The evolutionary history of meiotic genes: early origins by duplication and subsequent losses

Arthur William Pightling

Copyright 2011 Arthur Pightling

This dissertation is available at Iowa Research Online: http://ir.uiowa.edu/etd/2960

Recommended Citation
THE EVOLUTIONARY HISTORY OF MEIOTIC GENES: EARLY ORIGINS BY DUPLICATION AND SUBSEQUENT LOSSES

by

Arthur William Pightling

An Abstract
Of a thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Biology in the Graduate College of The University of Iowa

May 2011

Thesis Supervisor: Associate Professor John M. Logsdon, Jr.
Meiosis is necessary for sexual reproduction in eukaryotes. Genetic recombination between non-sister homologous chromosomes is needed in most organisms for successful completion of the first meiotic division. Proteins that function during meiotic recombination have been studied extensively in model organisms. However, less is known about the evolution of these proteins, especially among protists. We searched the genomes of diverse eukaryotes, representing all currently recognized supergroups, for 26 genes encoding proteins important for different stages of interhomolog recombination. We also performed phylogenetic analyses to determine the evolutionary relationships of gene homologs. At least 23 of the genes tested (nine that are known to function only during meiosis in model organisms) are likely to have been present in the Last Eukaryotic Common Ancestor (LECA). These genes encode products that function during: i) synaptonemal complex formation; ii) interhomolog DNA strand exchange; iii) Holliday junction resolution; and iv) sister-chromatid cohesion. These data strongly suggest that the LECA was capable of these distinct and important functions during meiosis. We also determined that several genes whose products function during both mitosis and meiosis are paralogs of genes whose products are known to function only during meiosis. Therefore, these meiotic genes likely arose by duplication events that occurred prior to the LECA.

The Rad51 protein catalyzes DNA strand exchange during both mitosis and meiosis, while Dmc1 catalyzes interhomolog DNA strand exchange only during meiosis. To study the evolution of these important proteins, we performed degenerate PCR and extensive nucleotide and protein sequence database searches to obtain data from representatives of all available eukaryotic supergroups. We also performed phylogenetic analyses on the Rad51 and Dmc1 protein sequence data obtained to evaluate their utility as phylogenetic markers. We determined that evolutionary relationships of five of the six currently recognized eukaryotic supergroups are supported with Bayesian phylogenetic analyses. Using this dataset, we also identified ten amino acid residues that are highly conserved among Rad51 and Dmc1 protein sequences and, therefore, are likely to confer protein-specific functions. Due to the distributions of these residues, they are likely to have been present in the Rad51 and Dmc1 proteins of the LECA.
To address an important issue with the gene inventory method of scientific inquiry, we developed a heuristic metric for determining whether apparent gene absences are due to limitations of the sequence search regimen or represent true losses of genes from genomes. We collected RNA polymerase I (Pol I), Replication Protein A (RPA), and DNA strand exchange (SE) sequence data from 47 diverse eukaryotes. We then compared the numbers of apparent absences to a single measure of protein sequence length and sequence conservation (Smith-Waterman pairwise alignment (S-W) scores) obtained by comparing yeast and human protein sequence data. Using Poisson correlation regression to analyze the Pol I and RPA subunit datasets, we confirmed that S-W scores and apparent gene absences are correlated. We also determined that genes encoding products that are critical for interhomolog SE in model organisms (Rad52, Rad51, Dmc1, Rad54, and Rdh54) have been lost frequently during eukaryotic evolution. *Saccharomyces cerevisiae* null rad52, dmc1, rad54, and rdh54 mutant phenotypes are suppressed by rad51 overexpression or mutation. If *rad51* overexpression or mutation affects other eukaryotes in a similar fashion, this phenomenon may account for frequent losses of genes whose products are critical for the completion of meiosis in model organisms.

Finally, we place this work into greater context with a review of hypotheses for the selective forces and mechanisms that resulted in the origin of meiosis. The review and the data presented in this thesis provide the basis for a model of the origin of meiotic genes in which meiosis arose from mitosis by large-scale gene duplication, following a preadaptation that served to reduce increased numbers of chromosomes (from diploid to haploid) caused by erroneous eukaryotic cell-cell fusions.

Abstract Approved: _______________________________

Thesis Supervisor

Title and Department

Date
THE EVOLUTIONARY HISTORY OF MEIOTIC GENES: EARLY ORIGINS BY DUPLICATION AND SUBSEQUENT LOSSES

by

Arthur William Pightling

A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Biology in the Graduate College of The University of Iowa

May 2011

Thesis Supervisor: Associate Professor John M. Logsdon, Jr.
CERTIFICATE OF APPROVAL

PH.D. THESIS

This is to certify that the PH.D. thesis of

Arthur William Pightling

has been approved by the Examining Committee for the thesis requirement for the Doctor of Philosophy degree in Biology at the May 2011 graduation.

Thesis Committee:

John M. Logsdon, Jr., Thesis Supervisor

Stephen D. Hendrix

Robert E. Malone

Bryant F. McAllister

Hallie J. Sims
For my family
ACKNOWLEDGMENTS

I would like to express my gratitude towards the many people at the University of Iowa that have contributed to my growth as a scientist. I would like to thank my thesis advisor Dr. John Logsdon for his mentorship, time, unfettered access to his laboratory, and insightful feedback. I would especially like to thank him for allowing me to work on projects he conceived, while also extending me the freedom to develop concepts of my own. I am fortunate that he funded my participation in two workshops at which I learned new molecular phylogenetic and evolutionary analyses, in addition to techniques for collecting, identifying, and culturing eukaryotic microorganisms (protists). I am very grateful to the other members of my thesis advisory committee, Dr. Stephen Hendrix, Dr. Bryant McAllister, Dr. Robert Malone, and Dr. Hallie Sims for their invaluable time, patience, constructive criticism, and helpful advice. I would also like thank Dr. John Logsdon, Dr. Bryant McAllister, Dr. Josep Comeron, Dr. Ana Llopard, Dr. Jeff Klahn, and Dr. Maurine Neiman for the opportunity to teach the Evolution course with them nearly every semester in the past 6 years. My teaching experience has clearly contributed to my growth as an evolutionary biologist.

I am thankful for the contributions of my collaborators to my projects. I would like to thank Matthew Brockman for his enthusiasm and unfailing extensive computer support, without which most of this thesis would not have been possible. Cindy Brochu, Abram Doval, Nicole Adams, Lauren Stefaniak, and Nevin Sebastian are thanked for their technical assistance and sequencing. I would also like to acknowledge former and current members of the Logsdon Lab for illustrative discussions. I would especially like to thank Dr. Banoo Malik for training me initially in the lab and with phylogenetic analyses, for helpful discussions, for helpful comments on chapters 2 and 3 of this thesis, and for encouraging me to collaborate with her.
The research in this thesis would not have been possible without financial support from various sources. Funds for an honorable mention for the Sally Casanova Predoctoral Fellowship at the California State University enabled me to apply to the University of Iowa’s Ph.D. program, and purchase textbooks and phylogenetic software used in the initial years of my doctoral thesis projects to launch my research. During the summers of 2004 and 2005, the University of Iowa’s Avis Cone Graduate Summer Fellowships supported me while performing laboratory research, and in 2010 a University of Iowa Graduate College Graduate Summer Fellowship supported my bioinformatic research. In 2008 I received a travel award for participation in the Bodega Bay Applied Phylogenetics Workshop, and student travel awards from the International Society of Protistologists and International Society of Evolutionary Protistology for presenting my work as a talk at the Protist 2008 conference. I was also supported as a teaching assistant in fall 2004, fall and spring of 2006, 2007, 2008, 2009 and 2010 by the Biology Department at the University of Iowa. Otherwise my research has been supported by funding to my thesis advisor from the National Science Foundation, grants # MCB-0216702 and EF-0431117.
ABSTRACT

Meiosis is necessary for sexual reproduction in eukaryotes. Genetic recombination between non-sister homologous chromosomes is needed in most organisms for successful completion of the first meiotic division. Proteins that function during meiotic recombination have been studied extensively in model organisms. However, less is known about the evolution of these proteins, especially among protists. We searched the genomes of diverse eukaryotes, representing all currently recognized supergroups, for 26 genes encoding proteins important for different stages of interhomolog recombination. We also performed phylogenetic analyses to determine the evolutionary relationships of gene homologs. At least 23 of the genes tested (nine that are known to function only during meiosis in model organisms) are likely to have been present in the Last Eukaryotic Common Ancestor (LECA). These genes encode products that function during: i) synaptonemal complex formation; ii) interhomolog DNA strand exchange; iii) Holliday junction resolution; and iv) sister-chromatid cohesion. These data strongly suggest that the LECA was capable of these distinct and important functions during meiosis. We also determined that several genes whose products function during both mitosis and meiosis are paralogs of genes whose products are known to function only during meiosis. Therefore, these meiotic genes likely arose by duplication events that occurred prior to the LECA.

The Rad51 protein catalyzes DNA strand exchange during both mitosis and meiosis, while Dmc1 catalyzes interhomolog DNA strand exchange only during meiosis. To study the evolution of these important proteins, we performed degenerate PCR and extensive nucleotide and protein sequence database searches to obtain data from representatives of all available eukaryotic supergroups. We also performed phylogenetic analyses on the Rad51 and Dmc1 protein sequence data obtained to evaluate their utility as phylogenetic markers. We determined that evolutionary relationships of five of the six
currently recognized eukaryotic supergroups are supported with Bayesian phylogenetic analyses. Using this dataset, we also identified ten amino acid residues that are highly conserved among Rad51 and Dmc1 protein sequences and, therefore, are likely to confer protein-specific functions. Due to the distributions of these residues, they are likely to have been present in the Rad51 and Dmc1 proteins of the LECA.

To address an important issue with the gene inventory method of scientific inquiry, we developed a heuristic metric for determining whether apparent gene absences are due to limitations of the sequence search regimen or represent true losses of genes from genomes. We collected RNA polymerase I (Pol I), Replication Protein A (RPA), and DNA strand exchange (SE) sequence data from 47 diverse eukaryotes. We then compared the numbers of apparent absences to a single measure of protein sequence length and sequence conservation (Smith-Waterman pairwise alignment (S-W) scores) obtained by comparing yeast and human protein sequence data. Using Poisson correlation regression to analyze the Pol I and RPA subunit datasets, we confirmed that S-W scores and apparent gene absences are correlated. We also determined that genes encoding products that are critical for interhomolog SE in model organisms (Rad52, Rad51, Dmc1, Rad54, and Rdh54) have been lost frequently during eukaryotic evolution. Saccharomyces cerevisiae null rad52, dmc1, rad54, and rdh54 mutant phenotypes are suppressed by rad51 overexpression or mutation. If rad51 overexpression or mutation affects other eukaryotes in a similar fashion, this phenomenon may account for frequent losses of genes whose products are critical for the completion of meiosis in model organisms.

Finally, we place this work into greater context with a review of hypotheses for the selective forces and mechanisms that resulted in the origin of meiosis. The review and the data presented in this thesis provide the basis for a model of the origin of meiotic genes in which meiosis arose from mitosis by large-scale gene duplication, following a
preadaptation that served to reduce increased numbers of chromosomes (from diploid to haploid) caused by erroneous eukaryotic cell-cell fusions.
# TABLE OF CONTENTS

LIST OF TABLES .......................................................................................................................... x

LIST OF FIGURES ...................................................................................................................... xi

CHAPTER

1. GENERAL INTRODUCTION ....................................................................................... 1

   The origin of eukaryotes ....................................................................................... 3
   A comparison of mitotic and meiotic divisions ............................................... 6
   The origin and evolution of meiotic genes ................................................. 10
   Components of meiotic interhomolog DNA strand exchange .................. 14
   Current state of the eukaryotic phylogeny ............................................ 20
   Summary ........................................................................................................... 24

2. A PAN-EUKARYOTIC INVENTORY OF DNA STRAND EXCHANGE COMPONENTS REVEALS PATTERNS OF CONSERVATION AND LOSS ............................................................... 35

   Abstract ........................................................................................................ 35
   Introduction .................................................................................................. 36
   Methods ....................................................................................................... 39
       Data acquisition .................................................................................. 39
       Phylogenetic analyses ....................................................................... 41
       Inventory assembly ........................................................................... 41
   Results and discussion .............................................................................. 42
       Limits of sequence detection and distribution of strand exchange genes among eukaryotes ........................................................................ 44
       Suppressors of strand exchange component mutant phenotypes in *Saccharomyces cerevisiae* ........................................................................ 49
       Conclusions .......................................................................................... 51

3. PHYLOGENOMIC ANALYSIS OF RECA HOMOLOGS RAD51 AND DMC1 FROM ALL SUPERGROUPS PROVIDES EVIDENCE FOR MEIOSIS IN THE LAST COMMON ANCESTOR OF EUKARYOTES ........................................................................ 95

   Background .................................................................................................. 95
   Results and discussion .............................................................................. 99
       Phylogenetic analysis of Dmc1 .......................................................... 99
       Phylogenetic analysis of Rad51 ........................................................ 100
       Phylogenetic analyses of Rad51 and Dmc1 .................................. 100
       Characteristics of Rad51 and Dmc1 protein sequences .................. 104
       Conclusions ....................................................................................... 106
   Methods ...................................................................................................... 108
       Database searches ............................................................................. 108
       Degenerate PCR ............................................................................... 109
       Phylogenetic analyses ...................................................................... 111
4. **MEIOSIS-SPECIFIC GENES AROSE BY DUPLICATION PRIOR TO THE LAST COMMON ANCESTOR OF EUKARYOTES**

   Abstract ......................................................................................................... 132
   Introduction ................................................................................................... 133
   Results and discussion .................................................................................. 135
     Distributions of meiotic genes ............................................................... 135
     Assessment of distributions ................................................................... 138
     Case study: the *Spo11* genes ............................................................... 138
     Conclusions ........................................................................................... 139
   Methods ........................................................................................................ 140
     Database searches .................................................................................. 140
     Phylogenetic analyses ............................................................................ 142
     Inventory assembly ................................................................................ 142

5. **CONCLUDING REMARKS**......................................................................... 173

   Why meiosis? ................................................................................................ 173
   Meiosis arose from mitosis ........................................................................... 180
   A model for the evolution of meiotic DNA strand exchange genes .......... 184
   A model for the origin of meiosis ................................................................. 187
   Future directions ........................................................................................... 193

REFERENCES ................................................................................................................ 198
LIST OF TABLES

Table

2.1 DNA strand exchange component absences from eukaryotic groups ......................75

2.2 Protein sequence comparisons between *Saccharomyces cerevisiae* and *Homo sapiens* ......................................................................................................................83

2.3 Protein sequence comparisons between *Homo sapiens* and *Oryza sativa*........84

2.4 Protein sequence comparisons between *Oryza sativa* and *Saccharomyces cerevisiae* ..................................................................................................................85

2.5 Functions of strand exchange protein with *Saccharomyces cerevisiae* null mutant phenotypes and suppressors.................................................................88

2.6 The most complete genomes of the genera searched during this study with web addresses ...........................................................................................................93

3.1 Support for eukaryotic supergroups and first order groups from phylogenetic analyses of Rad51, Dmc1, and concatenated protein sequence data .................113

3.2 Degenerate primers and their positions .................................................................130

3.3 Proposed functions of residues identified during this study ...............................131

4.1 Proteins involved in four general categories of meiosis and their functions .......167

4.2 Observed numbers of sequence absences from 46 genomes, Smith-Waterman pairwise alignment scores, predicted numbers of absences, and the proportion of observed absences likely due to detection failures for 20 proteins that function during meiosis ..........................................................................................170

4.3 Genome sequence databases searched with web address and references ..........171
LIST OF FIGURES

Figure

1.1 Evolutionary relationships among prokaryotes, members of six currently recognized eukaryotic supergroups and Apusozoa according to multigene phylogenetic analyses .................................................................27

1.2 The three-kingdom tree of life with relative order of major events during eukaryotic evolution ........................................................................................................29

1.3 General schematic of mitosis and meiosis ...............................................................30

1.4 General model of interhomolog DNA strand exchange during meiosis .................32

1.5 A model for the origin of meiotic function by gene duplication .............................34

2.1 Phylogenetic distribution among eukaryotes of DNA strand exchange genes ......55

2.2 Unrooted phylogenetic tree of 47 Replication Protein A – 1 (RPA1) homologs ....57

2.3 Unrooted phylogenetic tree of 42 Replication Protein A – 2 (RPA2) homologs ....58

2.4 Unrooted phylogenetic tree of 36 Replication Protein A – 3 (RPA3) homologs ....59

2.5 Unrooted phylogenetic tree of 44 Replication Protein A – 1 (RPA1) homologs ....60

2.6 Unrooted phylogenetic tree of 29 Rad52 homologs ..............................................61

2.7 Unrooted phylogenetic tree of 46 Rad51 homologs ..............................................62

2.8 Unrooted phylogenetic tree of 41 Rad51 homologs ..............................................63

2.9 Unrooted phylogenetic tree of 34 Rad55 homologs ..............................................64

2.10 Unrooted phylogenetic tree of 42 Rad57 homologs ............................................65

2.11 Unrooted phylogenetic tree of 38 Rad57 homologs ............................................66

2.12 Unrooted phylogenetic tree of 34 Dmc1 homologs .............................................67

2.13 Unrooted phylogenetic tree of 38 Hop2 homologs .............................................68

2.14 Unrooted phylogenetic tree of 41 Mnd1 homologs .............................................69

2.15 Unrooted phylogenetic tree of 34 Mnd1 homologs .............................................70

2.16 Unrooted phylogenetic tree of 29 Rdh54 homologs ...........................................71

2.17 Unrooted phylogenetic tree of 34 Rad54 homologs ...........................................72

2.18 Unrooted phylogenetic tree of 13 Rad59 homologs ..........................................73
2.19 Unrooted phylogenetic tree of 46 sets of 13 concatenated strand exchange homologs

2.20 Multiple sequence alignment of RPA1 ssDNA binding domain (DBD-A) from 54 diverse eukaryotes

2.21 Multiple sequence alignment of RPA1 ssDNA binding domain (DBD-B) from 54 diverse eukaryotes

2.22 Multiple sequence alignment of RPA1 ssDNA binding domain (DBD-C) from 54 diverse eukaryotes

2.23 Multiple sequence alignment of RPA2 ssDNA binding domain (DBD-D) from 45 diverse eukaryotes

2.24 Multiple sequence alignment of RPA1 ssDNA binding domain (DBD-F) from 54 diverse eukaryotes

2.25 Multiple sequence alignment of RPA3 ssDNA binding domain (DBD-E) from 36 diverse eukaryotes

2.26 Phylogenetic distribution among eukaryotes of RNA Polymerase I core complex subunit genes

2.27 Number of detection failures for RNA polymerase I, RPA and SE proteins as predicted by Poisson regression analysis compared with observed numbers of detection failures

3.1 Graphic representation of Rad51 or Dmc1 gene sequence fragments amplified with degenerate PCR from representatives of four eukaryotic supergroups and Apusozoa relative to Saccharomyces cerevisiae Rad51 protein sequence

3.2 Unrooted phylogenetic tree of 47 Dmc1 homologs

3.3 Unrooted phylogenetic tree of 47 Dmc1 homologs with accession numbers

3.4 Unrooted phylogenetic tree of 54 Dmc1 and RadA homologs with accession numbers

3.5 Unrooted phylogenetic tree of 105 Rad51 and Dmc1 homologs

3.6 Unrooted phylogenetic tree of 112 Rad51, Dmc1, and RadA homologs

3.7 Unrooted phylogenetic tree of 157 Rad51, Dmc1 and RadA homologs

3.8 Unrooted phylogenetic tree of 52 Rad51 homologs

3.9 Unrooted phylogenetic tree of 58 Rad51 homologs with accession numbers

3.10 Unrooted phylogenetic tree of 65 Rad51 and RadA homologs with accession numbers

3.11 Unrooted phylogenetic tree of 40 Concatenated Rad51 and Dmc1 homologs
3.12 Unrooted phylogenetic tree of 40 Concatenated Rad51 and Dmc1 homologs with accession numbers (Dmc1/Rad51) .................................................................126

3.13 Protein sequence alignment of prokaryotic and eukaryotic recA orthologs with amino acids conserved among 158 protein sequences indicated.................127

3.14 p-distance matrix of prokaryotic and eukaryotic recA orthologs .........................129

4.1 Distribution of 20 homologs that function during meiosis among 46 eukaryotes representing all eukaryotic supergroups.........................................................143

4.2 Presence of 20 homologs that function during meiosis in the last eukaryotic common ancestor (LECA) inferred by their distribution among eukaryotic supergroups .............................................................................................................145

4.3 Unrooted phylogenetic tree of 50 eukaryotic Hop1 and Rev7 homologs..................146

4.4 Unrooted phylogenetic tree of 49 eukaryotic Rad21 and Rec8 homologs ..............147

4.5 Unrooted phylogenetic tree of 69 eukaryotic Spo11-1, Spo11-2, and Spo11-3 homologs with 6 archaebacterial Top6A homologs .........................................................149

4.6 Unrooted phylogenetic tree of 69 eukaryotic Spo11-1, Spo11-2, and Spo11-3 homologs .................................................................................................................151

4.7 Unrooted phylogenetic tree of 81 eukaryotic Rad51 and Dmc1 homologs with 6 archaebacterial RadA homologs .................................................................152

4.8 Unrooted phylogenetic tree of 81 eukaryotic Rad51 and Dmc1 homologs with....154

4.9 Unrooted phylogenetic tree of 82 eukaryotic Hop2 and Mnd1 homologs ..........155

4.10 Unrooted phylogenetic tree of 131 eukaryotic Mlh1, Mlh2, Mlh3, and Pms1 homologs with 4 archaebacterial MutL homologs .........................................................156

4.11 Unrooted phylogenetic tree of 131 eukaryotic Mlh1, Mlh2, Mlh3, and Pms1 homologs .................................................................................................................158

4.12 Unrooted phylogenetic tree of 113 eukaryotic Mer3, Brr2, and Slh1 homologs with 6 archaebacterial Ski2 homologs .................................................................159

4.13 Unrooted phylogenetic tree of 113 eukaryotic Mer3, Brr2, and Slh1 homologs....161

4.14 Unrooted phylogenetic tree of 183 eukaryotic Msh2, Msh3, Msh4, Msh5, and Msh6 homologs with 5 archaebacterial MutS homologs .........................................................163

4.15 Unrooted phylogenetic tree of 183 eukaryotic Msh2, Msh3, Msh4, Msh5, and Msh6 homologs .................................................................................................................165

4.16 Radial tree topologies of archaebacterial and eukaryotic homologs ....................166

4.17 Number of detection failures as predicted by Poisson regression analysis or RNA polymerase I and Replication Protein A subunits with observed numbers of detection failures for 18 meiotic genes .................................................................169
5.1 General model for the evolution of DNA strand exchange genes ..........................195
5.2 Alignment of conserved Rad51 and Dmc1 residues ...............................................196
5.3 Model for mitotic ploidy reduction in ancestral eukaryotes ........................................197
PREFACE

This thesis describes research on the origins and evolution of eukaryotic gene homologs involved in different stages of meiosis. Chapters 2, 3 and 4 of this thesis are written in the form of manuscripts intended for submission to peer-reviewed journals for publication in the very near future. The chapters are formatted according to the requirements of each journal. The exception is the references, which are formatted consistently throughout in the style of Molecular Biology and Evolution. When referring to gene names italics are used, the first letter is uppercase for archaeabacterial and eukaryotic genes (e.g. RadA and Rad51), genetic mutants of genes are presented in lower case (e.g. radA and rad51), and proteins are presented in normal font with the first letter in uppercase (e.g. RadA and Rad51).

The research presented in this thesis builds upon a now published phylogenomic study of 29 genes involved in meiosis in 5 of 6 currently recognized supergroups of eukaryotes (Malik et al. 2008), which itself is not included in this thesis. Dr. Banoo Malik was the primary author; I was the second author, followed by Lauren Stefaniak, Dr. Andrew Schurko, and Dr. John Logsdon. I cloned and sequenced Trichomonas vaginalis mutL homologs during my laboratory rotation project, and later analyzed homologs of the Rad52, recA and mutL gene families, contributed 40% of Figure S1 and Tables S1.1, S1.2 and S1.3, and helped revise the manuscript. Banoo advised me on the laboratory and phylogenetic techniques used in my contribution to this publication. Other specific contributions and acknowledgements for this project are detailed in the publication. The entire content of this thesis was initiated following from this project, or motivated by discussions with my advisor, Dr. John Logsdon, who provided advice and technical expertise throughout, with additional feedback over the years from former and current members of the Logsdon lab, as well as members of my supervisory committee.
Chapter 2 describes a bioinformatic study and is organized as a manuscript for submission to the journal *Molecular Biology and Evolution*. This chapter details the phylogenomic distributions of 13 genes encoding proteins that catalyze DNA strand exchange during interhomolog recombination among 47 diverse genera. This project builds upon the previously published study (Malik *et al.* 2008) by further investigating the distributions of *Rad51, Dmc1, Rad52, Hop2, Mnd1* and other “strand exchange” genes among 34 diverse eukaryotes. These data were also used to explore a heuristic metric for determining the limits of sequence detection versus bona fide gene loss. I began developing this project in 2006 during the “Writing in the Natural Sciences” graduate course offered by Dr. Stephen Hendrix, with feedback on the initial draft manuscript offered by Dr. Hendrix and my classmates Rebecca Hart-Schmidt, Mike Peglar, Banoo Malik, and Min Wu. The analyses in the current version of the manuscript really took shape immediately before, during and after my December 2009 two-week visit to New York to accompany my wife during her surgery, when I was a guest in Dr. Jane Carlton’s laboratory at New York University’s Department of Medical Parasitology. Helpful discussions with Dr. Malik, Dr. Carlton and Dr. Steven Sullivan led me to devise my criteria for selecting the genomes scrutinized in this chapter, and led me to further utilize NCBI’s BLAST tools to search local databases (that I built myself on my own computer) by PSI-tBLASTn and HMMer. I conceived the project, performed all the analyses, and am the primary author of the manuscript. My advisor, Dr. John Logsdon, is the senior author, provided advice on the research design and implementation, and revised the manuscript. My thesis committee members provided helpful comments throughout the project, including some time-consuming detailed technical suggestions and advice on statistical analyses of regression provided by Dr. Bryant McAllister and Dr. Stephen Hendrix.

