrons* will be mis-annotated as genes in assessments of microcolinearity [14,16–19], thereby predicting less genic colinearity than truly exists. Hence, reassessments of microcolinearity in plants are warranted now that the presence and properties of Pack-MULEs and *Helitrons* have been demonstrated.

Future studies will investigate the frequency with which plant mobile DNAs, especially Pack-MULEs and *Helitrons*, have contributed to the genetic repertoire and variable arrangement of angiosperm genomes. It is astounding that so much interesting structural novelty has been discovered in plants genomes, given that many fewer plant genomes have been subjected to comprehensive sequence analysis than have genomes from the prokaryotic, fungal and animal kingdoms. As more plant genomes are sequenced, more raw material for genome analysis will be generated, and a wealth of unexpected outcomes can be predicted.

Acknowledgements

The author thanks Ning Jiang and Susan Wessler for providing unpublished information, and Katrien Devos for assistance with the production of the manuscript. The preparation of this article was supported by National Science Foundation grant DBI031724.

References and recommended reading

Papers of particular interest, republished within the annual period of review, have been highlighted as:

- of special interest
 - of outstanding interest
1. Bonierbale MW, Plaisted RL, Tanksley SD: **RFLP maps based on a common set of clones reveals modes of chromosomal evolution in potato and tomato.** *Genetics* 1988, **120**:1287-1292.
 2. Hulbert SH, Richter TE, Axtell JD, Bennetzen JL: **Genetic mapping and characterization of sorghum and related crops by means of maize DNA probes.** *Proc Natl Acad Sci USA* 1990, **87**:4251-4255.
 3. Moore G, Devos KM, Wang Z, Gale MD: **Grasses, line up and form a circle.** *Curr Biol* 1995, **5**:737-739.
 4. Bennetzen JL, Freeling M: **Grasses as a single genetic system: genome composition, collinearity and compatibility.** *Trends Genet* 1993, **9**:259-261.
 5. Bennett MD, Leitch IJ: **Nuclear DNA amounts in angiosperms.** *Ann Bot (Lond)* 1995, **76**:113-176.
 6. SanMiguel P, Tikhonov A, Jin YK, Motchoulskaia N, Zakharov D, Melake-Berhan A, Springer PS, Edwards KJ, Lee M, Avramova Z, Bennetzen JL: **Nested retrotransposons in the intergenic regions of the maize genome.** *Science* 1996, **274**:765-768.
 7. Bennetzen JL: **Mechanisms and rates of genome expansion and contraction in flowering plants.** *Genetica* 2002, **115**:29-36.
 8. Shirasu K, Schulman AH, Lahaye T, Schulze-Lefert T: **A contiguous 66-kb barley DNA sequence provides evidence for reversible genome expansion.** *Genome Res* 2000, **10**:908-915.

9. Wicker T, Stein N, Albar L, Feuillet C, Schlagenhauf E, Keller B: **Analysis of a contiguous 211 kb sequence in diploid wheat (*Triticum monococcum* L.) reveals multiple mechanisms of genome evolution.** *Plant J* 2001, **26**:307-316.
10. Tikhonov AP, SanMiguel PJ, Nakajima Y, Gorenstein NM, Bennetzen JL, Avramova Z: **Colinearity and its exceptions in orthologous *adh* regions of maize and sorghum.** *Proc Natl Acad Sci USA* 1999, **96**:7409-7414.
11. Ramakrishna W, Dubcovsky J, Park YJ, Busso C, Emberton J, SanMiguel P, Bennetzen JL: **Different types and rates of genome evolution detected by comparative sequence analysis of orthologous segments from four cereal genomes.** *Genetics* 2002, **162**:1389-1400.
12. Boivin K, Acarkan A, Mbulu R-S, Clarenz O, Schmidt R: **The *Arabidopsis* genome sequence as a tool for genome analysis in Brassicaceae. A comparison of the *Arabidopsis* and *Capsella rubella* genomes.** *Plant Physiol* 2004, **135**:735-744.
13. Ilic K, SanMiguel PJ, Bennetzen JL: **A complex history of rearrangement in an orthologous region of the maize, sorghum and rice genomes.** *Proc Natl Acad Sci USA* 2003, **100**:12265-12270.
14. Lai J, Ma J, Swigonova Z, Ramakrishna W, Linton E, Llaca V, Tanyolac B, Park YJ, Jeong OY, Bennetzen JL, Messing J: **Gene loss and movement in the maize genome.** *Genome Res* 2004, **14**:1924-1931.
15. Mural RJ, Adams MD, Myers EW, Smith HO, Miklos GL, Wides R, Halpern A, Li PW, Sutton GG, Nadeau J *et al.*: **A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome.** *Science* 2002, **296**:1661-1671.
16. Feng Q, Zhang Y, Hao P, Wang S, Fu G, Huang Y, Li Y, Zhu J, Liu Y, Hu X *et al.*: **Sequence and analysis of rice chromosome 4.** *Nature* 2002, **420**:316-320.
17. Han B, Xue Y: **Genome-wide intraspecific DNA-sequence variations in rice.** *Curr Opin Plant Biol* 2003, **6**:134-138.
18. Fu H, Dooner HK: **Intraspecific violation of genetic colinearity and its implications in maize.** *Proc Natl Acad Sci USA* 2002, **99**:9573-9578.
19. Song R, Messing J: **Gene expression of a gene family in maize based on noncolinear haplotypes.** *Proc Natl Acad Sci USA* 2003, **100**:9055-9060.
20. Brunner S, Fengler K, Morgante M, Tingey S, Rafalski A:
 • **Evolution of DNA sequence nonhomologies among maize inbreds.** *Plant Cell* 2005, **17**:343-360.
 The authors compared BAC sequences from orthologous regions in two maize inbreds, Mo17 and B73, covering more than 2.8 Mb of DNA on four chromosomes. The data indicated that more than one-third of the sequences annotated as genes were absent in one of the two inbreds at the locations sequenced. The authors found that these 'missing genes' were usually gene fragments found in clusters that were oriented in the same direction.
21. Ma J, Bennetzen JL: **Recent rapid growth and divergence of the rice nuclear genome.** *Proc Natl Acad Sci USA* 2004, **101**:12404-12410.
 The authors used the African rice *Oryza glaberrima* as an outgroup in comparisons of orthologous DNA from two *Oryza sativa* subspecies, *japonica* and *indica*. Across more than 1.2 Mb of sequence among several chromosomal segments, deletions were found to outnumber insertions, but the much larger average size of insertions was seen to have led to genome growth of >2% and >6% for these respective *indica* and *japonica* varieties since their divergence from a shared ancestor. The manuscript details the natures, sizes, origins, frequencies and distributions of indels across this portion of the rice genome, and also demonstrates that the molecular clock for LTR-retrotransposons runs at least twice as fast as the rate of synonymous substitutions in genes.
22. Bennetzen JL, Coleman C, Liu C, Ma J, Ramakrishna W: **Consistent over-estimation of gene number in complex plant genomes.** *Curr Opin Plant Biol* 2004, **7**:732-736.
23. Adams KL, Wendel JF: **Polyploidy and genome evolution in plants.** *Curr Opin Plant Biol* 2005, **8**:135-141.
24. Nagamura Y, Inoue T, Antonio B, Shimano T, Kajiya H, Shomura A, Lin S, Kuboki Y, Harushima Y, Kurata N *et al.*: **Conservation of duplicated segments between rice chromosomes 11 and 12.** *Breeding Sci* 1995, **45**:373-376.
25. Bennetzen JL, Ma J, Devos KM: **Mechanisms of recent genome size variation in flowering plants.** *Ann Bot (Lond)* 2005, **95**:127-132.
26. SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL: **The paleontology of intergene retrotransposons in maize.** *Nat Genet* 1998, **20**:43-45.
27. Devos KM, Brown JK, Bennetzen JL: **Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*.** *Genome Res* 2002, **12**:1075-1079.
28. Ma J, Devos KM, Bennetzen JL: **Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice.** *Genome Res* 2004, **14**:860-869.
29. Wicker T, Yahiaoui N, Guyot R, Schlagenhauf E, Liu Z-D, Dubcovsky J, Keller B: **Rapid genome divergence at orthologous low molecular weight glutenin loci of the A and A^m genomes of wheat.** *Plant Cell* 2003, **15**:1186-1197.
30. Kirik A, Salomon S, Puchta H: **Species-specific double-strand break repair and genome evolution in plants.** *EMBO J* 2000, **19**:5562-5566.
31. McClintock B: **The significance of responses of the genome to challenge.** *Science* 1984, **226**:792-801.
32. Shaked H, Kashkush K, Ozkan H, Feldman M, Levy AA: **Sequence elimination and cytosine methylation are rapid and reproducible responses of the genome to wide hybridization and allopolyploidy in wheat.** *Plant Cell* 2001, **13**:1749-1759.
33. Madlung A, Comai L: **The effect of stress on genome regulation and structure.** *Ann Bot (Lond)* 2004, **94**:481-495.
34. Schmid KJ, Aquadro CF: **The evolutionary analysis of "orphans" from the *Drosophila* genome identifies rapidly diverging and incorrectly annotated genes.** *Genetics* 2001, **159**:589-598.
35. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES: **Sequencing and comparison of yeast species to identify genes and regulatory elements.** *Nature* 2003, **423**:241-254.
36. Gilbert W: **Why genes in pieces?** *Nature* 1978, **271**:501.
37. Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR: **Pack-MULE transposable elements mediate gene evolution in plants.** *Nature* 2004, **431**:569-573.
 The authors provide a careful and comprehensive analysis of the structure, frequency, distribution and properties of a class of transposable elements that they name Pack-MULEs. More than 3000 Pack-MULEs were predicted across the entire rice genome, with gene fragments from more than one thousand cellular genes. At least 5% of Pack-MULEs were found to be transcribed, and the fusion of fragments from different genes within Pack-MULEs, often with appropriate intron processing, was frequently observed.
38. Gupta S, Gallavotti A, Stryker GA, Schmidt RJ, Lal SK: **A novel class of Helitron-related transposable elements in maize contain portions of multiple pseudogenes.** *Plant Mol Biol* 2005, **57**:115-127.
 An insertion mutation in the maize *barren stalk1* gene was found to be caused by the insertion of a *Helitron*, and the authors also found that *Helitrons* containing different gene fragments are common in the maize genome.
39. Lai J, Li Y, Messing J, Dooner HK: **Gene movement by Helitron transposons contributes to the haplotype variability in maize.** *Proc Natl Acad Sci USA* 2005, **102**:9068-9073.
 The authors find that the apparent genic non-colinearity that they previously reported in the *bz1* region is caused by gene fragments acquired by *Helitrons*. This study also provides a rare example of an acquired gene fragment that is in the opposite orientation to most gene segments acquired by these and other *Helitrons*.
40. Morgante M, Brunner S, Pea G, Fengler K, Zuccolo A, Rafalski A: **Gene duplication and exon shuffling by helitron-like transposons generate intraspecific diversity in maize.** *Nat Genet* 2005, **37**:997-1002.
 Comparison of over 20 000 gene models (see Glossary) across B73 and Mo17 orthologous regions predicted that more than 20% are not shared.

Inspection of nine such non-colinear annotated genes found that eight were actually gene fragments within *Helitrons*. This result suggests that the maize genome will contain a minimum of several thousand *Helitrons* carrying fragments of cellular genes. Some, but not all, *Helitrons* were found to produce transcripts that fused fragments from different genes. Insertions of other transposable elements were found in some *Helitrons*, including one LTR-retrotransposon insertion dated to 2 mya, suggesting that these elements have been active over a very long time period.

41. Jin Y-K, Bennetzen JL: **Integration and nonrandom mutation of a plasma membrane proton ATPase gene fragment within the *Bs1* retroelement of maize.** *Plant Cell* 1994, **6**:1177-1186.
 42. Bureau T, White SE, Wessler SR: **Transduction of a cellular gene by a plant retroelement.** *Cell* 1994, **77**:479-480.
 43. Talbert LE, Chandler VL: **Characterization of a highly conserved sequence related to *Mutator* transposable elements in maize.** *Mol Biol Evol* 1988, **5**:519-529.
 44. Malik HS, Henikoff S, Eickbush TH: **Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses.** *Genome Res* 2000, **10**:1307-1318.
 45. Yu Z, Wright SI, Bureau TE: ***Mutator*-like elements in *Arabidopsis thaliana*: structure, diversity and evolution.** *Genetics* 2000, **156**:2019-2031.
 46. Bennetzen JL, Springer PS: **The generation of *Mutator* transposable element subfamilies in maize.** *Theor Appl Genet* 1994, **87**:657-667.
 47. Ramakrishna W, Emberton J, Ogden M, SanMiguel P, Bennetzen JL: **Sequence and physical map analyses of the maize *Rp1* complex uncovers numerous sites and unexpected mechanisms of local rearrangement.** *Plant Cell* 2002, **14**:3213-3223.
 48. Lal SK, Giroux MJ, Brendel V, Vallejos E, Hannah C: **The maize genome contains a *Helitron* insertion.** *Plant Cell* 2003, **15**:381-391.
 49. Kapitonov VV, Jurka J: **Rolling-circle transposons in eukaryotes.** *Proc Natl Acad Sci USA* 2001, **98**:8714-8719.
 50. Kapitonov VV, Jurka J: **Molecular paleontology of transposable elements in the *Drosophila melanogaster* genome.** *Proc Natl Acad Sci USA* 2003, **100**:6569-6574.
 51. Poulter RT, Goodwin TJ, Butler MI: **Vertebrate helitrons and other novel helitrons.** *Gene* 2003, **313**:201-212.
 52. Mahillon J, Chandler M: **Insertion sequences.** *Microbiol Mol Biol Rev* 1998, **62**:725-774.
 53. Brunner S, Pea G, Rafalski A: **Origins, genetic organization and transcription of a family of non-autonomous helitron elements in maize.** *Plant J* 2005, **43**:799-810.
- The authors describe a family of four closely related *Helitrons*. One element is found to be transcribed and produces a chimeric transcript that joins fragments from several predicted genes. This transcript was correctly spliced at most introns but employed alternative splicing at one normal intron and one chimeric intron. The structure of the elements in this family suggest how these non-autonomous *Helitrons* can accumulate both terminal and internal additional gene fragments by a serial acquisition process.
54. Meyer P: In *Annual Plant Reviews: Plant Epigenetics*. Edited by Meyer P. Oxford: Blackwell Publishing; 2005.



ELSEVIER

Conserved sequences and the evolution of gene regulatory signals

Mark D Adams

Studies of evolutionary conservation of gene regulatory signals have led to a paradox: extensive sequence similarity implies functional conservation in non-coding regions across mammalian species; however, this stands in contrast to our understanding of transcriptional regulatory sites composed of degenerate recognition sequences for transcription factors that can maintain functional equivalence despite considerable sequence divergence. The latter observation provides an explanation for the rapid evolution of new traits through the gain and loss of transcription factor binding sites that bring new genes under the control of an existing genetic regulatory network. The former observation might point to novel mechanisms of gene regulation and/or chromosome function that are currently unappreciated. Recent comparative genome analysis has highlighted extensive conserved sequences in mammalian genomes that are beginning to be functionally characterized.

Addresses

Department of Genetics, Center for Human Genetics, Center for Computational Genomics and Systems Biology, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, OH 44106, USA

Corresponding author: Adams, Mark D (markadams@case.edu)

Current Opinion in Genetics & Development 2005, **15**:628–633

This review comes from a themed issue on
Genomes and evolution
Edited by Stephen J O'Brien and Claire M Fraser

Available online 26th September 2005

0959-437X/\$ – see front matter
© 2005 Elsevier Ltd. All rights reserved.

DOI 10.1016/j.gde.2005.09.004

Introduction

Two fundamental features of genomes define the way in which heritable information is expressed to conduct an organism through development to a functioning adult: the complement of genes encoded in the genome; and the orderly pattern of expression of those genes. Considerable progress has been made in defining the protein-coding genes in the human genome, on the basis of finished genomic sequences [1], and automated [2,3] and manual [4] gene annotation programs. Mammalian genomes encode a remarkably consistent set of genes [5–7], with clear orthologs present for a large majority of these genes in human, mouse and rat. Where differences occur, they are primarily in paralogous duplications of certain gene families. Given that a similar set of genes exist in each

mammalian species, it has been proposed that regulatory differences might play an important role in defining the differences among the species [8].

The activity of a gene might be regulated at numerous stages of its expression, and each of these stages might be subject to evolutionary pressure. Mammalian gene regulatory mechanisms fall — currently — into one of three recognized categories: (i) regulation of transcription *in cis*, mediated by promoters and enhancers; (ii) regulation *in trans*, mediated by transcription factors binding to *cis* sites, RNA interference and microRNAs (miRNAs); and (iii) regulation on the basis of the modification state of the DNA (e.g. CpG methylation [9]) and how it is packaged (e.g. histone acetylation and methylation [10]). Of course, further modulation of gene activity occurs at the level of alternative splicing, RNA stability, translation efficiency, protein stability and protein activity, none of which will be considered here.

Study of the regulation of many genes over many years has led to the basic understanding of transcription initiation at core promoters, regulated by adjacent enhancers, silencers and insulators [11,12]. The promotion of transcription is not merely the result of a simple binary interaction of regulatory protein with DNA but the concerted effort of several factors that bind both DNA and one another [11–13]. The panoply of factors that control the amount, timing and location of expression are unknown for most genes. It is not currently possible to predict the set of genes regulated by a given transcription factor, let alone how amino acid substitutions in a transcription factor might affect its function, and we are far from being able to predict the pattern of expression of a gene simply from analysis of its flanking DNA sequence. Understanding the mechanisms of gene regulation, particularly during development, and how evolution of the pattern of gene regulation contributes to morphological and phenotypic differences among organisms are fundamentally important goals in the genome era. Comparative genomics is a powerful new tool to identify and characterize functional sequences using evolutionary analysis methods.

In this review, I discuss recent studies that address the importance of gene expression differences in species evolution and the potential role(s) of conserved non-coding sequences in gene regulation and genome function.

Gene expression as an evolutionary variable

Variation in patterns of gene expression can have evolutionary consequences. A particularly compelling example

of the impact of altered gene regulation comes from comparison of freshwater and marine populations of the threespine stickleback [14^{••}]. Reduced expression of the transcription factor gene *Pitx1* during development in freshwater threespine sticklebacks, compared with expression in marine populations, results in loss of the pelvic skeleton [14^{••}]. Although the molecular mechanism for the expression difference has not yet been deciphered, this provides a strong indication that regulatory changes can have a direct and rapid evolutionary impact. As another example, subtle variation in cardiac gene expression is associated with divergent metabolic output between groups of individuals of the killifish *Fundulus heteroclitus* [15]. Gene expression variation can affect quite complex social traits as well. Differences in expression of the vasopressin 1a receptor as a result of microsatellite polymorphism lead to marked differences in social behavior, including partner preference in related species of vole [16,17].

Studies of the evolution of gene regulatory patterns have been facilitated by the availability of whole-genome microarrays [18–20]. By treating gene expression levels as a quantitative trait in a segregating population, Schadt *et al.* [20] were able to map quantitative trait loci (QTLs) affecting the expression of more than 3000 mouse genes. The most statistically significant QTLs tended to be those that were coincident in location with the gene whose expression was monitored, suggesting that these QTLs were controlling expression through *cis* effects. By contrast, QTLs of more moderate significance tended to be using *trans* effects to mediate gene expression. Using a classic *cis-trans* test, Doss *et al.* [21] showed that a majority of the predicted *cis* QTLs are likely to be true positives, and nearly all map to regions that do not share single nucleotide polymorphism haplotypes [22,23].

Morley *et al.* [24[•]] used microarrays to determine gene expression levels of ~3500 human genes; these levels were treated as phenotypes that were genetically mapped in 14 large families, revealing over 1000 loci that contributed to the expression phenotype. Interestingly, a majority of these loci act *in trans* to the gene that is regulated. The fact that expression levels of genes can vary in populations and that these differences are heritable reinforces the idea that changes in expression level can have an evolutionary impact.

Distinguishing important functional differences in gene expression from benign or neutral changes is quite challenging [25^{••},26]. Khaitovich *et al.* [25^{••}] compared gene expression profiles in human, chimpanzee, orang-utan and macaque and concluded that expression differences have accumulated linearly with time, suggesting that global changes in gene expression patterns were not a major factor in primate evolution. The same group sub-

sequently introduced an evolutionary model of gene expression that attempts to distinguish neutral from accelerated changes and to improve the prediction of specific changes that have had evolutionary consequences [26].

Identifying functional sequences by comparative genomics approaches

The impact of evolution is most readily seen in the pattern of sequence conservation and variation among species. Sequencing of the mouse and rat genomes has revealed remarkable features of evolutionary conservation [5–7]. About 5% of each genome is under selective constraint, of which only about 1.5% is protein-coding. This conservation extends throughout mammalian species [27–30,31^{••}]. The importance of gene regulation is highlighted by the fact that two-thirds of the sequence conserved among mammals is not protein-coding. These 'conserved non-coding sequence elements' (CNEs) are largely unique in each genome (i.e. non-repetitive), and, to date, no readily recognizable clusters, classes or subdivisions have been defined that might be useful in further characterizing them.

Members of a subset of CNEs show extraordinary conservation. Bejarano *et al.* [32^{••}] described 481 ultra-conserved elements with perfect identity over more than 200 bp in human, mouse and rat. These elements are also highly conserved in non-mammalian vertebrates but not in invertebrates. Ultra-conserved elements tend to be located in the vicinity of genes important in developmental processes, suggesting that their exceedingly strong conservation might be owing to their having crucial roles in specifying the developmental program that is shared among vertebrates. Conversely, extensive conservation is not sufficient for defining functional importance: sequence comparison between mammals as divergent as human and mouse might fail to reveal shared functional sequences [33,34].

It is difficult to imagine what function of DNA requires strong conservation over scores of base pairs. Transcription factors and other DNA-binding proteins typically recognize only 6–12 bases and generally tolerate some degeneracy in recognition sequence. The information content of transcription factor binding sites is too low for computer algorithms to accurately predict which transcription factor binding site might be used *in vivo* [35]. Xie *et al.* [36[•]] have searched for short regulatory sequences in human promoters and 3' untranslated regions (UTRs) by examining 6–18-nucleotide motifs that are over-represented genome-wide in multiple alignments of human, mouse, rat and dog sequences. In promoter regions, most previously known transcription factor binding sites were identified, and over 100 new motifs were predicted to have *cis* regulatory activity.

Potential functions of conserved non-coding sequences

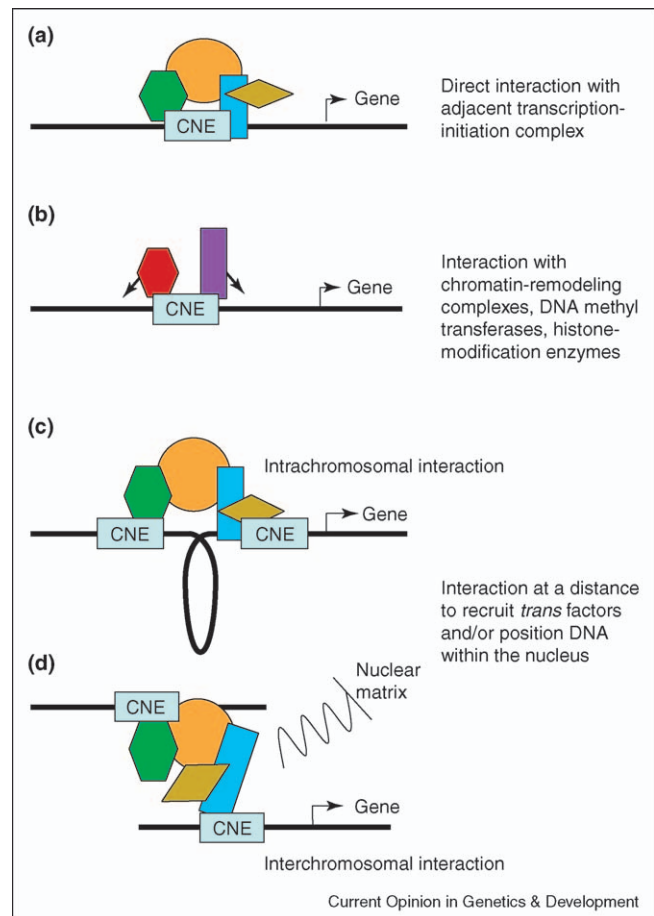
At least three explanations can be offered for the function of CNEs: (i) they represent traditional enhancers that are composed of collections of transcription factor binding sites upon which the complexity of the regulatory pattern has imposed a particularly strong constraint; (ii) they are involved in establishment and maintenance of local chromatin structure; and (iii) they are components of longer-range genome structure determinants that specify the spatial organization of chromosomes in the nucleus (Figure 1).