Chapter 3 is organized as a manuscript for submission to the journal *BioMednet Central – Evolutionary Biology*. I am the primary author, followed by Rebecca Hernan,
Dr. Nidhi Sahni and Dr. John Logsdon. The project builds further upon my advisor’s evolutionary analyses of Rad51 and Dmc1 protein sequences from animals, plants, and fungi initially published by Stassen, et al. (1997), and Dr. Logsdon’s unpublished work on protist Rad51 and Dmc1 genes conceived and begun during his own postdoctoral research. This chapter reports bioinformatic analyses of Rad51 and Dmc1 sequence data obtained from searches of public gene and genome sequence databases and with help of my co-authors by degenerate PCR experiments in the laboratory. I amplified and cloned 69% of the reported degenerate PCR products, oversaw the laboratory research of my undergraduate assistant, Rebecca Hernan, and Dr. Nidhi Sahni during her laboratory rotation, and I searched public databases, performed phylogenetic analyses and wrote the manuscript. Nidhi amplified and cloned 14% of the reported degenerate PCR products, and Rebecca amplified and cloned 14% of the degenerate PCR products. Nevin Sebastian amplified and cloned 3% of the degenerate PCR products, overseen by Dr. Andrew Schurko. Degenerate PCR products that I isolated for several organisms were superseded by the public release of genome sequence data, and so these sequenced PCR products are excluded from the chapter. DNA samples were obtained by collaboration with Jeff Cole and Dr. Robert Molestina at the American Type Culture Collection (ATCC, Manassas, VA) and Dr. Laura Katz and her assistant Jessica Grant (Smith College, Northampton MA). Research assistants Cindy Brochu, Abram Doval, Nicole Adams, Lauren Stefaniak and Nevin Sebastian sequenced the clones. Dr. John Logsdon conceived and initiated the project, developed the initial set of degenerate PCR primers, advised on degenerate PCR strategies and phylogenetic analysis, provided helpful discussion of research design and implementation and revised the manuscript.

In Chapter 4, a phylogenomic analysis of the distribution among 46 diverse eukaryotes of 20 genes whose products function during meiosis in model organisms is presented. The chapter is organized as a manuscript for the journal Molecular Biology and Evolution. It represents the culmination of the studies presented here and follows
from Dr. Banoo Malik’s doctoral research while she was in the Logsdon lab (Malik et al. 2008). I am the primary author of this chapter; Dr. Malik (now at Dalhousie University) is co-primary author, followed by Dr. John Archibald (Dalhousie University) and Dr. John Logsdon. Banoo’s thesis indicated that several genes encoding proteins that are known to function only during meiosis in model animals, fungi and plants actually arose early during eukaryotic evolution by gene duplication. I have expanded the taxonomic sampling to include more putatively basal lineages in the diverse eukaryotic groups, I learned and made use of several new applications for phylogenetic analysis and gene sequence search methods, and I wrote the manuscript. Banoo identified meiotic gene models in *Bigelowiella natans*, provided her initial multiple sequence alignments of meiotic proteins from 2008, helped with taxon selection and in considering key discussion points. *B. natans* is the first sequenced representative of the Rhizaria, the only eukaryotic supergroup for which we lacked genetic information in our previous phylogenomic analyses. Dr. John Archibald, his co-investigators (Dr. M.W. Gray, Dr. G.I. McFadden, Dr. P.J. Keeling, and Dr. C. Lane), and the Joint Genome Institute provided access to their data for the first Rhizarian genome sequence (of *Bigelowiella natans*) prior to its public release. Dr. John Logsdon advised Banoo and I on the research design and implementation, provided helpful discussion, and revised the manuscript.

My thesis committee members, Dr. Logsdon and Dr. Malik all provided helpful comments or discussion for Chapters 1 and 5.
CHAPTER 1
GENERAL INTRODUCTION

All known extant eukaryotes descended from an ancestor (Darwin 1859) that lived approximately 2.1 - 2.7 billion years ago, according to geochemical and fossil evidence (Han and Runnegar 1992; Brocks et al. 1999). Based upon the distributions of traits among eukaryotes, the last common ancestor was most likely a free-living, unicellular eukaryote that occupied moderate (mesophilic), aerobic environments and obtained nutrients by engulfing other organisms (phagocytosis) (Cavalier-Smith 2002a). Today, a wide variety of unicellular and multicellular eukaryotes are observed that live in diverse habitats (e.g. aerobic, anaerobic, extremophilic, and mesophilic) and fulfill many different lifestyles (e.g. symbiotic, free-living, sexual, and asexual) (Knoll 2003; Adl et al. 2005).

Remarkably, all extant eukaryotic lineages began their evolutionary journeys with the same genetic material (i.e. a common ancestral genome) that was subsequently shaped by random genetic mutations (Watson and Crick 1953) and natural selection (Darwin 1859). However, which genes were present within that ancestral genome and how those genes subsequently evolved are open questions. Elucidating the origins of genes that encode products responsible for important biological processes provides a means of comparing extant eukaryotes to their ancestors (Villeneuve and Hillers 2001). Although direct observation of the ancestor of extant eukaryotes is obviously impossible, inferring which genes were likely to have been present within its genome is possible (Dacks and Doolittle 2001). By comparing inferred suites of genes in the last common ancestor of eukaryotes to the suites of genes present in extant eukaryotes we can study the evolutionary histories of the genes themselves. In this way, we can gain insight into the origins and evolution of important biological reactions and we can begin to establish the order of events that occurred during the early evolution of eukaryotes (Roger 1999).
An approach to studying the origin and evolution of genes is to search for them within the genomes of diverse organisms (Dacks and Doolittle 2001). Among eukaryotes, animals, fungi, and plants are estimated to represent the global majority of named species (Fenchel and Finlay 2004). Currently six eukaryotic groups (supergroups) have been proposed on the basis of ultrastructural, genetic, and phylogenetic analyses (Figure 1.1) (Cavalier-Smith 2004; Baldauf 2008). Animals, fungi, and plants occupy only two eukaryotic supergroups (Opisthokonta and Archaeplastida), while protists (eukaryotic organisms with unicellular, colonial, filamentous, or parenchymatous organization that lack vegetative tissue differentiation, except for reproduction (Adl et al. 2005)) are present in all six eukaryotic supergroups and are the predominant or sole occupants of four of them (Amoebozoa, Chromalveolata, Excavata, and Rhizaria; see Current state of the eukaryotic phylogeny below) (Adl et al. 2005; Adl et al. 2007). Therefore, including diverse protists in evolutionary studies is important in order to sample the full breadth of eukaryotes (Ramesh, Malik, and Logsdon 2005).

The presence of orthologs (genes inherited from common ancestors) (Ridley 2004) among groups of eukaryotes implies that those genes were present in their last common ancestor (Villeneuve and Hillers 2001; Ramesh, Malik, and Logsdon 2005). If genes are detected in the genomes of representatives of all known eukaryotic groups then they are inferred to have been present in the last common ancestor of all known eukaryotes (Koonin 2010). Apparent absences of particular genes from the genomes of eukaryotes may be observed if either the gene arose later during eukaryotic evolution (after the evolutionary divergence of lineages from other eukaryotes) or it was subsequently lost (Dacks and Doolittle 2001). The interpretation of apparently missing genes depends upon our current understanding of the evolutionary relationships among eukaryotes (i.e. the eukaryotic phylogeny).
The origin of eukaryotes

All living organisms share characteristics which indicate they arose from a common cellular ancestor (Darwin 1859) or a population of ancestral cells promiscuously exchanging molecules (Doolittle et al. 2008). To name a few, all living things are cellular, using ATP for energy, DNA as the hereditary genetic material, a common genetic code (for the most part), and similar transcription and translation machinery including RNA (Griffiths et al. 2000; Knoll 2003). The phylogenetic tree of life, which depicts the genealogical relationships of all living organisms, is composed of three major branches (domains) occupied by eubacteria (Bacteria), archaebacteria (Archaea), and eukaryotes (Eucarya) (Figure 1.2) (Woese and Fox 1977; Woese, Kandler, and Wheelis 1990; Brown and Doolittle 1995). The position of the root of the tree of life, with Bacteria on one side and Archaea and Eucarya on the other, was determined by phylogenetic analyses (Gogarten et al. 1989; Iwabe et al. 1989). This tree topology, which proposes that eubacteria are the earliest-diverging forms of life (Pool 1990), is supported by fossil and biogeochemical data (Brocks et al. 1999; Knoll 2003). However, the relationship between Archaea and Eucarya is currently in dispute. The three domain hypothesis proposes that archaebacteria and eukaryotes are sisters (monophyletic groups that share a common ancestor) (Cavalier-Smith 1987a), while the eocyte hypothesis proposes that eukaryotes are not sisters of but arose from within the archaebacterial lineage (Crenarchaeota) (Lake et al. 1984). Compelling arguments put forward by Cavalier-Smith (1987 and 2002) point out that the most parsimonious interpretation of the distributions of homologous features among the three domains is that Archaea and Eucarya are sisters (Cavalier-Smith 1987a; Cavalier-Smith 2002c). However, a relatively recent set of robust phylogenetic analyses appear to support the eocyte hypothesis (Archibald 2008; Cox et al. 2008). As this issue is currently unresolved and the effect of this distinction to the following discussion is subtle, by convention, I will continue with the three domain model.
All extant eukaryotes share features that distinguish them from other forms of life and support their common ancestry (Maynard Smith and Szathmary 1995). The most relevant of these features to the subject of meiosis are linear chromosomes, contained within a nucleus that is part of an endomembrane system (which includes the nuclear envelope, endoplasmic reticulum, Golgi apparatus, and lysosomes) and an endoskeleton (Cavalier-Smith 2002a; Cavalier-Smith 2010). Eukaryotes also possess mitochondria (Roger 1999) (although some have highly derived forms of mitochondria called hydrogenosomes (Muller 1993) and mitosomes (Tovar, Fischer, and Clark 1999) instead) and many eukaryotic cells also contain photosynthetic plastids (Adl et al. 2005).

Mereshkowsky (1910) and Koso-Polyanski (1924) first proposed that mitochondria and chloroplasts are symbionts that arose from the engulfment of bacteria by an ancestral eukaryotic cell; this forgotten concept was later independently revived by Lynn Margulis (Sagan 1967; Maynard Smith and Szathmary 1995; Knoll 2003). It is now widely accepted (Embley and Martin 2006; Poole and Penny 2007) that mitochondria and chloroplasts are the endosymbiotic descendants of bacteria (α-proteo- and cyanobacteria, respectively) that were engulfed by eukaryotes (Margulis 1970; Gray and Doolittle 1982; Gray 1989). The most convincing support for the origins of organelles by endosymbioses comes from phylogenetic analyses that indicate genes from mitochondria or chloroplasts are more closely related to bacteria than to eukaryotes (Poole and Penny 2007). While endosymbioses of cyanobacteria have likely occurred multiple times after the divergence of eukaryotes from their last common ancestor (Yoon et al. 2004), the engulfment of α-proteobacteria probably occurred one time prior to the divergence of all extant eukaryotes (Roger 1999). These observations have led to hypotheses in which the nucleus is also proposed to be an endosymbiont, usually an archaebacterium (Lake and Rivera 1994; Horiike et al. 2001; Shinozawa, Horiike, and Hamada 2001).
If we compare the cytological and phylogenetic features of mitochondria and chloroplasts to the nucleus, several important differences are apparent (Poole and Penny 2007). Unlike mitochondria and chloroplasts, which have at least two different membranes (i.e. the original bacterial membrane surrounded by the eukaryotic endomembrane), the nuclear envelope is in dynamic continuity with the rest of the endomembrane apparatus (Margulis 1970). In addition, although endosymbioses of eubacterial and eukaryotic (secondary plastids) cells within eukaryotic host cells are well known, there are no known cases of archaebacterial intracellular endosymbionts (Poole and Penny 2007). Finally, although many genes within eukaryotic genomes have been identified as eubacterial or archaebacterial homologs (Koonin 2010), phylogenetic analyses consistently retrieve topologies in which eukaryotic genes form distinct monophyletic groups and do not arise from within eubacterial or archaebacterial groups (Poole and Penny 2007). These topologies differ from evolutionary relationships inferred from phylogenetic analyses of mitochondrial and plastid genes. The cytological and phylogenetic differences between these organelles are best explained by autogenous models of nuclear formation (Martin 1999).

According to the neomuran (“new walls”) hypothesis the most important event during eukaryotic evolution was the replacement of the peptidoglycan murein in eubacterial cell walls with N-linked glycoproteins in the common ancestor of archaebacteria and eukaryotes, resulting in a “…more flexible surface coat…” (Cavalier-Smith 2002a). Initially, this change may have provided resistance to antibiotics similar to penicillin that disrupted peptidoglycan synthesis (Maynard Smith and Szathmary 1995). Archaeabacteria may have substituted eubacterial acyl ester lipids with prenyl ether lipids, resulting in a new exoskeleton, while proto-eukaryotes retained the flexible surface (Cavalier-Smith 1987a; Cavalier-Smith 2002a). The eukaryotic membrane, along with a complex cytoskeleton, allowed for the evolution of a phagocytic lifestyle, in which particles (including other cells) are engulfed within a vacuole (Cavalier-Smith 1987a;
Cavalier-Smith 2002a). In short, subsequent invaginations resulted in the formation of the endomembrane system, including the nuclear envelope (Cavalier-Smith 1987a; Cavalier-Smith 1988; Cavalier-Smith 2002d; Cavalier-Smith 2010). Internal compartmentalization may have provided spatial and temporal control of transcription and translation and protected the genomic integrity of the proto-eukaryote (Cavalier-Smith 1987a; Cavalier-Smith 1988; Cavalier-Smith 2002d; Cavalier-Smith 2010). The origin of nucleated cells, which are dramatically different from eubacterial or archaeabacterial cells that have no endomembrane system or organelles, necessitated the evolution of distinctly eukaryotic modes of nuclear division (Maynard Smith and Szathmary 1995).

**A comparison of mitotic and meiotic divisions**

Among eukaryotes two types of nuclear division are possible, mitosis and meiosis (Figure 1.3) (Griffiths et al. 2000). The mitotic nuclear division, in which two genetically identical cells arise from one, is the only mode of replication for somatic or vegetative cells in multicellular organisms and serves as a form of asexual reproduction in unicellular organisms (Flemming 1878; Weismann, Parker, and Ronnfeldt 1893; Huxley 1942). The generalized meiotic nuclear division, during which the genetic content of the cell products is halved, is the sole source of the cells necessary for sexual reproduction (e.g. spores or gametes) in eukaryotes (Weismann, Parker, and Ronnfeldt 1893; Churchill 1970). This halving of the genetic material during meiosis serves to maintain appropriate numbers of chromosomes when cells are subsequently fused, restoring the parental state (Weismann, Parker, and Ronnfeldt 1893).

During both mitotic and pre-meiotic interphases, the genomes of cells are replicated once (the synthetic or S-phase) (John 1990). Although mitotic and pre-meiotic S-phases are similar in that the duplication of chromosomes results in paired, nearly identical, chromosome copies (sister chromatids), there are differences that distinguish them (John 1990; DePamphilis 1996). For example, in yeast, pre-meiotic S-
phase is 2-3 times longer than the mitotic S-phase in diploids (approximately 30 and 65 minutes, respectively) (Williamson et al. 1983). This phenomenon has also been witnessed in different animals, such as the newt Triturus (Callan 1972) and the fruit fly Drosophila (Chandley 1966). These differences may be attributed to variation in numbers and activation of replicon origins and the rate of replication fork migration (John 1990; DePamphilis 1996). It is interesting, however, that when cells undergoing pre-meiotic S-phase were removed from the anthers of Lilium and Trillium and placed in a culture medium, the cells successfully completed mitosis (Lima-de-Faria 1969; Ito and Takegami 1982), indicating that although pre-meiotic S-phase is different than mitotic S-phase, they are similar enough in these organisms that mitosis still proceeds (John 1990).

Following interphase, cells undergoing either mitotic or meiotic divisions enter a stage called prophase (John 1990; DePamphilis 1996; Griffiths et al. 2000). During mitotic prophase, the paired sister chromatids contract into a series of coils, packaging them for alignment along the metaphase plate and subsequent segregation to opposite sides of the nucleus later during the mitotic cell cycle (Figure 1.3 A. - ii.) (John 1990; Griffiths et al. 2000). Thus mitosis is distinguished by a single round of DNA replication followed by a single nuclear division (Flemming 1878). Because the numbers of homologous chromosome sets (ploidy – e.g. diploid or 2n) is maintained, the mitotic division is called an equational division (John 1990). In meiosis, two rounds of division (Meiosis I and Meiosis II) follow a single round of DNA replication (Figure 1.3 B.) (Churchill 1970). The first meiotic division results in the reduction of ploidy (diploid (2n) to haploid (n)) (John 1990), and, so, is called the reductional division, while the second meiotic division is equational because the haploid state is maintained (Weismann, Parker, and Ronnfeldt 1893; Churchill 1970). Thus, in a single diploid (2n) cell, meiosis yields four haploid (n) products (although in female meiosis not all products survive as gametes) (John 1990).
Meiotic prophase I is much longer and more complex than mitotic prophase, due primarily to the formation of bivalents (paired homologous chromosomes) during meiosis that do not form during mitosis (Figure 1.3 B. - iii). In addition, the formation of synaptonemal complexes (Carpenter 1987) and/or crossing over (chiasma) (Ruckert 1892; Janssens 1909) between homologous chromosomes are required in most (but not all) organisms for appropriate pairing and segregation (John 1990). However, it is possible that the binding of microtubules to chromosomes may be even more important than the formation of bivalents in determining the segregation patterns of the chromosomes (Simchen and Hugerat 1993). Sets of microtubular structures (spindles), arising from barrel-shaped organelles (centrioles) located at opposite sides of the nucleus, attach to protein structures (kinetochores) that are associated with sister chromatid contact points (centromeres) (Figure 1.3 A. – iii and B. – iii and iv) (John 1990; Griffiths et al. 2000). During mitosis, each sister chromatid is attached to an opposing set of microtubules, resulting in alignment of the chromosomes between two poles (Simchen and Hugerat 1993). These microtubules exert opposing forces toward the poles (John 1990; Simchen and Hugerat 1993). When sister chromatid cohesion lapses, sisters segregate to opposite poles (John 1990; Simchen and Hugerat 1993). In meiosis I, microtubules connect to only one sister chromatid per chromosome, on opposite sides of bivalents, when attachments between homologous chromosomes lapse, sisters co-segregate to opposite poles (John 1990; Simchen and Hugerat 1993). Thus the interactions of opposing microtubules with one or both sister chromatids (unipolar or bipolar attachment) determine whether reductive or equational divisions occur (Simchen and Hugerat 1993). The same process that is used for equational divisions during mitosis is used during meiosis I for reductive divisions, in the presence of bivalents, modified kinetochores, and persistent sister chromatid cohesion (Nicklas 1977). Indeed, yeast (Saccharomyces cerevisiae) cells that are in the process of meiosis can be transferred to a vegetative medium, resulting in diploid colonies with
recombined genetic markers (Sherman and Roman 1963). These cells likely formed bivalents and crossovers, (resulting in genetic recombination at high, meiotic levels) followed by mitotic (equational) divisions (Simchen and Hugerat 1993).

Given they are both equational divisions, it is tempting to conclude that the second meiotic divisions are the same as mitotic divisions (John 1990). It is true that the second meiotic division does not require many proteins necessary for completion of the first meiotic division, during which several specialized proteins known to function only during pairing of homologous chromosomes and formation of synaptonemal complexes and crossovers are necessary (Paques and Haber 1999). However, there are at least three important differences; 1) during mitosis, sister chromatids are associated along their entire lengths but, during meiosis II, sister chromatid cohesion is maintained only around centromeres, resulting in splayed chromosome arms; 2) paired sister chromatids in mitosis are genetically identical, while, in meiosis II, sister chromatids are not identical, due to genetic recombination; and 3) nuclei are diploid (2n) during mitosis but haploid (n) during meiosis II (John 1990; DePamphilis 1996). Therefore, although mitotic and meiotic equational divisions are, in principle, the same, they are not identical (John 1990).

In total, all of these observations indicate that, although mitosis and meiosis are very similar, hinting at a close evolutionary relationship, they are also distinguished by important functional differences. The apparent similarities presented here are confirmed by studies which indicate that many proteins necessary for the completion of mitosis are also important for completion of meiosis (Marcon and Moens 2005). Likewise, the presence of proteins known only to function during meiosis sheds light on an unparalleled evolutionary history. The functions of these proteins have been studied most often in animals, fungi, and plants (representing the eukaryotic supergroups Opisthokonta and Archaeplastida (Figure 1.1; and discussed further below), although other organisms, such as the ciliate Tetrahymena thermophila (Cole et al. 1997)
(Chromalveolata) and the amoeba *Entamoeba histolytica* (Lopez-Casamichana *et al.* 2008) (Amoebozoa) have also been studied. Though there are differences in meiosis among different eukaryotes, the proteins studied here are highly conserved in both sequence and function. Although not all eukaryotes have been studied, the similar functions of these proteins among four of the six currently recognized eukaryotic supergroups and the high degree of amino acid sequence conservation, strongly implies that the proteins fulfill the same functions in unstudied extant organisms. Furthermore, the inferred presence of genes encoding these proteins in the common ancestors of eukaryotes strongly implies that mitotic and meiotic functions were also occurring (Ramesh, Malik, and Logsdon 2005).

**The origin and evolution of meiotic genes**

Due to the presence of mitosis in all extant eukaryotes, it is widely accepted that this nuclear division was likely to have been present in their last common ancestor (Cavalier-Smith 1981b). Furthermore, it is widely accepted that genes encoding proteins that function during mitosis were present in the common ancestor of all extant eukaryotes (Eme *et al.* 2009; Wickstead, Gull, and Richards 2010). More contentious is the notion that meiosis, and genes encoding proteins that function during meiosis, were present in the ancestor of all extant eukaryotes (Malik *et al.* 2008). This is due primarily to the fact that, although mitosis has been observed in all extant eukaryotes, meiosis is not observed in some putatively asexual eukaryotes (Schurko and Logsdon 2008; Schurko, Neiman, and Logsdon 2009). Specifically, the apparent absences of meiosis and sexual reproduction during the lifecycles of *Giardia intestinalis*, *Trichomonas vaginalis*, and *Vairimorpha necatrix* led to the speculation that these organisms diverged prior to the origin of meiosis (Cavalier-Smith 1989). Molecular phylogenetic analyses of small ribosomal subunit and translation elongation factor EF-1 alpha nucleotide sequences yielded tree topologies in which these supposed “primitive” organisms were depicted as the earliest diverging eukaryotes (Leipe *et al.* 1993; Kamaishi *et al.* 1996; Hashimoto *et
These eukaryotic phylogenies appeared to support the Archezoa hypothesis (Cavalier-Smith 1989), in which organisms with presumed ancestral features and no observed meiosis emerged early during eukaryotic evolution. Such features include: prokaryote-like transcriptional apparatus (van Keulen et al. 1991a; van Keulen et al. 1991b) and the apparent absence of mitochondria (Tovar, Fischer, and Clark 1999). Complex organisms (i.e. animals, fungi, and plants), would form a “crown” at the top of the eukaryotic tree of life (Sogin, Elwood, and Gunderson 1986; Woese, Kandler, and Wheelis 1990; Brinkmann and Philippe 2007).

Subsequent phylogenetic studies with more sophisticated methods and data, using more realistic models of protein substitution, have revealed that the placement of organisms at the base of the original phylogenetic trees were the result of a statistical anomaly called long-branch attraction (Edlind et al. 1996; Keeling and Doolittle 1996; Hirt et al. 1997; Hirt et al. 1999; Felsenstein 2004). Although the possibility remains that *T. vaginalis* and *G. intestinalis* may be among the earliest-diverging eukaryotes, *V. necatrix* (a Microsporidian) is now known to be a fungus (Hirt et al. 1999). The presumption that the Archezoa lack mitochondria has also been proven erroneous (Roger 1999). Instead, derived and highly reduced forms of mitochondria (mitosomes and hydrogenosomes) have been discovered in representatives of each putatively early-diverging eukaryotic lineage (Germot, Philippe, and LeGuyader 1997; Tovar, Fischer, and Clark 1999; Tovar et al. 2003; van der Giezen 2009). However, sexual reproduction has yet to be observed among any of these lineages.

Direct observation of meiosis is often difficult or impossible with many diverse eukaryotes (Schurko and Logsdon 2008). However, we may determine if organisms have the potential to undergo meiosis by the presence of genes (Schurko and Logsdon 2008). We can also use the distribution of meiotic genes to infer their presence in the common ancestors of different eukaryotic groups (including the ancestor to all eukaryotes) (Dacks and Doolittle 2001; Villeneuve and Hillers 2001; Ramesh, Malik,
By using phylogenetic analysis it is possible to determine when meiosis-specific genes, and meiosis, arose and to determine if any eukaryotes diverged evolutionarily prior to the origins of these meiotic genes (Ramesh, Malik, and Logsdon 2005; Malik et al. 2008). That is, it is possible to determine if apparent gene absences are primitive or derived states (Dacks and Doolittle 2001). Some eukaryotes may have diverged prior to the origin of sexual reproduction in eukaryotes while others may utilize a primitive type of meiosis (Cleveland 1947; Cleveland 1956; Cavalier-Smith 1981b; Archetti 2004).

Previously, a set of “core meiotic recombination machinery” (Spo11, Rad50, Mre11, Dmc1, Rad51, Msh4, Msh5, and Mlh1) was defined by Villeneuve and Hillers (2001) as a collection of highly conserved orthologs present in animals, fungi, and plants (Villeneuve and Hillers 2001). This list includes some components known to function only during meiosis in model organisms (Spo11, Dmc1, Msh4, and Msh5) (Bishop et al. 1992; Lichten 2001; Snowden et al. 2004). Thus the authors rightly pointed out that at least three events were important for the evolution of meiosis: endogenous double-strand DNA breaks (Spo11 (Keeney, Giroux, and Kleckner 1997)), interhomolog DNA strand exchange (Dmc1 (Bishop et al. 1992)), and resolution of Holliday junctions as crossovers (Villeneuve and Hillers 2001). Furthermore, the distribution of these genes among animals, fungi, and plants implies that they arose prior to the divergence of the eukaryotes considered. Villeneuve and Hillers suggest, therefore, that the genes arose in the common ancestor of all eukaryotes (Villeneuve and Hillers 2001). However, given current hypotheses for rooting the eukaryotic phylogeny (Stechmann and Cavalier-Smith 2002; Stechmann and Cavalier-Smith 2003a; Cavalier-Smith 2010), only the placement of the root between the Bikonta and the Unikonta (Figure 1.1) would support the conclusion that the ancestor of animals, fungi, and plants is also the last common ancestor of eukaryotes. That is, if either members of the Metamonada or Discoba are the earliest-diverging eukaryotes (discussed further in
Current state of the eukaryotic phylogeny below) then the meiotic genes could have arisen after their divergence. Then, the meiotic genes would be present in the ancestor of animals, fungi, and plants but not the ancestor of all eukaryotes. More complete testing of the eukaryotes would need to be performed to determine if any organisms diverged prior to the origin of meiosis or if the last common ancestor of eukaryotes was capable of meiosis.