Several CNEs have been confirmed to have regulatory activity. Non-coding sequences — surrounding the developmental genes *SOX21*, *PAX6*, *HLXB9* and *SHH* — that are highly conserved between human and *Fugu rubripes* were tested for their ability to act as tissue-specific enhancers in zebrafish embryos [37**]. Putative enhancers were co-injected with a green fluorescent protein reporter gene coupled to a minimal promoter from the mouse β -globin gene. Out of twenty-five CNEs located near the four genes that were tested, twenty-three showed enhancer activity in one or more tissues [37**]. Poulin *et al.* [38] showed that a very highly conserved CNE adjacent to the mouse *Dach1* gene has enhancer activity. Transgenic mice carrying the CNE in a *lacZ* reporter construct showed expression that was specific to certain brain regions during development. Some modifications to the CNE abolished this activity, but, interestingly, two small insertions in the most conserved region did not detectably alter the enhancer function.

Not all active enhancers are well conserved, however. Plessy *et al.* [39*] found that only 11 of 104 experimentally validated enhancer sequences in mouse have sequence matches in the zebrafish genome. All eleven of the conserved enhancers function during embryonic development in mice, suggesting that they contribute to a common core of signals that lead to specification of the vertebrate body plan. It is likely that at least some of the remaining enhancers contribute to species-specific expression regulation. Anand *et al.* [40] demonstrated considerable divergence of *Hoxc8* enhancer activity between mouse and fish — when assayed in transgenic mouse embryos. They concluded that “remodeling of *Hox* regulatory elements in different species has played a significant role in generating morphological diversity.”

Are CNEs essential? As is the case with many conserved protein-coding genes, mice deficient in CNEs might show no phenotype. Nobrega [41*] examined the effect of deleting more than 1000 CNEs present in two large gene-deserts in mouse. Although it is impossible to assess the evolutionary impact of this deletion, the mice were viable and fertile, and no gross phenotypic changes were observable.

Figure 1



Potential roles of CNEs in gene regulation. Conserved non-coding sequence elements (CNEs) might have a diverse array of functions related to gene regulation and maintenance of chromosome structure. Some roles for CNEs include (a) acting as *cis*-regulatory elements; for example, as an enhancer, silencer or insulator to modulate transcription of an adjacent gene; (b) interacting with components of chromatin remodelling complexes, which create ‘open’ structures that promote active transcription or ‘closed’ structures that inhibit it; and acting in concert to bridge regulatory elements located at long distances, either (c) on the same chromosome or (d) on another chromosome.

Quite complex developmental patterns of gene expression can be conserved in the absence of detectable sequence similarity, suggesting considerable plasticity in how a regulatory program is encoded in DNA. Developmental expression of the homeobox gene *Otx* is remarkably similar in the ascidians *Halocynthia roretzi* and *Ciona intestinalis* despite lack of *cis* regulatory sequence conservation [42**]. Rather, the presence of a shared set of transcription factor binding sites, albeit in different number, order and orientation, seems sufficient to direct the regulatory program for this crucial gene. A similar situation has been seen in *Drosophila*, in which substantially diverged enhancers can direct identical and quite precise expression of the *even-skipped* gene in the blastoderm embryo [43]. A male-specific wing spot has evolved in

Drosophila biarmipes — but not in the closely related species *melanogaster* and *pseudoobscura* — as a result of a small number of mutations that create transcription factor binding sites and bring the *yellow* gene under the control of the Engrailed regulatory network, which is active in specifying wing development [44••]. Taken together, these studies imply that strict sequence conservation is not essential for maintaining regulatory function and that CNEs might not be primarily serving as tissue-specific enhancers.

Dermitzakis *et al.* [31••] showed that the number of CNEs and the extent of sequence conservation among species is not biased with respect to the location of genes, as would be expected if CNEs act in a *cis* regulatory fashion. Evidence that CNEs might be involved in longer-range determination of chromosome structure comes from an observed enrichment of nuclear matrix attachment sites in CNEs [45].

Perhaps even more compelling is the argument [31••] that CNEs might be involved in specification of the spatial location of chromosomes in the nucleus [46,47]. Gene regulation at the *β -globin* locus involves interaction of a locus control region located more than 50 kb away from the *β -globin* promoter [48,49]. Furthermore, physical interactions between promoters on different chromosomes have recently been shown to regulate the alternative expression of *IFN- γ* and *IL-5* in mice in the differentiation of naïve CD4⁺ T-cells to T_{H1} or T_{H2} cells [50••]. Recent advances in methods to study chromosome organization [48,51] will facilitate analysis of the role of CNEs in long-range regulatory activities.

Other contributors to regulatory diversity

Epigenetic mechanisms, including DNA methylation, histone modification, and binding of chromatin-remodeling factors, also directly influence rates of transcription [52]. In yeast, a consistent scheme (the so-called ‘histone code’) has been observed for the contribution of histone modification to transcriptional status [53]. Early experiments examining human and mouse chromosomes have also shown a consistent and conserved pattern, with increased histone H3 lysine 4 methylation associated with active transcription [54•]. The mechanism by which particular chromatin domains are established and maintained is not well known, but it appears that histone modification states are conserved between mouse and human in the absence of broad sequence similarity, so the DNA elements that specify chromatin structure might be either subtle or located at some distance from the chromatin domain.

Our ignorance of basic regulatory processes has been highlighted by the recent discovery of a large class of genes encoding small RNAs that perform a pervasive gene-regulatory role [55–57]. In zebrafish, much more

than 100 miRNAs are expressed late in development in a highly tissue-specific manner and might contribute to late events in cellular differentiation and tissue maintenance [58••]. miRNA genes are conserved in sequence and structure [59], but the extent to which evolution in these genes or their targets might contribute to heritable changes in gene regulation has not yet been addressed.

Conclusion

Complete and draft genome sequences have given us tantalizing clues about the extent of functional and potentially regulatory sequence in mammalian genomes. The pattern of evolution that is inferred from analysis of these sequences has led to the identification of tissue-specific enhancers, matrix attachment sites, and miRNA targets, but most conserved sequence remains unclassified. Thus, it seems that there might be several functional classes of conserved sequence involved not only in local gene-regulation but also in the spatial organization of the genome to promote coordinated gene expression.

It would be premature to speculate on which of the many modes of gene regulation are subject to the strongest evolutionary constraint, or which might contribute most to divergence among species, but the extent of conservation suggests unappreciated roles for a large amount of conserved sequence throughout the mammals.

Acknowledgements

I thank Gabrielle Nickel for comments on the manuscript, and Kimberly Bercaw for assistance in preparing the figure.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- of special interest
- of outstanding interest

1. International Human Genome Sequencing Consortium: **Finishing the euchromatic sequence of the human genome.** *Nature* 2004, **431**:931-945.
2. Wheeler DL, Church DM, Edgar R, Federhen S, Helmberg W, Madden TL, Pontius JU, Schuler GD, Schriml LM, Sequeira E *et al.*: **Database resources of the National Center for Biotechnology Information: update.** *Nucleic Acids Res* 2004, **32**:D35-D40.
3. Birney E, Andrews D, Bevan P, Caccamo M, Cameron G, Chen Y, Clarke L, Coates G, Cox T, Cuff J *et al.*: **Ensembl 2004.** *Nucleic Acids Res* 2004, **32 Database issue**:D468-470.
4. Ashurst JL, Chen CK, Gilbert JGR, Jekosch K, Keenan S, Meidl P, Searle SM, Stalker J, Storey R, Trevanion S *et al.*: **The Vertebrate Genome Annotation (Vega) database.** *Nucleic Acids Res* 2005, **33**:D459-D465.
5. Rat Genome Sequencing Project Consortium: **Genome sequence of the Brown Norway rat yields insights into mammalian evolution.** *Nature* 2004, **428**:493-521.
6. Mural RJ, Adams MD, Myers EW, Smith HO, Miklos GL, Wides R, Halpern A, Li PW, Sutton GG, Nadeau J *et al.*: **A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome.** *Science* 2002, **296**:1661-1671.
7. International Mouse Genome Sequencing Consortium: **Initial sequencing and comparative analysis of the mouse genome.** *Nature* 2002, **420**:520-562.

8. Wilson AC, Maxson LR, Sarich VM: **Two types of molecular evolution. Evidence from studies of interspecific hybridization.** *Proc Natl Acad Sci USA* 1974, **71**:2843-2847.
9. Jaenisch R, Bird A: **Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals.** *Nat Genet* 2003, **33**:245-254.
10. Margueron R, Trojer P, Reinberg D: **The key to development: interpreting the histone code?** *Curr Opin Genet Dev* 2005, **15**:163-176.
11. Levine M, Tjian R: **Transcription regulation and animal diversity.** *Nature* 2003, **424**:147-151.
12. Davidson EH: *Genomic Regulatory Systems: Development and Evolution.* San Diego, CA: Academic Press; 2001.
13. Murakami K, Kojima T, Sakaki Y: **Assessment of clusters of transcription factor binding sites in relationship to human promoter, CpG islands and gene expression.** *BMC Genomics* 2004, **5**:16.
14. Shapiro MD, Marks ME, Peichel CL, Blackman BK, Nereng KS, ● Jónsson B, Schluter D, Kingsley DM: **Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks.** *Nature* 2004, **428**:717-723.
- A reduction in expression of *Pitx1*, compared with the level of expression in marine isolates, leads to loss of pelvic structures in freshwater sticklebacks. There is strong evidence for a selective advantage of each structure in the respective population. Although the molecular mechanism for the change in expression is not yet known, this demonstrates that regulatory variation can cause substantial phenotypic changes while preserving essential developmental roles in other tissues.
15. Oleksiak MF, Roach JL, Crawford DL: **Natural variation in cardiac metabolism and gene expression in *Fundulus heteroclitus*.** *Nat Genet* 2005, **37**:67-72.
16. Hammock EA, Young LJ: **Microsatellite instability generates diversity in brain and sociobehavioral traits.** *Science* 2005, **308**:1630-1634.
17. Lim MM, Wang Z, Olazabal DE, Ren X, Terwilliger EF, Young LJ: **Enhanced partner preference in a promiscuous species by manipulating the expression of a single gene.** *Nature* 2004, **429**:754-757.
18. Rifkin SA, Kim J, White KP: **Evolution of gene expression in the *Drosophila melanogaster* subgroup.** *Nat Genet* 2003, **33**:138-144.
19. Enard W, Khaitovich P, Klose J, Zollner S, Heissig F, Giavalisco P, Nieselt-Struwe K, Muchmore E, Varki A, Ravid R *et al.*: **Intra- and interspecific variation in primate gene expression patterns.** *Science* 2002, **296**:340-343.
20. Schadt EE, Monks SA, Drake TA, Lusk AJ, Che N, Colinayo V, Ruff TG, Milligan SB, Lamb JR, Cavet G *et al.*: **Genetics of gene expression surveyed in maize, mouse and man.** *Nature* 2003, **422**:297-302.
21. Doss S, Schadt EE, Drake TA, Lusk AJ: **Cis-acting expression quantitative trait loci in mice.** *Genome Res* 2005, **15**:681-691.
22. Wade CM, Kulbokas EJ III, Kirby AW, Zody MC, Mullikin JC, Lander ES, Lindblad-Toh K, Daly MJ: **The mosaic structure of variation in the laboratory mouse genome.** *Nature* 2002, **420**:574-578.
23. Wiltshire T, Pletcher MT, Batalov S, Barnes SW, Tarantino LM, Cooke MP, Wu H, Smylie K, Santrosyan A, Copeland NG *et al.*: **Genome-wide single-nucleotide polymorphism analysis defines haplotype patterns in mouse.** *Proc Natl Acad Sci USA* 2003, **100**:3380-3385.
24. Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, ● Spielman RS, Cheung VG: **Genetic analysis of genome-wide variation in human gene expression.** *Nature* 2004, **430**:743-747.
- Variation in gene expression levels is, in part, heritable; this study mapped loci that contribute to heritable variation in expression of 3554 genes. As expected, many gene expression phenotypes were influenced by several factors, both *cis*- and *trans*-acting, although, interestingly, a majority were *trans*, in contrast to those reported by Schadt *et al.* [20].
25. Khaitovich P, Weiss G, Lachmann M, Hellmann I, Enard W, ● Muetzel B, Wirkner U, Ansoorge W, Paabo S: **A neutral model of transcriptome evolution.** *PLoS Biol* 2004, **2**:E132.
- Considerable variation in expression of orthologous genes is observed between species. The authors conclude that most of that variation can be accounted for by a model of neutral drift in which expression change increases linearly with evolutionary time, and thus might not be of functional significance. This increases the burden of proof required to claim that an expression difference has evolutionary consequences.
26. Khaitovich P, Paabo S, Weiss G: **Towards a neutral evolutionary model of gene expression.** *Genetics* 2005, **170**:929-939.
27. Margulies EH, Blanchette M, Haussler D, Green ED: **Identification and characterization of multi-species conserved sequences.** *Genome Res* 2003, **13**:2507-2518.
28. Thomas JW, Touchman JW, Blakesley RW, Bouffard GG, Beckstrom-Sternberg SM, Margulies EH, Blanchette M, Siepel AC, Thomas PJ, McDowell JC *et al.*: **Comparative analyses of multi-species sequences from targeted genomic regions.** *Nature* 2003, **424**:788-793.
29. Dermitzakis ET, Reymond A, Lyle R, Scamuffa N, Ucla C, Deutsch S, Stevenson BJ, Flegel V, Bucher P, Jongeneel CV *et al.*: **Numerous potentially functional but non-genic conserved sequences on human chromosome 21.** *Nature* 2002, **420**:578-582.
30. Dermitzakis ET, Reymond A, Scamuffa N, Ucla C, Kirkness E, Rossier C, Antonarakis SE: **Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs).** *Science* 2003, **302**:1033-1035.
31. Dermitzakis ET, Kirkness E, Schwarz S, Birney E, Reymond A, ● Antonarakis SE: **Comparison of human chromosome 21 conserved nongenic sequences (CNGs) with the mouse and dog genomes shows that their selective constraint is independent of their genic environment.** *Genome Res* 2004, **14**:852-859.
- Almost all non-coding sequences conserved between mouse and human are also conserved in the dog genome. The position of these conserved sequences is essentially random with respect to adjacent genes, and there is no preference for those closer to genes to be better conserved. The authors conclude that most of these sequences are probably not *cis*-regulatory elements, but might have a global role in genome function through long-distance (or distance-independent) *cis* or *trans* chromosomal interactions.
32. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, ● Mattick JS, Haussler D: **Ultraconserved elements in the human genome.** *Science* 2004, **304**:1321-1325.
- 481 perfectly identical sequences at least 200 bp in length were found in orthologous locations in the human, mouse and rat genome sequences; three-quarters of these are not in protein-coding regions of genes. Extraordinary conservation is seen in non-mammalian vertebrates but not in invertebrates. Ultra-conserved elements that don't overlap exons tend to be clustered around transcription factor genes and developmentally regulated genes.
33. Dermitzakis ET, Clark AG: **Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover.** *Mol Biol Evol* 2002, **19**:1114-1121.
34. Smith NG, Brandstrom M, Ellegren H: **Evidence for turnover of functional noncoding DNA in Mammalian genome evolution.** *Genomics* 2004, **84**:806-813.
35. Bulyk ML: **Computational prediction of transcription-factor binding site locations.** *Genome Biol* 2003, **5**:201.
36. Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, ● Lander ES, Kellis M: **Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals.** *Nature* 2005, **434**:338-345.
- Short sequence motifs that are more highly conserved among mouse, human, rat and dog, and are over-represented in 5' flanking sequence or 3'UTRs were identified. These motifs are heavily enriched for known functional elements, such as transcription factor binding sites. Analysis of 3'UTR motifs revealed extensive evidence of miRNA binding sites.
37. Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, ● Smith SF, North P, Callaway H, Kelly K *et al.*: **Highly conserved non-coding sequences are associated with vertebrate development.** *PLoS Biol* 2005, **3**:e7.

More than 1400 non-coding sequences that are highly conserved between *Fugu* and human were found, with a large majority in the vicinity of developmentally regulated genes. None have similarity in invertebrates. 23 out of 25 conserved sequences adjacent to *sox21*, *pax6*, *hlxB9* or *shh* genes functioned as tissue-specific enhancers in zebrafish embryos.

38. Poulin F, Nobrega MA, Plajzer-Frick I, Holt A, Afzal V, Rubin EM, Pennacchio LA: **In vivo characterization of a vertebrate ultraconserved enhancer.** *Genomics* 2005, **85**:774-781.
39. Plessy C, Dickmeis T, Chalmel F, Strahle U: **Enhancer sequence conservation between vertebrates is favoured in developmental regulator genes.** *Trends Genet* 2005, **21**:207-210.
- Only 10% of 104 experimentally defined enhancers in mouse are conserved at the sequence level in zebrafish. The conserved enhancers are all located adjacent to genes that are active during embryonic development, possibly indicating that the high sequence conservation reflects extensive constraint on binding sites that integrate the complex signals necessary for specifying a vertebrate body plan.
40. Anand S, Wang WC, Powell DR, Bolanowski SA, Zhang J, Ledje C, Pawashe AB, Amemiya CT, Shashikant CS: **Divergence of Hoxc8 early enhancer parallels diverged axial morphologies between mammals and fishes.** *Proc Natl Acad Sci USA* 2003, **100**:15666-15669.
41. Nobrega MA, Zhu Y, Plajzer-Frick I, Afzal V, Rubin EM: **Megabase deletions of gene deserts result in viable mice.** *Nature* 2004, **431**:988-993.
- Two gene deserts spanning ~1 Mb and ~2 Mb and containing more than 1200 non-coding sequences conserved between humans and rodents were deleted in mice. Although the long-term evolutionary impact of these deletions cannot be known, there were no observable phenotypic differences in mice carrying homozygous deletions, and expression of neighboring genes was only marginally affected.
42. Oda-Ishii I, Bertrand V, Matsuo I, Lemaire P, Saiga H: **Making very similar embryos with divergent genomes: conservation of regulatory mechanisms of Otx between the ascidians *Halocynthia roretzi* and *Ciona intestinalis*.** *Development* 2005, **132**:1663-1674.
- A complex pattern of *Otx* expression is shared in development between these two chordate species, despite a lack of regulatory region sequence similarity. A similar array of transcription factor binding sites is present in each *Otx* enhancer region, but in varying numbers, position and orientation, suggesting that degeneracy of binding sites has enabled conservation of overall function in the presence of extensive sequence change.
43. Ludwig MZ, Palsson A, Alekseeva E, Bergman CM, Nathan J, Kreitman M: **Functional evolution of a cis-regulatory module.** *PLoS Biol* 2005, **3**:e93.
44. Gompel N, Prud'homme B, Wittkopp PJ, Kassner VA, Carroll SB: **Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in *Drosophila*.** *Nature* 2005, **433**:481-487.
- Regulation of the *yellow* gene, which controls pigmentation, has evolved to be responsive to *Engrailed* signaling in one *Drosophila* lineage. Mutations that created binding sites for transcription factors that are active in wing development have enabled this gene to co-opt an existing regulatory network.
45. Glazko GV, Koonin EV, Rogozin IB, Shabalina SA: **A significant fraction of conserved noncoding DNA in human and mouse consists of predicted matrix attachment regions.** *Trends Genet* 2003, **19**:119-124.
46. Cremer T, Kupper K, Dietzel S, Fakan S: **Higher order chromatin architecture in the cell nucleus: on the way from structure to function.** *Biol Cell* 2004, **96**:555-567.
47. Misteli T: **Spatial positioning; a new dimension in genome function.** *Cell* 2004, **119**:153-156.
48. Carter D, Chakalova L, Osborne CS, Dai YF, Fraser P: **Long-range chromatin regulatory interactions in vivo.** *Nat Genet* 2002, **32**:623-626.
49. Tolhuis B, Palstra RJ, Splinter E, Grosveld F, de Laat W: **Looping and interaction between hypersensitive sites in the active β -globin locus.** *Mol Cell* 2002, **10**:1453-1465.
50. Spilianakis CG, Lalioti MD, Town T, Lee GR, Flavell RA: **Interchromosomal associations between alternatively expressed loci.** *Nature* 2005, **435**:637-645.
- The *Il7g* gene on mouse chromosome 10 physically interacts with the T_H2 locus control region on chromosome 11. This interaction is markedly reduced following differentiation of naive CD4⁺ T-cells to effector T_H1 or T_H2 cells.
51. Dekker J, Rippe K, Dekker M, Kleckner N: **Capturing chromosome conformation.** *Science* 2002, **295**:1306-1311.
52. Fazzari MJ, Grealley JM: **Epigenomics: beyond CpG islands.** *Nat Rev Genet* 2004, **5**:446-455.
53. Dion MF, Altschuler SJ, Wu LF, Rando OJ: **Genomic characterization reveals a simple histone H4 acetylation code.** *Proc Natl Acad Sci USA* 2005, **102**:5501-5506.
54. Bernstein BE, Kamal M, Lindblad-Toh K, Bekiranov S, Bailey DK, Huebert DJ, McMahon S, Karlsson EK, Kulbokas EJ III, Gingeras TR et al.: **Genomic maps and comparative analysis of histone modifications in human and mouse.** *Cell* 2005, **120**:169-181.
- Chromatin immunoprecipitation was used to assess the methylation pattern of histone H3 lysines 4, 9, and 14 on human chromosomes 12 and 22, and on several selected orthologous mouse loci. Transcription start sites and actively transcribed regions show distinctive histone methylation patterns that are conserved in mouse and human. Hox gene clusters have a different pattern that spans several genes.
55. Pasquinelli AE, Hunter S, Bracht J: **MicroRNAs: a developing story.** *Curr Opin Genet Dev* 2005, **15**:200-205.
56. Lee R, Feinbaum R, Ambros V: **A short history of a short RNA.** *Cell* 2004, **116**:S89-92, 81 p following S96.
57. Ruvkun G, Wightman B, Ha I: **The 20 years it took to recognize the importance of tiny RNAs.** *Cell* 2004, **116**:S93-96, 92 p following S96.
58. Wienholds E, Kloosterman WP, Miska E, Alvarez-Saavedra E, Berezikov E, de Bruijn E, Horvitz HR, Kauppinen S, Plasterk RH: **MicroRNA expression in zebrafish embryonic development.** *Science* 2005, **309**:310-311.
- More than 100 miRNAs are expressed in zebrafish development, most in a highly tissue-specific manner and, generally, later in development. The authors suggest that miRNAs function primarily in cell-type differentiation or maintenance of tissue identity rather than in earlier stages of cell fate determination.
59. Bentwich I, Avniel A, Karov Y, Aharonov R, Gilad S, Barad O, Barzilai A, Einat P, Einav U, Meiri E et al.: **Identification of hundreds of conserved and nonconserved human microRNAs.** *Nat Genet* 2005, **37**:766-770.

Comparative genomics as a tool in the understanding of eukaryotic transcriptional regulation

Julie E Baggs, Kevin R Hayes and John B Hogenesch

Comparative genomics approaches are having a remarkable impact on the study of transcriptional regulation in eukaryotes. Many eukaryotic genome sequences are being explored by new computational methods and high-throughput experimental tools such as DNA arrays and genome-wide location analyses. These tools are enabling efficient panning for common regulatory cassettes underlying fundamental biological processes, extending the use of existing techniques for the discovery of response elements to mammals, deciphering the transcriptional regulatory code in eukaryotes and providing the first global insights into a recently described post-transcriptional regulatory mechanism. Collectively, these approaches are rapidly expanding both our knowledge and our definition of transcriptional regulation.

Addresses

Department of Biomedical Sciences, The Scripps Research Institute, 5353 Parkside Drive RF1, Jupiter, FL 33458, USA

Corresponding author: Hogenesch, John B (hogenesch@scripps.edu)

Current Opinion in Genetics & Development 2005, **15**:634–639

This review comes from a themed issue on
Genomes and evolution
Edited by Stephen J O'Brien and Claire M Fraser

Available online 13th October 2005

0959-437X/\$ – see front matter

© 2005 Elsevier Ltd. All rights reserved.

DOI 10.1016/j.gde.2005.09.012

Introduction

The completion of dozens of eukaryotic genome sequence projects led to an initial wave of analyses focusing on the complete description of all protein-encoding genes (NCBI Genomic Biology, <http://www.ncbi.nlm.nih.gov/Genomes>). One surprise from this effort, the non-linear relationship between organismal complexity and gene number, has become widely accepted in recent years and, in turn, has fostered a second wave of comparative genome analyses [1]. These approaches take advantage of novel computational methods to analyze multiple genome sequences to understand gene regulation. Experimental biology is then employed to test these hypotheses in a systematic manner.

Here, we discuss recent investigations in which comparative genome analysis has been used to understand transcriptional regulation in eukaryotes (Figure 1). We start

by looking at how the study of global RNA expression, enabled by DNA chip analysis (see Glossary), is being employed across species to gain insight into fundamental biological processes. Next, we explore how recent advances in informatics, experimental biology, and comparative genomics are being used to define response elements in mammals, in addition to being integrated with large-scale experimental studies aimed at the definition of the regulatory code of eukaryotes. We extend our review to the combinatorial analysis of these response elements in higher-order structures called enhancers. Finally, we highlight the remarkable impact that comparative genomics has had on the description of a more complete description of microRNAs (miRNAs; see also Glossary), newly defined *trans*-factors that regulate gene expression at both the message and the protein level, and at their downstream target genes [2].

Comparing patterns of gene expression

In addition to complete genome sequences — or partly as a result of them — recent technological developments in the field of RNA dynamics have made it possible to measure global transcriptional output in eukaryotes on a single chip. This has enabled comparative studies aimed at analyzing global transcriptional output to provide insight into fundamental biological processes. For example, McCarroll *et al.* [3] used this approach to investigate transcriptional responses during aging in *Caenorhabditis elegans* and *Drosophila melanogaster* and found a cassette of genes that are involved in metabolism, catabolism and DNA repair, and which turn on in adulthood in both organisms. The authors go on to explore other large-scale datasets and demonstrate the power of this approach in uncovering transcriptional signatures for several other fundamental biological processes. Khaitovich *et al.* [4^{*}] used cross-species expression analysis in primates to explore evolution of genomes and transcriptomes. By comparing the transcriptional output between chimpanzees and humans, they found that some tissues, such as the brain, displayed similar expression patterns whereas other tissues had greater inter-species differences, indicating that transcriptional output in these tissues can have differing rates of divergence. In addition, uniformly expressed genes were found to harbor fewer changes in amino acid sequence than do more-specifically expressed genes [4^{*}]. The authors conclude that this behavior is consistent with a model of neutral evolution with negative selection; however, they note that for X chromosome-expressed genes in testis, patterns suggestive of positive selection apply to both the expression signatures and the amino acid sequences [4^{*}]. In a second report by the same

Glossary

5'RACE experiments: Rapid amplification of 5' complementary DNA ends. A method to extend a cDNA clone by amplification of corresponding mRNA.

BLAT: BLAST-like alignment tool. A tool to find regions of similarity between sequences.

ChIP (chromatin immunoprecipitation)-on-chip: A method to identify nucleotide sequences bound by a specific protein such as a transcription factor. The targets are identified by hybridizing labeled DNA to arrays that harbor target sequences.

DNA chip analysis: A method for analyzing the expression of nucleic acid targets in parallel on glass arrays.

HeLa cells: Immortalized human epithelial cell line commonly used in laboratory experiments.

MicroRNAs: Small RNA molecules that are negative regulators of gene and protein expression.

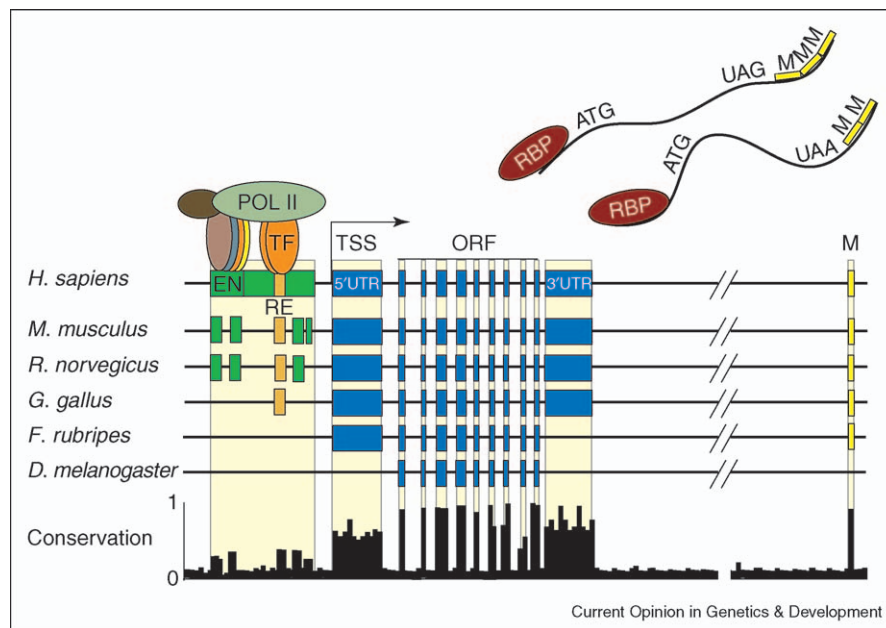
Positional weight matrix: A description of binding-site sequence frequency by position that can be used to evaluate new sequences for similarity to the matrix.

group [5], the authors use reporter assays in cell lines to examine some of the functional differences between the activities of human and chimpanzee promoters. The behavior of these promoter sequences in cell lines is remarkably poorly conserved, suggesting the possibility that rapidly accruing sequence differences in these regions contribute to expression differences [5]. These results show that the application of comparative genomics to the study of RNA expression can lead to insight into fundamental biological processes such as aging and evolution [6].

Response element discovery and the transcriptional regulatory code

Comparative genomics has long held the promise for the identification of response elements in eukaryotic genomes [7]. Initially, these searches were conducted with consensus sequences and positional weight matrices (see Glossary) and were confined to the detection of known elements. Later, *ab initio* approaches showed great potential for the identification of response elements in lower eukaryotic organisms [8,9]. The sheer size and complexity of mammalian genome sequences, however, has confounded the application of these approaches to the study of mammalian transcription [10]. Recently, robust methods to define common regulatory motifs have emerged in yeast and, later, in mammals [11,12,13^{**}]. Kellis *et al.* [12] explored three sequenced yeast strains for common regulatory motifs, identifying with both sensitivity and specificity 72 genome-wide elements, including most of the known response elements. In addition, they extended these analyses to the mammalian genome. By aligning and exploring the genome sequences of human, mouse, rat and dog, Xie *et al.* [13^{**}] identified 174 common regulatory motifs in mammalian promoters, including 72% of the known motifs in Transfac, a transcription factor database, and 105 novel motifs. Many of these elements show tissue-specific expression and proximity to transcriptional start sites, suggesting that they have a role in mediating binding of *bona fide trans*-factors. Several groups have taken advantage of comparative genomics

Figure 1



Depiction of conserved DNA elements that play an important role in transcriptional regulation. Elements proximal to the transcriptional start site (TSS) include both single response elements (REs) and organized groups of them in enhancer (EN) regions. Other elements in the 5' and 3' untranslated regions (UTRs) encode regions for post-transcriptional regulation, including sites for RNA binding proteins (RBPs) and sites for regulation by microRNAs (Ms). The microRNAs act as *trans*-factors and can have multiple targets.

and positional weight matrices to identify new target genes of known *trans*-factors, such as p53 and CREB [14,15]. This approach has recently been automated in an integrated system that combines cross-species analysis with co-expression information from DNA expression analysis [16]. Thus, the past two years of research in comparative genomics has extended *ab initio* response element discovery from lower eukaryotes to higher eukaryotes and has expanded the toolset for researchers in mammalian biology who are interested in discovering new targets of specific transcription factors.

One of the more compelling recent trends has been the development of large-scale experimental approaches that focus on elucidating transcriptional regulation at the network level. Genome-wide location analysis, or chromatin-immunoprecipitation on arrays (ChIP-on-chip; see also Glossary), has emerged as a powerful tool to determine the DNA target sequences of specific transcription factors [17,18]. Harbison *et al.* [19**] integrated genome-wide location analysis with comparative genomic methods to elucidate the transcriptional regulatory code of yeast. In this heroic study, the group performed ChIP-on-chip analysis with 203 different yeast transcription factors, determining thousands of unique interactions between these *trans*-factors and promoters. They merged this dataset with response elements predicted by six different element-discovery algorithms and phylogenetic footprinting data from six other yeast strains. This combined analysis was synthesized into a first version of the transcriptional regulatory code containing 3353 sites in 1296 promoters. This work resulted in the thorough characterization of known transcription factor–response element combinations and, more importantly, determined the sites bound by many previously uncharacterized factors. The authors went on to explore the arrangement of the binding sites within the promoter (promoter architecture) and developed models for transcriptional responses to environmental change [19**]. Another approach to determine the sites bound by specific transcription factors used DNA array technology to measure specific binding of *in vitro* labeled transcription factors to promoter sequences. This data was fed into an *ab initio* response-element prediction algorithm, and the new sequences were both retrospectively and experimentally validated [20**]. These approaches hold great promise for the construction of regulatory networks in eukaryotes, and when combined with other types of data (e.g. RNA dynamics experiments) they promise to build predictive models of transcriptional regulation that codify normal physiology and responses to environmental change.

Combinatorics: response elements and enhancers

More often than not, response elements act in association with other elements in larger structures known as enhancers, silencers or insulator sequences [21–23]. Recent

studies to identify DNA regions that function in endogenous message expression have focused on co-occurrence of response elements in global comparative genomics studies [12]. For example, Kreiman [24] used co-expression, prior knowledge of the characteristics of described response elements, and phylogenetic footprinting to determine known and unknown elements that contribute to specific expression patterns in *Drosophila* pattern formation and in human skeletal muscle. Comparative genomic studies used different species of flies to examine in detail the even-skipped (*eve*) stripe 2 enhancer, which comprises specific combinations of activator and repressor elements [25–27]. These studies have combined the latest generation of sequence comparison algorithms with elegant genetic studies aimed at elucidating regulatory mechanisms that contribute to complex expression patterns (reviewed in [28]). Early studies employed sequence comparisons with transgenic analysis and showed that, surprisingly, distantly related sequences from four fly strains, *Drosophila melanogaster*, *Drosophila yakuba*, *Drosophila erecta* and *Drosophila pseudoobscura*, yielded the same pattern of expression when fused to a reporter gene in transgenic *D. melanogaster* flies [29]. However, chimeric enhancers constructed from two species failed to lead to normal expression *in vivo*, suggesting that these enhancer sequences do not function analogously [30]. This suggestion was recently supported by reverse-genetic studies that demonstrated the importance of both spatial and temporal gene expression in addition to dosage [31**]. An enhancer from the more distantly related *D. pseudoobscura* was able to functionally rescue *D. melanogaster* flies lacking this region, whereas the enhancer from *D. erecta*, a less divergent strain, was not [31**]. In sum, these studies indicate that enhancer sequences have evolved different solutions to the problem of dosage and spatio-temporal patterning, using the same (or similar) complement of *trans*-factors.

MicroRNAs: novel post-transcriptional regulators

Perhaps the greatest recent impact that comparative genomics has had on the understanding of eukaryotic transcriptional regulation is its application to the study of miRNAs and their mechanisms of gene regulation. miRNAs can act as *trans*-factors to degrade mRNAs by precise or imprecise base-pairing and can also act to inhibit translation through poorly understood mechanisms [32]. miRNAs were first identified in forward genetic screens in *C. elegans* [33] and have since been found in all Metazoa examined. These genetic screens have the tremendous advantage of relating the existence of a miRNA to a given biological phenotype (for example, the identification and characterization of *lin-4* and *let-7* mutants led to the determination that miRNAs play crucial roles in developmental timing in worms [33,34]). However, forward genetics approaches are laborious, time consuming and, as yet, cannot be directed toward one particular class of

entities, such as miRNAs. Comparative genomics has been used effectively as a tool in plants, flies and mammals to discover and catalog a more complete collection of miRNAs. In general, these methods take advantage of computer models of known miRNA structures and search several genome sequences for candidate miRNAs. Lai *et al.* [35] used such an approach to identify 48 novel miRNAs in *Drosophila*, validated the expression of 24 of these and concluded that flies probably contain ~110 miRNA genes [35]. Similarly, Lim *et al.* [36] used computational methods in addition to molecular identification and validation to determine a more complete set of *C. elegans* miRNAs. The vast majority of these are conserved across different strains of worm; a quarter are conserved in humans; and more than a third are differentially expressed during development, highlighting their potential importance in this process [36]. Computational methods have also been used to greatly expand the known repertoire of miRNA genes in vertebrates. Examination of human and *Fugu* consensus structures and phylogenetically conserved stem loop structures found 188 known and novel miRNAs [37]. Extrapolating from these data, the authors estimate an upper limit of no more than 255 miRNAs in humans [37]. A different computational approach was taken during the study by Xie *et al.* [13**], in which they aligned human, mouse, rat and dog genome sequences to define a lexicon of common regulatory motifs. In addition to defining response elements, this work resulted in the identification of 129 new predicted miRNAs, and in validation studies the expression of half of their test set was confirmed. Lastly, a BLAT-based searching method (see Glossary) using known miRNA sequences found 35 additional human and 45 additional mouse miRNAs [38]. Thus, comparative genomics has led to the rapid increase in our catalog of miRNAs in several eukaryotic organisms.

Comparative genomics has also played an indispensable role in determining the targets of miRNAs. By comparing *Arabidopsis* and rice genomes, Rhoades *et al.* [39] were able to use the homology of miRNA sequences to their target sequences and identified potential regulatory targets for 14 of the 16 miRNAs that had been described at that point. These initial studies exploring the *Arabidopsis* and *Oryza* genomes were extended to enable the identification of 83 putative miRNA genes, the prediction of their target genes, and the use of independent microarray data to show anti-correlation between the expression of miRNAs and these targets [40*]. Finally, three of these microRNA–target pairs were confirmed experimentally by 5'RACE experiments (rapid amplification of 5' complementary DNA ends; see also Glossary) [40*]. A similar approach was used to determine the targets of known *Drosophila* miRNAs, by comparing 3'UTR (untranslated region) sequences from *D. melanogaster*, *D. pseudoobscura* and *Anopheles* [41]. More recently, the genome sequences of seven *Drosophila* species were used to investigate

miRNA targets, predicting that each miRNA targets 54 messages on average [42]. Several groups have performed similar studies on vertebrates. Lewis *et al.* [43] have used computational methods, taking advantage of the human, mouse, rat, and fish genomes, to predict ~400 targets of conserved vertebrate miRNAs; some of these targets were experimentally validated using reporter assays in HeLa cells (see Glossary). In a second study, cross-species analysis of human, mouse, rat and fish genomes identified more than 2000 miRNA targets, suggesting that 10% of all mammalian genes are regulated by these mechanisms [44*]. More recently, Xie *et al.* [13**] performed a comprehensive analysis of 3'UTRs and suggested that about 20% of all mammalian genes are regulated by miRNAs [13**]. Collectively, these studies have offered the first comprehensive analyses of miRNA targets, highlighting the power of comparative genomics. Future work incorporating larger-scale experimental studies will be required to confirm the validity and importance of this regulation.

Conclusion

The past two years have witnessed an explosion of activity in the application of comparative genomics to the understanding of transcriptional regulation. Insight into global organization of transcription is being gathered from cross-species comparisons of RNA expression data. Robust methods to discover response elements, once only possible in less complex organisms, are now being enabled by comparative genomics and applied to mammalian genomes. These methods are also now being integrated with large-scale experimental strategies aimed at defining the regulatory code of eukaryotes for the first time. Finally, comparative genomics has played the lead role in defining the larger complement of miRNAs and their target sequences. The next few years should see improvements in these methods and an increase in their power as additional genome sequences are made available. Genomes, they're not just for finding genes anymore.

Acknowledgements

We acknowledge our many colleagues whose work in the field of comparative genomics and transcription we were unable to discuss, but whose work helped guide this review.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- of special interest
- of outstanding interest

1. Miller W, Makova KD, Nekrutenko A, Hardison RC: **Comparative genomics**. *Annu Rev Genomics Hum Genet* 2004, **5**:15-56.
2. Zamore PD, Haley B: **Ribo-gnome: the big world of small RNAs**. *Science* 2005, **309**:1519-1524.
3. McCarroll SA, Murphy CT, Zou S, Pletcher SD, Chin CS, Jan YN, Kenyon C, Bargmann CI, Li H: **Comparing genomic expression**

patterns across species identifies shared transcriptional profile in aging. *Nat Genet* 2004, **36**:197-204.

4. Khaïtovich P, Hellmann I, Enard W, Nowick K, Leinweber M, Franz H, Weiss G, Lachmann M, Paabo S: **Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees.** *Science* 2005, **309**:1850-1854.
- The authors analyze differences in gene expression in several tissues of human and chimpanzee to explore evolution of genomes and transcriptomes.
5. Heissig F, Krause J, Bryk J, Khaïtovich P, Enard W, Paabo S: **Functional analysis of human and chimpanzee promoters.** *Genome Biol* 2005, **6**:R57.
 6. Enard W, Khaïtovich P, Klose J, Zollner S, Heissig F, Giavalisco P, Nieselt-Struwe K, Muchmore E, Varki A, Ravid R *et al.*: **Intra- and interspecific variation in primate gene expression patterns.** *Science* 2002, **296**:340-343.
 7. Hardison RC, Oeltjen J, Miller W: **Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome.** *Genome Res* 1997, **7**:959-966.
 8. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC: **Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment.** *Science* 1993, **262**:208-214.
 9. Roth FP, Hughes JD, Estep PW, Church GM: **Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation.** *Nat Biotechnol* 1998, **16**:939-945.
 10. Wasserman WW, Palumbo M, Thompson W, Fickett JW, Lawrence CE: **Human-mouse genome comparisons to locate regulatory sites.** *Nat Genet* 2000, **26**:225-228.
 11. Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterston R, Cohen BA, Johnston M: **Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting.** *Science* 2003, **301**:71-76.
 12. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES: **Sequencing and comparison of yeast species to identify genes and regulatory elements.** *Nature* 2003, **423**:241-254.
 13. Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M: **Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals.** *Nature* 2005, **434**:338-345.
- This study uses comparative analysis of mammalian genomes to identify common regulatory motifs in promoters and 3'UTRs. Both novel motifs and previously characterized transcription factor binding sites were identified in promoter regions. Analysis of 3'UTRs also identified conserved motifs, half of which are associated with miRNAs.
14. Hoh J, Jin S, Parrado T, Edington J, Levine AJ, Ott J: **The p53MH algorithm and its application in detecting p53-responsive genes.** *Proc Natl Acad Sci USA* 2002, **99**:8467-8472.
 15. Zhang X, Odom DT, Koo SH, Conkright MD, Canettieri G, Best J, Chen H, Jenner R, Herbolsheimer E, Jacobsen E *et al.*: **Genome-wide analysis of cAMP-response element binding protein occupancy, phosphorylation, and target gene activation in human tissues.** *Proc Natl Acad Sci USA* 2005, **102**:4459-4464.
 16. Ho Sui SJ, Mortimer JR, Arenillas DJ, Brumm J, Walsh CJ, Kennedy BP, Wasserman WW: **oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes.** *Nucleic Acids Res* 2005, **33**:3154-3164.
 17. Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO: **Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF.** *Nature* 2001, **409**:533-538.
 18. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E *et al.*: **Genome-wide location and function of DNA binding proteins.** *Science* 2000, **290**:2306-2309.
 19. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J *et al.*: **Transcriptional regulatory code of a eukaryotic genome.** *Nature* 2004, **431**:99-104.

The authors combine comparative genomic analyses with genome-wide location data to identify conserved yeast-sequence elements bound by transcriptional regulators, thus generating a map of the transcriptional regulatory code for yeast.

20. Mukherjee S, Berger MF, Jona G, Wang XS, Muzzey D, Snyder M, Young RA, Bulyk ML: **Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays.** *Nat Genet* 2004, **36**:1331-1339.
- The authors use protein binding microarrays, a new DNA microarray-based technology, to map transcription factor binding sites of known DNA-binding proteins. Comparative sequence analysis of identified binding sites and experimental validation suggests this method can accurately identify *in vivo* binding sites.
21. Banerji J, Rusconi S, Schaffner W: **Expression of a β -globin gene is enhanced by remote SV40 DNA sequences.** *Cell* 1981, **27**:299-308.
 22. Bell AC, West AG, Felsenfeld G: **Insulators and boundaries: versatile regulatory elements in the eukaryotic genome.** *Science* 2001, **291**:447-450.
 23. Brand AH, Breeden L, Abraham J, Sternglanz R, Nasmyth K: **Characterization of a 'silencer' in yeast: a DNA sequence with properties opposite to those of a transcriptional enhancer.** *Cell* 1985, **41**:41-48.
 24. Kreiman G: **Identification of sparsely distributed clusters of cis-regulatory elements in sets of co-expressed genes.** *Nucleic Acids Res* 2004, **32**:2889-2900.
 25. Andrioli LP, Vasisth V, Theodosopoulou E, Oberstein A, Small S: **Anterior repression of a *Drosophila* stripe enhancer requires three position-specific mechanisms.** *Development* 2002, **129**:4931-4940.
 26. Arnosti DN, Barolo S, Levine M, Small S: **The eve stripe 2 enhancer employs multiple modes of transcriptional synergy.** *Development* 1996, **122**:205-214.
 27. Small S, Blair A, Levine M: **Regulation of even-skipped stripe 2 in the *Drosophila* embryo.** *EMBO J* 1992, **11**:4047-4057.
 28. Markstein M, Levine M: **Decoding cis-regulatory DNAs in the *Drosophila* genome.** *Curr Opin Genet Dev* 2002, **12**:601-606.
 29. Ludwig MZ, Patel NH, Kreitman M: **Functional analysis of eve stripe 2 enhancer evolution in *Drosophila*: rules governing conservation and change.** *Development* 1998, **125**:949-958.
 30. Ludwig MZ, Bergman C, Patel NH, Kreitman M: **Evidence for stabilizing selection in a eukaryotic enhancer element.** *Nature* 2000, **403**:564-567.
 31. Ludwig MZ, Palsson A, Alekseeva E, Bergman CM, Nathan J, Kreitman M: **Functional evolution of a cis-regulatory module.** *PLoS Biol* 2005, **3**:e93.
- Together with the studies by Small *et al.* [27] and Markstein *et al.* [28], the authors use several *Drosophila* species to study the evolution of enhancer sequences that regulate the *even-skipped* gene.
32. Filipowicz W: **RNAi: the nuts and bolts of the RISC machine.** *Cell* 2005, **122**:17-20.
 33. Lee RC, Feinbaum RL, Ambros V: **The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*.** *Cell* 1993, **75**:843-854.
 34. Reinhart BJ, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC, Rougvie AE, Horvitz HR, Ruvkun G: **The 21-nucleotide *let-7* RNA regulates developmental timing in *Caenorhabditis elegans*.** *Nature* 2000, **403**:901-906.
 35. Lai EC, Tomancak P, Williams RW, Rubin GM: **Computational identification of *Drosophila* microRNA genes.** *Genome Biol* 2003, **4**:R42.
 36. Lim LP, Lau NC, Weinstein EG, Abdelhakim A, Yekta S, Rhoades MW, Burge CB, Bartel DP: **The microRNAs of *Caenorhabditis elegans*.** *Genes Dev* 2003, **17**:991-1008.
 37. Lim LP, Glasner ME, Yekta S, Burge CB, Bartel DP: **Vertebrate microRNA genes.** *Science* 2003, **299**:1540.
 38. Weber MJ: **New human and mouse microRNA genes found by homology search.** *FEBS J* 2005, **272**:59-73.

39. Rhoades MW, Reinhart BJ, Lim LP, Burge CB, Bartel B, Bartel DP: **Prediction of plant microRNA targets.** *Cell* 2002, **110**:513-520.
40. Wang XJ, Reyes JL, Chua NH, Gaasterland T: **Prediction and identification of *Arabidopsis thaliana* microRNAs and their mRNA targets.** *Genome Biol* 2004, **5**:R65.
The authors present a computational method to predict *Arabidopsis* miRNAs and their target messages, using *Arabidopsis* and *Oryza sativa* genome comparison. New and previously known miRNAs were tested for expression, and several predicted miRNA targets were validated experimentally.
41. Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS: **MicroRNA targets in *Drosophila*.** *Genome Biol* 2003, **5**:R1.
42. Grun D, Wang YL, Langenberger D, Gunsalus KC, Rajewsky N: **microRNA target predictions across seven *Drosophila* species and comparison to mammalian targets.** *PLoS Comput Biol* 2005, **1**:e13.
43. Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP, Burge CB: **Prediction of mammalian microRNA targets.** *Cell* 2003, **115**:787-798.
44. John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS: **Human microRNA targets.** *PLoS Biol* 2004, **2**:e363.
Using a prediction algorithm that relies on cross-species comparison, the authors identify potential targets of all known and predicted human miRNAs.



ELSEVIER

Short, local duplications in eukaryotic genomes

Elizabeth E Thomas

Short, local duplications lead to an increase in the local copy number of a 1–100 bp sequence motif. They are usually unstable and evolve rapidly. When they involve a functional sequence such as a transcription factor binding site or a protein–protein interaction domain, they can drive phenotypic diversity. Short, local duplications have been implicated in the dramatic morphological differences among different dog breeds, and in the differences in social structure between two sister species of voles. Several human diseases and disorders are also caused by this class of duplication, which encompasses microsatellites, minisatellites and doublets.

Addresses

Bauer Center for Genomics Research, Harvard University, 7 Divinity Avenue, Cambridge, MA 02138, USA

Corresponding author: Thomas, Elizabeth E
(ethomas@cgr.harvard.edu)

Current Opinion in Genetics & Development 2005, 15:640–644

This review comes from a themed issue on
Genomes and evolution
Edited by Stephen J O'Brien and Claire M Fraser

Available online 7th October 2005

0959-437X/\$ – see front matter
© 2005 Elsevier Ltd. All rights reserved.