Additional studies tested the specific hypotheses that *G. intestinalis* and *T. vaginalis* diverged prior to the origin of meiosis by determining which genes that encode products necessary for completion of meiosis (Ramesh, Malik, and Logsdon 2005; Malik et al. 2008). In total; 29 genes (9 meiosis-specific) were studied among eukaryotes representing five of the six currently recognized eukaryotic supergroups. Several genes tested are present in *G. intestinalis* and *T. vaginalis* (21 and 27, respectively), including the meiosis-specific genes (6 and 8, respectively). These studies indicate that *G. intestinalis* and *T. vaginalis* are not candidates for ancient asexuality.

Although previous studies have failed to produce a candidate lineage for ancient asexuality or primitive meiosis, they provided the basis for addressing open questions about when and how meiosis arose and how it subsequently evolved (Ramesh, Malik, and Logsdon 2005; Malik et al. 2008). That is, rather than using a candidate organism approach, in which different species that may represent early-diverging eukaryotic lineages are studied, here we have employed a gene-centric approach in which the distribution and phylogenetic analyses of genes are the focus. Thus the goals of this thesis were not to detect meiosis or to identify putative ancient asexual organisms *per se* but, instead, to study the evolution of meiotic genes in their own right. Specifically, I address the following questions: 1) Were meiotic gene present in the genome of the last eukaryotic common ancestor?; 2) By what genetic mechanisms did meiotic genes arise?; 3) Could the products that meiotic genes encoded in the last eukaryotic ancestor have
functioned during meiosis?; and 4) How have the suites of meiotic genes observed in various extant eukaryotic genomes evolved? The answers to these questions will contribute to our understanding of the origin and evolution of meiosis and, more generally, the evolution of eukaryotes.

Components of meiotic interhomolog DNA strand exchange

To add resolution to our view of the evolution of complex processes that occur during meiosis, the genes encoding proteins central to interhomolog DNA strand exchange are the focus of this thesis. Interhomolog recombination (X in Figure 1.3 – B – iii) can result in gene conversion and/or crossing-over that greatly increases the efficacy of natural selection (Rice and Chippindale 2001; Agrawal 2006; Otto and Gerstein 2006). That is, it produces novel combinations of genes (Figure 1.3 – B – v, black and grey regions of chromosomes) that, when combined with other products of meiosis (e.g. fertilization), enabling eukaryotes to respond evolutionarily to changing environments more rapidly than asexual organisms (Fisher 1930; Muller 1932; Van Valen 1973). Although many benefits of recombination are observed at the population level, its origin and persistence during the evolution of eukaryotes is due, more likely, to the selective benefits of the appropriate pairing and segregation of homologous chromosomes in maintaining genomic integrity (Kleckner 1996; Villeneuve and Hillers 2001; Cavalier-Smith 2002d). In most cases, interhomolog DNA strand exchange (Figure 1.3 – B – iii) is necessary for the formation of bivalents and correct segregation of homologous chromosomes to opposite poles during meiosis I (Moore and Orr-Weaver 1998). Following these observations, it has been declared that “…the very essence of sex is meiotic recombination.” (Villeneuve and Hillers 2001). Different models of meiotic recombination have been proposed, such as the synthesis-dependent strand annealing and double-strand break repair models (Paques and Haber 1999). In each of these models, the interhomolog DNA strand exchange reaction that occurs
between homologous chromosomes during meiosis I (Figures 1.3 and 1.4) is central to meiotic interhomolog DNA strand exchange (Paques and Haber 1999). Thus, the very essence of meiotic recombination is the interhomolog DNA strand exchange reaction. Therefore, the best way to gain a more complete understanding of the evolution of genes involved in meiosis and to detect any sort of “primitive” meiosis is to study the interhomolog DNA strand exchange reaction; the components involved in interhomolog DNA strand exchange during meiosis are the main focus of this thesis.

Genetic recombination between sister chromatids is important for repair of DNA double-strand (dsDNA) breaks that may be caused by mutagens or by collapsed or damaged replication forks (John 1990). During entry into meiosis, dsDNA cuts introduced by meiosis-specific proteins (Spo11-1 or Spo11-2) are repaired with recombination between either sister chromatids or homologous chromosomes (Keeney, Giroux, and Kleckner 1997; Hartung et al. 2002). In animals, fungi, and plants, several genes whose products are important for both sister chromatid and interhomolog DNA strand exchange (Rad52, Rad59, Rad51, Rad55, Rad57, Rad54, and Rdh54) and some that are known to function only during meiosis in model organisms (Dmc1, Hop2, and Mnd1) have been studied extensively (see Table 2.5 and the description of Figure 1.4 for references). A general model of the interactions of thirteen proteins that function during interhomolog DNA strand exchange is presented (Figure 1.4). This model illustrates four important steps: i) formation of a Rad51/Dmc1-ssDNA pre-synaptic filament on a 3’ ending DNA strand (A-D); ii) capture of a DNA duplex by the pre-synaptic filament (E and F); iii) search by the pre-synaptic filament for regions of DNA duplex homology (F.); and iv) invasion of DNA duplex by the pre-synaptic filament and D-loop formation (G) (Filippo, Sung, and Klein 2008). Studies in which searches for genes that encode Rad51, Rad52, Dmc1, Hop2, and Mnd1 proteins among diverse eukaryotes (i.e. all eukaryotic supergroups except Rhizaria; Figure 1.1) indicate that, due to their distributions, these strand exchange components must have arisen very
early during eukaryotic evolution (Villeneuve and Hillers 2001; Ramesh, Malik, and Logsdon 2005; Malik et al. 2008).

Interestingly, some of these genes have not been found within the genomes of many diverse, less well-studied, eukaryotes (Ramesh, Malik, and Logsdon 2005; Malik et al. 2008). These apparent absences may be due either to true losses of genes from genomes or they may represent instances of non-detection (i.e. type II error). Previous investigation of the distributions of genes cannot easily distinguish between these possibilities (Villeneuve and Hillers 2001; Ramesh, Malik, and Logsdon 2005; Malik et al. 2008). This limitation inhibits studies of the distribution of genes among eukaryotes since different suites among diverse organisms cannot be verified bioinformatically and functional studies may be difficult or impossible to perform, especially with large numbers of eukaryotes. To address this issue, the distributions of 13 genes that encode meiotic interhomolog DNA strand exchange components (Figure 1.4) among 47 diverse eukaryotes (representing all supergroups except Rhizaria, Figure 1.1) were determined. In addition, these data were useful for the development of a heuristic metric for determining the likelihood that observed absences represent true losses of genes from genomes. For the first time, we were able to assess our confidence in the suites of genes observed in diverse, relatively unstudied, eukaryotic genomes. This new insight allowed us to study patterns in the distributions of strand exchange genes across eukaryotes and to formulate an evolutionary hypothesis explaining these patterns. This project is presented in Chapter 2 of this thesis.

Of particular interest are the eukaryotic *Rad51* (Shinohara, Ogawa, and Ogawa 1992) and *Dmc1* (Bishop et al. 1992) genes whose products catalyze homologous DNA strand exchange during genetic recombination (Paques and Haber 1999). Both *Rad51* and *Dmc1* are related (orthologous (Ridley 2004)) to the eubacterial *recA* and the archaeabacterial *RadA* genes, whose products function during homologous recombination and DNA repair (Cox 1993; Clark and Sandler 1994; Camerini-Otero
and Hsieh 1995; Sandler et al. 1996). The master recombinase (Rad51) forms right-handed helical filaments on single-stranded and double-stranded DNA (Conway et al. 2004) during repair of all double-strand breaks (Shinohara, Ogawa, and Ogawa 1992). The Dmc1 proteins function similarly, promoting interhomolog strand exchange only during meiosis in model organisms (Figure 1.1) (Bishop et al. 1992). Among animals, fungi, and plants, rad51 mutants experience reduced recombination, resulting in decreased resistance to mutagens, and diminished sporulation or fertility, while mutations in vertebrates cause embryonic lethality (Bishop 1994; Bleuyard, Gallego, and White 2006). In animals, fungi, and plants, dmc1 mutants reduce or eliminate homologous recombination during meiosis (Bishop et al. 1999; Tsubouchi and Roeder 2003). Given the evolutionary relationships of eukaryotic Rad51 and Dmc1 genes to eubacterial recA and archaeabacterial RadA genes and the central role of DNA strand exchange catalysis to the DNA damage repair in all organisms and meiosis in eukaryotes, elucidating the evolutionary histories of Rad51 and Dmc1 is important to understanding the origin and evolution of meiosis.

Previous studies, including the one presented in Chapter 2, in which the distributions of Rad51 and Dmc1 genes were studied, were limited by the availability of genome sequence data from diverse eukaryotic lineages. We have concluded that these genes arose very early during eukaryotic evolution but not whether they were present in the last common ancestor of all eukaryotes. Determining more accurately when Rad51 and Dmc1 arose during eukaryotic evolution and whether there are any other organisms or groups of organisms that lack these genes may provide valuable insight into the evolution of interhomolog DNA strand exchange during meiosis. The presence of the Dmc1 gene may also serve as proxy for the presence of meiosis itself (Ramesh, Malik, and Logsdon 2005; Malik et al. 2008). That is, although the absence of Dmc1 does not indicate that meiosis is absent, the presence of a functional copy indicates that meiosis is likely to be present (Schurko and Logsdon 2008). If the Dmc1 gene is found in
representatives of all eukaryotic groups we can infer that their ancestor also possessed a
\textit{Dmc1} gene (Dacks and Doolittle 2001; Koonin 2010) and was likely to have been
capable of meiosis (Villeneuve and Hillers 2001; Ramesh, Malik, and Logsdon 2005;

In Chapter 3, using a combination of extensive searches of gene and protein
sequence repositories and degenerate PCR, we demonstrate that both \textit{Rad51} and \textit{Dmc1}
genes are present in genomes of organisms representing all known eukaryotic
supergroups (Figure 1.1) and were, therefore, likely to have been present in the genome
of the last eukaryotic common ancestor. To understand the importance of specific
amino acids to the functions of Rad51 and Dmc1, we aligned protein sequence data
from all known eukaryotic supergroups and identified amino acids that are highly
conserved among them. We also identified several amino acid residues that may confer
Rad51- or Dmc-specific functions, due to their conservation in one set of proteins but
not the other, and that were likely to have been present in the last common ancestor of
all extant eukaryotes. Collectively, these data imply that \textit{Rad51} and \textit{Dmc1} genes
present within the genome of the last common ancestor of eukaryotes encoded proteins
that functioned during both mitosis and meiosis.

Finally, although the distributions of genes among diverse eukaryotes allow us
to infer \textit{when} meiotic genes may have arisen, more analyses are necessary to determine
\textit{how} they arose. What known genetic mechanisms yielded genes that encode products
that function during meiosis? Phylogenetic studies of meiotic genes can provide insight
into their evolutionary histories and may inform their origins (Ramesh, Malik, and
Logsdon 2005; Malik \textit{et al.} 2008). For example, studies indicate that Rad51 and \textit{Dmc1}
genes are paralogs (genes resulting from gene duplication events (Ridley 2004))
and \textit{Dmc1} genes are orthologous to both the eubacterial \textit{recA} and the archaeabacterial
\textit{RadA} genes that are known to be important for DNA damage repair in prokaryotes
(Marcon and Moens 2005), two genes, one encoding proteins that are known to catalyze DNA strand exchange during both mitosis and meiosis (Rad51) and the other encoding proteins that are known to catalyze interhomolog DNA strand exchange only during meiosis in model organisms (Dmc1), arose from a single gene that most likely encoded products involved in DNA damage repair early during eukaryotic evolution (Figure 1.5) (Ramesh, Malik, and Logsdon 2005; Lin et al. 2006; Malik et al. 2008). Thus we can study the evolutionary histories of these genes to determine when meiosis may have arisen during eukaryotic evolution and whether any organisms diverged prior to the duplication event yielding the meiosis-specific gene (Dmc1) (Figure 1.5).

Similarly, this pattern has been observed in studies of the Spo11 paralogs (Malik et al. 2007). The Spo11-1 and Spo11-2 genes encode meiosis-specific products (Atcheson et al. 1987; Keeney, Giroux, and Kleckner 1997; Hartung et al. 2002) that are paralogs of the Spo11-3 gene (Malik et al. 2007) whose products function only during vegetative growth DNA in Arabidopsis thaliana (Hartung and Puchta 2001; Sugimoto-Shirasu et al. 2002; Yin et al. 2002). The Spo11 homologs are orthologous to the archaebacterial Top6A gene (Atcheson et al. 1987), a type II topoisomerase that functions to separate replicated chromosomes (Bergerat et al. 1997; Nichols et al. 1999; Corbett and Berger 2003). Like Rad51 and Dmc1, the evolutionary history of Spo11 genes is similar to the model presented in Figure 1.5 (Malik et al. 2007). Indeed, Malik (2007) demonstrated that many meiosis-specific genes fit this pattern (Malik 2007). We hypothesized, therefore, that meiosis may have arisen in toto by large-scale gene duplications, early during eukaryotic evolution. Consistent with this hypothesis, it has been shown that large-scale gene duplication events may have occurred early during eukaryotic evolution (Zhou, Lin, and Ma 2010).

Due primarily to the availability of more sensitive gene sequence search methods (see Chapter 4 Methods), more realistic models of protein substitution (e.g. LG) for phylogenetic analyses, and unprecedented access to genome sequence data for
diverse eukaryotic groups, we were able to extend the previous study on the origins of meiosis-specific genes by duplication to include representatives of all known eukaryotic supergroups (Figure 1.1). We determined the eukaryote-wide distributions of twenty genes that encode products that perform five important functions during meiosis: 1) pairing of homologous chromosomes; 2) sister chromatid cohesion; 3) dsDNA cuts; 4) interhomolog DNA strand exchange; and 5) Holliday junction resolution (Table 4.1). Eighteen out of 20 genes tested fit the pattern presented in Figure 1.5. Furthermore, given their phylogenetic distributions among eukaryotes groups, these paralogs are inferred to have been present in the ancestor of all extant eukaryotes.

Current state of the eukaryotic phylogeny

An important motivation for the studies presented in Chapters 3 and 4 was to determine which components were likely to have been present in the common ancestor to eukaryotes. In order to correctly interpret the distributions of genes among eukaryotes and the phylogenetic analyses of their products, an accurate understanding of the evolutionary relationship of eukaryotes is required. The following discussion provides the appropriate framework for such studies.

Eukaryotes can be divided into at least six “supergroups” (Opisthokonta, Amoebozoa, Archaeplastida, Chromalveolata, Rhizaria, and Excavata) and at least one group of unclassified organisms (Apusozoa) on the basis of phenotypic, ultrastructural, and phylogenetic studies (Figure 1.1) (Baldauf 2003; Simpson and Roger 2004; Roger and Hug 2006; Baldauf 2008; Cavalier-Smith 2010; Roger and Simpson 2009). Two major divisions of eukaryotes (Unikonta and Bikonta) are recognized (Cavalier-Smith 2002a). The Unikonta (Opisthokonta + Amoebozoa) are named for the ancestral possession of a single flagellum and are distinguished by the fusion of three genes that encode enzymes that synthesize pyrimidine nucleotides (Cavalier-Smith 2002a; Stechmann and Cavalier-Smith 2002). The Bikonta (named for the presence of two flagella in their last common ancestor) share a similar two gene fusion (dihydrofolate
reductase and thymidylate synthase) (Stechmann and Cavalier-Smith 2002; Stechmann and Cavalier-Smith 2003b; Stechmann and Cavalier-Smith 2003a). Recently, phylogenetic analyses including sequence data from representatives of an unclassified group of eukaryotes (Apusozoa) have challenged the monophyly of Unikonta by retrieving topologies in which Apusozoa and Opisthokonta are closely related (Cavalier-Smith and Chao 2010; Parfrey et al. 2010). The presence of the previously mentioned two gene fusion and two flagella would appear to support the placement of Apusozoa within Bikonta (Stechmann and Cavalier-Smith 2002). These conflicting data make the inclusion of representative Apusozoa important for determining which genes may have been present in the last common ancestor of eukaryotes, for they may very well be the earliest-diverged eukaryotes.

While the monophyly of some proposed eukaryotic supergroups is widely accepted, other groups remain controversial. The supergroup Chromalveolata is composed of two smaller groups, Chromista (Cryptomonads, Haptophytes, and Stramenopila) and Alveolata whose ancestor engulfed and enslaved red algae (secondary endosymbiosis) (Figure 1.1) (Cavalier-Smith 1981a; Fast et al. 2001; Yoon et al. 2002). Chromalveolates may be sister to Archaeplastida (Cavalier-Smith 2003a). However, whether Chromalveolata are truly monophyletic has been disputed (Parfrey et al. 2006; Parfrey et al. 2010). Although evolutionary relationships inferred from phylogenetic analyses of plastid genes support monophyly, phylogenetic support for monophyly from topologies determined from analyses of nuclear genes is tenuous (Parfrey et al. 2006). In addition, recent phylogenetic analyses indicate that Rhizaria share a more recent common ancestor with the Stramenopiles and Alveolates, prompting calls for the placement of Rhizaria within the Chromista (Burki et al. 2010; Burki et al. 2007; Hackett et al. 2007; Cavalier-Smith 2010). The relationships among Cryptomonads and Haptophytes to other chromalveolates are also known as phylogenetic analyses are somewhat conflicting and ambiguous (Patron, Inagaki, and
Keeling 2007; Burki, Shalchian-Tabrizi, and Pawlowski 2008; Reeb et al. 2009). What we do know is that the evolutionary relationships among chromalveolates and rhizarians are for more complicated than previously supposed. Therefore, including exemplar chromalveolates and rhizarians is important to these studies.

The supergroup Excavata (Discoba and Metamonada) was proposed on the basis of a ventral feeding groove (Figure 1.2) (Simpson and Patterson 1999). The Excavate hypothesis remains controversial due to conflicting evolutionary relationships implied by phylogenetic analyses. While some analyses support the monophyly of excavates (Hampl et al. 2009; Parfrey et al. 2010) others refute it (Parfrey et al. 2006), retrieving polyphyletic groups instead (a group with multiple ancestors).

Recall that apparent absences of mitochondria from excavate taxa (*Trichomonas vaginalis* and *Giardia intestinalis*) and a primitive looking small subunit rRNA sequence in *G. intestinalis* initially supported the notion that the excavates are the earliest-diverging eukaryotes (Figure 1.2) (Cavalier-Smith 1987b). However, more recent studies show that relics of mitochondria (mitosomes and hydrogenosomes) are found in *T. vaginalis* and *G. intestinalis* (Muller 1993; Tovar, Fischer, and Clark 1999). These observations make the status of these excavates as the earliest-diverged eukaryotes questionable. It has also been proposed recently that the Euglenozoa (Discoba) are the extant representatives of the earliest-diverging eukaryotes as they lack an origin recognition complex and four genes (*Tom40, CenpA, Smc5* and *Smc6*) that are thought to be present in all other eukaryotic groups (Cavalier-Smith 2010). The point is that the question of whether Excavata are monophyletic is tied to the determination of the earliest-diverging eukaryotes. If any excavates are the earliest-diverging eukaryotes, the root of the eukaryotic phylogeny lies between either Discoba and/or Metamonada and all other eukaryotes (Figure 1.1). This placement of the root requires that Excavata are paraphyletic (a group including an ancestor and some, but not all, of its descendants (Ridley 2004)). Excavates are only monophyletic if none of them are
the earliest-diverging eukaryotes, and that is currently unknown. Therefore, including representative Metamonada and Discoba is important for inferring which genes were present in the last common ancestor of all eukaryotes.

The root of eukaryotes has also been proposed to lie between Unikonta and Bikonta on the basis of the ultrastructural data discussed previously (Figure 1.1) (Stechmann and Cavalier-Smith 2002). However, the reliability of these features for placement of the root of eukaryotic life has recently been called into question by the possible evolutionary relationship of the Apusozoa to Opisthokonta discussed previously (Cavalier-Smith 2010). Placement of the root on the eukaryotic phylogeny remains one of the most vexing questions in molecular phylogenetics. It seems most likely that, despite the presence of highly derived mitochondria (mitosomes), *G. intestinalis* is still the best candidate for the earliest-diverged eukaryote. Whatever the answer, the ability to polarize the eukaryotic phylogeny will have significant impacts upon studies of the ancestor of eukaryotes.

Since the eukaryotic phylogeny remains incompletely resolved, genes must be detected in *all* eukaryotes to infer its presence in the ancestor to *all* eukaryotes. However, this approach is certain to underestimate the numbers of components present in the last common ancestor of eukaryotes, due to subsequent gene losses. That is, if genes were lost in some lineages than we cannot, without knowing how they are related, determine whether the genes were present in the ancestor of eukaryotes. Therefore, searches for components among diverse eukaryotes combined with phylogenetic analyses that add resolution to the eukaryotic phylogeny are necessary (Cavalier-Smith 2010). In addition, this method is always prone to the discovery of putative early-diverging eukaryotes. That is, the last eukaryotic common ancestor is defined by our knowledge of extant eukaryotes and if new lineages are discovered then inferences regarding the common ancestor of extant eukaryotes will obviously need to be reassessed. David Patterson (1999) has estimated that there are approximately 220
known genera whose evolutionary relationships have yet to be completely resolved (Patterson 1999). Although most of those unclassified genera have ultrastructural identities similar to eukaryotes with well resolved evolutionary relationships (Patterson 1999), indicating that the full breadth of eukaryotes has probably been discovered, the possibility that one (or more) may represent previously unknown lineages always exists.

We performed phylogenetic analyses upon Rad51 and Dmc1 protein sequence data to determine if they are effective markers for resolving the eukaryotic phylogeny. These data are presented in Chapter 3. Products of Rad51 and Dmc1 gene sequences are well conserved among animals, fungi, and plants, with a great degree of similarity and retention of functional motifs (Stassen et al. 1997). The Rad51 gene is present in the genomes of all but one eukaryote studied (G. intestinalis) and both Rad51 and Dmc1 genes are present in single-copy in most organisms (Ramesh, Malik, and Logsdon 2005; Malik et al. 2008). These qualities make Rad51 and Dmc1 protein sequences attractive markers for phylogenetic reconstruction. Since Rad51 and Dmc1 genes are paralogs, we also attempted to determine which eukaryotes are the earliest-diverging by reciprocally rooting them (paralogous rooting) (Gogarten et al. 1989; Iwabe et al. 1989; Schlegel 1994; Brown and Doolittle 1995; Baldauf, Palmer, and Doolittle 1996). Although we failed to positively place the root of eukaryotes using these methods, Rad51 and Dmc1 protein sequence data was useful for resolving five of six eukaryotic supergroups and several first order groups (Table 3.1).

**Summary**

The origin of meiosis is likely to have been one of the most important events in eukaryotic evolution. The effects of this event can be observed at the genetical, cytological, organismal, and population levels of eukaryotic biology. Indeed, meiosis and sexual reproduction may have provided the genetic grist which, when subsequently acted upon by natural selection, resulted in the rapid evolution of the diverse eukaryotic lineages observed today. This thesis presents a body of work in which the distributions
of meiotic genes and phylogenetic analyses of the proteins they encode were used to study the origin and evolution of meiotic genes.

The study presented in Chapter 2 shows that at least eight genes whose products are known to be involved in interhomolog DNA strand exchange during meiosis arose very early during eukaryotic evolution. In addition, we applied a heuristic metric to determine if apparent gene absences are due to limitations of the gene sequence search regimen or to bona fide absences of genes from genomes. These analyses indicate that some genes are detected far less frequently than predicted and are likely to indicate true gene losses (Figure 2.1). Interestingly, some organisms have retained all of the genes tested (e.g. *Saccharomyces cerevisiae* and *Homo sapiens*) and others have retained relatively few (e.g. *Caenorhabditis elegans*) (Figure 2.1). Based upon these observations we propose a general hypothesis in which overexpression or mutation of Rad51 gene may allow other components to be lost due to relaxed selection (Chapter 5). In addition, genes that encode components known to function only in complexes may be vulnerable to loss when another component of the complex is lost.

The study presented in Chapter 3 focuses on the eukaryotic RecA homologs Rad51 and Dmc1. Rad51 protein functions during both mitotic DNA repair and during meiosis, while Dmc1 catalyzes interhomolog DNA strand exchange only during meiosis in model organisms. We collected nucleotide and protein sequence data from databases and by using degenerate PCR. The dataset contains Rad51 and Dmc1 protein sequences from all available eukaryotic supergroups. Therefore, Rad51 and *Dmc1* genes were likely present in the ancestor to all extant eukaryotes. We also analyzed an alignment of 98 Rad51 and 51 Dmc1 protein sequences to determine which amino acid residues are conserved and, therefore, might have conferred Rad51- or Dmc1-specific activities in the common ancestor of eukaryotes (Figure 3.13). We found 18 sites among Rad51 protein sequences and 15 sites among Dmc1 that are completely conserved and likely to have been present in the ancestor of eukaryotes. In addition, we detected 10 sites that
are highly conserved in one protein and conserved but different in the other. These residues are likely to facilitate Rad51- or Dmc1-specific activities; their distributions indicate that these functions were present in the common ancestor of eukaryotes.

The study presented in Chapter 4 was designed to determine the distribution of genes that encode proteins that function during different stages of interhomolog DNA strand exchange in model organisms: 1) synaptonemal complex formation; 2) interhomolog strand exchange; 3) sister chromatid cohesion; and 4) resolution of Holliday junctions as crossovers. We studied the distributions of 20 genes: 10 of that encode proteins that are known to function only during meiosis in model organisms. We determined that 19 of the genes tested are likely to have been present in the common ancestor of eukaryotes. Furthermore, phylogenetic analyses of the protein sequences indicate that all of the putative meiosis-specific genes arose by gene duplication and that they are often paralogs of genes that encode products which function during general DNA damage repair in mitotic cells.