DOI 10.1016/j.gde.2005.09.008

Introduction

Duplicative processes are central to eukaryotic genome evolution. This is evident both in the abundance of large duplications, which can duplicate single genes, genomic regions or entire genomes, and in the abundance of short duplications, which can increase the local copy number of transcription factor binding sites, peptide motifs and subunits of gene structure. Here, I focus on the latter class of duplications and their role in genome biology and evolution. There are three classes of short, local duplications: microsatellites, minisatellites and doublets (see Figure 1).

Microsatellites are tandem arrays made up of many copies of a short repeating unit that is 1–9 bp in length. They are relatively simple sequences, and the exact copy number of the array can change rapidly. They are found in both coding and non-coding parts of genes. Some microsatellites are binding sites for transcription factors, facilitating DNA–protein interactions; others form homopeptide tracts within proteins and mediate protein–protein interactions.

Minisatellites are very similar to microsatellites, but with longer repeating units, typically 10–100 bp in length. They have at least two copies of the repeating unit, but might have hundreds of copies. Their higher information content makes them more likely to duplicate regulatory or protein-coding sequences.

Non-tandem or ‘nearby’ doublets are a recently described type of short, local duplication. They are similar to minisatellites, but with the repeating units spaced apart rather than in tandem. Only pairs of repeats have been studied, and the two copies of the repeat tend to be about 1 kb apart.

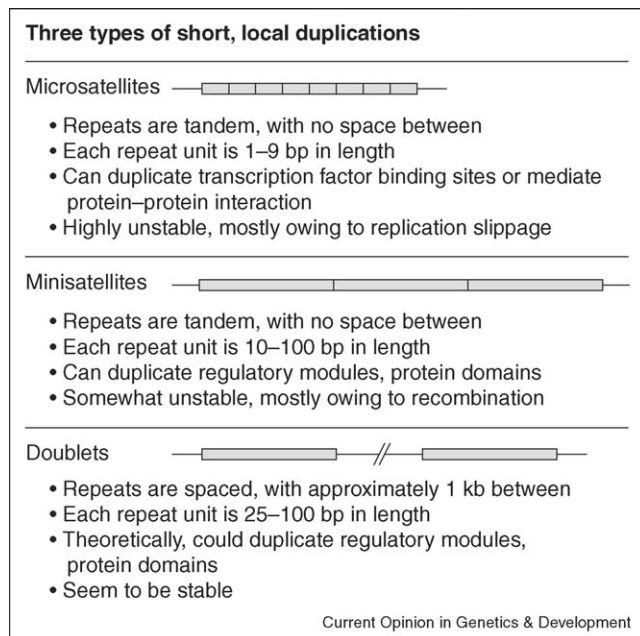
Microsatellites

Microsatellites are tandem arrays composed of multiple copies of a short repeat unit (generally 1–9 bp). They are the most unstable type of short, local duplications. Long microsatellites, with high numbers of copies of the repeating unit, are more abundant in eukaryotic genomes than would be expected by chance [1]. The length of the repeating unit, its sequence, and the number of consecutive repeat units affect the instability of microsatellites [2,3]. Point mutations that disrupt the periodic pattern increase microsatellite stability. It has been suggested that several mechanisms that involve replication, recombination and/or repair might lead to expansion and contraction of microsatellites, but it is not clear which are the most important [2–5]. In general, microsatellites are both highly abundant and highly variable.

It is clear that the expansion of microsatellites in either coding or non-coding sequences can have deleterious effects, and at least 13 different human diseases have been associated with the expansion of coding CTG/CAG or CCG/CGG repeats [2,3,6]. However, microsatellite expansions can have more subtle phenotypic effects. In prokaryotes, microsatellites have been found that modulate the expression or amino acid sequence of virulence factors, thereby accelerating the evolutionary rate of these genes [7,8**].

The domesticated dog has undergone strong artificial selection for a variety of morphological traits. Fondon and Garner [9**] hypothesized that gene-associated microsatellites have probably contributed to the resulting phenotypic variation. They looked at orthologous microsatellites in both human and dog and found that, overall, the microsatellites of both species were equally likely to have their periodic patterns interrupted by point mutations. In a subset of 17 developmental genes known or suspected to be involved in morphological development, dog microsatellites are considerably less likely to have

Figure 1



Three types of short, local duplications.

point mutations than those of human, indicating that they might have undergone recent changes in length, and that they are less stable than the orthologous human repeats. The authors sequenced the repeats from 92 different dog breeds and found that 15 of the 17 genes contain at least two microsatellite alleles. In the *Alx-4* gene, they found a repeat allele, in the Great Pyrenees breed, associated with the presence of an extra digit on both hind paws. It appears that alleles of this gene with differences in microsatellite length differ both in their ability to bind the LEF-1 protein and in their expression pattern in the limb bud [9[•],10]. The authors also found a pair of variable microsatellites in the *Runx-2* gene. The relative lengths of these two microsatellites are correlated with quantitative measures of skull morphology. These coding microsatellites, and others like them, might have played an important role in generating the phenotypic diversity currently seen in the domesticated dog.

Although many coding microsatellites have been implicated in human disease [6,8^{••},11], homopeptide tracts only seem to cause problems after a threshold amount of expansion — typically, to at least 30 copies [2,3,6,11] — and homopeptide tracts are found in many eukaryotic proteins [11,12]. Homopeptide tracts are not uniformly distributed throughout all functional classes of proteins; they are enriched in transcription factors, DNA-binding proteins and, more generally, proteins involved in large protein–protein or protein–nucleic acid complexes. This enrichment in certain functional groups suggests that

these tracts have a functional role. In some cases, it is known that the repeats are important in mediating protein–protein interactions [11–13]. An unstable microsatellite that mediates the ability of a protein to form protein–protein interactions can lead to individual differences in the wiring of the protein–interaction network, facilitating network evolution [12].

Microsatellites can also affect a gene in a regulatory fashion. Prairie and montane voles are closely related species that vary both in their social structure and in their regulatory microsatellite content. Prairie voles are social, biparental and form life-long pair bonds. Montane voles are socially indifferent, lack paternal care and are promiscuous. One of the differences between these two species is variation in the length of a microsatellite upstream of the *vasopressin 1a receptor* gene (*avpr1a*). A luciferase reporter assay indicated that changes in this microsatellite modify the gene expression level in a cell-dependent manner [14]. Hammock *et al.* [15^{••},16] looked at the intraspecific variations in the lengths of this microsatellite in prairie voles. They found that males with the longer allele of the microsatellite had increased expression of *avpr1a* in specific areas of the brain, that they more frequently licked and groomed their pups, and that they showed increased preference for their partners over stranger females. This microsatellite appears to have facilitated rapidly evolving changes in mammalian behavior and social structure, mediated by changes in gene expression.

To identify variable microsatellites that might affect human gene expression, Iglesias *et al.* [17[•]] scanned through the human genome for all microsatellites within 10 kb upstream of annotated genes. They checked for polymorphisms in this group and found 51 polymorphic microsatellites. The most interesting is the class of CTC microsatellites, which tend to be within 1 kb of the transcription start site. The class of genes whose members have at least five tandem copies of a CTC sequence is significantly enriched for TATA-less genes, indicating that these repeats might be important in the initiation of transcription for this type of gene.

To study the role of tandem duplication in regulatory evolution, Sinha and Siggia [18] looked at 76 experimentally validated *cis*-regulatory modules in both *Drosophila melanogaster* and *Drosophila yakuba*. They found that more base pairs of variation between the two species were caused by insertions or deletions (indels) than by point mutations, and that indels tend to coincide with microsatellites. They also found a significant overlap between transcription factor binding sites and microsatellites. Taken together, these results suggest that microsatellites have played an important role in the evolution of the *Drosophila* genome and have impacted the evolution of regulatory sequences.

The inherent instability of microsatellite sequences, combined with the ability of some microsatellites to alter protein–protein or DNA–protein interactions, makes them an important contributor to quantitative genetic variation [8^{••},19].

Minisatellites

Minisatellites are similar to microsatellites, but with a longer repeat unit, typically 10–150 bp in length. They can have as few as two copies of the repeat unit. Minisatellites are more stable than microsatellites, and expand and contract as a result of recombination more frequently than they do by replication slippage. Minisatellites vary a lot in their stability, and some work has been done to identify hypervariable minisatellites. Parameters that increase their rates of polymorphism include GC content and the purity of the repeating units [20].

Minisatellites have been implicated in genomic imprinting, chromosome pairing, and fragile chromosome sites [21]. They can be found within or near genes, and in some cases they are known to affect gene function by modifying either gene regulation or protein sequence. Many proteins contain tandem protein domains, and the simplest explanation for this phenomenon is that they are derived from historical, minisatellite, tandem duplications [22–24]. Minisatellites enable a gene to amplify an aspect of its function by adding additional copies of a functional motif, domain or exon.

One example of this is seen in the *FLO1* gene in *Saccharomyces cerevisiae* [25^{••}]. This gene encodes a cell-surface adhesin that enables the organism to adhere to other organisms and to its environment. Within the gene is a coding minisatellite with a repeat unit of approximately 100 bp. The copy number of the repeat varies between wild yeast strains. The authors tested a spectrum of constructs with different copy numbers and found that long alleles correlate with increased adherence between the organism and either a polystyrene substrate or another yeast cell. Presence of this variable, coding minisatellite can lead to evolutionarily rapid changes in adherence, important in maintaining a balance between staying in a rich environment and exploring a new environment. In pathogenic fungi, modulating adhesion factors is important both in populating a host and in evading the immune system.

One example of a functional human minisatellite polymorphism is found in the promoter of the *reduced folate carrier (RFC)* gene, which is important in cellular uptake of members of the antifolate class of cancer drugs [26]. There is a minisatellite within the promoter region of this gene, and the repeat unit contains binding sites for two different transcription factors. Alleles with more repeat units have been shown to drive a statistically significant increase in gene expression *in vitro*, indicating that this

polymorphism is a potentially important predictor of patient response to antifolate cancer drugs.

The *TNFRSF11A* (*tumour necrosis factor receptor superfamily member 11a precursor*) gene seems to be a hot spot for the creation of two-copy minisatellites. Three related skeletal disorders have been traced back to tandem duplications in the same region of the gene [27–30]. Each of these insertions is of a different length, with a different repeat unit. All of the repeat units have the same 3' boundary and overlap the signal peptide of the RANK (receptor activator of nuclear factor- κ) protein, which is encoded by this gene. At least two of the insertions have been shown to prevent cleavage of the signal peptide, leading to increased activation of the downstream signaling pathway [27].

Minisatellites are more stable than microsatellites, and their greater information content enables them to duplicate more complex functional modules. In a similar manner to microsatellites, their expansion and contraction can lead to either subtle or drastic phenotypic changes.

Doublets

Doublets are a recently described type of short, local duplication [31^{••}]. Doublets have been defined as pairs of identically duplicated sequence, at least 25 bp in length, found in a background of unique sequence at distances of between 100 bp and 10 kb, but typically tightly concentrated around 1 kb. The longest exact repeated sequence observed is approximately 300 bp, but most are much shorter. They appear to be distinct from long, segmental duplications.

In a similar fashion to microsatellites and minisatellites, doublets should be able to amplify modular functions of a gene. The doublet-creation process seems to work by copying a short patch of DNA from one region and inserting it within about 1 kb of the original sequence. Paired, spaced duplications matching the characterization of doublets are seen in a variety of eukaryotes, from plants to nematodes to humans, suggesting that the doublet-creation process is ubiquitous.

In an earlier study, Achaz *et al.* [32] found similarly spaced paired duplications across a variety of eukaryotic genomes and proposed that these duplications arise as tandem duplications that are later dispersed by chromosomal rearrangements [32,33]. Comparative sequence analysis of human doublets indicates that this is not the case; rather, the second copy of the repeat unit is inserted at a distance from the original without affecting the intervening sequence [31^{••}].

These types of short, local duplications could give rise to the scrambled patterns of transcription factor binding

sites seen in some comparative promoter analyses. The promoter of the *hunchback* gene in insects is one example of a promoter in which the functional motifs have had their relative orientation and spacing changed over time, as is evident in a comparison of motif sequences and locations in the *hunchback* promoters of *Musca domestica* and *Drosophila melanogaster* [34].

There are two significant ways in which doublets differ from microsatellites and minisatellites in their evolutionary potential and dynamics. First, no sequence signature has been found that leads to the creation of a doublet; it seems that any sequence could potentially be duplicated to become a doublet. Second, there is no evidence that doublets are unstable; the doublet creation process leads to an increase in the local copy number of a sequence without introducing instability.

Conclusions

Microsatellites, minisatellites and doublets are three classes of short, local duplications that can modulate the local copy number of a functional sequence. Both microsatellites and minisatellites have been shown to cause intraspecific phenotypic variations by altering either protein coding sequence or gene regulation, thereby providing a substrate for evolution. It is not yet clear whether doublets play a similarly important role.

Acknowledgements

EE Thomas thanks Nathan Srebro and Dana Pe'er for helpful comments. EE Thomas is funded by the NIGMS.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- of special interest
 - of outstanding interest
1. Dieringer D, Schlotterer C: **Two distinct modes of microsatellite mutation processes: evidence from the complete genomic sequences of nine species.** *Genome Res* 2003, **13**:2242-2251.
 2. Lenzmeier BA, Freudenreich CH: **Trinucleotide repeat instability: a hairpin curve at the crossroads of replication, recombination, and repair.** *Cytogenet Genome Res* 2003, **100**:7-24.
 3. Cleary JD, Pearson CE: **The contribution of cis-elements to disease-associated repeat instability: clinical and experimental evidence.** *Cytogenet Genome Res* 2003, **100**:25-55.
 4. Richard G-F, Paques F: **Mini- and microsatellite expansions: the recombination connection.** *EMBO Rep* 2000, **1**:122-126.
 5. Cleary JD, Pearson CE: **Replication fork dynamics and dynamic mutations: the fork-shift model of repeat instability.** *Trends Genet* 2005, **21**:272-280.
 6. Cummings CJ, Zoghbi HY: **Trinucleotide repeats: mechanisms and pathophysiology.** *Annu Rev Genomics Hum Genet* 2000, **1**:281-328.
 7. van Belkum A, Scherer S, van Alphen L, Verbrugh H: **Short-sequence DNA repeats in prokaryotic genomes.** *Microbiol Mol Biol Rev* 1998, **62**:275-293.
 8. Li Y-C, Korol AB, Fahima T, Nevo E: **Microsatellites within genes: structure, function, and evolution.** *Mol Biol Evol* 2004, **21**:991-1007.
A great review of coding and regulatory microsatellites in both prokaryotes and eukaryotes. It addresses their impact in different parts of genes, the mechanisms by which they evolve, and the evolutionary consequences.
 9. Fondon JW, Garner HR: **Molecular origins of rapid and continuous morphological evolution.** *Proc Natl Acad Sci USA* 2004, **101**:18058-18063.
The authors find differences between dog and human in the repeat purity of microsatellites within important developmental genes. They investigate polymorphic microsatellites in dog *Alx-4* and *Runx-2* and associate both genes with phenotypic diversity.
 10. Boras K, Hamel PA: **Alx4 binding to LEF-1 regulates N-CAM promoter activity.** *J Biol Chem* 2002, **277**:1120-1127.
 11. Faux NG, Bottomley SP, Lesk AM, Irving JA, Morrison JR, Garcia de la Banda M, Whisstock JC: **Functional insights from the distribution and role of homopeptide repeat-containing proteins.** *Genome Res* 2005, **15**:537-551.
 12. Hancock JM, Simon M: **Simple sequence repeats in proteins and their significance for network evolution.** *Gene* 2004, **345**:113-118.
 13. Alba MM, Guigo R: **Comparative analysis of amino acid repeats in rodents and humans.** *Genome Res* 2004, **14**:549-554.
 14. Hammock EA, Young LJ: **Functional microsatellite polymorphism associated with divergent social structure in vole species.** *Mol Biol Evol* 2004, **21**:1057-1063.
 15. Hammock EAD, Young LJ: **Microsatellite instability generates diversity in brain and sociobehavioral traits.** *Science* 2005, **308**:1630-1634.
The authors find a link between length of a microsatellite in the promoter of the *avpr1a* gene and intraspecific differences in male prairie vole behavior.
 16. Hammock EAD, Lim MM, Nair HP, Young LJ: **Association of vasopressin 1a receptor levels with a regulatory microsatellite and behavior.** *Genes Brain Behav* 2005, **4**:289-301.
 17. Iglesias AR, Kindlund E, Tammi M, Wadelius C: **Some microsatellites may act as novel polymorphic cis-regulatory elements through transcription factor binding.** *Gene* 2004, **341**:149-165.
The authors identify all microsatellites within 10 kb upstream of human genes. They also identify a polymorphic subset, and then use gel shift assays to identify protein-microsatellite interactions.
 18. Sinha S, Siggia ED: **Sequence turnover and tandem repeats in cis-Regulatory modules in Drosophila.** *Mol Biol Evol* 2005, **22**:874-885.
 19. Kashi Y, King D, Soller M: **Simple sequence repeats as a source of quantitative genetic variation.** *Trends Genet* 1997, **13**:74-78.
 20. Denoeud F, Vergnaud G, Benson G: **Predicting human minisatellite polymorphism.** *Genome Res* 2003, **13**:856-867.
 21. Vergnaud G, Denoeud F: **Minisatellites: mutability and genome architecture.** *Genome Res* 2000, **10**:899-907.
 22. Marcotte EM, Pellegrini M, Yeates TO, Eisenberg D: **A census of protein repeats.** *J Mol Biol* 1999, **293**:151-160.
 23. Andrade MA, Perez-Iratxeta C, Ponting CP: **Protein repeats: structures, functions, and evolution.** *J Struct Biol* 2001, **134**:117-131.
 24. Ponting CP, Mott R, Bork P, Copley RR: **Novel protein domains and repeats in Drosophila melanogaster: Insights into structure, function, and evolution.** *Genome Res* 2001, **11**:1996-2008.
 25. Verstrepen KJ, Jansen A, Lewitter F, Fink GR: **Intragenic tandem repeats generate functional variability.** *Nat Genet* 2005, **37**:986-990.
The authors find an enrichment of protein-coding microsatellites and minisatellites in cell-wall proteins in *Saccharomyces cerevisiae* and study the phenotypic consequences of changes in minisatellite length in the *Flo1* gene.

26. Whetstone JR, Witt TL, Matherly LH: **The human reduced folate carrier gene is regulated by the AP2 and Sp1 transcription factor families and a functional 61-base pair polymorphism.** *J Biol Chem* 2002, **277**:43873-43880.
27. Hughes AE, Ralston SH, Marken J, Bell C, MacPherson H, Wallace RGH, van Hul W, Whyte MP, Nakatsuka K, Hovy L, Anderson DM: **Mutations in TNFRSF11A, affecting the signal peptide of RANK, cause familial expansile osteolysis.** *Nat Genet* 2000, **24**:45-48.
28. Johnson-Pais TL, Singer FR, Bone HG, McMurray CT, Hansen MF, Leach RJ: **Identification of a novel tandem duplication in exon 1 of the TNFRSF11A gene in two unrelated patients with familial expansile osteolysis.** *J Bone Miner Res* 2003, **18**:376-380.
29. Nakatsuka K, Nishizawa Y, Ralston SH: **Phenotypic characterization of early onset Paget's disease of bone caused by a 27 bp duplication in the TNFRSF11A gene.** *J Bone Miner Res* 2003, **18**:1381-1385.
30. Whyte MP, Mills BG, Reinus WR, Podgornik MN, Roodman GD, Gannon FH, Eddy MC, McAlister WH: **Expansile skeletal hyperphosphatasia: a new familial metabolic bone disease.** *J Bone Miner Res* 2000, **15**:2330-2344.
31. Thomas EE, Srebro N, Sebat J, Navin N, Healy J, Mishra B, Wigler M: **Distribution of short paired duplications in mammalian genomes.** *Proc Natl Acad Sci USA* 2004, **101**:10349-10354.
- The authors introduce doublets and show that they are distinct from other types of repeats.
32. Achaz G, Coissac E, Viari A, Netter P: **Analysis of intrachromosomal duplications in yeast *Saccharomyces cerevisiae*: a possible model for their origin.** *Mol Biol Evol* 2000, **17**:1268-1275.
33. Achaz G, Netter P, Coissac E: **Study of intrachromosomal duplications among the eukaryote genomes.** *Mol Biol Evol* 2001, **18**:2280-2288.
34. Hancock JM, Shaw PJ, Bonneton F, Dover GA: **High sequence turnover in the regulatory regions of the developmental gene *hunchback* in insects.** *Mol Biol Evol* 1999, **16**:253-265.

Free journals for developing countries

The WHO and six medical journal publishers have launched the Access to Research Initiative, which enables nearly 70 of the world's poorest countries to gain free access to biomedical literature through the Internet.

Gro Harlem Brundtland, director-general for the WHO, said that this initiative was 'perhaps the biggest step ever taken towards reducing the health information gap between rich and poor countries'.

For more information, visit www.healthinternet.net



ELSEVIER

Chromosomal sex-determining regions in animals, plants and fungi

James A Fraser and Joseph Heitman

The independent evolution of sex chromosomes in many eukaryotic species raises questions about the evolutionary forces that drive their formation. Recent advances in our understanding of these genomic structures in mammals in parallel with alternate models such as the monotremes, fish, dioecious plants, and fungi support the idea of a remarkable convergence in structure to form large, non-recombining regions with discrete evolutionary strata. The discovery that evolutionary events similar to those that have transpired in humans have also occurred during the formation of sex chromosomes in organisms as divergent as the plant *Silene*, the fungus *Cryptococcus* and the fish medaka highlights the importance of future studies in these systems. Such investigation will broaden our knowledge of the evolution and plasticity of these ubiquitous genomic features underlying sexual dimorphism and reproduction.

Addresses

Departments of Molecular Genetics and Microbiology, Medicine, and Pharmacology and Cancer Biology, Howard Hughes Medical Institute, Duke University Medical Center, Durham, NC 27710, USA

Corresponding author: Heitman, Joseph (heitm001@duke.edu)

Current Opinion in Genetics & Development 2005, **15**:645–651

This review comes from a themed issue on
Genomes and evolution
Edited by Stephen J O'Brien and Claire M Fraser

Available online 22nd September 2005

0959-437X/\$ – see front matter

© 2005 Elsevier Ltd. All rights reserved.

DOI 10.1016/j.gde.2005.09.002

Introduction

A long-considered fundamental question of genomic biology is how different species have evolved the complex structures recognized as sex chromosomes. This question has been driven by the observations that the sex chromosomes of different organisms vary wildly, with multiple examples indicating the independent evolution of systems that have ultimately undergone a convergence of structure and function — the evolution of heteromorphic structures that share multiple features and are the cornerstones of sex determination.

Almost 40 years ago, Ohno proposed how animal sex chromosomes might have been forged [1]. In this model, the original master sex-determining gene was autosomal

and was only later captured onto the incipient sex chromosome in conjunction with other genes encoding sex-related functions. Building on this more discrete genomic foundation, the evolution of distinct sex chromosomes was then proposed to have occurred by suppression of recombination — by mechanisms including inversions — leading to the divergence of large genomic tracts and the emergence of heteromorphic sex chromosomes. Gene capture and suppression of recombination, therefore, punctuate sex chromosome evolution, because differentiation of the nascent sex chromosomes is only accelerated when recombination is decreased.

Because synonymous substitutions accumulate at a linear rate over time, comparison of their occurrence in sex chromosome genes reveals how long such genes have been diverging from a common ancestor. These analyses can, therefore, be used to date evolutionary events, such as inversions, that have suppressed recombination and driven the formation of genomic sex-determining regions, ushering in a new era in the understanding of these complex genomic regions.

In this review, we focus on the recent advances in our understanding of sex chromosome evolution, and on the remarkable similarities present in highly divergent model organisms.

From humans to birds

The evolutionary history of the human sex chromosomes is a paradigmatic example of the formation of genomic sex-determining regions. In humans, ~300 million years ago the ancestral Y chromosome was an autosome on which recombinational suppression was limited to a small portion of the chromosome around the sex-determining *SRY* gene [1,2]. The male-specific ~50–60 Mb Y chromosome subsequently evolved by chromosomal rearrangements, gene conversion, duplication and degeneration over ~300 million years [2,3]. Analysis of synonymous substitution rates between X- and Y-linked gene homologues reveals five temporal clusters, or gene strata, representing the sequential acquisition of genes to the male-specific region [3,4••]. Other genes have been lost through a process of Y degradation caused by the absence of recombination in the male-specific region [5]. All that remains of once extensive recombination between the X and the Y are two small, remnant pseudoautosomal regions.

Unlike the Y chromosome, the human X is highly stable and is colinear with the canine X, and it only shows a low

level of rearrangement in rodent lineages, in agreement with predictions that the human X is identical to the putative ancestral eutherian X chromosome [4**,6]. Genomic comparisons have also identified steps in the evolution of this more stable sex chromosome, including the translocation of new genetic material (the 'X-added region') from an autosome before the radiation of the eutherians ~105 million years ago. This material has since been engulfed by the sex-specific region and become highly degenerate on the Y. Also, there is evidence for transposition of non-pseudoautosomal X sequences to the Y within the past ~4.7 million years, again with the subsequent degradation of the Y copy, highlighting the rapid degeneration of sequences on this very different structure [4**]. This degeneration of the Y chromosome in the absence of recombination is emphasized by the observation that the X chromosome now only shares ~25 genes with the Y outside of the pseudoautosomal recombining regions.

The sequence of the human X chromosome still holds many secrets to be revealed. An example of this is the recently discovered relationship between retroposed genes and this sex chromosome. A genome-wide survey has shown that the X chromosome has generated and recruited a disproportionately high number of functional retroposed genes in both human and mouse, possibly owing to differences in chromatin structure relative to that in other chromosomes [7,8].

Are the human X and Y representative of sex chromosome evolution? Recent studies in the mouse have confirmed the presence of equivalent strata, indicating that these structures might be common on eutherian sex chromosomes [9]. However, these species are relatively closely related. In birds, a ZZ/ZW chromosomal system evolved during the past 300 million years and governs the establishment of sexual identity in a fashion opposite to that in humans; in birds, females are the heterogametic sex (ZW) and males the homogametic (ZZ). The bird Z and W sex chromosomes evolved from a different pair of autosomes than did the human X and Y: the chicken Z has homology to human chromosome 9, whereas the human X has homology to chicken chromosomes 1 and 4 [4**,10,11]. Nevertheless, these systems share several similarities: the X and Z are both large and gene-rich, whereas their Y and W counterparts are small and gene-poor. Both the Y and the W maintain gene families involved in sex determination by gene conversion of amplicons [12,13]. Moreover, the forces that created these structures were similar, and analysis of the chicken genome unveiled the existence of at least two evolutionary strata on the Z chromosome [14].

Until recently, it was thought that bird and eutherian sex chromosomes evolved independently. However, landmark studies of the duck-billed platypus (*Ornithorhynchus anatinus*) dispelled this notion, unifying sex chromosome

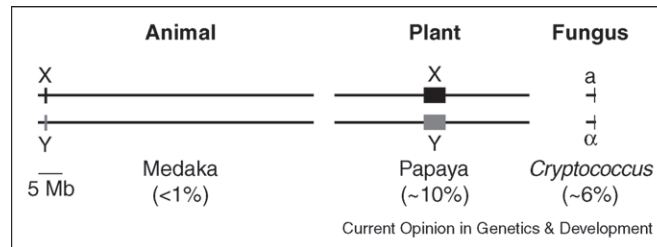
evolutionary models. Mammals diverged from birds ~300–350 million years ago [15]. The platypus is a monotreme, and diverged from the mammalian lineage ~210 million years ago. This was the earliest offshoot from the mammalian lineage, ~30 million years before the divergence of the marsupial and placental mammals [16], providing a unique evolutionary vantage point. In this unusual Australian mammal, the male karyotype is 21 pairs of chromosomes and 10 unpaired sex chromosomes. These ten chromosomes share short regions of homology that enable them to form an $X_1Y_1X_2Y_2X_3Y_3X_4Y_4X_5Y_5$ multivalent chain at male meiosis, adopting an alternating pattern to segregate into $X_1X_2X_3X_4X_5$ - and $Y_1Y_2Y_3Y_4Y_5$ -bearing sperm [17,18**].

The unique translocation chain in the platypus has been hypothesized to have formed in a step-by-step fashion, beginning with a translocation between a heteromorphic sex chromosome pair and an autosome, which subsequently continued by serial sex chromosome-autosome translocations [18**]. The location of homologues of human X-linked genes in the platypus genome provides evidence of a further translocation, suggesting that sex chromosome rearrangement in monotremes is ongoing [19]. How does this fascinating system shed light on the evolution of sex chromosomes in other animals? In the multivalent chain, the largest chromosome, which lies at one end of the chain, contains homologues of multiple genes from the human X, whereas a much smaller chromosome that lies near the other end has homology to the bird Z chromosome [18**]. Although, to date, no sex chromosomes have been sequenced to completion, these data indicate a previously unforeseen link between mammalian and bird sex chromosome evolution that will be elaborated further as more data becomes available.

A simpler view of sex chromosomes

Although animal sex chromosomes are fascinating, they are ancient and large and, therefore, more difficult to study. Characteristics of other, less developed systems provide further insights (Figure 1). For example, a high repeat content is a common feature in sex chromosomes from many species. This is highlighted in certain vole species, in which the sex chromosomes have expanded to unusual size. These 'giant sex chromosomes' are at their most extreme in the European field vole *Microtus agrestis*, in which the X chromosome represents nearly 20% of the genome [20], and this enlargement is the result of the rapid accumulation of repetitive sequences to form large heterochromatic blocks [21]. Smaller sex chromosomes also yield informative insights: for example, studies in the fish medaka (*Oryzias latipes*) reveal a sex chromosome at an earlier evolutionary stage (<10 million years) [22]. Here, the Y chromosome was formed by a duplication of the Y-specific DM domain gene, *DMY* (also known as *Dmrt1*) [23,24], which served as the seed for the formation of an ~250 kb male-specific region that is rich in repe-

Figure 1



Proto-sex chromosomes in animals, plants and fungi. Early genomic sex-determining regions identified in animals such as the fish medaka (*Oryzias latipes*), plants such as papaya (*Carica papaya*) and fungi such as *Cryptococcus* reveal that the early stages of sex chromosome evolution occur through common mechanisms in these very different eukaryotes. The sex-specific regions are labeled (X and Y, or a and α), and the approximate proportion of the proto-sex chromosome occupied by these regions is given as a percentage.

titive elements and which represents less than 1% of the chromosome upon which it resides [25]. Similarly, in the three-spine stickleback (*Gasterosteus aculeatus*), the master sex-determining locus maps to a small, repeat-rich sex-specific region on one end of chromosome 19, in a region that became sex-specific less than 10 million years ago [26,27]. Therefore, these fish represent a unique system in which to study the early steps of sex chromosome evolution in animals.

How rapidly can a species develop discrete sex chromosomes? Discoveries in *Drosophila* species have indicated that this can occur over a relatively short evolutionary time-span. Comparison of *Drosophila melanogaster* [28] and *Drosophila pseudoobscura* [29] uncovered that the original Y chromosome has become an autosome in this divergent lineage, and that a new chromosome has evolved to take its place [30]. Although the precise mechanism behind this genomic plasticity is not entirely clear, comparison with other closely related *Drosophila* species has revealed that the process has occurred in less than 18 million years.

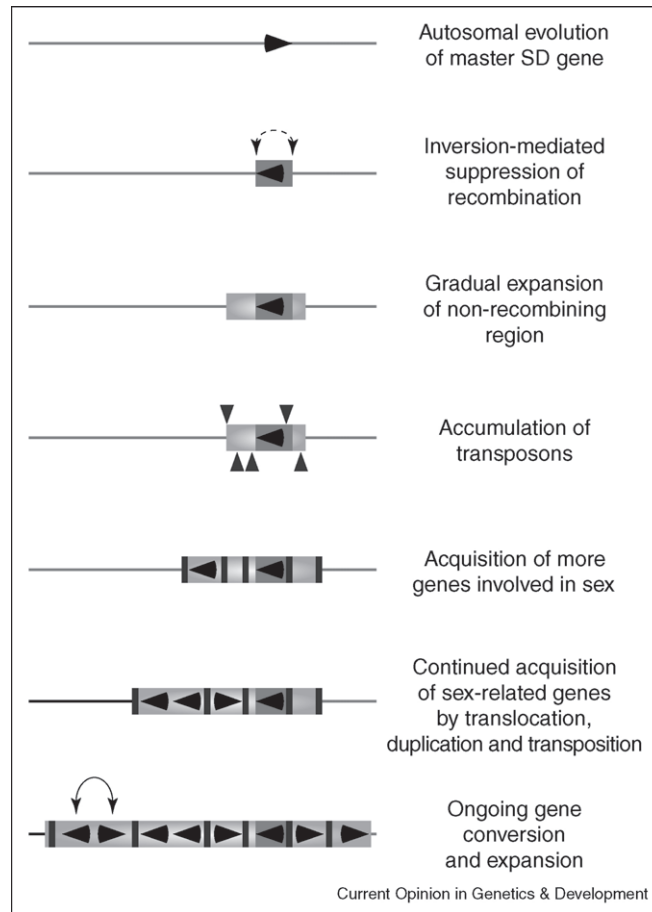
These mechanisms of sex chromosome evolution are widespread in other multicellular eukaryotes (Figure 2). Studies in papaya (*Carica papaya*) have identified primitive, repeat-rich sex chromosomes that bear a 4.4 Mb male-specific region on one chromosome (~10% of its length; Figure 1) [31,32]. However, much more complex structures have been observed in other plants. Perhaps best understood at this time is the structure of the sex chromosomes of the dioecious plants belonging to the genus *Silene*. Studies of X- and Y-linked alleles of several genes in *Silene latifolia*, *Silene dioica* and *Silene diclinis* have shown that these heteromorphic sex chromosomes have evolutionary strata — similar to eutherian chromosomes but, in this case, three strata rather than the eutherian five — indicating evolution by a similar mechanism of gradual expansion of sex-specific chromosomal regions to produce fully fledged sex chromosomes [33,34].

All of these sex chromosome models involve diploid eukaryotes. However, work in the dioecious liverwort *Marchantia polymorpha* has shown that the evolution of sex chromosomes is not dependent on a diploid genome. This haploid plant also has heteromorphic sex chromosomes — in this case, the Y is larger than the X — although individuals bear only one chromosome or the other. Studies of the Y have revealed that male-specific repeats occupy at least a quarter of the 10 Mb chromosome, interspersed with copies of a sex-specific RING-finger gene family. Therefore, although no more detail is known about this structure, it clearly shares similarities with the Y chromosome of eutherians [35].

Revelations from fungal mating-type loci

Are these chromosomal sex-determining features also present in the haploid, genetically amenable fungi, which have much smaller genomes? In contrast to animals and plants, fungal cell-type identity and sexual cycle are orchestrated by a more restricted chromosomal region, known as the mating type (*MAT*) locus. However, in several cases, clear parallels can be drawn between the structures of *MAT* and of animal sex chromosomes [36,37]. An example is the basidiomycete *Cryptococcus*, a ubiquitous human fungal pathogen with a bipolar mating-type system involving haploid a and α cells. Sequencing of the *Cryptococcus MAT* locus revealed it is unusually large (>100 kb), contains >20 genes and is composed almost entirely of divergent alleles of a common gene set [38,39]. *MAT* represents ~6% of the chromosome upon which it resides [40] and, therefore, represents a proportionately larger sex-specific region than do some 'sex chromosomes' in animals such as the medaka (Figure 1). The unusual structure of the *Cryptococcus MAT* locus has been characterized in a and α isolates of a *Cryptococcus* species-cluster diverging over 40 million years. Although the size and content of *MAT* is conserved, gene positions within *MAT* are highly rearranged, possibly either as a consequence or as a driving force of speciation. Using techniques similar to those applied to studies of animal and plant sex chromosomes, including analysis of

Figure 2



Birth of a sex chromosome. Studies in a wide variety of eukaryotes have revealed that sex chromosomes evolve by a series of common evolutionary mechanisms. Starting with a master sex-determining (SD) gene on an autosome, the chromosome becomes sex-specific through a series of evolutionary events including, but not restricted to, inversion-mediated suppression of recombination, acquisition of additional genes involved in the sexual cycle by transposition and translocation, and accumulation of transposable elements.

gene order, phylogeny and synonymous substitution rates, the evolutionary steps that fashioned *MAT* were deduced. First, *MAT* expanded by the acquisition of sex-related genes into two unlinked clusters on different autosomes. Next, these independent regions fused to form a single, larger sex-determining genomic structure in one mating type. The opposite mating type was then converted to an equivalent structure by gene conversion or recombination between the linked and the unlinked sex-determining regions. These steps are highlighted by the presence of four evolutionary gene strata similar to those seen in mammals, birds and plants. Finally, *MAT* was then subjected to ongoing intra- and inter-allelic gene conversion, and inversions that suppress recombination, thereby giving rise to the extant structures of today [39••].

Therefore, the evolution of *MAT* in *Cryptococcus* shows remarkable parallels to the evolution of the human Y chromosome (Table 1) [39••]. As with animal and plant

sex chromosomes, *MAT* is rich in repeated elements, with transposons and other repeats representing ~25% of the structure, roughly five times more common than their occurrence in other genomic regions. Beyond these features common to most sex chromosomes, *MAT* has even further similarity to the human Y. Analysis of the phylogeny of each gene in the *MAT* locus identified ancient boundaries of the expanding sex-determining regions where the part of a gene on one side of the boundary is *MAT*-specific and part on the other is not — a similar scenario to the *amelogenin* locus, which spans an ancient pseudoautosomal boundary in primates [41]. And finally, as with duplicated genes involved in fertility on the human Y chromosome [12], the *Cryptococcus* pheromone genes are present in multiple copies and arranged in inverted repeats that are maintained by intra-allelic gene conversion, enabling gene repair in the absence of a paired chromosome for recombinational repair. In short, the *Cryptococcus MAT* locus bears many of the hallmarks of the evolutionary events responsible for the formation of

Table 1

Conserved features of sex-chromosome evolution in animals, plants, and fungi.

1	Master sex determinants emerged on autosomes and served as the origin of a non-recombining sex chromosome or <i>MAT</i> locus.
2	The sex-determining region is composed of gene strata of distinct evolutionary ages, which might have been rearranged.
3	Coherence of gene function in the sex-determining genomic locus.
4	Sex-unique genes are duplicated in inverted repeats, enabling repair by intrachromosomal recombination.
5	These regions of the genome are sheltered from recombination and, therefore, accumulate transposable elements at a higher level than elsewhere in the genome, which might contribute to evolution and rearrangement.
6	New genetic material is introduced by translocations from autosomes, creating sex-specific alleles of genes not previously linked to sex.

the human Y chromosome, and all within a structure that is only a fraction of the size (Table 1).

Cryptococcus represents one of several examples of convergent evolution of complex genomic sex-determining regions in fungi. In the basidiomycete *Ustilago hordei*, the two small unlinked *MAT* loci — known as a tetrapolar mating system — normally found in species of this phylum are, instead, linked, forming a single larger locus — a bipolar system — that spans ~500 kb, more than one-sixth of the ~2.8 Mb chromosome on which it resides [42,43]. Again, this expansion of the sex-determining region in a bipolar system to occupy a substantial portion of a chromosome is a key feature shared with mammalian sex chromosomes. Furthermore, recombination is suppressed across the entire region, which is rich in transposons and repetitive elements (JW Kronstad, unpublished). A similar scenario has been observed in the ascomycete *Neurospora tetrasperma*, in which one allele of the mating-type locus has undergone a dramatic rearrangement and now occupies a large recombination-suppressed region of its host chromosome [44,45]. Finally, in the basidiomycete *Microbotryum violaceum*, the mating-type chromosomes are large (2.9 Mb and 3.5 Mb, for the A1 and A2 chromosomes, respectively), exhibit clear heteromorphism and are also rich in repetitive elements [36,46,47].

Therefore, the fungi represent a unique window on the evolution of sex chromosomes, and the genetic and genomic tractability of these organisms make them attractive systems in which to dissect further this complex, fascinating evolutionary drama played out on a genomic stage.

Conclusions

Our understanding of the evolutionary forces that drive the formation of sex chromosomes has grown dramatically over the past decade. Although the foundation of our knowledge has come from the study of these structures in humans, much has come from newer and less traditional systems. Our understanding of sex chromosome evolution is growing rapidly: ranging from the bewildering yet fascinating evolution of the ‘conga-line’ sex chromosomes of the platypus to the remarkably animal-like chromosomes in plants and, finally, to the small yet complex fungal mating-type loci.

Although we can certainly expect further enlightenment of this process from more traditional sources, the complexity of the story of sex chromosome evolution is gaining chapters from some very unexpected and unlikely sources. Further genomic information, whether from the platypus genome project, the fungal genome initiative or elsewhere, will continue to illuminate this fascinating saga of convergent genomic evolution in highly divergent eukaryotes.

Acknowledgements

This work was supported by the National Institutes for Health and the National Institute of Allergy and Infectious Disease R01 grant AI50113 to JH.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- of special interest
- of outstanding interest

1. Ohno S: *Sex Chromosomes and Sex-linked Genes*. New York: Springer-Verlag; 1967.
 2. Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, Repping S, Pyntikova T, Ali J, Bieri T *et al.*: **The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes.** *Nature* 2003, **423**:825-837.
 3. Lahn BT, Page DC: **Four evolutionary strata on the human X chromosome.** *Science* 1999, **286**:964-967.
 4. Ross MT, Grafham DV, Coffey AJ, Scherer S, McLay K, Muzny D, **Platzer M, Howell GR, Burrows C, Bird CP *et al.*: The DNA sequence of the human X chromosome.** *Nature* 2005, **434**:325-337.
- This analysis of the structure of the human X chromosome sequence reports the presence of an additional evolutionary strata, traces multiple additional events in the divergence of X and Y, and provides insight into the evolution of sex chromosomes in many other animals.
5. Charlesworth B: **The evolution of sex chromosomes.** *Science* 1991, **251**:1030-1033.
 6. Hillier LW, Miller W, Birney E, Warren W, Hardison RC, Ponting CP, Bork P, Burt DW, Groenen MA, Delany ME *et al.*: **Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution.** *Nature* 2004, **432**:695-716.
 7. Emerson JJ, Kaessmann H, Betran E, Long M: **Extensive gene traffic on the mammalian X chromosome.** *Science* 2004, **303**:537-540.
 8. Khil PP, Oliver B, Camerini-Otero RD: **X for intersection: retrotransposition both on and off the X chromosome is more frequent.** *Trends Genet* 2005, **21**:3-7.
 9. Sandstedt SA, Tucker PK: **Evolutionary strata on the mouse X chromosome correspond to strata on the human X chromosome.** *Genome Res* 2004, **14**:267-272.

10. Nanda I, Shan Z, Scharlt M, Burt DW, Koehler M, Nothwang H, Grutzner F, Paton IR, Windsor D, Dunn I *et al.*: **300 million years of conserved synteny between chicken Z and human chromosome 9.** *Nat Genet* 1999, **21**:258-259.
11. Fridolfsson AK, Cheng H, Copeland NG, Jenkins NA, Liu HC, Raudsepp T, Woodage T, Chowdhary B, Halverson J, Ellegren H: **Evolution of the avian sex chromosomes from an ancestral pair of autosomes.** *Proc Natl Acad Sci USA* 1998, **95**:8147-8152.
12. Rozen S, Skaletsky H, Marszalek JD, Minx PJ, Cordum HS, Waterston RH, Wilson RK, Page DC: **Abundant gene conversion between arms of palindromes in human and ape Y chromosomes.** *Nature* 2003, **423**:873-876.
13. Backstrom N, Cepplitis H, Berlin S, Ellegren H: **Gene conversion drives the evolution of *HINTW*, an ampliconic gene on the female-specific avian W chromosome.** *Mol Biol Evol* 2005, **22**:1992-1999.
14. Handley LJ, Cepplitis H, Ellegren H: **Evolutionary strata on the chicken Z chromosome: implications for sex chromosome evolution.** *Genetics* 2004, **167**:367-376.
15. Kumar S, Hedges SB: **A molecular timescale for vertebrate evolution.** *Nature* 1998, **392**:917-920.
16. Woodburne MO, Rich TH, Springer MS: **The evolution of tribospheny and the antiquity of mammalian clades.** *Mol Phylogenet Evol* 2003, **28**:360-385.
17. Rens W, Grutzner F, O'Brien PC, Fairclough H, Graves JA, Ferguson-Smith MA: **Resolution and evolution of the duck-billed platypus karyotype with an X₁Y₁X₂Y₂X₃Y₃X₄Y₄X₅Y₅ male sex chromosome constitution.** *Proc Natl Acad Sci USA* 2004, **101**:16257-16261.
18. Grutzner F, Rens W, Tsend-Ayush E, El-Mogharbel N, O'Brien PC, Jones RC, Ferguson-Smith MA, Marshall Graves JA: **In the platypus a meiotic chain of ten sex chromosomes shares genes with the bird Z and mammal X chromosomes.** *Nature* 2004, **432**:913-917.
- This landmark sex chromosome evolution study reports the characterization of a 10 sex chromosome system (5X and 5Y) in the platypus, and provides a crucial link between the models of sex chromosome evolution in placental mammals and birds.
19. Waters PD, Delbridge ML, Deakin JE, El-Mogharbel N, Kirby PJ, Carvalho-Silva DR, Graves JA: **Autosomal location of genes from the conserved mammalian X in the platypus (*Ornithorhynchus anatinus*): implications for mammalian sex chromosome evolution.** *Chromosome Res* 2005, **13**:401-410.
20. Nanda I, Neitzel H, Sperling K, Studer R, Epplen JT: **Simple GATCA repeats characterize the X chromosomal heterochromatin of *Microtus agrestis*, European field vole (*Rodentia, Cricetidae*).** *Chromosoma* 1988, **96**:213-219.
21. Marchal JA, Acosta MJ, Nietzel H, Sperling K, Bullejos M, Diaz de la Guardia R, Sanchez A: **X chromosome painting in *Microtus*: origin and evolution of the giant sex chromosomes.** *Chromosome Res* 2004, **12**:767-776.
22. Kondo M, Nanda I, Hornung U, Schmid M, Scharlt M: **Evolutionary origin of the medaka Y chromosome.** *Curr Biol* 2004, **14**:1664-1669.
23. Nanda I, Kondo M, Hornung U, Asakawa S, Winkler C, Shimizu A, Shan Z, Haaf T, Shimizu N, Shima A *et al.*: **A duplicated copy of *DMRT1* in the sex-determining region of the Y chromosome of the medaka, *Oryzias latipes*.** *Proc Natl Acad Sci USA* 2002, **99**:11778-11783.
24. Matsuda M, Nagahama Y, Shinomiya A, Sato T, Matsuda C, Kobayashi T, Morrey CE, Shibata N, Asakawa S, Shimizu N *et al.*: ***DMY* is a Y-specific DM-domain gene required for male development in the medaka fish.** *Nature* 2002, **417**:559-563.
25. Naruse K, Tanaka M, Mita K, Shima A, Postlethwait J, Mitani H: **A medaka gene map: the trace of ancestral vertebrate proto-chromosomes revealed by comparative gene mapping.** *Genome Res* 2004, **14**:820-828.
26. Peichel CL, Ross JA, Matson CK, Dickson M, Grimwood J, Schmutz J, Myers RM, Mori S, Schluter D, Kingsley DM: **The master sex-determination locus in threespine sticklebacks is on a nascent Y chromosome.** *Curr Biol* 2004, **14**:1416-1424.
- The authors describe the use of genome-wide linkage mapping to identify a nascent sex chromosome in the fish *Gasterosteus aculeatus*. The sex-specific region is a small region on the end of one chromosome and is not cytogenetically visible.
27. Filatov D: **Evolutionary genetics: stickleback's view of sex chromosome evolution.** *Heredity* 2005, **94**:275-276.
28. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF *et al.*: **The genome sequence of *Drosophila melanogaster*.** *Science* 2000, **287**:2185-2195.
29. Richards S, Liu Y, Bettencourt BR, Hradecky P, Letovsky S, Nielsen R, Thornton K, Hubisz MJ, Chen R, Meisel RP *et al.*: **Comparative genome sequencing of *Drosophila pseudoobscura*: chromosomal, gene, and cis-element evolution.** *Genome Res* 2005, **15**:1-18.
30. Carvalho AB, Clark AG: **Y chromosome of *D. pseudoobscura* is not homologous to the ancestral *Drosophila*.** *Science* 2005, **307**:108-110.
- This study describes a comparison of Y-linked genes between *D. melanogaster* and *D. pseudoobscura*, revealing the evolution of a new Y in *D. pseudoobscura* within the past 18 million years, corresponding with the old Y becoming autosomal.
31. Liu Z, Moore PH, Ma H, Ackerman CM, Ragiba M, Yu Q, Pearl HM, Kim MS, Charlton JW, Stiles JI *et al.*: **A primitive Y chromosome in papaya marks incipient sex chromosome evolution.** *Nature* 2004, **427**:348-352.
- This study describes an analysis of a primitive sex chromosome in *Carica papaya*. The male-specific region is only ~10% of the chromosome, providing evidence for the autosomal origin of these structures.
32. Charlesworth D: **Plant evolution: modern sex chromosomes.** *Curr Biol* 2004, **14**:R271-R273.
33. Nicolas M, Marais G, Hykelova V, Janousek B, Laporte V, Vyskot B, Mouchiroud D, Negritiu I, Charlesworth D, Moneger F: **A gradual process of recombination restriction in the evolutionary history of the sex chromosomes in dioecious plants.** *PLoS Biol* 2005, **3**:e4.
- A milestone in the study of plant sex chromosomes, this study revealed the existence of evolutionary strata on the sex chromosomes in multiple *Silene* species, showing these structures can evolve by a similar process in animals and plants.
34. Filatov DA: **Evolutionary history of *Silene latifolia* sex chromosomes revealed by genetic mapping of four genes.** *Genetics* 2005, **170**:975-979.
35. Okada S, Sone T, Fujisawa M, Nakayama S, Takenaka M, Ishizaki K, Kono K, Shimizu-Ueda Y, Hanajiri T, Yamato KT *et al.*: **The Y chromosome in the liverwort *Marchantia polymorpha* has accumulated unique repeat sequences harboring a male-specific gene.** *Proc Natl Acad Sci USA* 2001, **98**:9454-9459.
36. Hood ME, Antonovics J, Koskella B: **Shared forces of sex chromosome evolution in haploid-mating and diploid-mating organisms: *Microbotryum violaceum* and other model organisms.** *Genetics* 2004, **168**:141-146.
37. Fraser JA, Heitman J: **Evolution of fungal sex chromosomes.** *Mol Microbiol* 2004, **51**:299-306.
38. Lengeler KB, Fox DS, Fraser JA, Allen A, Forrester K, Dietrich FS, Heitman J: **Mating-type locus of *Cryptococcus neoformans*: a step in the evolution of sex chromosomes.** *Eukaryot Cell* 2002, **1**:704-718.
39. Fraser JA, Diezmann S, Subaran RL, Allen A, Lengeler KB, Dietrich FS, Heitman J: **Convergent evolution of chromosomal sex-determining regions in the animal and fungal kingdoms.** *PLoS Biol* 2004, **2**:e384.
- The authors describe the molecular dissection of the steps in the evolution of the unusually large mating-type locus in the pathogenic fungus *Cryptococcus*, revealing many similarities to the sex chromosomes of animals and plants.
40. Loftus BJ, Fung E, Roncaglia P, Rowley D, Amedeo P, Bruno D, Vamathevan J, Miranda M, Anderson IJ, Fraser JA *et al.*: **The genome of the basidiomycetous yeast and human**

- pathogen *Cryptococcus neoformans*. *Science* 2005, **307**:1321-1324.
41. Iwase M, Satta Y, Hirai Y, Hirai H, Imai H, Takahata N: **The *amelogenin* loci span an ancient pseudoautosomal boundary in diverse mammalian species.** *Proc Natl Acad Sci USA* 2003, **100**:5258-5263.
 42. Lee N, Bakkeren G, Wong K, Sherwood JE, Kronstad JW: **The mating-type and pathogenicity locus of the fungus *Ustilago hordei* spans a 500-kb region.** *Proc Natl Acad Sci USA* 1999, **96**:15026-15031.
 43. Bakkeren G, Kronstad JW: **Linkage of mating-type loci distinguishes bipolar from tetrapolar mating in basidiomycetous smut fungi.** *Proc Natl Acad Sci USA* 1994, **91**:7085-7089.
 44. Gallegos A, Jacobson DJ, Raju NB, Skupski MP, Natvig DO: **Suppressed recombination and a pairing anomaly on the mating-type chromosome of *Neurospora tetrasperma*.** *Genetics* 2000, **154**:623-633.
 45. Jacobson DJ: **Blocked recombination along the mating-type chromosomes of *Neurospora tetrasperma* involves both structural heterozygosity and autosomal genes.** *Genetics* 2005. DOI: 10.1534/genetics.105.044040.
 46. Hood ME: **Dimorphic mating-type chromosomes in the fungus *Microbotryum violaceum*.** *Genetics* 2002, **160**:457-461.
 47. Hood ME: **Repetitive DNA in the automictic fungus *Microbotryum violaceum*.** *Genetica* 2005, **124**:1-10.