Together, the results of these studies have culminated in general models for the origin and subsequent evolution of meiotic genes that are presented in Chapter 5.
Figure 1.1: Evolutionary relationships among prokaryotes, members of six currently recognized eukaryotic supergroups and Apusozoa according to multigene phylogenetic analyses. Relationships that are well supported in the literature have solid branches while unsupported or conflicting relationships are represented by dotted lines. Although the monophyly of Rhizaria, Stramenopila, and Alveolata is well supported, the relationships among Archaeplastida, Cryptomonads, and Haptophytes within the photosynthetic ‘megagroup’ have not been established. Current hypotheses for placement of the root of eukaryotes are shown. (Baldauf and Palmer 1993; Baldauf et al. 2000; Stechmann and Cavalier-Smith 2002; Cavalier-Smith and Chao 2003a; Cavalier-Smith and Chao 2003b; Cavalier-Smith and Chao 2003c; Stechmann and Cavalier-Smith 2003b; Stechmann and Cavalier-Smith 2003a; Simpson and Roger 2004; Cavalier-Smith and Chao 2006; Kim, Simpson, and Graham 2006; Burki et al. 2007; Moreira et al. 2007; Burki, Shalchian-Tabrizi, and Pawlowski 2008; Yoon et al. 2008; Reeb et al. 2009; Roger and Simpson 2009; Cavalier-Smith and Chao 2010; Parfrey et al. 2010)
Figure 1.2: The three-kingdom tree of life with relative order of major events during eukaryotic evolution. A simplified universal tree of life as determined by phylogenetic analyses of ribosomal RNA nucleotide sequence data is presented (Woese and Fox 1977; Gogarten et al. 1989; Iwabe et al. 1989; Woese, Kandler, and Wheelis 1990; Brown and Doolittle 1995). Neither branch lengths nor distances between arrows depict specific amounts of time but indicate only the proposed relative order of events. The dashed arrows are intended to illustrate two competing hypotheses; 1) the Archezoa hypothesis, in which some organisms diverged prior to the origin of meiosis (or some meiotic functions) and 2) that meiosis (including all currently known steps) was present in the last common ancestor of eukaryotes. The events are taken primarily from Cavalier-Smith’s “neomuran” hypothesis (Cavalier-Smith 1987a; Cavalier-Smith 1988; Cavalier-Smith 2002d; Cavalier-Smith 2002c; Cavalier-Smith 2002a; Cavalier-Smith 2010). However, Cavalier-Smith seems to imply that meiosis arose at the same time as mitosis, whereas meiosis is shown here to have arisen after mitosis. In Eucarya, one dashed branch is intended to indicate putative primitive eukaryotes (Archezoa (Cavalier-Smith 1989)) and the other branch represents all other eukaryotes.
**Figure 1.3: General schematic of mitosis and meiosis.**

**A. Mitosis** – Chromosomes in a diploid (2n = 2) cell (i) (chromosomes are shown here condensed for convenience but are, in reality, unwound, appearing as threads) replicate (yielding 2 x 2n products) and condense, sister chromatids are tightly associated (ii). Pairs of sister chromatids (chromosomes) line up on the metaphase plate and microtubules bind kinetochores of both sister chromatids (iii) in preparation for the mitotic (equational) division (yielding 2n products) (iv).

**B. Meiosis** - Chromosomes in a diploid (2n = 2) cell (i) (chromosomes are shown here condensed for convenience but are, in reality, unwound, appearing as threads) replicate (yielding 2 x 2n products) and condense, sister chromatids are tightly associated (ii). Homologous chromosomes pair, creating bivalents, synaptonemal complexes form (grey bars), interhomolog DNA recombination (crossing-over) occurs (chiasmata are indicated with an X; only one crossover event is shown but at least one event per chromosome arm occurs in most organisms studied), and microtubules bind only one sister chromatid per pair (iii). This is followed by the first meiotic (reductional) division (2 x n). Pairs of non-identical sister chromatids (chromosomes), with chromosome arms splayed and only the centromeres tightly associated, align (iv) for the second meiotic (equational) division, yielding four non-identical haploid products (n) (v).

This image is adapted from (Schurko, Neiman, and Logsdon 2009) with permission. Additional details were provided by (John 1990; Simchen and Hugerat 1993; Kleckner 1996).
Mitosis

A.

i. $2n$

ii. $2x2n$

iii. $2x2n$

iv. $2n$

Interphase

Mitosis

B.

i. $2n$

ii. $2x2n$

iii. $2x2n$

iv. $2x n$

v. $n$

Interphase

Meiosis I

Meiosis II
Figure 1.4: General model of interhomolog DNA strand exchange during meiosis.

This model (based upon details from studies of animals, fungi, and plants) presents interactions of 13 proteins and illustrates four steps of interhomolog DNA strand exchange during meiosis: formation of a pre-synaptic filament on a 3’ ending DNA strand (A-D), capture of a DNA duplex by the pre-synaptic filament (E and F), search by the pre-synaptic filament for regions of DNA duplex homology (F), and invasion of the DNA duplex by the pre-synaptic filament and D-loop formation (G). Components with blue labels are known to function only during meiosis in model organisms. Exact stoichiometry is not implied. The interactions between Rad51 proteins, Rad52 and 59 proteins, and single-stranded DNA (A and B), the formation of Rad52/Rad59 heteroheptamers (A-C), and extension of a Rad51-ssDNA nucleoprotein filament by the Dmc1 protein (C and D) are speculative. (Brill and Stillman 1991; Bishop et al. 1992; Kadyk and Hartwell 1992; Milne and Weaver 1993; Bishop 1994; Bai and Symington 1996; Noble and Guthrie 1996; Klein 1997; Nishinaka et al. 1998; Petukhova, Stratton, and Sung 1998; Shinozaka et al. 1998; Arbel, Zenvirth, and Simchen 1999; Bai, Davis, and Symington 1999; Bishop et al. 1999; Chen et al. 1999; Paques and Haber 1999; Petukhova et al. 1999; Borts, Chambers, and Abdullah 2000; Muniyappa, Anuradha, and Byers 2000; Shinozaka et al. 2000; Davis and Symington 2001; Gasior et al. 2001; Masson and West 2001; Bochkareva et al. 2002; Brush 2002; Fortin and Symington 2002; Kianitsa, Solinger, and Heyer 2002; Krejci et al. 2002; Miyagawa et al. 2002; Pellegrini et al. 2002; Solinger, Kianitsa, and Heyer 2002; Symington 2002; Tsubouchi and Roeder 2002; Cox 2003; Davis and Symington 2003; Sugawara, Wang, and Haber 2003; Anuradha and Muniyappa 2004a; Anuradha and Muniyappa 2004b; Bishop and Zickler 2004; Chen et al. 2004; Dudas and Chovanec 2004; Grishchuk et al. 2004; Krogh and Symington 2004; Sehorn et al. 2004; Ishibashi et al. 2005; Sauvageau et al. 2005; Bleuyard, Gallego, and White 2006; Chi et al. 2006; Enomoto et al. 2006; Flaus et al. 2006; Fung et al. 2006; Henry et al. 2006; Holzen et al. 2006; Ishibashi, Kimura, and Sakaguchi 2006; Cox 2007; Feng et al. 2007; Nimmonkar et al. 2007; Chen, Yang, and Pavletich 2008; Filippo, Sung, and Klein 2008; Lopez-Casamichana et al. 2008; Mozlin, Fung, and Symington 2008; Octobre et al. 2008; Pannunzio, Manthey, and Bailis 2008; Sarai et al. 2008; Chang et al. 2009; Fung, Mozlin, and Symington 2009; Kudoh et al. 2009; Sakaguchi et al. 2009; Seong et al. 2009; Latypov et al. 2010; Okorokov et al. 2010; Szekvolgyi and Nicolas 2010).
• RPA binds ssDNA, preventing secondary structure formation
• Rad51 and Rad52/59 present as heptamers

• Rad52/59 recruits Rad51
• Rad52/59-Rad51 complex binds RPA-ssDNA complex
• RPA's displaced by Rad52/59

• Rad51 binds ssDNA, forming presynaptic filament
• Dmc1 binds, extending presynaptic filament
• Rad55/57 heterodimer mediates filament assembly

• Presynaptic filament extension displaces remaining RPA
• Hop2/Mnd1 heterodimer stabilizes presynaptic filament

• Hop2/Mnd1 heterodimer captures dsDNA

• Hop2/Mnd1 heterodimer stabilizes interactions between homologous DNA sequences

• Rad54 and Rdh54 stimulate D-loop formation and may remove recombinational intermediates
Figure 1.5: A model for the origin of meiotic function by gene duplication. This model hypothesizes that gene duplication events yielding meiosis-specific components mark the origins of their respective meiotic functions. In addition, it suggests that some organisms (most likely protists) diverged prior to the gene duplication events and, therefore, prior to the origins of some meiotic functions. Some organisms may have primitive meiosis or none at all as the ancestral state.
CHAPTER 2
A PAN-EUKARYOTIC INVENTORY OF DNA STRAND EXCHANGE COMPONENTS REVEALS PATTERNS OF CONSERVATION AND LOSS

Abstract:
Recombination is critical for repair of DNA double-strand breaks, and the DNA strand exchange (SE) reaction is central to recombination. We present a phylogenetic inventory of ten SE component proteins (Rad52, Rad59, Rad51, Rad55, Rad57, Dmc1, Hop2, Mnd1, Rad54, and Rdh54) among 47 genera representing five eukaryotic supergroups. We aligned SE protein sequences, verified their homology by phylogenetic analyses, and used these alignments to create hidden Markov model (HMM) profiles and position-specific scoring matrices (PSSM), which we used to further scrutinize public nucleotide sequence databases. Phylogenetic analyses of all the resulting sequences confirmed orthology of the evolutionarily diverse SE component proteins. Eight of ten SE proteins (Rad52, Rad51, Rad55, Rad57, Dmc1, Hop2, Mnd1, and Rdh54) are present in five of six eukaryotic supergroups and were likely present in the common ancestor of extant eukaryotes. An evolutionary analysis of the heterotrimeric Replication Protein A complex (RPA1, RPA2, and RPA3) is also presented. Since RPA subunit protein sequences and their single-stranded DNA binding domains are well conserved, apparent absences of RPA-coding genes from genomes most likely result from detection failures due to limitations of the search regimen. To validate the approach, we fitted a Poisson regression model to the numbers of observed RNA Polymerase I (Pol I) subunit detection failures. We then compared the numbers of RPA subunit detection failures observed to the numbers predicted by the Pol I regression analysis. The results demonstrate that the frequencies of RPA subunit detection failures and their Smith-Waterman alignment scores are strongly correlated. We then applied this approach to the SE proteins by comparing the numbers of detection failures for SE components, given their Smith-
Waterman scores. Detection failures of six proteins (Rad52, Rad59, Rad51, Dmc1, Rad54, and Rdh54) occurred more frequently than predicted, indicating the likely loss of these genes from some completely sequenced genomes. The inferred losses of these genes can be explained if compensatory changes (e.g. overexpression or functional mutations) of Rad51 suppress SE component mutant phenotypes.

Introduction:

In eukaryotes, meiosis is necessary for sexual reproduction (Weissmann, Parker, and Ronnfeldt 1893; Churchill 1970). During meiosis, a single round of genome-wide DNA replication is followed by two nuclear divisions (reductional and equational) (Churchill 1970). A diploid organism typically produces four haploid cells that combine with other haploid products of meiosis (e.g. spores and gametes) (Weissmann, Parker, and Ronnfeldt 1893). In this manner, the chromosomes of organisms are recombined while maintaining the appropriate numbers of chromosomes (Weissmann, Parker, and Ronnfeldt 1893; Cavalier-Smith 2002d). Although there are important differences between meiosis and mitosis, during which one nuclear division follows one round of DNA replication (Flemming 1878), many proteins that function during meiosis also function during mitosis (Marcon and Moens 2005). The pairing of non-sister homologous chromosomes during the first (reductional) division, followed by their segregation to opposite spindle poles, is unique to meiosis (Simchen and Hugerat 1993; Paques and Haber 1999; Dudas and Chovanec 2004; Krogh and Symington 2004; Filippo, Sung, and Klein 2008). However, the second (equational) division that occurs during meiosis is similar (though not identical) to the single equational division of mitosis, during which sister chromatids segregate to opposite spindle poles (Nicklas 1977).

Genetic recombination between homologous chromosomes is essential in most organisms for appropriate pairing and segregation during the reductional division of meiosis (Moore and Orr-Weaver 1998; Paques and Haber 1999; Dudas and Chovanec 2004; Krogh and Symington 2004; Filippo, Sung, and Klein 2008). The importance of
homologous recombination may also be observed at the population level as gene conversions and/or cross-over events may occur that increase the efficacy of natural selection (Fisher 1930; Muller 1932; Hill and Robertson 1966), allowing eukaryotes to respond evolutionarily to changing environments (Van Valen 1973; Rice and Chippindale 2001; Agrawal 2006; Otto and Gerstein 2006).

Several models of recombination have been proposed, such as the double-strand break repair, synthesis-dependent strand annealing, and break-induced replication (Paques and Haber 1999; Dudas and Chovanec 2004; Krogh and Symington 2004; Filippo, Sung, and Klein 2008). Central to all of these models is the DNA strand exchange (SE) reaction, in which 3’ ends of single stranded DNA (ssDNA) invade intact DNA duplexes (Paques and Haber 1999; Dudas and Chovanec 2004; Krogh and Symington 2004; Filippo, Sung, and Klein 2008). Double-strand DNA breaks (DSB) in mitotic cells are generally caused by mutagens and collapsed or damaged replication forks (Paques and Haber 1999; Dudas and Chovanec 2004; Krogh and Symington 2004; Filippo, Sung, and Klein 2008). During meiosis, DSBs are introduced by the Spo11 transesterase, followed by resection of the 5’ strand by nuclease activity (Lichten 2001; Krogh and Symington 2004). Several proteins important for SE activity have been studied in animals, fungi, and plants with genetics, molecular biology, and biochemistry (Brush 2002; Krogh and Symington 2004; Sakaguchi et al. 2009); however, less is known about the origins or evolution of SE components.

An approach to studying the evolution of genes is to search for and compare them among diverse eukaryotes (Dacks and Doolittle 2001). The presence of orthologs among groups of eukaryotes indicates that the genes must have been present in their last common ancestor, while absences might represent either ancestral or derived states (Villeneuve and Hillers 2001; Ramesh, Malik, and Logsdon 2005). It is important to include diverse protists in order to estimate when genes most likely arose during eukaryotic evolution (Ramesh, Malik, and Logsdon 2005). Previous analyses indicate
that some SE proteins (Rad52, Rad51, Dmc1, Hop2, and Mnd1) likely arose very early during eukaryotic evolution (Ramesh, Malik, and Logsdon 2005; Malik et al. 2008). These studies provided a much needed “snapshot” of the distribution of these components in representative animals, fungi, and plants and some (mainly parasitic) protists, yet more specific conclusions regarding the evolutionary histories of SE components could not be made due to limited availability of genome sequence data from diverse lineages within these groups (Adl et al. 2005). In addition, several important mediator proteins that interact with Rad51 or Dmc1 recombinase proteins were excluded from the prior analyses. Here, we present an expanded inventory of the SE machinery (Rad52, Rad51, Dmc1, Hop2, and Mnd1 studied previously, and Rad59, Rad55, Rad57, Rad54, Rhd54, RPA1, RPA2 and RPA3 (Krogh and Symington 2004; Sakaguchi et al. 2009)) with broad taxonomic sampling (47 eukaryotes representing five eukaryotic supergroups (Adl et al. 2005)).

This study also addresses the ambiguous interpretation of apparent gene absences, which has been and important limitation in prior phylogenetic analyses (Ramesh, Malik, and Logsdon 2005; Malik et al. 2008). In addition, the data collected during this study was used to address the issue of the ambiguous interpretation of apparent gene absences, which has been an important limitation to prior phylogenetic inquiries (Villeneuve and Hillers 2001; Ramesh, Malik, and Logsdon 2005; Malik et al. 2008). Current approaches do not distinguish between instances of non-detection that result from (i) failures of the search methods employed, or (ii) true absences (e.g. losses) of genes from completely sequenced genomes (Villeneuve and Hillers 2001; Ramesh, Malik, and Logsdon 2005; Malik et al. 2008). We describe a heuristic metric for detecting potential gene absences that can be applied to a broad range of diverse eukaryotes (Adl et al. 2005). The distribution of Replication Protein A complex subunits (RPA1, RPA2, RPA3) provides an empirical basis for determining the limits of sequence detection. RPA subunits are known to function only as heterotrimeric complexes that bind ssDNA and interact with
Rad52 proteins during recruitment of Rad51 proteins to pre-synaptic filaments in animals, fungi, and plants (Brill and Stillman 1991; Sakaguchi et al. 2009). In addition, we searched for ten RNA Polymerase I subunits (Kuhn et al. 2007) among the 47 taxa studied (Adl et al. 2005) and compared their distributions to the RPA and SE protein datasets. We determined that absences of at least four SE proteins in our inventory (Rad51, Dmc1, Rad54, and Rdh54) most likely represent true gene losses.

**Methods:**

**Data acquisition**

Keyword searches (e.g. *Saccharomyces cerevisiae* Rad51) of the National Center for Biotechnology Information (NCBI, www.ncbi.nlm.nih.gov/) non-redundant protein sequence database retrieved SE protein sequences of RPA1, RPA2, RPA3, Rad52, Rad59, Rad51, Rad55, Rad57, Dmc1, Hop2, Mnd1, Rad54, and Rdh54 (Krogh and Symington 2004; Sakaguchi et al. 2009) from representatives of animals, fungi, and plants. We also searched the clusters of euKaryotic Orthologous Groups of proteins (KOGs) database for each protein (Tatusov et al. 2003). The identities of retrieved protein sequences were initially verified by evaluating the results of bi-directional searches with the tBLASTn (Altschul et al. 1997) option of the Basic Local Alignment Search Tool (BLAST), in which the translated non-redundant nucleotide database is searched using a protein query.

The set of protein (and protein-coding) sequences collected in this manner were subsequently used as queries to search additional protein, nucleotide, and expressed sequence tag (EST) databases at NCBI and other public genome sequence databases by BLASTp, tBLASTn, or BLASTn (Table 2.6). Searches were performed for all homologous protein-coding sequences available between December 2009 and June 2010. In an effort to identify apparently missing homologous sequences from distantly-related organisms, additional searches were performed using protein sequence queries from organisms likely to share more recent common ancestors. For example, *Trypanosoma*
bucei protein sequences were used as additional queries for searches of sequences for a closely related kinetoplastid protist, Leishmania major (Adl et al. 2005). Identities of sequences were again confirmed with bi-directional BLASTp, BLASTx and tBLASTn searches.

When multiple sequences were found for a species, only the most complete open reading frame or protein prediction was retained for our analyses. If no previously annotated protein sequence was available in a database, we annotated the nucleotide sequences manually, using Sequencher v4.5 (Genecodes, Ann Arbor, MI). Exons were identified with reference to multiple protein sequence alignments, inferred translations from BLASTx pairwise comparisons to the NCBI protein sequence database, and the locations of putative intron splice donor and acceptor sites (Griffiths et al. 2000).

Multiple amino acid sequence alignments were calculated using MUSCLE v3.7 (Edgar 2004) and observed with BioEdit v7.0.5.3 (Hall 1999).

To further scrutinize publicly available genome sequence data for the presence of SE protein-coding genes, we created local databases of nucleotide and predicted protein sequences for completely sequenced (Sanger sequence coverage of 8x and greater or sequenced from end-to-end) genomes and searched them using HMMER v2.3.2 (Sonnhammer et al. 1998) and tBLASTn (Altschul et al. 1997). Multiple sequence alignments of homologous amino acid sequences positively identified by reciprocal BLAST searches and phylogenetic analysis were used to calculate hidden markov models (HMM, global and local) with HMMER v2.3.2 and position specific scoring matrices (PSSM) using a local installation of the suite of NCBI BLAST programs. These HMM files were then used to search protein sequence data with HMMER, and the PSSM files used to search protein and nucleotide sequence data for homologs using PSI-BLAST and tBLASTn.
Phylogenetic analyses

We aligned all protein sequences of potential eukaryotic orthologs using MUSCLE v3.7 (Edgar 2004), manually edited them by removing ambiguously aligned columns and gaps in BioEdit v7.0.5.3 (Hall 1999), and performed phylogenetic analyses on the multiple protein sequence alignment. Optimal protein substitution models and parameters were determined for each alignment independently with Modelgenerator v0.8 (Keane et al. 2006). Constant sites were excluded from analyses. Phylogenetic trees were calculated using PhyML v3.0 (Guindon et al. 2009) for 1000 replicates, and PhyloBayes v3. (Lartillot, Lepage, and Blanquart 2009), which used at least two independent chains in which maximum differences observed across all bipartitions were less than 0.10, an indicator that the chains have good convergence (Lartillot, Lepage, and Blanquart 2009). Every other tree after burnins (selected to minimize the differences across all bipartitions) was used to calculate consensus tree topologies in Phylobayes (Lartillot, Lepage, and Blanquart 2009). Analyses were also performed by reciprocally rooting all paralogs (e.g. Rad54 and Rdh54) for positive identification (Ridley 2004). All strand exchange protein sequence alignments were concatenated end-to-end using BioEdit v7.0.5.3 (Hall 1999). Both unpartitioned analyses and analyses partitioned to each protein in the concatenated dataset were performed with RAxML v7.2.7 (Stamatakis, Ludwig, and Meier 2005) for 1000 replicates at the CIPRES Science Gateway v3.0 (Miller et al. 2009).

Inventory assembly

Genes were determined to be present in an organism when putative orthologs were discovered and identified with bi-directional BLAST and phylogenetic analyses (Figures 2.2-2.18) (Ramesh, Malik, and Logsdon 2005; Malik et al. 2008). Protein sequence data for genes that encode all SE proteins (including RPA subunits) in Homo sapiens, Saccharomyces cerevisiae, and Oryza sativa (or its relative) were aligned and their Smith-Waterman pairwise alignment scores (Smith and Waterman 1981) were
calculated with the PRSS/PRFX tool (http://fasta.bioch.virginia.edu/fasta_www2/fasta_www.cgi?rm=shuffle) (Tables 2.3 and 2.4). In some cases protein sequences were either not available for one representative or we were unable to align them properly (Tables 2.3 and 2.4). We also determined Smith-Waterman pairwise alignment scores (Smith and Waterman 1981) for protein sequences from genes that encode RNA Polymerase I proteins (A190, A135, AC40, AC19, AC12.2, Rpb5, Rpb6, Rpb8, Rpb10, and Rpb12 (Kuhn et al. 2007)) from *H. sapiens* and *S. cerevisiae* (Table 2.2). Poisson regression analyses (Allison 1999) were calculated on counts of detection failures among 34 genomes and their respective Smith-Waterman pairwise alignment scores (Smith and Waterman 1981) for the RNA Polymerase I dataset and a combined RNA Polymerase I/RPA1-3 dataset using the genmod procedure in SAS v. 9.2 (SAS Institute Inc., Cary, NC). Graphs were created from the resulting parameter estimates and Wald 90% confidence limits using Microsoft Excel 2010, with the observed numbers of detection failures superimposed for comparison. In addition, parameter estimates and Wald 90% confidence limits from regression analyses were used to calculate the predicted numbers of detection failures given protein Smith-Waterman pairwise alignment scores listed in Tables 2.2-2.4 (Allison 1999).

**Results and discussion:**

An inventory of the presence of 13 component proteins predicted to catalyze the DNA strand exchange (SE) reaction (RPA1, RPA2, RPA3, Rad52, Rad59, Rad51, Rad55, Rad57, Dmc1, Hop2, Mnd1, Rad54, and Rdh54 (Krogh and Symington 2004; Sakaguchi *et al.* 2009)) among 47 diverse eukaryotes is presented here (Figure 2.1) (Adl *et al.* 2005). This inventory includes representatives of five of the six currently recognized eukaryotic supergroups (Adl *et al.* 2005) (Excavata, Chromalveolata, Archaeplastida, Opisthokonta, and Amoebozoa; but not Rhizaria). Completed genome sequence and other nucleotide databases (including ESTs) were rigorously searched using HMM profiles and PSSMs created from phylogenetically verified amino acid sequences
with HMMER v2.3.2 (Sonnhammer et al. 1998), PSI-BLAST, and tBLASTn (Altschul et al. 1997) (Figure 2.1). Identities of putative orthologs were confirmed with phylogenetic analyses (Ramesh, Malik, and Logsdon 2005; Malik et al. 2008) performed with PhyML v3.0 (Guindon et al. 2009) and Phylobayes v3.1 (Lartillot, Lepage, and Blanquart 2009) (Figures 2.2-2.18). Although phylogenetic analysis of several SE proteins yielded poorly resolved phylogenies, the resolution is sufficient to establish orthology (Ramesh, Malik, and Logsdon 2005; Malik et al. 2008).

To determine if either short amino acid sequences or substitution rate heterogeneity are causing phylogenetic artifacts (Felsenstein 2004), we analyzed the concatenated SE protein alignments (Figure 2.19) (Rokas et al. 2003). The concatenated protein sequence analyses strongly support the monophyly of several known groups: Opisthokonta, Chloroplastida, Stramenopila, Apicomplexa, Metamonada, Discoba and Amoebozoa (Adl et al. 2005). However, the unsupported topology, which places some Excavata with ciliates and Amoebozoa, are most likely due to methodological issues, such as long-branch attraction (Felsenstein 2004), and are unlikely to indicate cases of lateral gene transfer (Syvanen 1985).

Ten SE components (RPA1, RPA2, Rad52, Rad51, Rad55, Rad57, Dmc1, Hop2, Mnd1 and Rdh54 (Krogh and Symington 2004; Sakaguchi et al. 2009)) are present in every supergroup tested (Figure 2.1) (Adl et al. 2005). Therefore, they most likely arose very early during eukaryotic evolution, prior to the divergence of nearly all known eukaryotes (Dacks and Doolittle 2001; Villeneuve and Hillers 2001; Ramesh, Malik, and Logsdon 2005; Malik et al. 2008). The absence of the Rad54 gene from the eukaryotic supergroups Excavata and Amoebozoa indicates that it may have arisen later, after the divergence of Excavata or Amoebozoa from other eukaryotes (Figure 2.1 and Table 2.1) (Dacks and Doolittle 2001; Villeneuve and Hillers 2001; Ramesh, Malik, and Logsdon 2005; Malik et al. 2008). We detected the Rad59 gene in Opisthokonta and Amoebozoa that form the metagroup Unikonta (Cavalier-Smith 2002a); the most parsimonious
explanation is that *Rad59* arose more recently during eukaryotic evolution than the last eukaryotic common ancestor, possibly in the last common ancestor of Unikonta (Dacks and Doolittle 2001; Villeneuve and Hillers 2001; Ramesh, Malik, and Logsdon 2005; Malik *et al.* 2008).