ELSEVIER

Genomic inferences from Afrotheria and the evolution of elephants

Alfred L Roca and Stephen J O'Brien

Recent genetic studies have established that African forest and savanna elephants are distinct species with dissociated cytonuclear genomic patterns, and have identified Asian elephants from Borneo and Sumatra as conservation priorities. Representative of Afrotheria, a superordinal clade encompassing six eutherian orders, the African savanna elephant was among the first mammals chosen for whole-genome sequencing to provide a comparative understanding of the human genome. Elephants have large and complex brains and display advanced levels of social structure, communication, learning and intelligence. The elephant genome sequence might prove useful for comparative genomic studies of these advanced traits, which have appeared independently in only three mammalian orders: primates, cetaceans and proboscideans.

Addresses

Laboratory of Genomic Diversity, Basic Research Program, SAIC-Frederick and National Cancer Institute, Frederick, MD 21702, USA

Corresponding author: Roca, Alfred L (roca@ncifcrf.gov)

Current Opinion in Genetics & Development 2005, 15:652–659

This review comes from a themed issue on
Genomes and evolution
Edited by Stephen J O'Brien and Claire M Fraser

Available online 13th October 2005

0959-437X/\$ – see front matter

© 2005 Elsevier Ltd. All rights reserved.

DOI 10.1016/j.gde.2005.09.014

Introduction: genomics and the Afrotheria

A leading criterion used to select mammalian taxa for whole-genome sequencing is the role of comparative genomics in understanding the human genome [1,2]. Regions of crucial function in genes and non-coding elements tend to be conserved across taxa [1,2]. Aligning human DNA sequences to those of divergent taxa might facilitate the identification of genomic regions of importance to human health and disease, which can be targets for further study [1,2]. Eutherian (placental) mammalian orders comprise four primary superordinal groups (Figure 1a), the most basal of which includes elephants, hyraxes, sirenians (dugongs and manatees), aardvarks, elephant shrews, golden moles, and tenrecs [3,4]. The superordinal group was designated Afrotheria (see Glossary), or 'African beasts', because its constituent

orders had differentiated within Africa [3]. Afrotheria diverged from other eutherian mammals approximately 104 million years ago, coincident with the tectonic isolation of Africa from other continents [3,4,5]. The divergent position of Afrotheria among superordinal eutherian clades (Figure 1a) suggested that representative species of Afrotheria should be included among the first set of mammals selected for whole-genome sequencing [1]; the African savanna elephant (*Loxodonta africana*) and tenrec (Figure 1a) were selected.

Several additional factors make an elephant the ideal candidate within Afrotheria for whole-genome sequencing (Figure 1b). Elephants comprise the best-studied afrotherian order [6], with more research articles and DNA sequences generated than for all other afrotherian taxa combined. Elephants have a unique morphology [1], with derived structures such as their trunk and tusks, and lungs adapted for snorkeling, and well-developed senses of olfaction, hearing and touch [6,7,8]. Their long lifespans may be of interest for research on longevity, and they are subject to diseases, including anthrax, herpesvirus, orthopoxvirus and tuberculosis, that have analogues in humans or livestock [6,7,8]. Endogenous repetitive elements [9,10] and remarkable cytonuclear genomic patterns [11] have been characterized. Elephants have large brains, which enable them to exhibit complex systems of communication and social interaction, and relatively high levels of learning and memory [6,12,13]. The three extant elephant species are endangered [6,7,14,15], and genomic sequencing of one species will generate markers useful in conservation genetics for all three. The African forest elephant (*Loxodonta cyclotis*) was a candidate less ideal for sequencing: it is absent from zoos, precluding sample collection; it is relatively unstudied [6]; and it displays high genetic diversity [11,16,17], which complicates genome assembly. The African savanna elephant (*L. africana*) was selected over the Asian elephant (*Elephas maximus*) because the former has been studied more extensively using nuclear DNA markers [11,16,17]. The *L. africana* individual at the San Diego Zoo selected for sequencing derived from the Kruger National Park population, which displays the low genetic diversity typical of savanna elephants [11,16,17]. Initially, two-fold coverage of the genome is being generated, although the elephant should prove a strong candidate for high-quality complete genome assembly, which would enable comprehensive studies of repetitive elements, rearrangements, and changes in copy numbers [18].

Glossary

Afrotheria: A superordinal clade of mammals including the six eutherian orders Proboscidea (elephants), Sirenia (dugongs and manatees), Hyracoidea (hyraxes), Macroscelidea (elephant shrews), Tubulidentata (aardvarks) and Afrosoricida (tenrecs and golden moles).

Conplastic: Describes lineages, such as selectively bred strains of laboratory rodents, in which the nuclear genome from one species (or strain) has been crossed onto the cytoplasm of a different species (or strain). This occurs after a male from the first species mates with a female of the second species, followed by ten or more generations of recurrent backcrossing of female hybrid offspring to non-hybrid males of the first species.

Fission–fusion society: A flexible form of social organization in which fusion into larger groups or fission into smaller groups occurs depending on the activity of the group or the season of year.

Haldane’s rule: “When in the F₁ offspring of two different animal races one sex is absent, rare, or sterile, that sex is the heterozygous [heterogametic] sex.” Originally formulated by JBS Haldane in 1922 [51].

Metagenomic: Studying the genomes of multiple organisms by treating a microbial or environmental community as a single entity. Used in the context of ancient DNA, it describes a genetic library that includes a mix of DNA from the target fossil genome and from contaminating bacterial or environmental DNA.

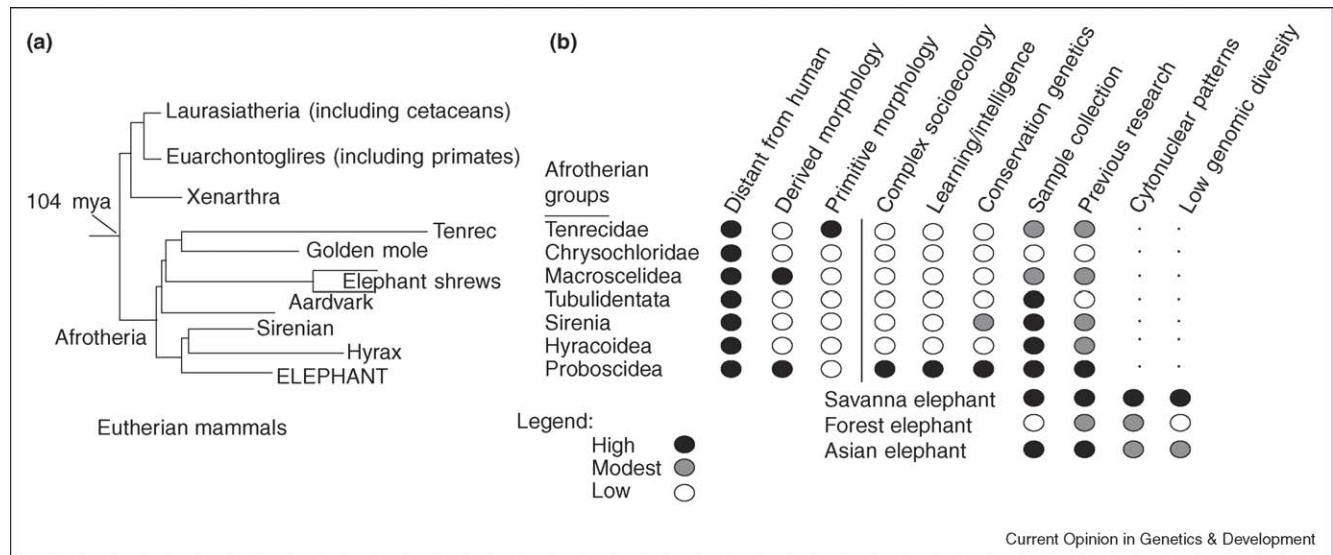
When savanna met forest: cytonuclear genomic dissociation and the African elephant species question

Although recent studies have identified two species of elephant in Africa, conservation strategies for African elephants traditionally treated them as a single species [15[•]]. Relative to savanna elephants, the elephants in

Africa’s tropical forests are smaller, with straighter and thinner tusks, rounded ears, and marked differences in skull morphology [14]. There is little overlap in the geographic range occupied by the two species (Figure 2); the forest elephant inhabits the tropical forests of Africa, whereas the savanna elephant lives in savanna, bush and lightly forested regions [14]. The forest elephant is more of a browser and frugivore, lives in smaller groups and communicates with lower-frequency vocalizations [14]. A comprehensive morphological comparison of 295 African elephant skulls of known provenance in museums concluded that forest and savanna elephants “rank as perfectly distinct species, with absolute differentiation between them” [19], and with only a relatively narrow habitat contact zone in which intermediate morphologies were detected. A second morphological survey [20] that sought to question the forest–savanna elephant split nonetheless produced concordant results, with 46 of 48 skulls readily identified as forest or savanna elephants following the geographic distribution recognized for the two groups [14]. The two morphologically intermediate specimens were from regions (Kanyatsi and Katanga) [20] where a few hybrids might be expected under the two-species model [14,19,21].

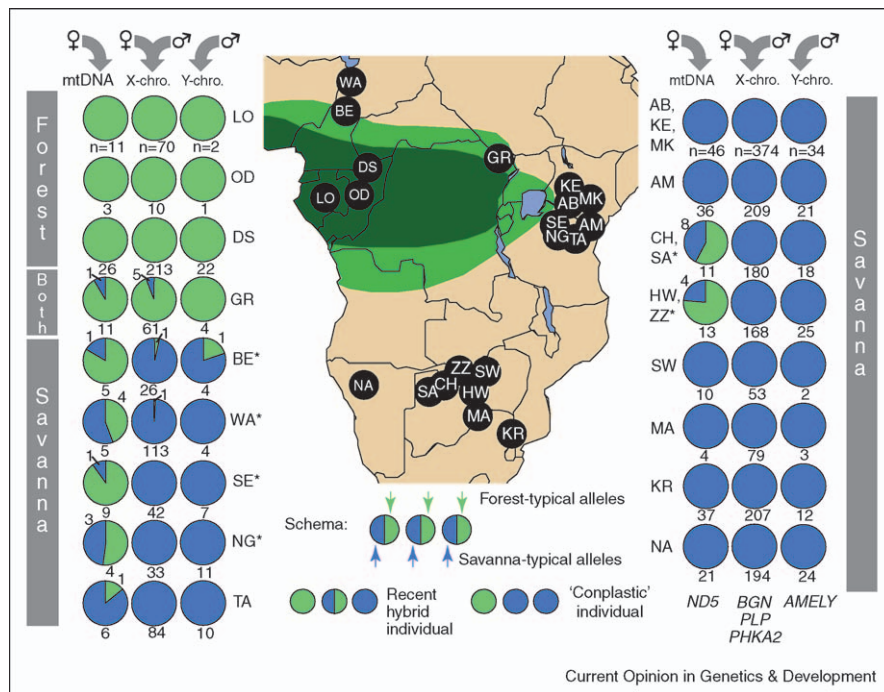
Genetic studies using nuclear DNA markers provided solid evidence indicating that forest and savanna elephants comprise two distinct species of African elephant. Genetic distances based on nuclear gene sequences suggested that the two species diverged from a common

Figure 1



Selecting taxa within Afrotheria for whole genome sequencing. **(a)** Molecular phylogeny of four superordinal clades of placental mammals; including Afrotheria, which separated from other mammals approximately 104 million years ago (mya) [3,4^{**},5]. Inter-ordinal relationships and branch lengths are depicted with permission (Nature Publishing Group) [5]. **(b)** The African savanna elephant was initially suggested for whole-genome sequencing as a representative of Afrotheria useful for annotation of the human genome, and because of its unique morphology [1]. There are a variety of other criteria (right of vertical line) that also make the African savanna elephant the ideal candidate among the Afrotheria for whole-genome sequencing.

Figure 2



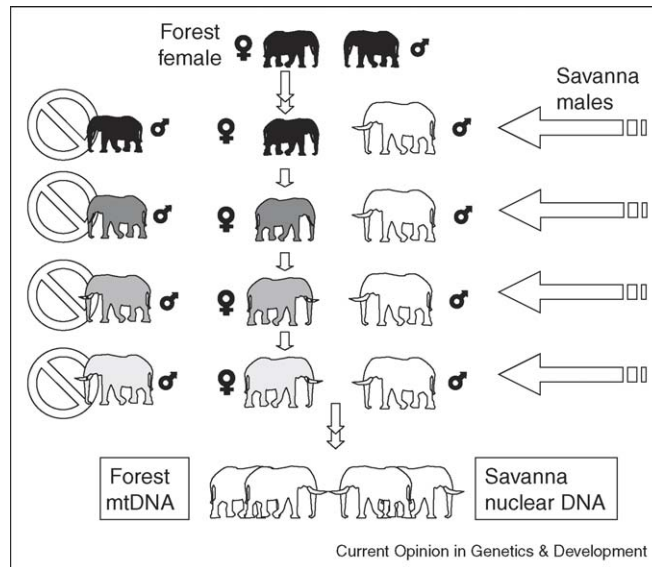
Cytonuclear genomic dissociation in African elephant species. Populations from which elephant DNA samples were sequenced for multiple genetic markers are shown, with savanna elephant habitats colored beige, tropical forests dark green, and intermediate habitats light green [11^{**},21]. Pie charts summarize results of genetic markers that are inherited maternally (left pie charts), biparentally (center) or paternally (right). With very few exceptions, alleles were distinct between forest and savanna elephants for the Y-chromosome and biparentally transmitted markers. However, some savanna populations with only savanna-typical nuclear markers carried mtDNA typical of forest elephants (locale codes labeled with an asterisk have significantly higher levels of forest-typical mtDNA than of forest-typical nuclear genes) [11^{**}]. This could not be accounted for by recent hybrids, which would be expected to show a mix of forest and savanna alleles (inset) but, instead, indicates a conplastic pattern (inset) caused by repeated unidirectional hybridization (see Figure 3) [11^{**}]. The unusual cytonuclear pattern might suggest areas of research that would merit future investigation, such as whether low fertility among some zoo elephants could be associated with a conplastic cytonuclear configuration; or whether there might be a role for parasites in excluding savanna elephants from tropical forests [49]. Reproduced with permission (Nature Publishing Group) [11^{**}]. Locales: forest populations: DS, Dzanga Sangha; LO, Lope; OD, Odzala; savanna populations: AB, Aberdares; AM, Amboseli; BE, Benoue; CH, Chobe; HW, Hwange; KE, Central Kenya; KR, Kruger; MA, Mashatu; MK, Mount Kenya; NA, Namibia; NG, Ngorongoro; SA, Savuti; SE, Serengeti; SW, Sengwa; TA, Tarangire; WA, Waza; and ZZ, Zambezi; and the mixed habitat zone of GR, Garamba.

ancestor approximately three million years ago [11^{**},16]. Multiple fixed nucleotide-site differences were detected between forest and savanna elephants, which also grouped into two reciprocally monophyletic clades [16]. Computation of F_{ST} indicated that a remarkably high proportion (94%) of the genetic variation among African elephants was a result of differences between forest and savanna elephants, and only 6% was caused by intra-group genetic variation [16]; this was later verified using microsatellites (forest-savanna $R_{ST} = 0.90$) [17]. Analysis of nuclear gene haplotypes [11^{**},16] provided further strong support for species-level distinction. An autosomal gene (*GBA*) haplotype found in 96% of savanna elephant chromosomes was completely absent from forest populations, whereas two autosomal gene (*CHRNA1* and *VIM*) insertion-deletion variants common in forest elephants across the Congolian forest were absent in savanna populations [16]. For three X-linked genes, *BGN*, *PHKA2* and

PLP, a total of 1762 (99.9%) chromosome sequences out of 1764 inspected in savanna elephants carried haplotypes completely absent in elephants from tropical forest locales, whereas 100% of 293 chromosome segments examined in forest populations had forest-typical haplotypes (Figure 2, middle pie-charts) [11^{**}]. Similarly, two distinctive Y-chromosome lineages (Figure 2, right pie-charts) were detected among males ($n = 205$), and these lineages split into forest and savanna elephant clades with only a single, exceptional male [11^{**}]. Thus, only a small number of forest-savanna elephant hybrids were detected using nuclear genes [11^{**},16,17].

Mitochondrial DNA (mtDNA) markers in the same populations also form two highly divergent clades [11^{**},20], but these did not always correspond to the nuclear genotype of the elephants that carried them (Figure 2) [11^{**},16,17]. In some savanna populations, even among

Figure 3



Generation of conplastic cytonuclear patterns in savanna elephants. The cytonuclear dissociation apparent in African elephant species (Figure 2) could only result from inter-species hybridization between forest elephant females and savanna elephant males [11**]. As habitats changed, large savanna males (unshaded) would gain access to forest females (shaded black), enabling hybridization to occur. Given that reproductive success among male elephants depends largely on body size [6*,22,23], recurrent backcrossing would occur between hybrid (intermediate shading) females and large savanna males, which would out-compete the smaller forest or hybrid males. The forest elephant component of the nuclear genome would be diluted and replaced, although herds would retain maternally inherited forest-typical mtDNA haplotypes. Other deleterious effects of hybridization that differentially harm male hybrids (Haldane's rule) could have also played a role.

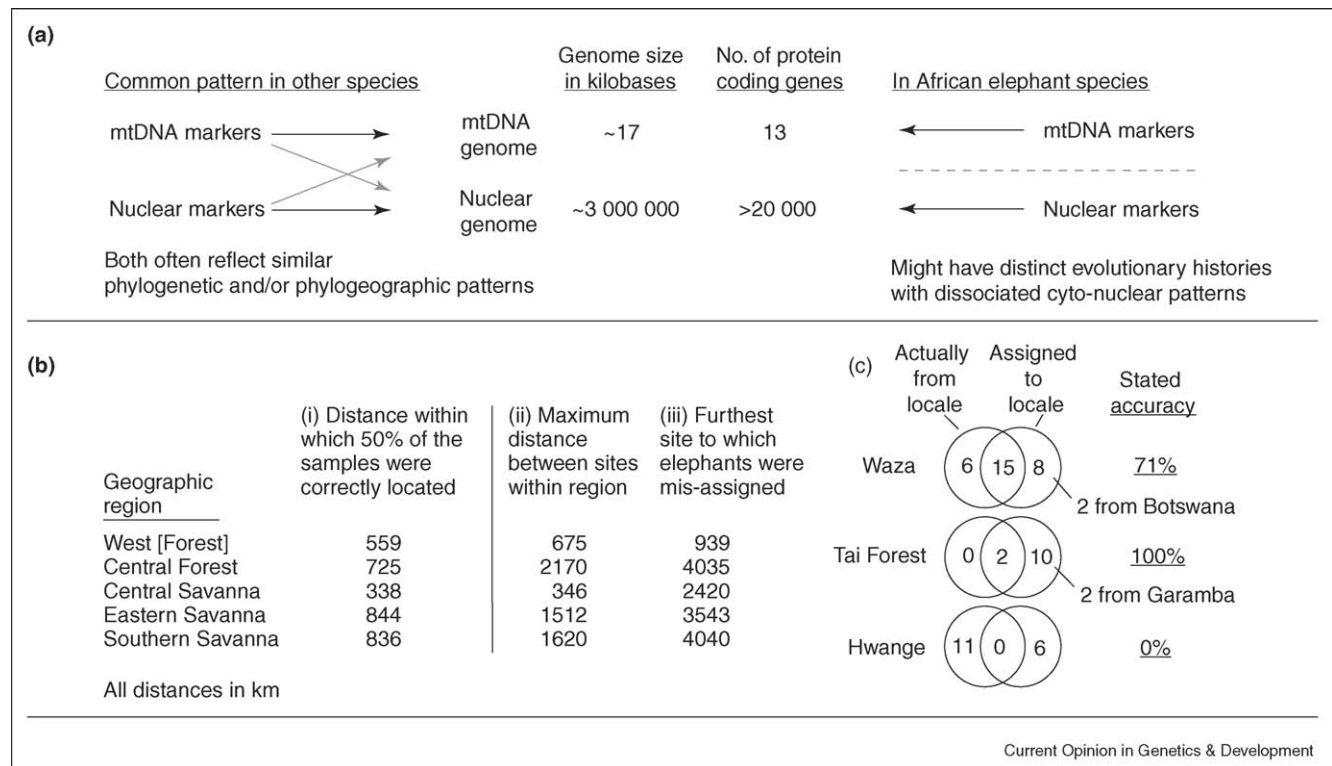
those geographically distant from present-day forest habitats, many individuals with only savanna-specific nuclear genotypes nonetheless carried mtDNA typical of forest elephants (Figure 2, left pie charts) [11**]. The pattern cannot be attributed to recent hybridization [20], because recent hybrids would display a mix of forest and savanna nuclear gene alleles (Figure 2, inset). Instead, a 'conplastic' (see Glossary) pattern is present, in which savanna elephants without forest-typical nuclear alleles nonetheless carry forest-typical mtDNA (Figure 2, inset). The conplastic cytonuclear pattern could only be interpreted as resulting from a three-step process of (i) species isolation and divergence; (ii) ancient inter-species hybridization between forest elephant females and savanna elephant males; and (iii) low reproductive success for hybrid males (Figure 3) [11**]. This scenario is plausible if one considers that fully grown savanna bulls are much larger than forest bulls, with no size overlap between the oldest males [14,19], and that reproductive success in male elephants depends largely on body size [6*,22,23]. Recurrent backcrossing of female hybrid elephants to large savanna males, out-competing the smaller forest or hybrid males, would dilute and replace the forest elephant component of the nuclear genome but would retain the maternally inherited forest-typical mtDNA haplotypes (Figure 3) [11**]. Other deleterious effects of hybridization that differentially harm male hybrids (Haldane's rule; see Glossary) might have also played a role.