Limits of sequence detection and distribution of strand exchange genes among eukaryotes

We selected the three Replication Protein A (RPA) subunits (Sakaguchi *et al.* 2009) for phylogenetic comparison with the other SE proteins and with the subunits of RNA Polymerase I (Kuhn *et al.* 2007), with the goal of establishing a threshold for detection of their component proteins. Replication Protein A, a complex composed of RPA1 (70kDa), RPA2 (32kDa), and RPA3 (14kDa) subunits, binds ssDNA (Brill and Stillman 1991; Bochkareva *et al.* 2002; Brush 2002; Ishibashi *et al.* 2005; Ishibashi, Kimura, and Sakaguchi 2006; Chang *et al.* 2009; Sakaguchi *et al.* 2009). In humans, the RPA heterotrimer is critical DNA metabolic pathways, such as DNA replication, DNA repair, recombination, cell cycle, and DNA damage checkpoints (Zou *et al.* 2006). In yeast, RPA has been indicted in DNA replication, repair, and recombination (Wold 1997). RPA is also necessary for DNA damage repair in plants but may not be critical for DNA replication and homologous recombination in *Oryza sativa* (Ishibashi *et al.* 2005; Ishibashi, Kimura, and Sakaguchi 2006; Kimura and Sakaguchi 2006; Sakaguchi *et al.* 2009). During meiosis the RPA complex recruits Rad52 proteins during pre-synaptic filament formation in animals, fungi, and plants (Davis and Symington 2003; Krogh and Symington 2004; Sakaguchi *et al.* 2009). Two conserved domains (DBD-A and DBD-B) allow RPA1 proteins to bind ssDNA, preventing formation of secondary structures of DNA that inhibit SE (Wold 1997; Brush 2002). RPA1 monomers may bind ssDNA weakly (approximately 8 nucleotides) but binding of the RPA1 subunit interaction motif (DBD-C) with RPA2/RPA3 heterodimers (RPA2 DBD-D, RPA3 DBD-E) causes conformational changes that result in stable interactions (approximately 30 nucleotides)
(Bochkareva et al. 2002). The role of RPA3, which has a single binding motif (DBD-D), is currently unclear. In *Saccharomyces cerevisiae* RPA1 binds only to RPA2/RPA3 heterodimers (Sakaguchi et al. 2009). Inspection of amino acid sequences among animals, fungi, and plants indicates that RPA1 proteins sequences are longer and more conserved than RPA2 (Tables 2.2 and 2.3). In addition, the binding domains of RPA1 and RPA2 protein sequences (DBD - A-D and F) appear to be well conserved among all of the eukaryotes studied here (Figures 2.20-2.24). RPA3 has the least conserved protein sequences of the three subunits (Figure 2.25 and Tables 2.2-2.4), possibly attesting to differences in ssDNA binding among eukaryotes (Sakaguchi et al. 2009). Various RPA complexes have been observed but there are no known functions of any component outside of the trimerization core (Bochkareva et al. 2002). Also, there are no known mutant phenotype suppressors for any of the three components (Table 2.5). Together, these data indicate that RPA trimerization is likely to be required for successful SE in all eukaryotes. Therefore, apparently missing *RPA* subunit genes are best explained by limitations of the search methods used (*i.e.* type II error). We propose that a correlation exists between the numbers of *RPA* subunit gene sequence detection failures and their respective protein sequence amino acid lengths and degrees of conservation (Table 2.2); this possibility is explored further below.

RPA1 protein sequences were obtained for all 47 organisms studied here, while genes that encode RPA2 and RPA3 homologs were not always identified in each study organism (Figures 2.1). Among the 34 organisms included in this study with at least one genome sequence of 8.0x whole-genome shotgun sequencing coverage, there are five *RPA2* and ten *RPA3* gene sequence detection failures. *RPA1* and *RPA2* genes were likely present in the last eukaryotic common ancestor due to their presence in every supergroup tested (Villeneuve and Hillers 2001; Ramesh, Malik, and Logsdon 2005; Malik et al. 2008). The apparent absence of *RPA3* genes from the Amoebozoa indicates that this subunit may have arisen later during eukaryotic evolution (Dacks and Doolittle 2001),
after the divergence of Amoebozoa from other eukaryotes. However, the Amoebozoa are not candidates as the earliest-diverging eukaryotes by current hypotheses for rooting the evolutionary tree of eukaryotes (Cavalier-Smith 2002a; Stechmann and Cavalier-Smith 2002; Roger and Simpson 2009; Cavalier-Smith 2010), making it more likely that RPA3 was lost in the common ancestor of *D. discoideum* and *Entamoeba* (Dacks and Doolittle 2001; Adl *et al.* 2005).

To test the hypothesis that the numbers of RPA subunit sequence detection failures correlate with their protein sequence lengths and degrees of conservation, we compared RPA subunit protein sequence data to homologs of the ten RNA Polymerase I (Pol I) subunit protein sequences (A190, A135, AC40, AC19, AC12.2, Rpb5, Rpb6, Rpb8, Rpb10, and Rpb12 (Kuhn *et al.* 2007)) (Figures 2.3 and 2.26) by determining Smith-Waterman pairwise alignment (S-W) scores (Smith and Waterman 1981). We selected the S-W algorithm to score protein sequence length and conservation due to its ability to apply a similarity measure to protein sequences of variable lengths (Smith and Waterman 1981). S-W scores were calculated from pairwise alignment of protein sequences encoded by genes from *Homo sapiens* and *S. cerevisiae* genomes. Yeast and human gene products were selected for S-W score assessment on the basis that each genome contains all of the genes tested, providing consistency between comparisons. Poisson regression analysis (Cameron and Trivedi 1998) was performed on the numbers of sequence absences observed among 34 organisms with at least one genome sequence per supergroup of 8.0x whole-genome shotgun sequencing coverage, although *Cyanidioschyzon merolae* and *Encephalitozoon cuniculi* were included on the basis that their genomes have been sequenced from end-to-end and, therefore, are complete (Katinka *et al.* 2001; Matsuzaki *et al.* 2004) (Figure 2.1). *Homo sapiens* is also included since all SE, RPA and RNA Pol I component proteins were detected within its genome, although the reference human genome sequence has less than 8.0x sequence coverage (Venter *et al.* 2001). Regression analysis indicates that RNA Pol I S-W scores are good
predictors of the numbers of observed RNA Pol I subunit detection failures ($p = 0.0017$) (Figure 2.27-a.) (Allison 1999). The numbers of observed RPA subunit detection failures relative to their S-W scores are similar to the expected numbers of detection failures, as predicted by the 90% confidence interval for the regression of RNA Pol I data (Figure 2.27-b.). To increase the numbers of proteins included in the regression analyses, we then performed Poisson regression analyses on a combined RPA and RNA Pol I dataset (Figure 2.27-c.). Regression analysis of the combined dataset indicates that S-W scores are good predictors of RPA/ RNA Pol I subunit detection failures ($p < 0.0001$).

We detected eight SE components (Rad51, Rad55, Rad57, Dmc1, Hop2, Mnd1, and Rdh54) in all eukaryotic supergroups tested (Figure 2.1, black rows). Therefore, these components are all likely to have been present in the last eukaryotic common ancestor (Dacks and Doolittle 2001). We failed to detect the Rad54 gene in the Amoebozoa tested (Figure 2.1 and Table 2.1); our search for Rad54 genes was conducted in all available Amoebozoa sequence data (individual EST, nucleotide, and protein sequence submissions and incomplete genomes) and none was discovered. In addition, the Rad54 gene may be absent from Excavata; however, the genome coverages of Trichomonas vaginalis and several Leishmania species are below 8.0x whole-genome shotgun sequencing (Carlton et al. 2007; Aslett et al. 2010), reducing our confidence in this conclusion. The Rad59 gene appears to have arisen later during eukaryotic evolution, as it is present in only two supergroups (Opisthokonta and Amoebozoa) that form a monophyletic “metagroup” (Unikonta) (Figure 2.1) (Malik et al. 2008). The last common ancestor to extant eukaryotes may have had all of the components necessary for homologous recombination, despite the possible absence of Rad54 and Rad59 (Bai and Symington 1996; Klein 1997; Arbel, Zenvirth, and Simchen 1999).

We then compared the numbers of observed SE component detection failures to those predicted by our analysis of the RNA Pol I/RPA dataset (Figure 2.27-c. and Table 2.2). Sequence detection failures for four SE components (Rad55, Rad57, Hop2, and
Mnd1) are within the 90% confidence interval of the predicted failure range. Interestingly, Rad55 and Rad57 proteins form heterodimers that stabilize Rad51-DNA filaments, and they are not known to function as either monomers or homodimers (Bleuyard, Gallego, and White 2006; Filippo, Sung, and Klein 2008). The same is true of Hop2 and Mnd1 proteins that stabilize Dmc1-DNA filaments (Chen et al. 2004; Henry et al. 2006). Therefore, the apparent absence of one of these proteins in the presence of the other as in *Theileria annulata*, *Thalassiosira pseudonana*, and *Phaeodactylum tricornutum* may be due to type II errors. Similarly, the absences of the genes that encode Hop2 and Mnd1 proteins and Dmc1 proteins may indicate that joint losses have occurred (*e.g.* *Drosophila* sp., *Caenorhabditis* sp., *Neurospora crassa*, *Gibberella zeae*, and *Ustilago maydis*) (Figure 2.1 and Table 2.1). Six SE components (Rad52, Rad59, Rad51, Dmc1, Rad54, and Rdh54) were detected less frequently than predicted among the taxa tested (Figure 2.2-c. and Table 2.2). Observed numbers of detection failures are 2-4 times higher than predicted for Rad52 and Rad59 genes and no sequence detection failures of Rad51, Dmc1, Rad54, and Rdh54 genes are predicted (Table 2.2).

S-W scores may not adequately predict the numbers of detection failures due to either variation in genome coverage, or the true absence of genes from a genome. In order to minimize the effects of variation in genome quality, only organisms with completed genomes were used for this analysis (Figure 2.1) (Malik et al. 2008). Poisson regression analysis of the RPA/RNA Pol I dataset (Figure 2.27) indicates that there is a strong correlation between S-W scores and the numbers of absences observed (p < 0.0001) (Allison 1999). So, the effect of variation in genome quality is likely to be negligible among organisms with completed genomes. The quality of gene sequence annotations and true absences of genes from completed genomes are the most likely causes of a reduction in correlation between S-W scores and the numbers of detection failures. As mentioned previously, the Rad59 gene likely arose later during eukaryotic evolution (Malik et al. 2008), explaining some of the detection failures observed (Dacks
and Doolittle 2001). Similarly, some of the sequence detection failures of the *Rad54* gene may be due to its emergence later during eukaryotic evolution (Dacks and Doolittle 2001) (after the divergence of Excavata and/or Amoebozoa (Adl et al. 2005)), although failures to detect the *Rad54* gene among Chromalveolata, Archaeplastida, and Opisthokonta are most likely due to subsequent losses. All apparent absences of Rad51, *Dmc1*, and *Rdh54* genes from the 34 genomes tested could indicate true gene losses. Despite the presence and inferred importance of SE component proteins in the earliest common ancestor of eukaryotes, lineage-specific losses seem pervasive; only some animals and fungi have retained all SE component proteins studied here.

Suppressors of strand exchange component mutant phenotypes in *Saccharomyces cerevisiae*

By interpreting our observed distributions of SE component proteins in comparison with functional studies of SE components in *S. cerevisiae*, it is possible that overexpression or mutation of Rad51 gene could suppress the mutation or loss of other SE protein-coding genes (Table 2.5). Extragenic suppressors of mutant phenotypes in *S. cerevisiae* are known for most of the SE components studied here except Rad51 (absent only in *Giardia intestinalis*) and RPA1-3 (Table 2.5). Overexpression of Rad51 gene in *S. cerevisiae* suppresses *rad52, dmc1, rad55, rad57, hop2*, and *mnd1 mutant* phenotypes (Milne and Weaver 1993; Klein 1997; Krejci et al. 2002; Tsubouchi and Roeder 2003; Henry et al. 2006; Schild and Wiese 2009). In addition, *S. cerevisiae rad51* mutants demonstrate decreased recombination and reduced viability (Bishop 1994; Tsuzuki et al. 1996; Bleuyard, Gallego, and White 2006). These characteristics may account for the infrequent loss of Rad51 gene from eukaryotes.

In fungi, *rad52* mutants are perhaps the most deleterious of the SE machinery, but *Rad52* genes may be absent from as many as 18 of 47 genomes, representing four of the five eukaryotic supergroups in our study (Feng et al. 2007). This apparent contradiction may be explained if compensatory changes in *Rad51* gene overcome the inhibitory
effects of single stranded DNA-RPA complexes in nature (Milne and Weaver 1993; Krejci et al. 2002).

Although critical for meiosis in many organisms studied, several organisms appear to be missing genes that encode Dmc1 (Bishop 1994; Bishop et al. 1999; Tsubouchi and Roeder 2003). However, in *S. cerevisiae* Rad51 protein is capable of completing homologous strand exchange during meiosis when Rad51 gene is overexpressed and does so without the assistance of Hop2 or Mnd1 proteins, which work in concert with Dmc1 (Table 2.2) (Tsubouchi and Roeder 2003). In addition, *dmc1* mutant phenotypes may be suppressed with high copy numbers of the *Rad54* gene, reducing the increased numbers of Rad51 foci that form, as in *S. cerevisiae* (Bishop 1994; Bishop et al. 1999).

We cannot distinguish between the possibility that absences of *Rad55* and *Rad57* genes in our analyses are real or that their apparent absences are artifacts of the search methods used. However, if *Rad51* gene expression is increased, it is possible that enough protein is available for successful pre-synaptic filament formation despite the destabilizing effects that *rad55* and *rad57* mutations may have on recombination, as in *S. cerevisiae* (Fung, Mozlin, and Symington 2009). This observation is consistent with the suppression of *rad55* and *rad57* mutant phenotypes by compensatory changes in the *Rad52* gene, encoding products that recruit Rad51 to form pre-synaptic filaments in *S. cerevisiae* (Milne and Weaver 1993). The numbers of *Hop2* and *Mnd1* gene absences may also be due to failures of the search methods used, however it is interesting that compensatory changes in Rad51 gene also suppresses *hop2* or *mnd1 mutant* phenotypes in *S. cerevisiae*, possibly by creating more Dmc1 foci (Bishop 1994).

Alternatively, the elimination of some Rad51 protein functions suppresses *S. cerevisiae* mutant phenotypes of *rad54* and *rdh54* mutants (Klein 1997). The detrimental effects of *rad54* and *rdh54* mutants are almost certainly the result of accumulation of Rad51 proteins on DNA, since mutants of *rad52, rad51, rad55*, and *rad57* suppress these
phenotypes, eliminating or reducing the number of Rad51 proteins bound to ssDNA in *S. cerevisiae* (Klein 1997).

Finally, the *Rad59* gene, a paralog of *Rad52*, encodes a protein that appears to overlap functionally with Rad52 but cannot suppress a *rad52* mutant phenotype in *S. cerevisiae* (Bai and Symington 1996). The *rad59* mutant confers the most benign mutant phenotype, mildly defective mitotic recombination and decreased resistance to ionizing radiation, which is suppressed by *Rad52* overexpression (Davis and Symington 2001; Davis and Symington 2003; Pannunzio, Manthey, and Bailis 2008), possibly leading to frequent losses.

Conclusions

We found that 10 of 13 strand exchange reaction components (*RPA1, RPA2, Rad52, Rad51, Rad55, Rad57, Dmc1, Hop2, Mnd1, and Rad54* (Krogh and Symington 2004; Sakaguchi *et al.* 2009)) are present in all of the eukaryotic supergroups (Adl *et al.* 2005) scrutinized and thus are likely to have been present in the last common ancestor to extant eukaryotes (Figure 2.1) (Dacks and Doolittle 2001). It is possible that one component (*Rad54*) may have arisen later during eukaryotic evolution (Dacks and Doolittle 2001), after the divergence of either Amoebozoa or Excavata (if either are the earliest diverging eukaryotes) (Adl *et al.* 2005). It is likely that *Rad59* arose later during eukaryotic evolution (Dacks and Doolittle 2001), after the divergence of the Unikonta (composed of Opisthokonta and Amoebozoa) (Cavalier-Smith 2002a) from other eukaryotes. *Rad54* (Petukhova, Stratton, and Sung 1998; Petukhova *et al.* 1999; Kiianitsa, Solinger, and Heyer 2002) and *Rad59* (Bai and Symington 1996; Davis and Symington 2003; Pannunzio, Manthey, and Bailis 2008) proteins are both thought to function primarily during sister chromatid or intrachromosomal recombination.

The requirement for trimerization of RPA protein subunits appears to be conserved among eukaryotes, as heterotrimers are observed among animals, fungi, and plants (Sakaguchi *et al.* 2009). The presence of RPA1 subunits in every organism studied
here strongly implies that RPA2 and RPA3 must also be present in these organisms; thus any apparent absences are inferred to be the result of search method detection limits (Figure 2.1). Detection of nucleotide and protein sequences is influenced by the length or degree of sequence conservation in different organisms (Pevsner 2009). The Smith-Waterman pairwise alignment algorithm (S-W) scores protein sequence data using a similarity measure that incorporates protein sequence length (Smith and Waterman 1981). Thus, we hypothesized that the number of detection failures is correlated with protein S-W scores. We tested this hypothesis with searches of RNA Pol I core complex subunits (Figure 2.26) (Kuhn et al. 2007). In addition, we determined RNA Pol I subunit S-W scores with pairwise alignments of human and S. cerevisiae gene (Figures 2.1 and 2.26). Poisson regression analyses (Allison 1999) indicate that there is a strong correlation between S-W scores and the number of undetected sequences among RNA Pol I proteins (Figure 2.27-a and Table 2.2). Furthermore, the number of detection failures predicted by the RNA Pol I regression analysis for the RPA subunits is similar to the observed numbers of RPA subunit detection failures (Table 2.2) (Allison 1999). These analyses indicate that absences among RPA components are likely due to failures of detection and may not represent true losses. We then combined the RNA Pol I and RPA and performed additional regression analyses (Figure 2.27-b.) and compared the numbers of predicted detection failures relative to S-W scores for the remaining SE components (Figure 2.27-c and Table 2.2). More detection failures were observed than predicted by the Pol I/RPA data for six SE components (Rad52, Rad59, Rad51, Dmc1, Rad54, and Rdh54), these absences may represent true losses of genes from genomes.

Complicating the inference of the early origins of SE component proteins are the frequent absences of SE protein-coding genes observed among diverse eukaryotes (Figure 2.1 and Table 2.1). Only eight organisms (all Opisthokonts (Adl et al. 2005) encode all of the SE proteins studied: Homo sapiens, Mus mus, Gallus gallus, Xenopus laevis, Danio rerio, Nematostella vectensis, Saccharomyces cerevisiae, and
Kluyveromyces lactis. The organism with the fewest SE protein-coding genes (Caenorhabditis elegans with only four) is also an Opisthokont (Figure 2.1) (Adl et al. 2005). Therefore, animals represent the greatest range in the number of SE proteins encoded. The common ancestor to the putatively early diverging eukaryotes (Trichomonas vaginalis and Giardia intestinalis) (Woese, Kandler, and Wheelis 1990) most likely had at least nine SE genes present in its genome (RPA1, RPA2, RPA3, Rad52, Rad51, Dmc1, Rad57, Hop2, and Mnd1).

In nature, frequent losses of SE protein coding genes may be facilitated by mutations in Rad51 gene (Schild and Wiese 2009). Overexpression or mutation of Rad51 genes suppresses the mutant phenotypes of several SE protein coding genes (rad52, rad59, rad55, rad57, dmc1, hop2, mnd1, rad54, and rdh54) in S. cerevisiae (Table 2.5). Rad51 mutations could result in the relaxation of selection on SE proteins, leaving the genes that encode them vulnerable to loss (Nei and Kumar 2000; Ridley 2004). Furthermore, when component proteins function in complexes, such as Hop2 and Mnd1 (Henry et al. 2006), loss of the gene encoding one component may expedite the loss of the gene encoding its partner protein, i.e. the genes that encode obligate complexes likely evolved together by coevolution and rely on one another to have functional value (Goh et al. 2000). As the Hop2-Mnd1 heterodimer is known only to function during meiosis, interacting with Dmc1-DNA filaments (Chen et al. 2004), the loss of the Dmc1 gene may hasten the loss of both Hop2 and Mnd1 genes. Alternately, it is imaginable that Hop2 and Mnd1 proteins could interact with Rad51 proteins in organisms missing Dmc1 (e.g. Encephalitozoon cuniculi and Paramecium tetraurelia) (Figure 2.1).

Although the protein components of the strand exchange reaction appear ubiquitously across eukaryotes (Adl et al. 2005), likely present in their last common ancestor (Dacks and Doolittle 2001), the manner in which SE proceeds may vary greatly due to subsequent loss of a few component proteins. All extant eukaryotes inherited a
complex of SE machinery that has been differentially retained over evolutionary time since the last eukaryotic common ancestor.
Figure 2.1: Phylogenetic distribution among eukaryotes of DNA strand exchange genes. The names of genera studied are listed. Asterisks indicate organisms with completed genomes (on the basis that they have at least one isolate genome-sequence with 8.0x whole-genome shotgun coverage or a genome that was sequenced from end-to-end). Supergroups are presented in black rows with a summary of the genes deduced to be present in their common ancestor and grey rows provide summaries of major Opisthokont lineages. Light grey columns designate RPA1-3, which were used to determine the thresholds of detection. Meiosis-specific proteins are presented in dark grey columns. Symbols: ‘+’ indicates sequence was found and phylogenetically verified, ‘(-)’ indicates that sequence was not found and may be outside the threshold of detection, blank spaces indicate sequences were not found and the genome project has less than the equivalent of 8.0x Sanger whole genome shotgun coverage, ‘-‘ indicates sequence was not found, is within the calculated threshold of detection and the genome project has ≥ 8.0X coverage. The tree is a cartoon that summarizes current literature (Simpson, Inagaki, and Roger 2006; Baldauf 2008; Burki, Shalchian-Tabrizi, and Pawlowski 2008; Kolisko et al. 2008; Timmermans et al. 2008; Minge et al. 2009; Reeb et al. 2009; Shadwick et al. 2009).
<table>
<thead>
<tr>
<th></th>
<th>Rpa1</th>
<th>Rpa2</th>
<th>Rpa3</th>
<th>Rad52</th>
<th>Rad59</th>
<th>Rad51</th>
<th>Rad55</th>
<th>Rad57</th>
<th>Del1</th>
<th>Hop1</th>
<th>Mid1</th>
<th>Rad54</th>
<th>Rad55</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EXCAVATA</strong></td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Giardia*</td>
<td>+</td>
<td>(-)</td>
<td>(-)</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Trichomonas</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Trypanosoma</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Leishmania</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Naegleria</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td><strong>CHROMALVEOLATA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plasmodium</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Theileria*</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Cryptosporidium*</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Tetrahymena*</td>
<td>+</td>
<td>(-)</td>
<td>(-)</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Paramecium*</td>
<td>(-)</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Thalassiosira*</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Phaeodactylum*</td>
<td>+</td>
<td>+</td>
<td>(-)</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Phytophthora*</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td><strong>OPISTHOCONTA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Homo*</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Mus*</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Monodelphis*</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Gallus*</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Xenopus*</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Uvario*</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Strongylocentrotus*</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Aeodes*</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Urosiphitella*</td>
<td>+</td>
<td>+</td>
<td>(+)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Caenothaellid*</td>
<td>+</td>
<td>+</td>
<td>(+)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Apis*</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Tribolium*</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Nematoscelis*</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Trichoplax*</td>
<td>+</td>
<td>+</td>
<td>(-)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Monosiga*</td>
<td>+</td>
<td>+</td>
<td>(+)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td><strong>AMOEBOZOA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dictyostelium*</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Entamoeba*</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td><strong>OPHYTOPHYTA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Saccharomyces*</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Kluyveromyces*</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Candida albicans*</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Neurospora*</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Gibberella*</td>
<td>+</td>
<td>+</td>
<td>(+)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Magnaporthe*</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Aspergillus*</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Schizosacch. *</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Coprinus*</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Ustilago*</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Encephalitozoon*</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td><strong>AMOEBOZOA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diplodinium*</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Licyclostemum*</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Linolaedua*</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>
Figure 2.2: Unrooted phylogenetic tree of 47 Replication Protein A – 1 (RPA1) homologs. Trees were estimated with maximum likelihood and Bayesian inference (LG+I+G+F) from 506 aligned amino acids. Opisthokonta are highlighted in purple, Amoebozoa in blue, Archaeplastida in green, Chromalveolata in orange, and Excavata in brown. GenBank Geninfo Identifiers are given for all sequences unless otherwise noted (e.g. “jgi”) refers to the Joint Genome Institutes public sequence databases. The consensus topology of 2 Phylobayes chains is shown. Numbers at the nodes indicate support from 1000 PhyML bootstrap replicates followed by the posterior probability estimated using Phylobayes.
Figure 2.3: Unrooted phylogenetic tree of 42 Replication Protein A – 2 (RPA2) homologs. Trees were estimated with maximum likelihood and Bayesian inference (LG+G) from 158 aligned amino acids. Opisthokonta are highlighted in purple, Amoebozoa in blue, Archaeplastida in green, Chromalveolata in orange, and Excavata in brown GenBank Geninfo Identifiers are given for all sequences unless otherwise noted (e.g. “jgi”) refers to the Joint Genome Institutes public sequence databases. The consensus topology of 2 Phylobayes chains is shown. Numbers at the nodes indicate support from 1000 PhyML bootstrap replicates followed by the posterior probability estimated using Phylobayes.
Figure 2.4: Unrooted phylogenetic tree of 36 Replication Protein A – 3 (RPA3) homologs. Trees were estimated with maximum likelihood and Bayesian inference (LG+G) from 79 aligned amino acids. Opisthokonta are highlighted in purple, Amoebozoa in blue, Archaeplastida in green, Chromalveolata in orange, and Excavata in brown. GenBank Geninfo Identifiers are given for all sequences unless otherwise noted (e.g. “jgi”) refers to the Joint Genome Institutes public sequence databases. The consensus topology of 2 Phylobayes chains is shown. Numbers at the nodes indicate support from 1000 PhyML bootstrap replicates followed by the posterior probability estimated using Phylobayes.
Figure 2.5: Unrooted phylogenetic tree of 44 Replication Protein A – 1 (RPA1) homologs. Trees were estimated with maximum likelihood and Bayesian inference (LG+I+G+F) from 506 aligned amino acids. Opisthokonta are highlighted in purple, Amoebozoa in blue, Archaeplastida in green, Chromalveolata in orange, and Excavata in brown. GenBank Geninfo Identifiers are given for all sequences unless otherwise noted (e.g. “jgi”) refers to the Joint Genome Institutes public sequence databases. The consensus topology of 2 Phylobayes chains is shown. Numbers at the nodes indicate support from 1000 PhyML bootstrap replicates followed by the posterior probability estimated using Phylobayes.
Figure 2.6: Unrooted phylogenetic tree of 29 Rad52 homologs. Trees were estimated with maximum likelihood and Bayesian inference (LG+G) from 127 aligned amino acids. Opisthokonta are highlighted in purple, Amoebozoa in blue, Archaeplastida in green, Chromalveolata in orange, and Excavata in brown. GenBank Geninfo Identifiers are given for all sequences unless otherwise noted (e.g. “jgi”) refers to the Joint Genome Institutes public sequence databases. The consensus topology of 2 Phylobayes chains is shown. Numbers at the nodes indicate support from 1000 PhyML bootstrap replicates followed by the posterior probability estimated using Phylobayes.
Figure 2.7: Unrooted phylogenetic tree of 46 Rad51 homologs. Trees were estimated with maximum likelihood and Bayesian inference (LG+G) from 312 aligned amino acids. Opisthokonta are highlighted in purple, Amoebozoa in blue, Archaeplastida in green, Chromalveolata in orange, and Excavata in brown. GenBank Geninfo Identifiers are given for all sequences unless otherwise noted (e.g. “jgi”) refers to the Joint Genome Institutes public sequence databases. The consensus topology of 2 Phylobayes chains is shown. Numbers at the nodes indicate support from 1000 PhyML bootstrap replicates followed by the posterior probability estimated using Phylobayes.
Figure 2.8: Unrooted phylogenetic tree of 41 Rad51 homologs. Trees were estimated with maximum likelihood and Bayesian inference (LG+G) from 312 aligned amino acids. Opisthokonta are highlighted in purple, Amoebozoa in blue, Archaeplastida in green, Chromalveolata in orange, and Excavata in brown. GenBank Geninfo Identifiers are given for all sequences unless otherwise noted (e.g. “jgi”) refers to the Joint Genome Institutes public sequence databases. The consensus topology of 2 Phylobayes chains is shown. Numbers at the nodes indicate support from 1000 PhyML bootstrap replicates followed by the posterior probability estimated using Phylobayes.
Figure 2.9: Unrooted phylogenetic tree of 34 Rad55 homologs. Trees were estimated with maximum likelihood and Bayesian inference (LG+G) from 125 aligned amino acids. Opisthokonta are highlighted in purple, Amoebozoa in blue, Archaeplastida in green, Chromalveolata in orange, and Excavata in brown. GenBank Geninfo Identifiers are given for all sequences unless otherwise noted (e.g. “jgi”) refers to the Joint Genome Institutes public sequence databases. The consensus topology of 2 Phylobayes chains is shown. Numbers at the nodes indicate support from 1000 PhyML bootstrap replicates followed by the posterior probability estimated using Phylobayes.
Figure 2.10: Unrooted phylogenetic tree of 42 Rad57 homologs. Trees were estimated with maximum likelihood and Bayesian inference (LG+G) from 119 aligned amino acids. Opisthokonta are highlighted in purple, Amoebozoa in blue, Archaeplastida in green, Chromalveolata in orange, and Excavata in brown. GenBank Geninfo Identifiers are given for all sequences unless otherwise noted (e.g. “jgi”) refers to the Joint Genome Institutes public sequence databases. The consensus topology of 2 Phylobayes chains is shown. Numbers at the nodes indicate support from 1000 PhyML bootstrap replicates followed by the posterior probability estimated using Phylobayes.
Figure 2.11: Unrooted phylogenetic tree of 38 Rad57 homologs. Trees were estimated with maximum likelihood and Bayesian inference (LG+G) from 119 aligned amino acids. Opisthokonta are highlighted in purple, Amoebozoa in blue, Archaeplastida in green, Chromalveolata in orange, and Excavata in brown. GenBank Geninfo Identifiers are given for all sequences unless otherwise noted (e.g. “jgi”) refers to the Joint Genome Institutes public sequence databases. The consensus topology of 2 Phylobayes chains is shown. Numbers at the nodes indicate support from 1000 PhyML bootstrap replicates followed by the posterior probability estimated using Phylobayes.
Figure 2.12: Unrooted phylogenetic tree of 34 Dmc1 homologs. Trees were estimated with maximum likelihood and Bayesian inference (LG+G) from 312 aligned amino acids. Opisthokonta are highlighted in purple, Amoebozoa in blue, Archaeplastida in green, Chromalveolata in orange, and Excavata in brown. GenBank Geninfo Identifiers are given for all sequences unless otherwise noted (e.g. “jgi”) refers to the Joint Genome Institutes public sequence databases. The consensus topology of 2 Phylobayes chains is shown. Numbers at the nodes indicate support from 1000 PhyML bootstrap replicates followed by the posterior probability estimated using Phylobayes.
Figure 2.13: **Unrooted phylogenetic tree of 38 Hop2 homologs.** Trees were estimated with maximum likelihood and Bayesian inference (LG+G) from 105 aligned amino acids. Opisthokonta are highlighted in purple, Amoebozoa in blue, Archaeplastida in green, Chromalveolata in orange, and Excavata in brown. GenBank Geninfo Identifiers are given for all sequences unless otherwise noted (e.g. “jgi”) refers to the Joint Genome Institutes public sequence databases. The consensus topology of 2 Phylobayes chains is shown. Numbers at the nodes indicate support from 1000 PhyML bootstrap replicates followed by the posterior probability estimated using Phylobayes.
**Figure 2.14: Unrooted phylogenetic tree of 41 Mnd1 homologs.** Trees were estimated with maximum likelihood and Bayesian inference (LG+G) from 137 aligned amino acids. Opisthokonta are highlighted in purple, Amoebozoa in blue, Archaeplastida in green, Chromalveolata in orange, and Excavata in brown. GenBank Geninfo Identifiers are given for all sequences unless otherwise noted (e.g. “jgi”) refers to the Joint Genome Institutes public sequence databases. The consensus topology of 2 Phylobayes chains is shown. Numbers at the nodes indicate support from 1000 PhyML bootstrap replicates followed by the posterior probability estimated using Phylobayes.
Figure 2.15: Unrooted phylogenetic tree of 34 Mnd1 homologs. Trees were estimated with maximum likelihood and Bayesian inference (LG+G) from 134 aligned amino acids. Opisthokonta are highlighted in purple, Amoebozoa in blue, Archaeplastida in green, Chromalveolata in orange, and Excavata in brown. GenBank Geninfo Identifiers are given for all sequences unless otherwise noted (e.g. “jgi”) refers to the Joint Genome Institutes public sequence databases. The consensus topology of 2 Phylobayes chains is shown. Numbers at the nodes indicate support for
Figure 2.16: Unrooted phylogenetic tree of 29 Rdh54 homologs. Trees were estimated with maximum likelihood and Bayesian inference (LG+G) from 495 aligned amino acids. Opisthokonta are highlighted in purple, Amoebozoa in blue, Archaeplastida in green, Chromalveolata in orange, and Excavata in brown. GenBank Geninfo Identifiers are given for all sequences unless otherwise noted (e.g. “jgi”) refers to the Joint Genome Institutes public sequence databases. The consensus topology of 2 Phylobayes chains is shown. Numbers at the nodes indicate support for
Figure 2.17: Unrooted phylogenetic tree of 34 Rad54 homologs. Trees were estimated with maximum likelihood and Bayesian inference (LG+G) from 495 aligned amino acids. Opisthokonta are highlighted in purple, Amoebozoa in blue, Archaeplastida in green, Chromalveolata in orange, and Excavata in brown. GenBank Geninfo Identifiers are given for all sequences unless otherwise noted (e.g. “jgi”) refers to the Joint Genome Institutes public sequence databases. The consensus topology of 2 Phylobayes chains is shown. Numbers at the nodes indicate support from 1000 PhyML bootstrap replicates followed by the posterior probability estimated using Phylobayes.
Figure 2.18: Unrooted phylogenetic tree of 13 Rad59 homologs. Trees were estimated with maximum likelihood and Bayesian inference (LG+G) from 102 aligned amino acids. Opisthokonta are highlighted in purple, Amoebozoa in blue, Archaeplastida in green, Chromalveolata in orange, and Excavata in brown. GenBank Geninfo Identifiers are given for all sequences unless otherwise noted (e.g. “jgi”) refers to the Joint Genome Institutes public sequence databases. The consensus topology of 2 Phylobayes chains is shown. Numbers at the nodes indicate support for
Figure 2.19: Unrooted phylogenetic tree of 46 sets of 13 concatenated strand exchange homologs. Trees were estimated with partitioned maximum likelihood inference from 3084 aligned amino acids. Opisthokonta are highlighted in purple, Amoebozoa in blue, Archaeplastida in green, Chromalveolata in orange, and Excavata in brown. The best tree from 1000 replicates is shown.
Table 2.1: DNA strand exchange component absences from eukaryotic groups.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Eukaryotic group</th>
</tr>
</thead>
</table>
| Rad52 | **Alveolata** (*Plasmodium, Theileria, Cryptosporidium, Tetrahymena, Paramecium*)  
 **Viridiplantae** (*Arabidopsis, Oryza, Physcomitrella, Chlamydomonas, Ostreococcus*)  
 **Endopterygota** (*Aedes, Drosophila, C. elegans, Apis, Tribolium*) |
| Rad59 | **Endopterygota** (*Aedes, Drosophila, C. elegans, Apis, Tribolium*)  
 Most Fungi (except Saccharomycetales – *S. cerevisiae, Kluyveromyces, Candida albicans*) |
| Rad55 | **Excavates** (*T. vaginalis, G. intestinalis*)  
 Chromista (*Thalassiosira, Phaeodactylum, Phytophthora*) |
| Dmc1 | **Bacillariophyta** (*Thalassiosira, Phaeodactylum*)  
 **Diptera** (*Aedes, Drosophila, C. elegans, Apis*)  
 **Sordariomycetes** (*Neurospora, Gibberella, Magnaporthe*) |
| Hop2 | **Bacillariophyta** (*Thalassiosira, Phaeodactylum*)  
 **Sordariomycetes** (*Neurospora, Gibberella, Magnaporthe*) |
| Mnd1 | **Sordariomycetes** (*Neurospora, Gibberella, Magnaporthe*) |
| Rad54 | **Ciliophora** (*Tetrahymena, Paramecium*) |
| Rdh54 | **Excavates** (*T. vaginalis, G. intestinalis*)  
 Most Chromalveolata (except Ciliophora – *Tetrahymena, Paramecium*)  
 **Embryophyta** (*Arabidopsis, Oryza*) |
Figure 2.20: Multiple sequence alignment of RPA1 ssDNA binding domain (DBD-A) from 54 diverse eukaryotes. Genus names for Excavata are highlighted with brown, Chromalveolata with orange, Archaeplastida with green, Opisthokonta with purple, and Amoebozoa with blue. Shaded columns indicate amino acids are 75% identical.
Figure 2.21: Multiple sequence alignment of RPA1 ssDNA binding domain (DBD-B) from 54 diverse eukaryotes. Genus names for Excavata are highlighted with brown, Chromalveolata with orange, Archaeplastida with green, Opisthokonta with purple, and Amoebozoa with blue. Shaded columns indicate amino acids are 75% identical.
Figure 2.22: Multiple sequence alignment of RPA1 ssDNA binding domain (DBD-C) from 54 diverse eukaryotes. Genus names for Excavata are highlighted with brown, Chromalveolata with orange, Archaeplastida with green, Opisthokonta with purple, and Amoebozoa with blue. Shaded columns indicate amino acids are 75% identical.
Figure 2.23: Multiple sequence alignment of RPA2 ssDNA binding domain (DBD-D) from 45 diverse eukaryotes. Genus names for Excavata are highlighted with brown, Chromalveolata with orange, Archaeplastida with green, Opisthokonta with purple, and Amoebozoa with blue. Shaded columns indicate amino acids are 75% identical.
Figure 2.24: Multiple sequence alignment of RPA1 ssDNA binding domain (DBD-F) from 54 diverse eukaryotes. Genus names for Excavata are highlighted with brown, Chromalveolata with orange, Archaeplastida with green, Opisthokonta with purple, and Amoebozoa with blue. Shaded columns indicate amino acids are 75% identical.
Figure 2.25: Multiple sequence alignment of RPA3 ssDNA binding domain (DBD-E) from 36 diverse eukaryotes. Excavata are highlighted with brown, Chromalveolata with orange, Archaeplastida with green, and Opisthokonta with purple. Shaded regions indicate amino acids are 75% identical.
Table 2.2: Protein sequence comparisons between *Saccharomyces cerevisiae* and *Homo sapiens*.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Length (aa)</th>
<th>Identity</th>
<th>S-W score</th>
<th>Number undetected observed</th>
<th>Number undetected expected (90%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A190</td>
<td>1664</td>
<td>0.38</td>
<td>3611</td>
<td>0</td>
<td>0.01</td>
</tr>
<tr>
<td>A135</td>
<td>1203</td>
<td>0.44</td>
<td>3185</td>
<td>0</td>
<td>0.02</td>
</tr>
<tr>
<td>AC40</td>
<td>335</td>
<td>0.46</td>
<td>1038</td>
<td>0</td>
<td>1.38</td>
</tr>
<tr>
<td>AC19</td>
<td>215</td>
<td>0.44</td>
<td>612</td>
<td>0</td>
<td>3.24</td>
</tr>
<tr>
<td>AC12.2</td>
<td>131</td>
<td>0.65</td>
<td>382</td>
<td>1</td>
<td>5.13</td>
</tr>
<tr>
<td>Rpb5</td>
<td>142</td>
<td>0.45</td>
<td>364</td>
<td>3</td>
<td>5.32</td>
</tr>
<tr>
<td>Rpb6</td>
<td>70</td>
<td>0.74</td>
<td>354</td>
<td>1</td>
<td>5.43</td>
</tr>
<tr>
<td>Rpb8</td>
<td>146</td>
<td>0.36</td>
<td>289</td>
<td>0</td>
<td>6.18</td>
</tr>
<tr>
<td>Rpb10</td>
<td>125</td>
<td>0.36</td>
<td>257</td>
<td>4</td>
<td>6.59</td>
</tr>
<tr>
<td>Rpb12</td>
<td>70</td>
<td>0.41</td>
<td>13</td>
<td>5</td>
<td>10.73</td>
</tr>
<tr>
<td>RPA1</td>
<td>621</td>
<td>0.32</td>
<td>1161</td>
<td>0</td>
<td>1.08</td>
</tr>
<tr>
<td>RPA2</td>
<td>273</td>
<td>0.26</td>
<td>255</td>
<td>5</td>
<td>6.62</td>
</tr>
<tr>
<td>RPA3</td>
<td>110</td>
<td>0.14</td>
<td>73</td>
<td>10</td>
<td>9.52</td>
</tr>
<tr>
<td>Rad52</td>
<td>471</td>
<td>0.24</td>
<td>550</td>
<td>12</td>
<td>2.96</td>
</tr>
<tr>
<td>Rad59</td>
<td>238</td>
<td>0.10</td>
<td>93</td>
<td>27</td>
<td>11.12</td>
</tr>
<tr>
<td>Rad51</td>
<td>400</td>
<td>0.66</td>
<td>1538</td>
<td>1</td>
<td>0.17</td>
</tr>
<tr>
<td>Rad55</td>
<td>406</td>
<td>0.17</td>
<td>98</td>
<td>12</td>
<td>10.96</td>
</tr>
<tr>
<td>Rad57</td>
<td>460</td>
<td>0.26</td>
<td>295</td>
<td>5</td>
<td>6.19</td>
</tr>
<tr>
<td>Dmc1</td>
<td>334</td>
<td>0.54</td>
<td>1178</td>
<td>10</td>
<td>0.48</td>
</tr>
<tr>
<td>Hop2</td>
<td>218</td>
<td>0.21</td>
<td>161</td>
<td>8</td>
<td>9.13</td>
</tr>
<tr>
<td>Mnd1</td>
<td>219</td>
<td>0.28</td>
<td>328</td>
<td>5</td>
<td>5.63</td>
</tr>
<tr>
<td>Rad54</td>
<td>898</td>
<td>0.45</td>
<td>1984</td>
<td>10</td>
<td>0.05</td>
</tr>
<tr>
<td>Rdh54</td>
<td>958</td>
<td>0.36</td>
<td>1524</td>
<td>15</td>
<td>0.18</td>
</tr>
</tbody>
</table>

Note: Yeast RNA Polymerase I, Replication Protein A, and strand exchange component amino acid lengths, their identities to human, Smith-Waterman scores, and the observed numbers of absences among 34 taxa with at least 8.0x whole-genome shotgun sequencing coverage (except for RPA3 and Rad59 in which *H. sapiens* was compared to *Candida albicans*) are shown. Proteins in bold function only during meiosis in model organisms.
Table 2.3: Protein sequence comparisons between *Homo sapiens* and *Oryza sativa*.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Length (aa)</th>
<th>Identity</th>
<th>S-W score</th>
<th>Number undetected</th>
</tr>
</thead>
<tbody>
<tr>
<td>RPA1</td>
<td>616</td>
<td>0.34</td>
<td>1292</td>
<td>0</td>
</tr>
<tr>
<td>RPA2</td>
<td>270</td>
<td>0.30</td>
<td>340</td>
<td>5</td>
</tr>
<tr>
<td>RPA3</td>
<td>121</td>
<td>0.21</td>
<td>81</td>
<td>10</td>
</tr>
<tr>
<td>Rad52</td>
<td>399</td>
<td>0.37</td>
<td>343</td>
<td>12</td>
</tr>
<tr>
<td>Rad51</td>
<td>339</td>
<td>0.69</td>
<td>1585</td>
<td>1</td>
</tr>
<tr>
<td>Rad55</td>
<td>280</td>
<td>0.26</td>
<td>199</td>
<td>12</td>
</tr>
<tr>
<td>Rad57</td>
<td>346</td>
<td>0.36</td>
<td>464</td>
<td>5</td>
</tr>
<tr>
<td><strong>Dmc1</strong></td>
<td>340</td>
<td>0.63</td>
<td>1321</td>
<td>10</td>
</tr>
<tr>
<td><strong>Hop2</strong></td>
<td>217</td>
<td>0.38</td>
<td>461</td>
<td>8</td>
</tr>
<tr>
<td><strong>Mnd1</strong></td>
<td>205</td>
<td>0.42</td>
<td>539</td>
<td>5</td>
</tr>
<tr>
<td>Rad54</td>
<td>747</td>
<td>0.47</td>
<td>1849</td>
<td>10</td>
</tr>
<tr>
<td>Rdh54</td>
<td>910</td>
<td>0.40</td>
<td>1638</td>
<td>15</td>
</tr>
</tbody>
</table>

Note: The lengths of *Homo sapiens* protein sequences, identities to *Oryza sativa* protein sequences, Smith-Waterman scores (except for RPA3 and Rdh54 which were compared to *Physcomitrella patens*, and Rad52 which was compared to *Cyanidioschyzon merolae*), observed numbers of absences among 34 taxa with at least 8.0x whole-genome shotgun sequencing coverage are shown. Proteins in bold function only during meiosis in model organisms.
Table 2.4: Protein sequence comparisons between *Oryza sativa* and *Saccharomyces cerevisiae*.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Length (aa)</th>
<th>Identity</th>
<th>S-W score</th>
<th>Number undetected</th>
</tr>
</thead>
<tbody>
<tr>
<td>RPA1</td>
<td>656</td>
<td>0.31</td>
<td>1037</td>
<td>0</td>
</tr>
<tr>
<td>RPA2</td>
<td>279</td>
<td>0.24</td>
<td>275</td>
<td>5</td>
</tr>
<tr>
<td>RPA3</td>
<td>106</td>
<td>0.13</td>
<td>Unalignable</td>
<td>10</td>
</tr>
<tr>
<td>Rad52</td>
<td>318</td>
<td>0.42</td>
<td>394</td>
<td>12</td>
</tr>
<tr>
<td>Rad51</td>
<td>339</td>
<td>0.65</td>
<td>1455</td>
<td>1</td>
</tr>
<tr>
<td>Rad55</td>
<td>280</td>
<td>0.33</td>
<td>80</td>
<td>12</td>
</tr>
<tr>
<td>Rad57</td>
<td>290</td>
<td>0.31</td>
<td>229</td>
<td>5</td>
</tr>
<tr>
<td>Dmc1</td>
<td>344</td>
<td>0.54</td>
<td>1125</td>
<td>10</td>
</tr>
<tr>
<td>Hop2</td>
<td>227</td>
<td>0.23</td>
<td>160</td>
<td>8</td>
</tr>
<tr>
<td>Mnd1</td>
<td>207</td>
<td>0.28</td>
<td>255</td>
<td>5</td>
</tr>
<tr>
<td>Rad54</td>
<td>980</td>
<td>0.43</td>
<td>1562</td>
<td>10</td>
</tr>
<tr>
<td>Rdh54</td>
<td>1122</td>
<td>0.35</td>
<td>1277</td>
<td>15</td>
</tr>
</tbody>
</table>

Note: The lengths of *Oryza sativa* protein sequences, identities to *Saccharomyces cerevisiae* protein sequences, Smith-Waterman scores (except Rad52, Rad55, and Rdh54 which were compared to *Cyanidioschyzon merolae*, *Chlamydomonas reinhardtii*, and *Physcomitrella*, respectively), observed numbers of absences among 34 taxa with at least 8.0x whole-genome shotgun sequencing coverage are shown. Proteins in bold function only during meiosis in model organisms.
Figure 2.26: Phylogenetic distribution among eukaryotes of RNA Polymerase I core complex subunit genes. The names of genera, the numbers of completed or nearly completed genome projects available for those genera, and the whole genome shotgun equivalent coverage of the most complete genome project is listed, except for Oryza, Mus, and Kluyveromyces, which were unavailable, and Cyanidioschyzon and E. cuniculi, which were sequenced from end to end with BAC and PCR. Grey regions indicate subunits shared by RNA Polymerase II or III. Supergroups are presented with white text on black background with a summary of the genes present. Symbols: ‘+’ indicates sequence was found and phylogenetically verified, ‘(-)’ indicates that sequence was not found and may be outside the calculated threshold of detection, blank spaces indicate sequences were not found and the genome project has less than the equivalent of 8.0X whole genome shotgun coverage. The tree is a cartoon that summarizes current literature (Simpson, Inagaki, and Roger 2006; Baldauf 2008; Burki, Shalchian-Tabrizi, and Pawlowski 2008; Kolisko et al. 2008; Timmermans et al. 2008; Minge et al. 2009; Reeb et al. 2009; Shadwick et al. 2009).
<table>
<thead>
<tr>
<th>Classification</th>
<th>Example Organisms</th>
<th>Relative Size</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EXCAVATA</strong></td>
<td>Giardia (3)</td>
<td>(11.3X)</td>
</tr>
<tr>
<td></td>
<td>Trichomonas (1)</td>
<td>(7.2X)</td>
</tr>
<tr>
<td></td>
<td>Trypanosoma (4)</td>
<td>(8.0-10.0X)</td>
</tr>
<tr>
<td></td>
<td>Leishmania (9)</td>
<td>(8.0-10.0X)</td>
</tr>
<tr>
<td></td>
<td>Naegleria (1)</td>
<td>(8.6X)</td>
</tr>
<tr>
<td><strong>CHROMALVEOLATA</strong></td>
<td>Plasmodium (8)</td>
<td>(8.0X)</td>
</tr>
<tr>
<td></td>
<td>Theileria (2)</td>
<td>(8.0X)</td>
</tr>
<tr>
<td></td>
<td>Cryptosporidium (4)</td>
<td>(13X)</td>
</tr>
<tr>
<td></td>
<td>Tetrahymena (19)</td>
<td>(9.1X)</td>
</tr>
<tr>
<td></td>
<td>Paramecium (1)</td>
<td>(8.0X)</td>
</tr>
<tr>
<td></td>
<td>Chlamydomonas (1)</td>
<td>(8.8X)</td>
</tr>
<tr>
<td><strong>OPISTHOKONTA</strong></td>
<td>Arabidopsis (2)</td>
<td>(8.0X)</td>
</tr>
<tr>
<td></td>
<td>Oryza (1)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Physcomitrella (1)</td>
<td>(8.1X)</td>
</tr>
<tr>
<td></td>
<td>Chlamydomonas (12)</td>
<td>(8.5X)</td>
</tr>
<tr>
<td></td>
<td>Ostreococcus (4)</td>
<td>(8.8X)</td>
</tr>
<tr>
<td><strong>AMOEBOZOA</strong></td>
<td>Tribolium (1)</td>
<td>(7.0X)</td>
</tr>
<tr>
<td></td>
<td>Nematostella (1)</td>
<td>(7.8X)</td>
</tr>
<tr>
<td><strong>FUNGI</strong></td>
<td>Saccharomyces (9)</td>
<td>(10.2X)</td>
</tr>
<tr>
<td></td>
<td>Kluyveromyces (3)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Candida albicans (2)</td>
<td>(10.0X)</td>
</tr>
<tr>
<td></td>
<td>Neurospora (3)</td>
<td>(8.6X)</td>
</tr>
<tr>
<td></td>
<td>Gibberella (2)</td>
<td>(10.0X)</td>
</tr>
<tr>
<td></td>
<td>Magnaporthes (1)</td>
<td>(7.0X)</td>
</tr>
<tr>
<td></td>
<td>Aspergillus (7)</td>
<td>(8.9X)</td>
</tr>
<tr>
<td></td>
<td>Schizosaccharomyces (4)</td>
<td>(11.8X)</td>
</tr>
<tr>
<td></td>
<td>Coprinus (1)</td>
<td>(10.0X)</td>
</tr>
<tr>
<td></td>
<td>Ustilago (1)</td>
<td>(10.0X)</td>
</tr>
<tr>
<td></td>
<td>Encephalitozoon (1)</td>
<td></td>
</tr>
<tr>
<td><strong>OPISTHOKONTA</strong></td>
<td>Homo (3)</td>
<td>(5.1X)</td>
</tr>
<tr>
<td></td>
<td>Mus (2)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Monodelphis (1)</td>
<td>(6.8X)</td>
</tr>
<tr>
<td></td>
<td>Gallus (1)</td>
<td>(6.6X)</td>
</tr>
<tr>
<td></td>
<td>Xenopus (2)</td>
<td>(7.7X)</td>
</tr>
<tr>
<td></td>
<td>Danio (1)</td>
<td>(10X)</td>
</tr>
<tr>
<td></td>
<td>Strongylacentrotus (1)</td>
<td>(8.0X)</td>
</tr>
<tr>
<td></td>
<td>Aedes (1)</td>
<td>(7.6X)</td>
</tr>
<tr>
<td></td>
<td>Drosophila (12)</td>
<td>(8.9X)</td>
</tr>
<tr>
<td></td>
<td>Caenorhabditis (2)</td>
<td>(10.0X)</td>
</tr>
<tr>
<td></td>
<td>Apis (1)</td>
<td>(7.5X)</td>
</tr>
<tr>
<td></td>
<td>Tribolium (1)</td>
<td>(7.0X)</td>
</tr>
<tr>
<td></td>
<td>Nematozoea (1)</td>
<td>(7.8X)</td>
</tr>
<tr>
<td></td>
<td>Trichoplax (1)</td>
<td>(8.1X)</td>
</tr>
<tr>
<td></td>
<td>Monosiga (1)</td>
<td>(8.4X)</td>
</tr>
<tr>
<td><strong>AMOEBOZOA</strong></td>
<td>Dicotyledon (2)</td>
<td>(8.3X)</td>
</tr>
<tr>
<td></td>
<td>Entamoeba (5)</td>
<td>(8.0X)</td>
</tr>
</tbody>
</table>
Figure 2.27: Number of detection failures for RNA Polymerase I, RPA and SE proteins as predicted by Poisson regression analysis compared with observed numbers of detection failures. (a.) Poisson regression analyses were performed using the numbers of failures to detect RNA Polymerase I subunits (A190, A135, AC40, AC19, AC12.2, Rpb5, Rpb6, Rpb8, Rpb10, and Rpb12) among 34 genera with at least one genome of 8.0X whole-genome shotgun sequencing coverage (or sequenced from end-to-end) relative to their Smith-Waterman scores. The predicted numbers of failures relative to Smith-Waterman scores (black dots) are plotted with Wald 90% confidence limits (green dots). The observed numbers of RNA Polymerase I subunit detection failures are indicated with open circles. (b.) The numbers of Replication Protein A (RPA1-3) subunit detection failures observed (open circles) compared with the Poisson regression predictions obtained from analyses of the RNA Polymerase I dataset. (c.) The observed numbers of detection failures among strand exchange components (Rad59, Rad52, Rad51, Rad55, Rad57, Dmc1, Hop2, Mnd1, Rad54, and Rdh54) (open circles) compared with Poisson regression predictions calculated from a combined RNA Polymerase I and Replication Protein A dataset.
Table 2.5: *Saccharomyces cerevisiae* strand exchange gene mutant phenotypes, suppressors, and meiotic functions of their products.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Mutant Phenotype</th>
<th>Meiotic Function</th>
<th>Suppressor</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mitosis</td>
<td>Meiosis</td>
<td></td>
</tr>
<tr>
<td><strong>Rpa1</strong>&lt;br&gt;<strong>Rpa2</strong>&lt;br&gt;<strong>Rpa3</strong></td>
<td>Abnormal growth, and arrest during S- or M- phases (Brill and Stillman 1991), UV and MMS sensitive, and deficient in homologous recombination (Umezu et al. 1998)</td>
<td><em>rpa1</em> mutations result in reduced sporulation efficiency, severely reduced spore viability, and defective recombination (Soustelle et al. 2002)</td>
<td>Form heterotrimeric complexes that bind ssDNA and recruit Rad52 to the Rpa-ssDNA complex (Firmenich, Elias-Arnanz, and Berg 1995; Gasior et al. 1998; Hays et al. 1998)</td>
</tr>
<tr>
<td><strong>Rad52</strong></td>
<td>Increased sensitivity to ionizing radiation (Game and Mortimer 1974; Saeki, Machida, and Nakai 1980) and reduced spontaneous recombination (Petes, Malone, and Symington 1991)</td>
<td>Reduced ability to sporulate, greatly reduced spore viability (Game and Mortimer 1974), and reduced meiotic recombination (Petes, Malone, and Symington 1991)</td>
<td>Forms heptamers that mediate displacement of RPA from ssDNA and recruits Rad51 (Shinohara, Ogawa, and Ogawa 1992; Milne and Weaver 1993; Hays, Firmenich, and Berg 1995; Sung 1997; Octobre et al. 2008)</td>
</tr>
<tr>
<td><strong>Rad59</strong></td>
<td>Increased sensitivity to IR and mildly defective recombination (Bai and Symington 1996; Davis and Symington 2001; Davis and Symington 2003)</td>
<td>Slightly reduced sporulation efficiency and spore viability (Bai and Symington 1996)</td>
<td>Forms homomeric rings or heteromeric rings with Rad52, functions partially overlap with Rad52, may stimulate or augment Rad52 functions (Bai and Symington 1996; Davis and Symington 2001; Pannunzio, Manthey, and Bailis 2008)</td>
</tr>
</tbody>
</table>
Table 2.5: *Saccharomyces cerevisiae* strand exchange gene mutant phenotypes, suppressors, and meiotic functions of their products. - continued