The promise of elephant conservation genomics

The persistence of residual forest elephant mitochondria in savanna elephant herds (Figures 2 and 3) would render evolutionary interpretations based on mtDNA misleading [11**]. The African elephant cytonuclear dissociation (Figures 2 and 3) [11**] implies that applying species concepts to elephant mtDNA sequence trees [20] would amount to an error of confusing the topology of a singularly unrepresentative gene tree for that of the species tree, because elephant mtDNA markers can mislead when they do not parallel the pattern present in the nuclear genome (see Figure 4a) [11**]. Nonetheless, mtDNA markers remain useful for some applications. For example, Nyakaana and colleagues [24,25] used mtDNA to show that gene flow in elephants is largely male-mediated. They also demonstrated that the typical social structure of elephants, in which females remain with their maternal family groups, was disrupted in Uganda by intensive poaching that led surviving elephants to form novel family units of unrelated females [26].

Asian elephants also display unusual mtDNA phylogeographic patterns, with the mtDNA haplotypes carried by elephants in Sundaland (Borneo, Sumatra and the Malay Peninsula) similar to those carried by many Sri Lankan elephants, whereas geographically intermediate mainland

Figure 4



The promise and pitfalls of elephant conservation genomics. **(a)** Genetic markers can mislead if they are unrepresentative of the diversity present in the nuclear genome, especially if used to assess overall population structure. In many species, mtDNA and nuclear markers imply similar phylogenetic or phylogeographic patterns (left); however, the conplastic cytonuclear pattern in many savanna elephants (Figures 2 and 3) means that mtDNA markers can be representative of a different evolutionary history from that of the nuclear genome. The nuclear genome includes >100 000-fold more sequence and contains >1000-fold more coding genes than the mitochondrial genome [18,50]; thus, the utility of mtDNA for describing the systematics or overall population structure of elephants can be quite circumscribed. **(b,c)** Wasser *et al.* [36] have proposed the use of microsatellites to pinpoint the provenance of ivory. Their method must overcome four difficulties: (1) the geographic clustering of sampling sites in some regions, notably western forest and central savanna (ii) might confer a misleading accuracy to the method. For example, 50% of elephants in the central savanna region were assigned to within ~338 km of their actual origin (i) [36], but this distance must be placed in context by noting that the only two locales included as part of the region (Benoue [BE] and Waza [WA] in Cameroon; Figure 2) are only ~346 km apart [15*]; (2) many elephants were grossly mis-assigned to a location very distant from their true provenance (iii) [36]. For example, savanna elephants from as far away as Botswana were assigned to Waza, and 10% of Garamba elephants were assigned to the most distant forest locale in the study, the Tai Forest of Côte d'Ivoire (Venn diagrams) [36]; (3) the percent accuracy of assignment calculated by Wasser *et al.* [36] does not reflect the lack of precision caused by elephants being mis-assigned to a locale from elsewhere. For example, 100% accuracy is calculated for the Tai Forest, because the two Tai Forest elephants were correctly assigned. Yet ten elephants from other locales were also incorrectly assigned to the Tai Forest (Venn diagrams) [36]; (4) finally, better sampling could actually diminish the accuracy of assignments. The addition of more nearby locales might increase the mis-assignment of elephants, because geographically close elephants will tend to have similar genotypes. Thus, for Hwange (HW), with many nearby sampled locales surrounding it (same as in Figure 2) all elephants were mis-assigned by Wasser *et al.* (Venn diagram) [36].

populations mostly carry mtDNA haplotypes that form a distinctive (≥ 1.2 million years) clade [27–29]. Proposed explanations for the unusual mtDNA phylogeography include the retention and subsequent loss of haplotypes in some regions, the mixing of two expanding populations that had been isolated in the Pleistocene, or the separation and hybridization of two ancestral elephant species [27,28]. Substantial trade in Asian elephants might have also affected the phylogeographic patterns [27], although the widespread belief that wild elephant populations in Borneo derived from introduced domestic animals was

contradicted by recent genetic studies showing that their mtDNA haplotypes are distinctive [30*]. A different subclade of haplotypes is carried by Sumatran elephants; thus, each of the two island populations has undergone lengthy isolation, forms an evolutionarily significant unit and is a priority for conservation [27,30*]. Given that mtDNA is often not reflective of overall population structure among African elephants (Figure 4a) [11**], a microsatellite survey of Asian elephants would appear to be crucial for determining conservation priorities among mainland and Sri Lankan populations.

The elephant genome sequence will facilitate development of additional nuclear markers useful for a variety of important endeavors: for estimating elephant population sizes through multi-locus genotyping of dung samples [31]; for determining the sex of individuals and calculating sex ratios in a population [31]; or for identifying conservation priorities. Microsatellite analysis suggests that African savanna elephants in the Sudanian vegetation region [21] may be considered a conservation priority, because Cameroon elephants are genetically different from eastern and southern savanna populations [17], while savanna populations west of Cameroon should be at least as distinctive. Analysis of mtDNA sequences had initially suggested that forest and savanna West African elephants together might comprise a separate species of *Loxodonta* [31,32]. However, this contention was undermined by the discovery that some Central African elephants carried mtDNA haplotypes similar to those of West African elephants [33], and it was further weakened after mtDNA patterns in other savanna locales proved to be incongruent with overall population structure (Figure 2) [11]. Morphological analyses have been inconclusive, and the West African dataset of elephant skulls in museums is believed to be unrepresentative of elephants in the region [34]. Resolution of the genetic status of savanna and forest West African populations remains a major goal for elephant conservation genetics, especially given that West Africa has the smallest elephant numbers and most-highly fragmented habitats of any African region [15].

DNA has been successfully extracted from small amounts of ivory anywhere along the length of a tusk [35], and microsatellite genotyping has been proposed for determining the provenance of illicit ivory [36]. However, although nuclear genetic markers accurately and precisely differentiate among the three extant elephant species [11,16,17], and in a more limited way identify broad regional geographic patterns within species [17], attempts to further pinpoint the provenance of ivory through spatial smoothing of genotypes [36] are of uncertain practical utility, as a result of several important caveats (see Figure 4b).

Conclusion: trumpeting the future and the past

One spectacular recent advance in genomic studies involved the construction of ancient DNA 'metagenomic' (partly contaminated; see also Glossary) libraries using two 40 000 year-old fossils of the extinct cave bear (*Ursus spelaeus*), demonstrating that the libraries have the potential for 10-fold coverage of the entire genome of a fossil animal [37]. Using the savanna elephant genome sequence as a comparative standard, extinct taxa across four proboscidean families could be candidates for metagenomic sequencing, if specimens with intact DNA are

located [7,8]. An excellent candidate for metagenomic sequencing is the woolly mammoth, *Mammuthus primigenius*, which is in the same family (Elephantidae) as the three living elephant species. Mammoths are the only extinct proboscidean for which mitochondrial and nuclear DNA have been sequenced by multiple laboratories, with conflicting reports of its phylogenetic relationship to living elephants [8,9].

The elephant genome sequence might shed light on the evolution of large brains and advanced behaviors. Elephants have larger absolute brain weights and cortical volumes than do humans or other terrestrial mammals, with only marginally fewer cortical neurons than humans [12,13]; relative to humans, though, myelinated fiber thickness is much lower in elephants and cetaceans, giving them reduced information processing capacity [13]. Although not all studies concluded that elephants have high intelligence [38], wild and captive elephants have been known to engage in tool use [39,40] and communicate within and between social groups using tactile, chemical and vocal means [41–43]. Elephants can modify their vocalizations in response to auditory experience, suggesting a flexible and open communication system [43]. Older matriarchs can recognize the contact calls of approximately 100 others in the population, learn to readily discriminate between familiar and unfamiliar conspecifics, and gear the defensive behavioral response of their family groups accordingly [44,45]. Older matriarchs lead larger family groups [46] that are reproductively more successful than those headed by inexperienced females [44]. Elephants have a multi-level, fission–fusion social structure (see Glossary), being the first non-human animal in which four hierarchical tiers of social organization have been rigorously demonstrated [46]. It has even been suggested that social trauma such as early disruption of attachment can affect the physiology, behavior and 'culture' of elephants, much as post-traumatic stress disorder affects humans [47]. The genetics of large brains and advanced traits is not well understood; however, elephants, primates and cetaceans [48] comprise the only distinct mammal lineages (Figure 1a) in which such advanced traits could be studied on a comparative basis [4,12,13]. It should soon be possible to examine candidate genes for these traits in elephant and cetacean genomes, or to scan across all mammalian genomes for genes showing evidence of selection in the elephant, cetacean and primate lineages but not in other mammals. The advanced intelligence of humans — relatively speaking — might result from the combination and enhancement of properties found in non-human animals rather than from unique properties, making comparative studies appropriate [13]. Although in recent years modern genetic methods have taught us much about the evolutionary history of elephants, future genomic studies might use elephants to study the traits that make us human.

Acknowledgements

We thank N Georgiadis, MC Howell, and other colleagues, agencies and governments that provided samples or assistance in our primary work. We thank R Ruggiero and the US Fish and Wildlife Service African Elephant Conservation Fund. This publication has been funded in part with Federal funds from the National Cancer Institute, National Institutes of Health, under Contract No. N01-CO-12400. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government. This research was supported in part by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- of special interest
 - of outstanding interest
1. O'Brien SJ, Eizirik E, Murphy WJ: **Genomics. On choosing mammalian genomes for sequencing.** *Science* 2001, **292**:2264-2266.
 2. Thomas JW, Touchman JW, Blakesley RW, Bouffard GG, Beckstrom-Sternberg SM, Margulies EH, Blanchette M, Siepel AC, Thomas PJ, McDowell JC *et al.*: **Comparative analyses of multi-species sequences from targeted genomic regions.** *Nature* 2003, **424**:788-793.
 3. Springer MS, Cleven GC, Madsen O, de Jong WW, Waddell VG, Amrine HM, Stanhope MJ: **Endemic African mammals shake the phylogenetic tree.** *Nature* 1997, **388**:61-64.
 4. Springer MS, Murphy WJ, Eizirik E, O'Brien SJ: **Placental mammal diversification and the Cretaceous-Tertiary boundary.** *Proc Natl Acad Sci USA* 2003, **100**:1056-1061.
- Current molecular dating methods are used to estimate the divergence date of various mammalian lineages, including Afrotheria and the orders that comprise it.
5. Roca AL, Bar-Gal GK, Eizirik E, Helgen KM, Maria R, Springer MS, O'Brien SJ, Murphy WJ: **Mesozoic origin for West Indian insectivores.** *Nature* 2004, **429**:649-651.
 6. Sukumar R: **The Living Elephants: Evolutionary Ecology, Behavior, and Conservation.** Oxford: Oxford University Press; 2003.
- A comprehensive review of scientific literature on the extant elephants.
7. Shoshani J, Tassy P: **The Proboscidea: Evolution and Palaeoecology of Elephants and their Relatives.** Oxford; New York: Oxford University Press; 1996.
 8. Shoshani J, Tassy P: **Advances in proboscidean taxonomy & classification, anatomy & physiology, and ecology & behavior.** *Quaternary Int* 2005, **126-28**:5-20.
 9. Greenwood AD, Lee F, Capelli C, DeSalle R, Tikhonov A, Marx PA, MacPhee RD: **Evolution of endogenous retrovirus-like elements of the woolly mammoth (*Mammuthus primigenius*) and its relatives.** *Mol Biol Evol* 2001, **18**:840-847.
 10. Nikaido M, Nishihara H, Hukumoto Y, Okada N: **Ancient SINES from African endemic mammals.** *Mol Biol Evol* 2003, **20**:522-527.
 11. Roca AL, Georgiadis N, O'Brien SJ: **Cytonuclear genomic dissociation in African elephant species.** *Nat Genet* 2005, **37**:96-100.
- The authors established that some savanna elephants have a conplastic cytonuclear genomic structure, with an exclusively savanna elephant nuclear DNA genotype, but a residual forest elephant mtDNA haplotype.
12. Cozzi B, Spagnoli S, Bruno L: **An overview of the central nervous system of the elephant through a critical appraisal of the literature published in the XIX and XX centuries.** *Brain Res Bull* 2001, **54**:219-227.
 13. Roth G, Dicke U: **Evolution of the brain and intelligence.**
 - *Trends Cogn Sci* 2005, **9**:250-257.

A review of the independent appearance of large brains in vertebrate lineages, including elephants, which suggests that human intelligence

results from combination and enhancement of properties found in non-human animals, rather than from unique properties.

14. Grubb P, Groves CP, Dudley JP, Shoshani J: **Living African elephants belong to two species: *Loxodonta africana* (Blumenbach, 1797) and *Loxodonta cyclotis* (Matschie, 1900).** *Elephant* 2000, **2**:1-4.
15. Blanc JJ, Thouless CR, Hart HT, Dublin HT, Douglas-Hamilton I,
 - Craig CG, Barnes RFW: **African Elephant Status Report 2002.** Gland, Switzerland: IUCN; 2003.

Excellent and regularly updated country-by-country estimates of the population size and geographic range of African elephants.
16. Roca AL, Georgiadis N, Pecon-Slattery J, O'Brien SJ: **Genetic evidence for two species of elephant in Africa.** *Science* 2001, **293**:1473-1477.
17. Comstock KE, Georgiadis N, Pecon-Slattery J, Roca AL, Ostrander EA, O'Brien SJ, Wasser SK: **Patterns of molecular genetic variation among African elephant populations.** *Mol Ecol* 2002, **11**:2489-2498.
18. International Human Genome Sequencing, Consortium: **Finishing the euchromatic sequence of the human genome.** *Nature* 2004, **431**:931-945.
19. Groves CP, Grubb P: **Do *Loxodonta cyclotis* and *L. africana* interbreed?** *Elephant* 2000, **2**:4-7.
20. Debruyne R: **A case study of apparent conflict between molecular phylogenies: the interrelationships of African elephants.** *Cladistics* 2005, **21**:31-50.
21. White F: **The Vegetation of Africa.** Paris: UNESCO; 1983.
22. Slotow R, van Dyk G, Poole J, Page B, Klocke A: **Older bull elephants control young males.** *Nature* 2000, **408**:425-426.
23. Poole JH: **Signals and assessment in African elephants: evidence from playback experiments.** *Anim Behav* 1999, **58**:185-193.
24. Nyakaana S, Arctander P, Siegmund HR: **Population structure of the African savannah elephant inferred from mitochondrial control region sequences and nuclear microsatellite loci.** *Heredity* 2002, **89**:90-98.
25. Nyakaana S, Arctander P: **Population genetic structure of the African elephant in Uganda based on variation at mitochondrial and nuclear loci: evidence for male-biased gene flow.** *Mol Ecol* 1999, **8**:1105-1115.
26. Nyakaana S, Abe EL, Arctander P, Siegmund HR: **DNA evidence for elephant social behaviour breakdown in Queen Elizabeth National Park, Uganda.** *Anim Conserv* 2001, **4**:231-237.
27. Fleischer RC, Perry EA, Muralidharan K, Stevens EE, Wemmer CM: **Phylogeography of the Asian elephant (*Elephas maximus*) based on mitochondrial DNA.** *Evolution Int J Org Evolution* 2001, **55**:1882-1892.
28. Fernando P, Pfrender ME, Encalada SE, Lande R: **Mitochondrial DNA variation, phylogeography and population structure of the Asian elephant.** *Heredity* 2000, **84**:362-372.
29. Vidya TN, Fernando P, Melnick DJ, Sukumar R: **Population differentiation within and among Asian elephant (*Elephas maximus*) populations in southern India.** *Heredity* 2005, **94**:71-80.
30. Fernando P, Vidya TN, Payne J, Stuewe M, Davison G,
 - Alfred RJ, Andau P, Bosi E, Kilbourn A, Melnick DJ: **DNA analysis indicates that Asian elephants are native to Borneo and are therefore a high priority for conservation.** *PLoS Biol* 2003, **1**:E6.

This study establishes that Borneo elephants are genetically distinctive and not the descendants of escaped domestic elephants imported to the island.
31. Eggert LS, Eggert JA, Woodruff DS: **Estimating population sizes for elusive animals: the forest elephants of Kakum National Park, Ghana.** *Mol Ecol* 2003, **12**:1389-1402.
32. Eggert LS, Rasner CA, Woodruff DS: **The evolution and phylogeography of the African elephant inferred from mitochondrial DNA sequence and nuclear microsatellite markers.** *Proc Biol Sci* 2002, **269**:1993-2006.

33. Debruyne R, Van Holt A, Barriel V, Tassy P: **Status of the so-called African pygmy elephant (*Loxodonta pumilio* [NOACK 1906]): phylogeny of Cytochrome *b* and mitochondrial control region sequences.** *C R Biol* 2003, **326**:687-697.
The authors deny a distinctive genetic status for West African elephants and for the semi-legendary pygmy elephant.
34. Groves CP: **What are the elephants of West Africa?** *Elephant* 2000, **2**:7-8.
35. Comstock KE, Ostrander EA, Wasser SK: **Amplifying nuclear and mitochondrial DNA from African elephant ivory: a tool for monitoring the ivory trade.** *Conserv Biol* 2003, **17**:1840-1843.
This study demonstrates that DNA can be extracted from anywhere along an elephant tusk, permitting genetic analysis using nuclear and mitochondrial markers, even from tusks kept at ambient temperatures for decades.
36. Wasser SK, Shedlock AM, Comstock K, Ostrander EA, Mutayoba B, Stephens M: **Assigning African elephant DNA to geographic region of origin: applications to the ivory trade.** *Proc Natl Acad Sci USA* 2004, **101**:14847-14852.
37. Noonan JP, Hofreiter M, Smith D, Priest JR, Rohland N, Rabeder G, Krause J, Dettler JC, Paabo S, Rubin EM: **Genomic sequencing of Pleistocene cave bears.** *Science* 2005, **309**:597-599.
This study shows that ancient DNA from Pleistocene fossils can be cloned into 'metagenomic' libraries, which might revolutionize the study of nuclear genes from extinct taxa.
38. Nissani M, Hoefler-Nissani D, Lay UT, Htun UW: **Simultaneous visual discrimination in Asian elephants.** *J Exp Anal Behav* 2005, **83**:15-29.
39. Chevalier-Skolnikoff S, Liska JO: **Tool use by wild and captive elephants.** *Anim Behav* 1993, **46**:209-219.
40. Hart BL, Hart LA, McCoy M, Sarath CR: **Cognitive behaviour in Asian elephants: use and modification of branches for fly switching.** *Anim Behav* 2001, **62**:839-847.
41. Langbauer WR: **Elephant communication.** *Zoo Biol* 2000, **19**:425-445.
42. McComb K, Reby D, Baker L, Moss C, Sayialel S: **Long-distance communication of acoustic cues to social identity in African elephants.** *Anim Behav* 2003, **65**:317-329.
43. Poole JH, Tyack PL, Stoeger-Horwath AS, Watwood S: **Animal behaviour: elephants are capable of vocal learning.** *Nature* 2005, **434**:455-456.
44. McComb K, Moss C, Durant SM, Baker L, Sayialel S: **Matriarchs as repositories of social knowledge in African elephants.** *Science* 2001, **292**:491-494.
45. McComb K, Moss C, Sayialel S, Baker L: **Unusually extensive networks of vocal recognition in African elephants.** *Anim Behav* 2000, **59**:1103-1109.
The authors carry out rigorous quantitative analyses of four levels of social structure in an African elephant population.
46. Wittemyer G, Douglas-Hamilton I, Getz WM: **The socioecology of elephants: analysis of the processes creating multitiered social structures.** *Anim Behav* 2005, **69**:1357-1371.
47. Bradshaw GA, Schore AN, Brown JL, Poole JH, Moss CJ: **Elephant breakdown.** *Nature* 2005, **433**:807.
48. Whitehead H: **Cultural selection and genetic diversity in matrilineal whales.** *Science* 1998, **282**:1708-1711.
49. Watve MG, Sukumar R: **Asian elephants with longer tusks have lower parasite loads.** *Curr Sci* 1997, **72**:885-889.
50. Hauf J, Waddell PJ, Chalwatzis N, Joger U, Zimmermann FK: **The complete mitochondrial genome sequence of the African elephant (*Loxodonta africana*), phylogenetic relationships of Proboscidea to other mammals and D-loop heteroplasmy.** *Zoology* 2000, **102**:184-195.
51. Haldane JBS: **Sex ratio and unisexual sterility in hybrid animals.** *J Genet* 1922, **12**:101-109.



ELSEVIER

Sex chromosomes and sex determination in reptiles

Commentary

William S Modi¹ and David Crews²

Reptiles occupy a crucial position with respect to vertebrate phylogeny, having roamed the earth for more than 300 million years and given rise to both birds and mammals. To date, this group has been largely ignored by contemporary genomics technologies, although the green anole lizard was recently recommended for whole genome sequencing. Future experiments using flow-sorted chromosome libraries and high-throughout genomic sequencing will help to discover important findings regarding sex chromosome evolution, early events in sex determination, and dosage compensation. This information should contribute extensively toward a general understanding of the genetic control of development in amniotes.

Addresses

¹ SAIC Frederick, National Cancer Institute, Core Genotyping Facility, 8717 Grovemont Circle, Gaithersburg, MD 20877, USA

² Section of Integrative Biology, University of Texas, Austin, TX 78712, USA

Corresponding author: Modi, William S (modiw@mail.nih.gov)

Current Opinion in Genetics & Development 2005, 15:660–665

This review comes from a themed issue on
Genomes and evolution
Edited by Stephen J O'Brien and Claire M Fraser

Available online 7th October 2005

0959-437X/\$ – see front matter

© 2005 Elsevier Ltd. All rights reserved.

DOI 10.1016/j.gde.2005.09.009

Introduction

Reptiles are a familiar group of vertebrates, having existed for more than 300 million years. Although these animals reached their zenith during the Jurassic and Cretaceous periods, today they are represented by only four orders (turtles, crocodylians, squamates [snakes and lizards] and sphenodontians [tuatara]) (Figure 1). However, not only do these animals occupy a pivotal position on the phylogeny of vertebrates — they are the direct ancestor to birds and mammals — but they also possess several unique biological attributes that, if better understood, could contribute significantly to understanding basic evolutionary biology and the molecular mechanisms behind human health and disease. Recent technical advances in DNA sequencing have made whole genome sequencing possible for a variety of species. One mission of the ambitious National Human Genome Research Institute (NHGRI) is whole genome sequencing of various animal species. Among vertebrates, whole genome sequences are now available for chicken, *Xenopus* and

three species of fish, and the sequences of over 20 species of mammals are complete or underway (<http://www.genome.gov/11007951>). Interestingly, reptiles have remained impervious to the watchful eye of the NHGRI's comparative sequencing program. However, at NHGRI's request, the Reptile Genome Consortium met in April 2005 at Washington University's Genome Sequencing Center in St Louis. Their subsequent mission was to consult the scientific community and make recommendations to the Institute on which species should be considered for whole genome sequencing. Their recommendation was submitted to NHGRI as a 'White Paper' in July 2005, and the community overwhelmingly chose the green anole lizard, *Anolis carolinensis*, as its first target species, with the American alligator, garter snake and/or painted turtle to follow (<http://reptilegenome.com>). The green anole is an excellent choice, having been used as a model system for reptilian physiology, neurology and reproductive behavior for many years [1].

In addition to the whole genome sequencing of *Anolis* and subsequent species, parallel investigations into reptilian genomics will unlock many of the fascinating secrets that nature has bundled into these intriguing animals. This review summarizes research on reptilian sex chromosomes and sex determination, and recommends the preparation of flow-sorted chromosome-specific libraries (see Glossary). Such libraries will enable reconstruction of the evolutionary events involved in sex chromosome diversification and will provide the raw materials necessary to study the expression of specific genes involved in sex determination and dosage compensation.

Has the mechanism of sex determination evolved independently in different reptilian lineages?

Sex chromosomes are distinct from autosomes in that they differ in size, number, staining characteristics, and gene content when the two sexes are compared. Ohno's law [2] asserts that heteromorphic sex chromosomes originated from an autosomal ancestor following a mutation that conferred a sexual advantage. Additional sex-linked mutations in other genes then accumulated on the same homologue. Recombination between the primordial sex chromosomes was suppressed by chromosomal rearrangements such as inversions to preserve the block of sex-linked genes. The absence of recombination fostered the accumulation of mutations and repetitive sequences with subsequent 'heterochromatinization' of the sex-specific chromosome. Deletions of heterochromatin account for the smaller sizes usually observed for the Y or W chromo-

Glossary

Boids: The snake taxonomic family Boidae contains boa constrictors and pythons.

Diploid-triploid somatic mosaicism: The presence of different chromosome numbers in different adult tissues of the same individual. In this case, certain tissues have a diploid (2n) complement, whereas other tissues have a triploid (3n) complement.

Flow-sorted chromosome-specific libraries: A sample of genomic DNA enriched for a specific chromosome, prepared by separating a suspension of mixed chromosomes with fluorescence-activated cell sorter.