<table>
<thead>
<tr>
<th>Gene</th>
<th>Mutant Phenotype</th>
<th>Meiotic Function</th>
<th>Suppressor</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mitosis</td>
<td>Meiosis</td>
<td></td>
</tr>
<tr>
<td>Rad51</td>
<td>Increased sensitivity to ionizing radiation (Saeki, Machida, and Nakai 1980) and reduced spontaneous recombination (Petes, Malone, and Symington 1991)</td>
<td>Decreased recombination, reduced spore viability (Petes, Malone, and Symington 1991), and failure to form Dmc1 foci (Bishop 1994)</td>
<td>Forms helical filaments on ss- and dsDNA, catalyzes strand exchange, causes ssDNA extension and dsDNA rotational transition, may recruit Dmc1 to the pre-synaptic filament during meiosis (Nishinaka et al. 1998; Krogh and Symington 2004; Lopez-Casamichana et al. 2008)</td>
</tr>
<tr>
<td>Dmc1</td>
<td>None (Bishop et al. 1992)</td>
<td>Defective recombination and accumulation of double-strand break recombination intermediates, failure to form normal synaptonemal complexes, and arrest late in prophase (Bishop 1994)</td>
<td>Meiosis-specific protein with function similar to Rad51 (Bishop et al. 1992; Bishop 1994; Bishop et al. 1999; Sehorn et al. 2004; Sauvageau et al. 2005)</td>
</tr>
</tbody>
</table>
Table 2.5: *Saccharomyces cerevisiae* strand exchange gene mutant phenotypes, suppressors, and meiotic functions of their products. - continued

<table>
<thead>
<tr>
<th>Gene</th>
<th>Mutant Phenotype</th>
<th>Meiotic Function</th>
<th>Suppressor</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mitosis</td>
<td>Meiosis</td>
<td></td>
</tr>
<tr>
<td><em>Mnd1</em></td>
<td>Increased sensitivity to ionizing radiation (Game and Mortimer 1974; Saeki, Machida, and Nakai 1980) and MMS (Klein 1997), reduced sister chromatid recombination (Petes, Malone, and Symington 1991), and accumulation of Rad51 foci (Arbel, Zenvirth, and Simchen 1999; Shinohara <em>et al.</em> 2000)</td>
<td>30-100% reduced spore viability (Game and Mortimer 1974)</td>
<td>rad52, rad51, rad55, rad57 functional mutations (Klein 1997)</td>
</tr>
<tr>
<td><em>Rdh54</em></td>
<td>Diploid-specific MMS sensitivity and reduced growth (Klein 1997)</td>
<td>Reduced sporulation and spore viability (Klein 1997)</td>
<td>rad52, rad51, rad55, rad57 functional mutations (Klein 1997)</td>
</tr>
</tbody>
</table>
Table 2.6: The most complete genomes of the genera searched during this study with web addresses.

<table>
<thead>
<tr>
<th>Taxon</th>
<th>Web address</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trichomonas</td>
<td><a href="http://trichdb.org/trichdb/">http://trichdb.org/trichdb/</a></td>
</tr>
<tr>
<td>Giardia</td>
<td><a href="http://giardiadb.org/giardiadb/">http://giardiadb.org/giardiadb/</a></td>
</tr>
<tr>
<td>Naegleria</td>
<td><a href="http://genome.jgi-psf.org/Naeg1Naeg1.home.html">http://genome.jgi-psf.org/Naeg1Naeg1.home.html</a></td>
</tr>
<tr>
<td>Trypanosoma</td>
<td><a href="http://tritrypdb.org/tritrypdb/">http://tritrypdb.org/tritrypdb/</a></td>
</tr>
<tr>
<td>Leishmania</td>
<td><a href="http://tritrypdb.org/tritrypdb/">http://tritrypdb.org/tritrypdb/</a></td>
</tr>
<tr>
<td>Plasmodium</td>
<td><a href="http://plasmodb.org/plasmo/">http://plasmodb.org/plasmo/</a></td>
</tr>
<tr>
<td>Theileria</td>
<td><a href="http://www.sanger.ac.uk/Projects/T_annulata/">http://www.sanger.ac.uk/Projects/T_annulata/</a></td>
</tr>
<tr>
<td>Cryptosporidium</td>
<td><a href="http://cryptodb.org/cryptodb/">http://cryptodb.org/cryptodb/</a></td>
</tr>
<tr>
<td>Tetrahymena</td>
<td><a href="http://www.ciliate.org/">http://www.ciliate.org/</a></td>
</tr>
<tr>
<td>Paramaecium</td>
<td><a href="http://paramecium.cgm.cnrs-gif.fr/">http://paramecium.cgm.cnrs-gif.fr/</a></td>
</tr>
<tr>
<td>Thalassiosira</td>
<td><a href="http://genome.jgi-psf.org/Thaps3/Thaps3.home.html">http://genome.jgi-psf.org/Thaps3/Thaps3.home.html</a></td>
</tr>
<tr>
<td>Phaeodactylum</td>
<td><a href="http://genome.jgi-psf.org/Phatr2/Phatr2.home.html">http://genome.jgi-psf.org/Phatr2/Phatr2.home.html</a></td>
</tr>
<tr>
<td>Phytophthora</td>
<td><a href="http://genome.jgi-psf.org/Physol1/Physol1_1.home.html">http://genome.jgi-psf.org/Physol1/Physol1_1.home.html</a></td>
</tr>
<tr>
<td>Arabidopsis</td>
<td><a href="http://www.arabidopsis.org/">http://www.arabidopsis.org/</a></td>
</tr>
<tr>
<td>Oryza</td>
<td><a href="http://www.plantgdb.org/OsG.DB/">http://www.plantgdb.org/OsG.DB/</a></td>
</tr>
<tr>
<td>Pyscomitrella</td>
<td><a href="http://genome.jgi-psf.org/Phypa1_1/Phypa1_1.home.html">http://genome.jgi-psf.org/Phypa1_1/Phypa1_1.home.html</a></td>
</tr>
<tr>
<td>Chlamydomonas</td>
<td><a href="http://genome.jgi-psf.org/Chlre4/Chlre4.home.html">http://genome.jgi-psf.org/Chlre4/Chlre4.home.html</a></td>
</tr>
<tr>
<td>Cyanidioschyzon</td>
<td><a href="http://merolae.biol.s.u-tokyo.ac.jp/">http://merolae.biol.s.u-tokyo.ac.jp/</a></td>
</tr>
<tr>
<td>Homo</td>
<td><a href="http://genome.ucsc.edu/cgi-bin/hgGateway">http://genome.ucsc.edu/cgi-bin/hgGateway</a></td>
</tr>
<tr>
<td>Mus</td>
<td><a href="http://uswest.ensembl.org/Mus_musculus/Info/Index">http://uswest.ensembl.org/Mus_musculus/Info/Index</a></td>
</tr>
<tr>
<td>Monodelphis</td>
<td><a href="http://www.broadinstitute.org/mammals/opossum">http://www.broadinstitute.org/mammals/opossum</a></td>
</tr>
<tr>
<td>Gallus</td>
<td><a href="http://genome.ucsc.edu/cgi-bin/hgGateway?org=chicken">http://genome.ucsc.edu/cgi-bin/hgGateway?org=chicken</a></td>
</tr>
<tr>
<td>Xenopus</td>
<td><a href="http://genome.jgi-psf.org/Xentr4/Xentr4.home.html">http://genome.jgi-psf.org/Xentr4/Xentr4.home.html</a></td>
</tr>
<tr>
<td>Danio</td>
<td><a href="http://www.sanger.ac.uk/Projects/D_rerio/">http://www.sanger.ac.uk/Projects/D_rerio/</a></td>
</tr>
<tr>
<td>Strongylocentrotus</td>
<td><a href="http://www.hgsc.bcm.tmc.edu/project-species-o-Strongylocentrotus%20purpuratus.hgsc?pageLocation=Strongylocentrotus%20purpuratus">http://www.hgsc.bcm.tmc.edu/project-species-o-Strongylocentrotus%20purpuratus.hgsc?pageLocation=Strongylocentrotus%20purpuratus</a></td>
</tr>
<tr>
<td>Aedes</td>
<td><a href="http://www.nd.edu/~dseverso/genome.html">http://www.nd.edu/~dseverso/genome.html</a></td>
</tr>
<tr>
<td>Drosophila</td>
<td><a href="http://flybase.org/blast/">http://flybase.org/blast/</a></td>
</tr>
<tr>
<td>Caenorhabditis</td>
<td><a href="http://www.wormbase.org/">http://www.wormbase.org/</a></td>
</tr>
<tr>
<td>Apis</td>
<td><a href="http://www.hgsc.bcm.tmc.edu/project-species-i-Apis%20mellifera.hgsc?pageLocation=Apis%20mellifera">http://www.hgsc.bcm.tmc.edu/project-species-i-Apis%20mellifera.hgsc?pageLocation=Apis%20mellifera</a></td>
</tr>
<tr>
<td>Tribolium</td>
<td><a href="http://www.hgsc.bcm.tmc.edu/project-species-i-Tribolium%20castaneum.hgsc?pageLocation=Tribolium%20castaneum">http://www.hgsc.bcm.tmc.edu/project-species-i-Tribolium%20castaneum.hgsc?pageLocation=Tribolium%20castaneum</a></td>
</tr>
<tr>
<td>Nematostella</td>
<td><a href="http://genome.jgi-psf.org/Nemve1/Nemve1.home.html">http://genome.jgi-psf.org/Nemve1/Nemve1.home.html</a></td>
</tr>
</tbody>
</table>
Table 2.6: The most complete genomes of the genera searched during this study with web addresses. Continued

<table>
<thead>
<tr>
<th>Taxon</th>
<th>Web address</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monosiga</td>
<td><a href="http://genome.jgi-psf.org/Monbr1/Monbr1.home.html">http://genome.jgi-psf.org/Monbr1/Monbr1.home.html</a></td>
</tr>
<tr>
<td>Trichoplax</td>
<td><a href="http://genome.jgi-psf.org/Triad1/Triad1.home.html">http://genome.jgi-psf.org/Triad1/Triad1.home.html</a></td>
</tr>
<tr>
<td>Saccharomyces</td>
<td><a href="http://www.yeastgenome.org/">http://www.yeastgenome.org/</a></td>
</tr>
<tr>
<td>Kluveromyces</td>
<td><a href="http://wwwgenome.jp/keggbin/show_organism?org=kla">http://wwwgenome.jp/keggbin/show_organism?org=kla</a></td>
</tr>
<tr>
<td>Candida albicans</td>
<td><a href="http://www.candidagenome.org/">http://www.candidagenome.org/</a></td>
</tr>
<tr>
<td>Magnaporthe</td>
<td><a href="http://www.broadinstitute.org/annotation/fungi/magnaporthe/">http://www.broadinstitute.org/annotation/fungi/magnaporthe/</a></td>
</tr>
<tr>
<td>Neurospora</td>
<td><a href="http://www.broadinstitute.org/annotation/genome/neurospora/MultiHom">http://www.broadinstitute.org/annotation/genome/neurospora/MultiHom</a></td>
</tr>
<tr>
<td></td>
<td><a href="http://www.broadinstitute.org/annotation/genome/fusarium_verticillioides/Info.html.html">http://www.broadinstitute.org/annotation/genome/fusarium_verticillioides/Info.html.html</a></td>
</tr>
<tr>
<td>Gibberella</td>
<td><a href="http://www.broadinstitute.org/annotation/genome/fusarium_verticillioides/Info.html.html">http://www.broadinstitute.org/annotation/genome/fusarium_verticillioides/Info.html.html</a></td>
</tr>
<tr>
<td>Aspergillus</td>
<td><a href="http://genome.jgi-psf.org/Aspn5/Aspn5.home.html">http://genome.jgi-psf.org/Aspn5/Aspn5.home.html</a></td>
</tr>
<tr>
<td>Schizosacc.</td>
<td><a href="http://www.sanger.ac.uk/Projects/S_pombe/">http://www.sanger.ac.uk/Projects/S_pombe/</a></td>
</tr>
<tr>
<td>Coprinus</td>
<td><a href="http://www.broadinstitute.org/annotation/genome/coprinus_cinereus/MultiHome.html">http://www.broadinstitute.org/annotation/genome/coprinus_cinereus/MultiHome.html</a></td>
</tr>
<tr>
<td>Ustilago</td>
<td><a href="http://www.broadinstitute.org/annotation/genome/ustilago_maydis/Home.html">http://www.broadinstitute.org/annotation/genome/ustilago_maydis/Home.html</a></td>
</tr>
<tr>
<td>Encephalitozoon</td>
<td><a href="http://www.genome.jp/keggbin/show_organism?org=ecu">http://www.genome.jp/keggbin/show_organism?org=ecu</a></td>
</tr>
<tr>
<td>Dictyostelium</td>
<td><a href="http://genome.jgi-psf.org/Dicp1/Dicp1.home.html">http://genome.jgi-psf.org/Dicp1/Dicp1.home.html</a></td>
</tr>
<tr>
<td>Entamoeba</td>
<td><a href="http://www.sanger.ac.uk/Projects/Comp_Entamoeba/">http://www.sanger.ac.uk/Projects/Comp_Entamoeba/</a></td>
</tr>
</tbody>
</table>
CHAPTER 3

PHYLOGENETIC ANALYSIS OF RECA HOMOLOGS
RAD51 AND DMC1 FROM ALL SUPERGROUPS
PROVIDES EVIDENCE FOR MEIOSIS IN THE LAST
COMMON ANCESTOR OF EUKARYOTES

Background:

Genetic recombination is necessary for repair of DNA double-strand breaks, introduced during replication or exposure to mutagens, in prokaryotes and eukaryotes (West 1992; Bishop 1994; Sandler et al. 1996). Among eukaryotes, recombination is also necessary for repair of DSBs introduced during meiosis to ensure accurate pairing and segregation of chromosomes to opposite spindle poles during the first meiotic division (Bishop et al. 1992; Grishchuk et al. 2004). Eubacterial recA, archaeabacterial RadA, and eukaryotic Rad51 and Dmc1 genes are orthologs whose products are important because they catalyze homologous DNA strand exchange during recombination (Stassen et al. 1997; Lin et al. 2006). Rad51-ssDNA nucleoproteins seek out homologous Rad51-dsDNA complexes, promoting DNA strand exchange (Krogh and Symington 2004). Dmc1 functions similarly, promoting interhomolog DNA strand exchange but only during meiosis in model organisms (Bishop et al. 1992; Paques and Haber 1999; Symington 2002; Krogh and Symington 2004). Saccharomyces cerevisiae and Arabidopsis thaliana rad51 mutants display increased sensitivity to DNA damaging agents and diminished sporulation or fertility, as a result of reduced mitotic recombination (Bishop 1994; Bleuyard, Gallego, and White 2006). Among vertebrates, rad51 mutants have a lethal phenotype, indicating a possible dependence upon recombination during growth and development (Tsuzuki et al. 1996). Homologous recombination during meiosis is reduced or eliminated among dmc1 animal, fungi, and plant mutants (Bishop et al. 1999; Tsubouchi and Roeder 2003). Available animal, fungal, and plant Rad51 and Dmc1 protein sequences are highly conserved, with a great
degree of similarity and retention of motifs (Stassen et al. 1997). However, less is known about Rad51 and Dmc1 among diverse protist lineages. It is necessary to include protists in studies of eukaryotic evolution as they embody the greatest breadth of eukaryotes and their genes may encode products with deviant functions (Sogin 1991; Dacks and Doolittle 2001). We present analyses of the distribution, molecular phylogenetic relationships, and characteristics of Rad51 and Dmc1 protein sequences from organisms representing all currently recognized eukaryotic supergroups - Opisthokonta, Amoebozoa, Excavata, Chromalveolata, Rhizaria, and Archaeplastida – and a currently unclassified group, the Apusozoa (Cavalier-Smith 2004; Adl et al. 2005; Baldauf 2008).

Previous studies confirmed the presence of Rad51 and Dmc1 in all but one eukaryotic supergroup, Rhizaria, indicating that they likely arose early during eukaryotic evolution (Ramesh, Malik, and Logsdon 2005; Lin et al. 2006; Malik et al. 2008). The monophyly of Rad51 and Dmc1 has been demonstrated previously with phylogenetic analyses (Komori et al. 2000; Ramesh, Malik, and Logsdon 2005; Lin et al. 2006; Malik et al. 2008). The observations that homologous recombination is central to meiosis (Paques and Haber 1999; Krogh and Symington 2004) and that Dmc1 catalyzes interhomolog DNA strand exchange only during the first meiotic prophase (Bishop et al. 1992) have led to the inference that the presence of a Dmc1 gene in an organism indicates that meiosis may occur (Ramesh, Malik, and Logsdon 2005). The existence of Dmc1 in the putative early diverging eukaryotes *Giardia intestinalis* and *Trichomonas vaginalis* has been cited as evidence of meiosis in the last common ancestor to eukaryotes (Ramesh, Malik, and Logsdon 2005; Lin et al. 2006; Malik et al. 2008). This view is supported by the presence of several other meiotic genes in *G. intestinalis* and *T. vaginalis* (Ramesh, Malik, and Logsdon 2005; Malik et al. 2008). However, the status of *G. intestinalis* and *T. vaginalis* as “primitive” eukaryotes is now dubious as different hypotheses for rooting the evolutionary tree of eukaryotes have been proposed (Cavalier-Smith 2002a; Stechmann and Cavalier-Smith 2002; Roger and Simpson 2009; Cavalier-
Smith 2010). The relatively recent morphological and molecular phylogenetic analyses of unclassified eukaryotes, such as the Apusozoa, further revives the prospect that some organisms may be primitively asexual, having diverged prior to the origin of Dmc1 genes and, perhaps, meiosis. In the absence of a clearly established earliest-diverging branch on the eukaryotic tree, it is necessary to include representatives of all known eukaryotic supergroups to address the question of whether Dmc1 genes and meiosis were present in their last common ancestor.

Rad51 and Dmc1 protein sequences are well conserved, approximately 350 amino acids long, and may be distinguished by inspection of multiple sequence alignments. In addition, duplications of Rad51 and Dmc1 genes appear rare and, where present, seem to have occurred recently during eukaryotic evolution (Maeshima et al. 1995; Kathiresan, Khush, and Bennett 2002; Ramesh, Malik, and Logsdon 2005; Malik et al. 2008). Only one absence of Rad51 genes, in G. intestinalis, has been confirmed (Ramesh, Malik, and Logsdon 2005). Rad51 and Dmc1 genes are themselves paralogs, which means that it might be possible to determine which eukaryotes represent the earliest-diverging lineages with reciprocal rooting (Gogarten et al. 1989; Iwabe et al. 1989; Iwabe et al. 1991). These characteristics make Rad51 and Dmc1 good candidates for phylogenetic analyses (Baldauf and Palmer 1993). Several studies have determined that Rad51 and Dmc1 nucleotide and amino acid sequences are useful phylogenetic markers, resolving relationships among animals, fungi, and plants (Stassen et al. 1997; Petersen and Seberg 2002; Petersen, Seberg, and Baden 2004). However, it is unknown whether Rad51 and/or Dmc1 protein sequence data will be useful for elucidating the relationships among eukaryotic supergroups, or for the placement of unclassified organisms within the eukaryotic tree of life.

We collected 99 Rad51 and 51 Dmc1 protein sequences (representing 97 and 50 genera, respectively) from six eukaryotic supergroups and Apusozoa. Among these sequences, degenerate PCR was used to isolate 21 new Rad51 sequences and 8 new
Dmc1 sequences from evolutionarily diverse representatives of the eukaryotic supergroups Rhizaria, Excavata, Chromalveolata, Amoebozoa, and also unclassified Apusozoa (*Ancyromonas* sp. and *T. trahens* sp.) for which genome sequence data were unavailable. All publically available nucleotide and protein sequence repositories were also searched for homologs in diverse eukaryotes. To ensure that the breadth of sampling was sufficient for a eukaryote-wide study of Rad51 and Dmc1, and given the abundance of sequences from some eukaryotic groups (Fungi, Metazoa, Chloroplastida, Kinetoplastida, and Apicomplexa), discrete datasets composed of exemplars were collected for some over-represented groups, while exhaustive sequence data searches were performed for all other groups (see Methods). Phylogenetic analyses revealed no clear cases of lateral gene transfer of *Rad51* or *Dmc1* genes, indicating that vertical transmission is the predominant (if not exclusive) mode of inheritance. In addition, phylogenetic analyses of Rad51 and Dmc1 amino acid sequences indicated support for five of the six currently proposed eukaryotic supergroups (Table 3.1).

We also scrutinized our alignments of all Rad51 and Dmc1 protein sequences obtained and compared them to archaeabacterial RadA and eubacterial RecA sequences. Rad51 and Dmc1 protein sequences are highly conserved across all eukaryotic groups, including functional motifs previously identified in archaeabacterial RadA protein sequences (Story, Weber, and Steitz 1992). In addition, we identify ten amino acid residues conserved across all eukaryotic supergroups, but not among prokaryotes, which may confer Rad51- and Dmc1-specific functions. Taken together, these data indicate that the functions of Rad51 and Dmc1 are likely to be conserved across all eukaryotes. Thus meiosis and mitosis most likely occurred in the last common ancestor of eukaryotes.
Results and discussion:

Phylogenetic analysis of Dmc1:

We analyzed the distribution of 51 Dmc1 genes from representatives of 50 genera; 42 of which were obtained from databases and 8 by degenerate PCR (Figures 3.1-3.6). Dmc1 is present in representatives of all six currently recognized supergroups and the unclassified Apusozoa. However, the distribution of the Dmc1 gene is uneven since it is not detected in the genomes of entire groups of organisms, such as Diptera, Sordariomycota, or Stramenopila (except for oomycetes). Failure to detect Dmc1 among more stramenopiles is most parsimoniously interpreted as a loss following the divergence of oomycetes (Brown and Sorhannus 2010). Dmc1 gene losses have been confirmed in a few organisms that are known to undergo meiosis (e.g. Caenorhabditis elegans and Drosophila melanogaster) (Orr-Weaver 1995; Zalevsky et al. 1999). Therefore, meiosis may be accomplished without Dmc1 proteins in some organisms and its absence does not necessarily indicate the absence of meiosis, since these sexual organisms have adapted to Dmc1 loss. However, since Dmc1 is known to function only during meiosis, it is likely that the presence of the Dmc1 gene indicates that meiosis occurs (Bishop et al. 1992; Proudfoot and McCulloch 2006).