Male- or female-permissive temperature: The ambient temperature that causes embryonic development of a bipotential gonad into a testis or ovary in species with temperature-dependent sex determination.

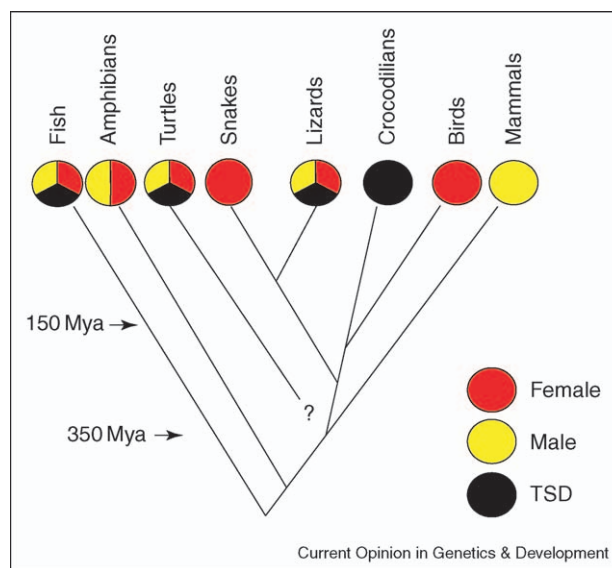
Subtractive hybridization analyses: A molecular biological procedure that compares RNA levels from different tissues in an attempt to identify transcripts that are over- or under-expressed in one tissue relative to in others.

Viperids: This family of poisonous snakes contains Old World vipers, such as puff adders, bushmasters and sand vipers.

somes compared with the X or Z chromosomes, respectively.

Reptiles exhibit some of the most extraordinary variability in sex chromosome structure and patterns of sex determination seen among vertebrates (Figure 1) [3]. For example, all crocodylians, the tuatara, most turtles and many lizards have temperature-dependent sex determination (TSD), in which adult anatomical sex is a function of the temperature at which eggs are incubated. Species that display TSD do not reveal karyotypic differ-

Figure 1



Vertebrate phylogeny illustrating sex determination modes in different taxa. "Female" and "Male" represent genetic sex determination with female and male heterogamety, respectively. TSD represents temperature-dependent sex determination. An unanswered question in contemporary reptilian phylogenomics regards the relationships of turtles to other reptiles.

ences between males and females, and the range in temperature that produces all males or all females can be as little as 1 °C [4].

By contrast, several turtles, some lizards and all snakes are subject to genetic sex determination (GSD), in which adult sex is chromosomally determined at the time of fertilization. At least two species of turtles and some lizards have male heterogamety (XY males and XX females), whereas other turtles, other lizards and all snakes have female heterogamety (ZZ males and ZW females). Other turtles are chromosomally monomorphic, and additional experiments are needed to determine if they have GSD or TSD. The ZW chromosomes of snakes reveal increased differentiation as one progresses from the phylogenetically primitive boids to the more advanced viperids (Figure 2a; see also Glossary). The origin of heteromorphic XY sex chromosomes in two species of turtles is thought to have occurred independently, and these same chromosomes appear as autosomes in other species of turtles with TSD (Figure 2b) [5,6].

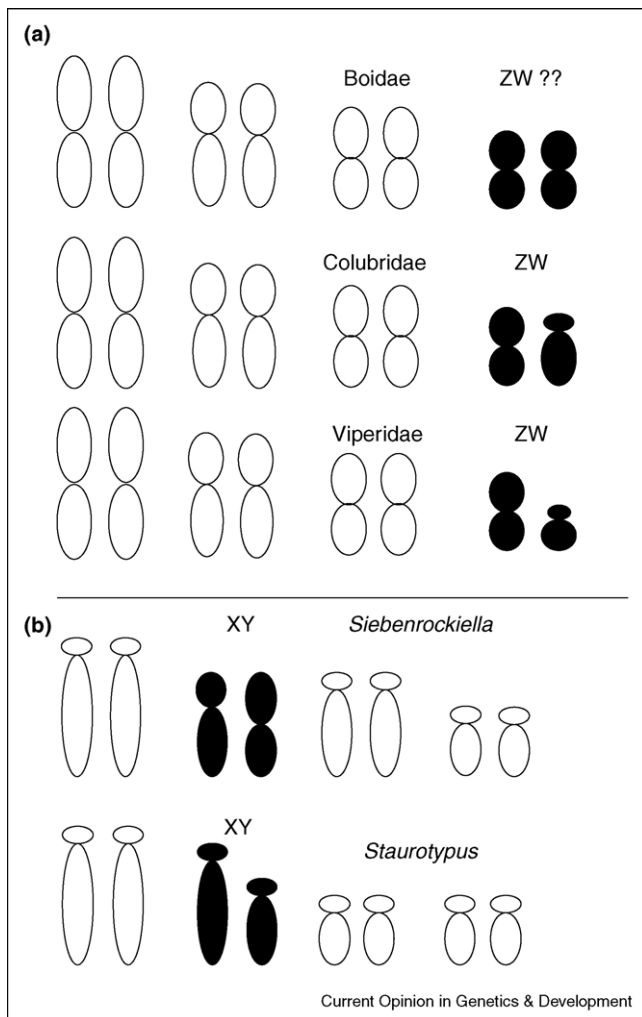
The variability seen among reptilian sex chromosomes suggests that sex chromosome and sex determination systems have evolved independently in different lineages. However, definitive molecular cytogenetic and gene mapping data for reptilian sex chromosomes, which would precisely define their evolutionary histories, are lacking. An important set of experiments would use flow cytometry [7] to prepare flow-sorted W, X, Y, and Z chromosome-specific libraries from several species with heteromorphic sex chromosomes (Table 1). In an effort to trace the origin of sex chromosome differentiation, these sex chromosome-specific libraries could then be used as hybridization probes in fluorescence *in situ* hybridization to metaphase chromosomes from various species having both TSD and GSD.

Identification of genes involved in vertebrate sex determination

A plausible hypothesis posits that sex-determining systems evolve by the retrograde addition of regulatory elements upstream of established developmental programs [8–11]. If this is accurate, then various sex-determining systems can be thought of as one evolutionarily conserved core network regulated by various taxon-specific upstream factors. For example, many of the same genes that are important in mammalian sex determination are found in other species and show strong similarity in their temporal patterns of expression during gonadogenesis (Figure 3).

Little is known about the cellular and molecular foundations of sex determination in reptiles. The most sought-after but least well-understood are the temperature-transduction mechanism(s) that initiate the TSD

Figure 2



Ideogrammatic depiction of heteromorphic sex chromosomes in two groups of reptiles. **(a)** Chromosomes from three families of snakes, illustrating the progressive differentiation of the ZW system when going from phylogenetically primitive to advanced taxa [2]. **(b)** Chromosomes from two genera of turtles, portraying what are believed to be independently derived XY sex chromosomes in each case [5,6].

pathway. One model species for the study of TSD is the red-eared slider, *Trachemys scripta* [4,12]. Although adult testis and ovary morphologies are highly conserved across amniotes, there are differences in the cellular events involved in gonadogenesis. Two notable differences between mouse and turtle include (i) the mechanism of sex cord formation; and (ii) the initial location and subsequent behavior of primordial germ cells in the gonad [13,14]. In turtles, both males and females form primitive sex cords before sexual differentiation of the gonads. By contrast, mouse testis cords appear to form *de novo* in the male gonad and not at all in the female gonad [15]. In turtles, primordial germ cells are initially located in the cortex and, subsequently, migrate into cord structures in the medullary region of the gonad at male-permissive temperature, but at female-permissive temperature (see Glossary) remain in the cortex. It is unclear whether recruitment of germ cells to sex cords is driven by supporting cells as it is in mammals or whether germ cells recruit and/or organize the formation of testis cords. Furthermore, the turtle orthologs (*tSox9*, *tWnt1*, *tSf1*, *tDmrt1*, *tWnt4*, *tDax1* and *tMis*) of the mammalian genes implicated in sex determination and differentiation are expressed early in the temperature-sensitive period. This suggests that incubation temperature is capable of engaging this core molecular cascade.

Details remain to be elucidated concerning the temporal patterns of gene expression, the cell types that express these genes, and the functions of their gene products. Nevertheless, this model predicts that the significant differences between turtles and mammals are in the upstream regulators, whereas the core downstream pathways mediating ovary and testis development are conserved. There are significant gaps in our knowledge of how temperature acts on target cells in the turtle gonad to influence gene expression, protein activity, mRNA stability and post-transcriptional events. There are similar gaps in our knowledge of mammalian sex determination. In both cases, the identity and action of genes and cells between the initial trigger (genetic or environmental) and the up-regulation of the core sex-determining pathway

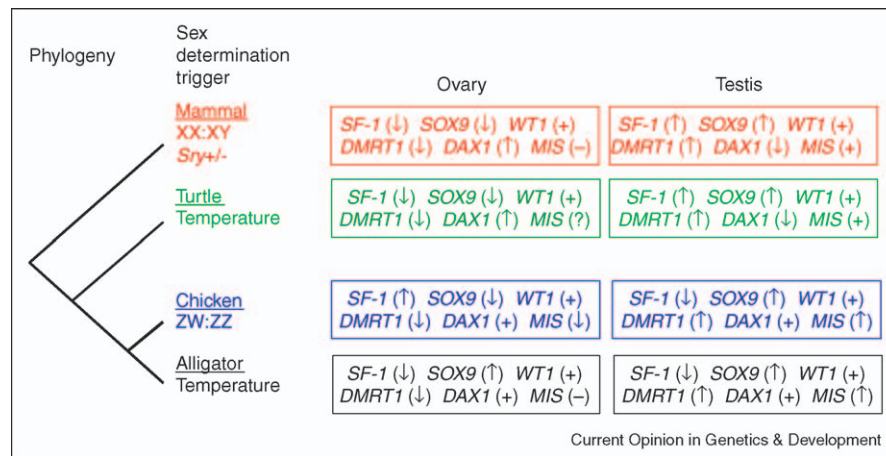
Table 1

Species of reptiles with heteromorphic sex chromosomes that could be used to generate flow-sorted, chromosome-specific, genomic DNA libraries.

Common name	Scientific name	Mode of sex determination*
Mexican musk turtle	<i>Staurotypus salvinii</i>	XX/XY
Asian black pond turtle	<i>Siebenrockiella crassicolis</i>	XX/XY
brown-roofed turtle	<i>Kachuga smithii</i>	ZZ/ZW
tiger whiptail lizard	<i>Cnemidophorus tigris</i>	XX/XY
house gecko lizard	<i>Gehyra australis</i>	ZZ/ZW
garter snake	<i>Thamnophis sirtalis</i>	ZZ/ZW
Russell's viper	<i>Daboia russellii</i>	ZZ/ZW

* XX/XY, refers to genetic sex determination with male heterogamety; ZZ/ZW, refers to genetic sex determination with female heterogamety.

Figure 3



Selected genes underlying differentiation of the genital ridge into an ovary or testis in amniote vertebrates. Phylogenetic relationships indicated on the left of the figure. In mammals and birds, gonadal sex is established by the genetic composition inherited at fertilization, a process known as genotypic sex determination (GSD). In some reptiles, gonadal sex depends, ultimately, on the temperature of the incubating egg, a process known as temperature-dependent sex determination (TSD). The trigger for gonad determination in mammals is the presence (+) or absence (-) of *Sry*; in birds the trigger is unknown but appears to be the Z chromosome to autosome ratio. Note that many of the same genes appear to be involved in gonadal differentiation for species exhibiting GSD (mammals and birds) and TSD (turtles and crocodylians). Note, also, that for these selected genes the patterns of expression appear to reflect phylogenetic relationships, with mammals being similar to turtles, and birds more similar to crocodylians. The regulatory mechanisms behind the expression patterns for most of these selected genes are currently being investigated, but timing of *SOX9* and *MIS* expression during testis development appears to fall along phylogenetic lines; in mammals and turtles, *SOX9* expression precedes *MIS* expression, whereas in alligator and bird the reverse pattern is seen. Finally, through manipulating the genetic, physical or chemical environment it is possible to modify gonadal sex in both GSD and TSD species. Abbreviations: DAX1, dosage-sensitive sex reversal-adrenal hypoplasia congenital critical region on the X chromosome; DM, DMRT1, doublesex- and mab3-related transcription factor one; MIS, Müllerian-inhibiting substance; SF-1, steroidogenic factor one; SOX9, SRY-related HMG box nine; SRY, sex-determining region on the Y chromosome; WT1, Wilm's tumor one. Plus (+) indicates presence, and minus (-) indicates absence. Up arrow (↑) indicates up-regulation, and down arrow (↓) indicates down-regulation.

remain obscure. *Trachemys scripta* offers unique advantages and opportunities for experimental manipulation and promises to provide insight into early events of testis and ovary determination.

In organisms subject to TSD, putatively conserved sex-determining pathways are triggered by unknown molecular mechanisms that respond to temperature during the middle third of incubation. Reconstructing the evolution of sex-determining mechanisms is a long-term goal and requires analysis of both conserved and taxon-specific components of vertebrate sex-determining pathways. Although the former is best achieved by examining the roles of known sex-determining genes in a variety of vertebrates, elucidation of mechanisms that are unique to particular sex-determining systems requires *de novo* screening in the organism in question. Subtractive hybridization analyses (see Glossary) have identified genes that differ in their expression in the gonad at male- and female-producing temperatures. Members of this set of promising candidate genes might be involved in sex determination upstream of conserved vertebrate sex-determining genes. Similarly, using a genomics approach, one could determine the nucleotide sequence of entire W and Y chromosomes isolated using flow cytometry. This

would produce a list of several hundred genes per chromosome. These sex-specific genes could then be analyzed using cDNA arrays, hopefully yielding a manageable subset of candidate genes, the expression patterns of which could then be studied in reptilian embryos of species having either GSD or TSD.

Dosage compensation in reptiles?

Species with heteromorphic sex chromosomes are faced with the dilemma of how to achieve equal levels of gene expression between the sexes when one sex has only one copy and the other sex has two copies of a particular chromosome. Dosage compensation allows for the differential expression of sex-linked genes [16]. This phenomenon has been most intensively studied in three systems in the animal kingdom: *Drosophila*, *Caenorhabditis elegans* and mammals (primarily human and mouse). Different processes characterize each of these three systems; however, there is one essential common feature: in all three cases, specialized complexes bind to the X chromosome of one sex, modify its chromatin conformation and regulate its transcription. Active areas of research include determination of the *cis*-acting site(s) on the X chromosome that initiate dosage compensation and elucidation of the subsequent downstream epigenetic events. Further

Figure 4

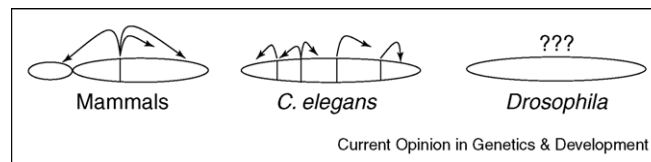


Diagram indicating the concept of 'spreading' in the process of dosage compensation in three model systems. In mammals, an important step in X chromosome inactivation is the 'coating' of the inactive X with *Xist* RNA, which is encoded by a single gene [22]. In *C. elegans*, the dosage compensation complex initially binds to multiple *cis*-acting recruitment sites on the X, and then spreads out along the entire chromosome [17]. Alternatively, spreading is not thought to occur in *Drosophila*; rather, a model involving hierarchical affinities of binding sites has been proposed [18].

studies of dosage compensation will provide greater insight into the mechanisms regulating chromatin domain organization, which is crucial toward understanding the regulation of gene expression in development [17].

In *Drosophila* males, transcription of the single X chromosome is doubled so that the amount of gene product equals that of XX females. Dosage compensation is carried out by a ribonucleoprotein complex called the dosage compensation complex (DCC) [18]. This assemblage contains six proteins: three male-specific lethals (MSL1, MSL2 and MSL3); males absent on the first (MOF); a histone acetyltransferase (HAT); and the JIL1 histone H3 kinase. In addition, *roX1* and *roX2*, two non-coding RNAs, are involved [19]. This DCC accumulates in the males' cells and coats the X chromosome. The histone H4 at lysine 16 on the X becomes acetylated, the chromatin becomes less compact and transcription is elevated. The actual mechanism of how histone modification affects transcriptional activation is unclear.

In *C. elegans*, genes on the XX chromosomes of hermaphrodites are downregulated so that their transcriptional output is equal to that of the genes in the XO male. This downregulation is initiated by a protein called SDC-2 (sex determination and dosage compensation defect-2), which assembles a collection of proteins including DPY-26 (Dumpy: shorter than wild type protein-26), DPY-27, SDC-3 and MIX-1 (mitosis- and X-associated protein-1), which, collectively, are known as the DCC [20]. The DCC is recruited to multiple *cis*-acting regions of the X chromosomes of hermaphrodites and spreads out along the chromosome from these initial binding sites to effect suppression of transcription on the entire chromosome (Figure 4) [17].

In mammals, dosage compensation is achieved by the transcriptional silencing of one complete X chromosome in all cells of the female body [21]. This process of X chromosome inactivation (Xi) takes place early in development [16]. It is mediated from a single X inactivation center (*Xic*), containing a *cis*-acting gene called *Xist*, which

encodes a non-coding RNA that coats the inactive X chromosome [22]. Subsequently, this process involves a series of epigenetic events, including methylation of both histones and nucleic acids, and histone hypoacetylation [23]. Furthermore, it has recently been shown that the PRC1 (protein regulator of cytokinesis 1) complex of polycomb proteins maintains the inactive state in somatic cells [24]. Finally, it is known that many human genes on the inactive X escape inactivation, and these escape genes are non-randomly distributed along the inactive X chromosome [25].

Given the extreme diversity in reptilian sex chromosome systems, one would predict that fundamentally distinct dosage-compensation systems exist in this group and that, by studying these species, novel molecular mechanisms controlling chromatin organization and gene expression in development might be discovered. In addition, we can ask whether dosage compensation occurs in polyploid reptiles such as the desert-grasslands whiptail lizard, a triploid parthenogenetic species descended from the hybrid union of two sexual species [26]; or the side-necked turtle, *Platemys platycephala*, a species having diploid–triploid somatic mosaicism (see Glossary) [27]. The study of dosage compensation has been restricted to model species because extensive genomic reagents are required for assessing gene expression of an entire chromosome. Current genomics technologies are poised to open this field to previously unstudied species. The nucleotide sequence of entire reptilian X and Z chromosomes would enable cDNA arrays and quantitative PCR to measure expression level differences between males and females [28]. Subsequent studies could determine whether novel mechanisms of gene silencing or expression are found during reptilian development.

Conclusions and future directions

Mother Nature created an excellent natural laboratory for studying the genetic control of development when she designed the sex chromosome and sex determination systems of living reptiles. Until now, progress in understanding these mysteries has been slow; however, the availability of contemporary genomics technologies such

as chromosome sorting, fluorescence *in situ* chromosome hybridization, high-throughput sequencing, subtractive hybridization and cDNA arrays are poised to bring about rapid increases in knowledge. As previous developmental genetic studies have shown, there are similarities and differences when mechanisms from various species are compared. In this vein, we can expect to learn not only reptile-specific processes but also findings that are much more general in nature. These broader, more general discoveries are of paramount interest because they will help us understand normal and abnormal development of various embryonic stages in different taxonomic groups.

Acknowledgements

Raymond Porter and Christina Shoemaker provided helpful comments. John Bickham critiqued the manuscript. This work was supported by National Institutes of Health (MH41770) and National Science Foundation (IBN2000126) grants to DC. This study has been funded in whole or in part with Federal funds from the National Cancer Institute, NIH, under contract number NO1 CO 124000. This research was supported in part by the Intramural Research Program of the NCI and NIH.

- Lovern MB, Holmes MM, Wade J: **The green anole (*Anolis carolinensis*): a reptilian model for laboratory studies of reproductive morphology and behavior.** *ILAR J* 2004, **45**:54-64.
- Ohno S: Sex chromosomes and Sex-linked Genes. Berlin: Springer; 1977.
- Valenzuela N, Lance L: Temperature-Dependent Sex Determination in Vertebrates. Washington: Smithsonian Books; 2004.
- Crews D, Bergeron JM, Bull JJ, Flores D, Tousignant A, Skipper JK, Wibbels T: **Temperature-dependent sex determination in reptiles: proximate mechanisms, ultimate outcomes, and practical applications.** *Dev Genet* 1994, **15**:297-312.
- Carr JL, Bickham JW: **Sex chromosomes of the Asian black pond turtle, *Siebenrockiella crassicollis* (Testudines: Emydidae).** *Cytogenet Cell Genet* 1981, **31**:178-183.
- Sites JW Jr, Bickham JW, Haiduk MW: **Derived X chromosome in the turtle genus *Staurotypus*.** *Science* 1979, **206**:1410-1412.
- Telenius H, Pelmeur AH, Tunnacliffe A, Carter NP, Behmel A, Ferguson-Smith MA, Nordenskjold M, Pfragner R, Ponder BA: **Cytogenetic analysis by chromosome painting using DOP-PCR amplified flow-sorted chromosomes.** *Genes Chromosomes Cancer* 1992, **4**:257-263.
- McLaren A: **Sex determination in mammals.** *Trends Genet* 1988, **4**:153-157.
- Graves JA: **The origin and function of the mammalian Y chromosome and Y-borne genes — an evolving understanding.** *Bioessays* 1995, **17**:311-320.
- Wilkins AS: **Moving up the hierarchy: a hypothesis on the evolution of a genetic sex determination pathway.** *Bioessays* 1995, **17**:71-77.
- Wilkins AS: **The Evolution of Developmental Pathways.** Sunderland, MA: Sinauer Associates, Inc; 2002.
- Crews D: **Sex determination: where environment and genetics meet.** *Evol Dev* 2003, **5**:50-55.
- Schmahl J, Yao HH, Pierucci-Alves F, Capel B: **Colocalization of WT1 and cell proliferation reveals conserved mechanisms in temperature-dependent sex determination.** *Genesis* 2003, **35**:193-201.
- Wibbels T, Bull JJ, Crews D: **Chronology and morphology of temperature-dependent sex determination.** *J Exp Zool* 1991, **260**:371-381.
- Tilmann C, Capel B: **Cellular and molecular pathways regulating mammalian sex determination.** *Recent Prog Horm Res* 2002, **57**:1-18.
- Marin I, Siegal ML, Baker BS: **The evolution of dosage-compensation mechanisms.** *Bioessays* 2000, **22**:1106-1114.
- Csankovszki G, McDonel P, Meyer BJ: **Recruitment and spreading of the *C. elegans* dosage compensation complex along X chromosomes.** *Science* 2004, **303**:1182-1185.
- Fagegaltier D, Baker BS: **X chromosome sites autonomously recruit the dosage compensation complex in *Drosophila* males.** *PLoS Biol* 2004, **2**:e341.
- Franke A, Baker BS: **The rox1 and rox2 RNAs are essential components of the compensasome, which mediates dosage compensation in *Drosophila*.** *Mol Cell* 1999, **4**:117-122.
- Dawes HE, Berlin DS, Lapidus DM, Nusbaum C, Davis TL, Meyer BJ: **Dosage compensation proteins targeted to X chromosomes by a determinant of hermaphrodite fate.** *Science* 1999, **284**:1800-1804.
- Lyon MF: **Gene action in the X-chromosome of the mouse (*Mus musculus* L.).** *Nature* 1961, **190**:372-373.
- Brown CJ, Ballabio A, Rupert JL, Lafreniere RG, Grompe M, Tonlorenzi R, Willard HF: **A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome.** *Nature* 1991, **349**:38-44.
- Brockdorff N: **X-chromosome inactivation: closing in on proteins that bind Xist RNA.** *Trends Genet* 2002, **18**:352-358.
- Hernandez-Munoz I, Lund AH, van der Stoep P, Boutsmas E, Muijters I, Verhoeven E, Nusinow DA, Panning B, Marahrens Y, van Lohuizen M: **Stable X chromosome inactivation involves the PRC1 Polycomb complex and requires histone MACROH2A1 and the CULLIN3/SPOP ubiquitin E3 ligase.** *Proc Natl Acad Sci USA* 2005, **102**:7635-7640.
- Carrel L, Willard HF: **X-inactivation profile reveals extensive variability in X-linked gene expression in females.** *Nature* 2005, **434**:400-404.
- Woolley SC, Sakata JT, Gupta A, Crews D: **Evolutionary changes in dopaminergic modulation of courtship behavior in *Cnemidophorus* whiptail lizards.** *Horm Behav* 2001, **40**:483-489.
- Bickham JW, Tucker PK, Legler JM: **Diploid-triploid mosaicism: an usual phenomenon in side-necked turtles (*Platemys platycephala*).** *Science* 1985, **227**:1591-1593.
- Craig IW, Mill J, Craig GM, Loat C, Schalkwyk LC: **Application of microarrays to the analysis of the inactivation status of human X-linked genes expressed in lymphocytes.** *Eur J Hum Genet* 2004, **12**:639-646.