Phylogenetic analyses of Dmc1 protein sequences consistently yield a single, distinct monophyletic group (Figures 3.5-3.7), indicating that the Dmc1 gene arose once during the evolutionary history of extant eukaryotes. Most organisms have a single copy of the Dmc1 gene within their genomes. Subsequent duplications of the Dmc1 gene appear to be rare, with recent duplications detected only in the genomes of G. intestinalis (Excavata) and Oryza sativa (Archaeplastida) (Kathiresan, Khush, and Bennett 2002; Ramesh, Malik, and Logsdon 2005; Malik et al. 2008). Interestingly, G. intestinalis is also the only organism with a confirmed absence of Rad51 gene from its genome (Ramesh, Malik, and Logsdon 2005), but whether these observations are related is currently unknown.
Phylogenetic analysis of Rad51:

We analyzed the phylogenetic relationships among 99 Rad51 protein sequences representing 97 genera (78 from databases and 21 inferred from degenerate PCR (data not shown), Figures 3.6-3.10). Rad51 genes were retrieved from the genomes of organisms representing every currently recognized eukaryotic supergroup and two Apusozoa (T. trahens and Ancyromonas). Unlike the Dmc1 gene, Rad51 gene appears to be present in most organisms, and so far is absent only from the genome of G. intestinalis. However, an extensive search for Rad51 in the genome of a related diplomonad, Spironucleus vortens (Jorgensen and Sterud 2007), was performed in which we explored all nucleotide, protein, and EST sequence databases and attempted to amplify Rad51 with degenerate PCR and no Rad51 gene sequences were recovered. Rad51 gene may have been lost prior to the divergence of G. intestinalis and S. vortens. Like the Dmc1 gene, duplications of Rad51 gene appear to be rare and relatively recent, with paralogs present only in Archaeplastida (Physcomitrella patens, Oryza sativa, and Zea mays), Xenopus laevis (Opisthokonta), and T. vaginalis (Excavata) (Maeshima et al. 1995; Stassen et al. 1997; Malik et al. 2008). One of the T. vaginalis Rad51 gene copies is a pseudogene, but both of the Xenopus Rad51 genes seem to encode functional products and are expressed (Maeshima et al. 1995; Malik et al. 2008). There are also no clear cases of Rad51 lateral transfer indicated, although Rad51 was discovered in the nucleomorph genomes of Bigelowiella (Rhizaria), and the cryptophytes Hemiselmis andersenii and Guillardia theta (both Chromalveolata) (Figure 3.6). Overall, our results show that Rad51 gene is thus only vertically transmitted, and arose once, prior to the divergence of extant eukaryotes.

Phylogenetic analyses of Rad51 and Dmc1:

Recently, many relationships among diverse eukaryotes have been determined by phylogenetic analyses performed on multiple concatenated protein sequences (Figure 3.8) (Burki and Pawlowski 2006; Kim, Simpson, and Graham 2006; Burki et al. 2007;
Moreira et al. 2007; Burki, Shalchian-Tabrizi, and Pawlowski 2008; Yoon et al. 2008; Reeb et al. 2009; Parfrey et al.). These eukaryotic phylogenies provide references for assessing the utility of individual nucleotide or protein sequence datasets as phylogenetic markers. We performed extensive phylogenetic analyses on Rad51 and Dmc1 individual and concatenated protein sequence datasets to test their phylogenetic utility (Figures 3.1-3.12 and Table 3.1).

The eukaryotic supergroup Opisthokonta (comprised of Animalia, Fungi and several protist groups) is unified by flat mitochondrial cristae and a 12-amino acid insertion in the translation elongation factor 1α (Baldauf and Palmer 1993; Adl et al. 2005; Steenkamp, Wright, and Baldauf 2006). Phylogenetic analyses typically provide strong support for topologies unifying animals and fungi, confirming these observations (Cavalier-Smith 1987c; Baldauf and Palmer 1993; Steenkamp, Wright, and Baldauf 2006). We obtained strong support with both maximum likelihood and Bayesian phylogenetic approaches for the monophyly of Metazoa and Fungi with Dmc1 and concatenated protein sequence alignments (Figures 1.2 and 3.5, Table 3.1). Although opisthokont unity was not formally observed for Rad51 protein sequence dataset which includes the Choanoflagellate, Monosiga brevicollis, this was a result of the likely erroneous placement of the Apusomonad, T. trahens, within this group (but, see below) (Figure 3.8) (Adl et al. 2005).

The Unikont hypothesis proposes that the eukaryotic supergroups Opisthokonta and Amoebozoa are monophyletic, on the basis that they ancestrally possessed a single flagellum (unlike the “bikont” Excavata, Archaeplastida, Chromalveolata, Rhizaria, and Apusozoa that mostly have two flagella) and three fused genes (carbamoyl-phosphate synthase, dihydroorotase, and aspartate carbamyl-transferase), the likely result of two rare gene fusion events (Cavalier-Smith 2002a; Stechmann and Cavalier-Smith 2002; Cavalier-Smith 2003a; Stechmann and Cavalier-Smith 2003b). Phylogenetic analyses have supported the “Unikont hypothesis” (Stechmann and Cavalier-Smith 2003b; Burki,
Shalchian-Tabrizi, and Pawlowski 2008), however, recent phylogenetic analyses retrieve topologies in which unclassified Apusozoa (*Ancyromonas* and *T. trahens*) are monophyletic and closely related to Opisthokonts (Kim, Simpson, and Graham 2006; Cavalier-Smith 2010). While none of our analyses retrieved topologies consistent with a common origin of the Apusozoa, *Ancyromonas* and *T. trahens*, our Bayesian analysis of the individual Rad51 protein sequences strongly supports their inclusion in the Unikont clade (Figure 3.8 and Table 3.1), and analysis of concatenated Rad51 and Dmc1 proteins moderately supports the inclusion of *Ancyromonas* in the Unikonts (Figure 3.11 and Table 3.1). In addition to having two emergent flagella (instead of one flagellum like other Unikonts), Apusozoa also lack the three-gene fusion. Instead, they share a fusion of two genes (dihydrofolate reductase and thymidylate synthase) that distinguishes Bikonts. Unikonts may, therefore, represent a polyphyletic group if Apusozoa are sisters to Opisthokonta (Figure 1.2) (Stechmann and Cavalier-Smith 2002; Stechmann and Cavalier-Smith 2003a).

On the basis of strongly supported topologies obtained with molecular phylogenetic analyses of many concatenated protein sequence alignments, a “megagroup” of predominantly photosynthetic eukaryotes has been proposed (including supergroups Archaeplastida, Chromalveolata, and Rhizaria) (Burki, Shalchian-Tabrizi, and Pawlowski 2008). The supergroup Chromalveolata was proposed to include secondarily photosynthetic eukaryotes (alveolates, stramenopiles, cryptomonads, and haptophytes) that obtained plastids by endosymbiosis with red algae (Cavalier-Smith 2002b; Cavalier-Smith 2003b; Janouskovec *et al.* 2010) (Figure 1.2). However, molecular phylogenetic analyses rarely support the monophyly of this group (Parfrey *et al.* 2006). Despite the complexities of developing the protein targeting system observed in nascent plastids, recent phylogenetic analyses suggest secondary photosynthesis evolved at least twice during eukaryotic evolution (Keeling 2010).
We included a subset of chromalveolates (stramenopiles and alveolates) in our analyses. Phylogenetic analysis of Rad51 and Dmc1 concatenated protein sequence dataset retrieved topologies consistent with the Chromalveolate hypothesis (Figure 3.11). The phylogenies of the individual Dmc1 and Rad51 protein sequences both retrieve discrepant topologies that support the grouping of stramenopiles with Chloroplastida, while red algae are most closely related to stramenopiles in the Dmc1 phylogeny (Figure 3.1), and to alveolates in Rad51 phylogeny, (Figure 3.8). These topologies could be the results of phylogenetic artifacts such as long-branch attraction (Felsenstein 2004). However, it is noteworthy that Bayesian analyses of our concatenated Rad51 and Dmc1 dataset strongly support the monophyly of alveolates, stramenopiles, and Rhodophyceae, and that Chloroplastida are grouped with Cercozoa (Rhizaria) (Figure 3.11). It has been hypothesized that the difficulties of resolving relationships among secondarily photosynthetic eukaryotes with multigene analyses may be due to the “mosaic” nature of their nuclear genomes as a result of endosymbiotic gene transfer, resulting in conflicting phylogenetic signals (Parfrey et al.). Our analyses of Rad51 and Dmc1 failed to support subgroups within the photosynthetic megagroup such as SAR, in which Stramenopila, Alveolata, and Rhizaria share a common ancestor, or the Archaeplastida, which all have plastids obtained by primary endosymbiosis of a cyanobacterium (Adl et al. 2005; Rodriguez-Ezpeleta et al. 2005; Burki, Shalchian-Tabrizi, and Pawlowski 2008; Parfrey et al.). However, we did observe support for the monophyly of Cercozoa (Rhizaria), stramenopiles and alveolates (Chromalveolata).

The eukaryotic supergroup Excavata (represented in our dataset by members of its subgroups Discoba and Metamonada) was proposed to describe organisms with suspension-feeding grooves (cytostomes) used to capture particles in a current produced by anterior flagella (Figure 1.2) (Simpson 2003; Adl et al. 2005). Excavates include organisms once considered to be among the earliest-diverging eukaryotes (e.g. Euglenozoa, T. vaginalis, and G. intestinalis), based upon so-called “primitive” features
(like the apparent absence of organelles such as mitochondria) and early phylogenetic analyses of small ribosomal subunit sequence data which retrieved topologies placing *T. vaginalis* and *G. intestinalis* at the base of eukaryotic trees (Woese, Kandler, and Wheelis 1990; Tovar *et al.* 2003; Adl *et al.* 2005; Cavalier-Smith 2010). However, more recent discoveries have cast doubt that they represent “primitive” eukaryotes. *G. intestinalis* and *T. vaginalis* do, indeed, possess highly derived mitochondria (mitosomes and hydrogenosomes, respectively), and their placement at the base of rooted eukaryotic phylogenetic trees were most likely caused by artifacts of the phylogenetic analysis (Tovar *et al.* 2003; Felsenstein 2004; van der Giezen, Tovar, and Clark 2005). Similarly, Microsporidia, were later determined to be fungi with mitosomes (Cavalier-Smith 1989; Hirt *et al.* 1999). If *T. vaginalis* and *G. intestinalis* (or any of Excavata) are the earliest-diverging eukaryotes, then Excavata would represent a paraphyletic group (a common ancestor plus some but not all of its descendants) whose members diverged separately at the base of the eukaryotic phylogenetic tree, *i.e.*, very early during the evolution of eukaryotes. However, recent phylogenetic analyses retrieve topologies that are consistent with the monophyly of Excavata (Burki *et al.* 2007; Burki, Shalchian-Tabrizi, and Pawlowski 2008; Hampl *et al.* 2009; Parfrey *et al.*). Our phylogenetic analysis of the Dmc1 protein sequence dataset also supports the monophyly of Excavata, although it is not resolved by Rad51 protein sequences (Figures 3.1 and 3.8). In an attempt to determine the earliest-diverging eukaryotic lineages we performed analyses in which one paralog was used to root the other, rather than assigning a root (Gogarten *et al.* 1989; Iwabe *et al.* 1989; Iwabe *et al.* 1991). However, the topologies retrieved with reciprocal rooting of Rad51 and Dmc1 protein sequence are poorly supported and discordant (Figures 3.5-3.7).

Characteristics of Rad51 and Dmc1 protein sequences:

We aligned Rad51 and Dmc1 protein sequences from representatives of all known eukaryotic supergroups and Apusozoa with representative archaeabacteria (*Nitrosopumilus*...
maritimus, Cenarchaeum symbiosum, Pyrobaculum islandicum, Candidatus Korarchaeum cryptofilum, Aeropyrum pernix, Nanoarchaeum equitans, and Methanocaldococcus fervens) and eubacteria (Bacillus amyloliquefaciens and Thermus thermophilus) (Figure 3.13). Visual inspection of the central domains responsible for recombinase activity of RecA, RadA, Rad51 and Dmc1 proteins indicates that the amino acid sequences are well conserved in all domains of life (Story, Weber, and Steitz 1992). Several motifs important for RecA function are highly conserved among eukaryotes. In addition, archaebacterial RadA sequences contain all of the described functional motifs (Chen et al. 2007); it is likely that these motifs were present in the common ancestor of archaeabacteria and eukaryotes, and thus were present in the last eukaryotic common ancestor.

Although Rad51 and Dmc1 perform very similar functions, Rad51 catalyzes DNA strand exchange during both mitosis and meiosis, while Dmc1 functions in interhomolog DNA strand exchange exclusively during meiosis. Specific interactions between Rad51 and Dmc1 with each other, other proteins, and DNA are required for successful completion of meiotic recombination (Krejci et al. 2001; Shin et al. 2003; Sugawara, Wang, and Haber 2003). However, the basis of these interactions remains largely unknown, especially for those interactions that distinguish Rad51 from Dmc1 function. We examined our multiple sequence alignments for conserved amino acid residues specific to Rad51 or Dmc1, which might confer Rad51- or Dmc1-specific activity. Comparison of the central domains of Rad51 and Dmc1 protein sequences from all of our representatives of six eukaryotic supergroups and Apusozoa indicate they are conserved, likely due to common ancestry and functional constraints (summarized in Figure 3.14). By identifying residues conserved in one protein but variable or different in the other, we can generate hypotheses for future functional studies. Comparing protein sequences from representatives of the entire breadth of eukaryotic diversity enables us to pinpoint residues fundamental to Rad51 or Dmc1 function.
To examine amino acid conservation, we analyzed an alignment of 98 Rad51 and 51 Dmc1 protein sequences from all eukaryotic supergroups and Apusozoa (Figure 3.13). The central domain (\textit{S. cerevisiae} Rad51 amino acid positions 90-397) was examined because it is conserved in all RecA homologs. All groups except Apusozoa and Rhizaria were represented at each amino acid position studied. Apusozoa were represented from \textit{S. cerevisiae} Rad51 amino acid positions 126-356 for the aligned Rad51 proteins and positions 188-324 for the aligned Dmc1 proteins; while Rhizaria were represented from positions 188-397 in the Dmc1 alignment. We identified 18 amino acids that are completely conserved among Rad51, and 15 completely conserved amino acids among Dmc1. Seven residues are present among at least 95% of Rad51 protein sequences, but are either different or variable among Dmc1 sequences, but among Dmc1, only three such sites were identified. We found no cases in which a residue is $\geq 95\%$ conserved in one protein dataset and a different residue conserved $\geq 95\%$ in the other dataset.

Studies in which the structures of RecA, RadA, Rad51, and Dmc1 have been analyzed have resulted in the identification of several important functional motifs and amino acid residues (Table 3.3) (Story, Weber, and Steitz 1992; Aihara \textit{et al.} 1999; Pellegrini \textit{et al.} 2002; Conway \textit{et al.} 2004; Chen \textit{et al.} 2007; Chen, Yang, and Pavletich 2008; Okorokov \textit{et al.} 2010). Residues identified with these methods are also highly conserved (often 100\%) in our sequence alignments. Five sites involved in ATP binding (G185, D219, E221, D280, and S281) and three sites involved in DNA binding (N325, G346, and G347) are present in all RecA, RadA, Rad51, and Dmc1 protein sequences studied here. However, specific interactions have not been proposed for several sites that we have determined are likely to be involved in Rad51- or Dmc1-specific activities (Table 3.3).

Conclusions:

We isolated 8 \textit{Dmc1} and 21 \textit{Rad51} genes with degenerate PCR from eukaryotes representing four of the six currently recognized supergroups (Amoebozoa, Excavata,
Chromalveolata, and Rhizaria) and the unplaced Apusozoa. In addition, we performed extensive searches of all publicly available nucleotide and amino acid sequence repositories, identified, and collected a total of 51 Dmc1 and 99 Rad51 sequences (representing 50 and 97 genera, respectively). Our phylogenetic analyses indicate support for all eukaryotic supergroups (Opisthokonta, Amoebozoa, Excavata, Chromalveolata, and Rhizaria) except Archaeplastida was observed during this study (Table 3.1). However, support was strongest for the supergroup Opisthokonta, which was retrieved with phylogenetic analysis of Dmc1, Rad51, and concatenated protein sequences. These results are consistent with previous studies in which the support for supergroups was assessed (Parfrey et al. 2006). Dmc1 appears to retrieve known relationships well when several protein sequences representing the greatest breadth of eukaryotes are available. Consistent with the predictions of Stassen, et al. (1997), our analyses of Rad51 proteins retrieve “somewhat anomalous” phylogenies, most likely due to substitution rate heterogeneity among taxa resulting in long-branch artifacts (Stassen et al. 1997; Felsenstein 2004). Analysis of Rad51 and Dmc1 concatenated protein sequence data provides better resolution of the evolutionary relationships of eukaryotes (Figure 3.11).

We aligned Rad51 and Dmc1 protein sequences from every eukaryotic supergroup and members of the currently unclassified Apusozoa with bacterial RecA and archaeabacterial RadA protein sequences (Figure 3.13). Previously identified (Sandler et al. 1996; Chen et al. 2007; Okorokov et al. 2010) functional motifs are present in all Rad51, Dmc1 and RadA proteins sampled, thus these motifs must have been present in Rad51 and Dmc1 sequences of the last eukaryotic common ancestor. Furthermore, we identified seven sites where the amino acids are conserved among Rad51 but not in Dmc1, and three sites where the amino acids are conserved among Dmc1 but not in Rad51. These amino acids are likely to be involved in functions that are specific to Rad51 or Dmc1 but not both. Given the conservation of these amino acids in protein sequences of diverse eukaryotes, they must have been present in the last eukaryotic
common ancestor as well. Thus, since both Rad51- and Dmc1-specific functions are likely to have been present in the last eukaryotic common ancestor, the hypothesis that Dmc1 was both present and functioning in a meiosis-specific role is supported by these results.

Methods:

Database searches:

Keyword searches (e.g. *S. cerevisiae* Rad51) of the National Center for Biotechnology Information (NCBI, www.ncbi.nlm.nih.gov/) protein sequence database retrieved Rad51 and Dmc1 protein sequences for representatives of animals, fungi, and plants (*Homo sapiens* (Rad51 – accession number NP_002866 and Dmc1 Q14565), *Saccharomyces cerevisiae* (Rad51 - CAA45563 and Dmc1 - AAA34571), and *Oryza sativa* (Rad51 - BAB85491 and Dmc1 - BAB85214) (Aboussekhra et al. 1992; Bishop et al. 1992; Collins et al. 2004; Sakane et al. 2008; Kudoh et al. 2009). In addition, the clusters of euKaryotic Orthologous Groups of proteins (KOGs) database for each protein were searched (Tatusov et al. 2003). Sequence identities were initially verified by evaluating the results of bi-directional searches with the tBLASTn (Altschul et al. 1997) option of the Basic Local Alignment Search Tool (BLAST), in which the translated nucleotide database is searched using a protein query. Rad51 and Dmc1 protein sequences collected in this manner were subsequently used as queries to search protein, nucleotide, and expressed sequence tag (EST) databases at NCBI, the Institute for Genomic Research (TIGR, www.tigr.org/tdb/euk, since moved to compbio.dfci.harvard.edu/tgi/protist.html), the Joint Genome Institute (JGI, genome.jgi-psf.org), the Canadian Protist EST Project (Taxonomically Broad Database, tbestdb.bcm.umontreal.ca), Michigan State University *Galdieria sulphuraria* Database (Weber et al. 2004; Barbier et al. 2005), genomics.msu.edu/galdieria) and the *Cyanidioschyzon merolae* Genome Project (Matsuzaki et al. 2004), merolae.biol.s.u-tokyo.ac.jp/blast/blast.html) with BLASTp, tBLASTn, and BLASTn, as necessary, for all
available Rad51 and Dmc1 sequences from January 2004 through April 2010. Due to the abundance of sequences from a few eukaryotic groups (Fungi, Metazoa, Chloroplastida, Kinetoplastida, and Apicomplexa), discrete datasets composed of exemplars were collected for these groups, while exhaustive sequence data searches were performed for all other groups, to ensure the breadth of sampling was sufficient for a eukaryote-wide study of Rad51 and Dmc1. In case sequences from distantly-related organisms were missed, additional searches were performed using protein sequence queries from organisms likely to share more recent common ancestors: e.g. *Trypanosoma brucei* (Rad51 CAA73605, Dmc1 XP_827266 (Berriman *et al.* 2005)) protein sequences were used as additional queries for searches of sequences for a closely related kinetoplastid protist, *Leishmania major*. Identities of sequences were again confirmed with bi-directional BLASTx and tBLASTn searches. When multiple sequences were found for a species, only the most complete open reading frame or protein prediction was retained. If no previously annotated protein sequence was available in a database (or, it was apparently incorrectly annotated on the basis of protein sequence alignments with other orthologs) then nucleotide sequences were annotated manually, using Sequencher v4.5 (Genecodes, Ann Arbor, MI). Exons were identified with the aid of inferred translations from BLASTx pairwise comparisons to the NCBI protein sequence database and the locations of putative intron splice donor and acceptor site sequences (*e.g.* G/GT to AG/G, although others may be observed among diverse eukaryotes). Additional comparisons of the inferred Rad51 and Dmc1 homologous amino acid sequences were performed with alignments created using MUSCLE v3.7 (Edgar 2004) and observed with BioEdit v7.0.5.3 (Hall 1999).

**Degenerate PCR:**

DNA samples were obtained by collaboration with Jeff Cole and Robert Molestina at the American Type Culture Collection (ATCC, Manassas, VA), mainly from xenic monoprotistan cultures. PCR amplifications were performed using degenerate
oligonucleotide primers (i.e. primers designed corresponding to highly conserved regions of protein sequence alignments which reflect the degeneracy of the genetic code, see Table 3.2 and Figure 3.13 arrows) synthesized by Integrated DNA Technologies (IDT, Coralville, IA). Degenerate PCR primers Forward 6 and 7 and Reverse 1 were designed by JML and the remaining degenerate primers were designed by AWP (Table 3.2). Gene fragments of Rad51 and Dmc1 homologs were amplified from total DNA by PCR from representatives of four eukaryotic supergroups and Apusozoa (Figure 3.2).

Amplifications utilized 0.03 U/1 MasterTag polymerase (5 Prime, Gaithersburg, MD) according to the manufacturer’s instructions, 0.002U/1 Stratagene Cloned Pfu (La Jolla, CA) (to increase yields), 0.5 - 1 ng total DNA, 0.25 mM each dNTP (Stratagene): 1.5 mM MgCl2, and 10 μM each primer. Reaction conditions were 95° C for 2 minutes followed by 40 cycles including denaturation at 94° C for 40 seconds, with replicates annealing at temperatures of 55° C, 60° C, or 65° C for 1 minute, extension at 72° C starting at 1.5 minutes, adding 6 seconds per cycle, and ending with 10 minutes at 72° C, in Eppendorf gradient Mastercyclers (Hamburg, Germany). Resulting PCR products were analyzed for size on 2% agarose gels by electrophoresis. Initially, eight degenerate primer combinations were tested for each sample. When necessary, additional primer combinations were applied or nested amplifications were performed using diluted (1:1000) PCR products. Subsequent amplifications extended coverage of target genes by primer walking, using exact-match primers vs. degenerate primers in all possible combinations. Amplicons for Perkinsus marinus Rad51 genes were obtained with exact-match primers designed from non-overlapping partial sequences (NCBI GenInfo numbers 126277177 and 126301963, Table 3.2). Selected amplicons were fractionated and excised from 0.5% NuSieve GTG: 0.5% low-melt agarose gels (BioWhittaker [Walkersville, MD], Fisher [Pittsburgh, PA]) at 4° C and 100 V for 40 minutes in 1x TAE buffer) and cloned directly into the pSC-A™ vector (StrataClone™ kit, Stratagene, La Jolla CA, USA). Positive clones were identified by PCR with T3 and T7 primers to verify the
presence of appropriately sized inserts (cycling conditions: 94° for 2 minutes followed by 30 cycles at 94° C for 1 minute, 57° C for 1 minute, and 72° C for 1.5 minutes, ending with 72° C for 5 minutes [Stratagene and Promega]). At least two clones per PCR product were isolated with FastPlasmid Mini kits (5 Prime, Gaithersburg, MD) and sequenced in each direction with ABI BigDye 3.1 reagents and T3 and T7 primers, on an ABI 3730 sequencer (Applied Biosystems [Foster City, CA]).

Nucleotide sequence data was assembled with Sequencher v4.5 (Genecodes, Ann Arbor, MI) and the identities were initially verified with BLASTx searches in NCBI. If either Rad51 or Dmc1 gene sequence fragments were isolated, but not both genes, then single sequences from four or five additional clones were obtained to detect the other paralog. In total, sequences generated from both strands for at least three clones per gene were obtained. Nucleotide sequences were annotated and inferred exons were translated to proteins as described above (Database Searches).

Phylogenetic analyses:

We aligned all potential eukaryotic Rad51 and Dmc1 protein sequences with archaebacterial RadA protein sequences using MUSCLE v3.7, manually edited them by removing ambiguously aligned columns and gaps in BioEdit v7.0.5.3 (Hall 1999; Edgar 2004), and performed phylogenetic analyses on the multiple sequence alignment. Optimal protein substitution models and parameters were determined for each alignment independently with Modelgenerator v0.85 (Keane et al. 2006). Analyses were performed with PhyML v3.0 (Guindon et al. 2009) for 1000 replicates, and PhyloBayes v3.1 (Lartillot, Lepage, and Blanquart 2009), which used at least two independent converged chains in which maximum differences observed across all bipartitions were less than 0.10. Every other tree after burnins (selected to minimize the differences across all bipartitions) was used to calculate consensus tree topologies. Only sequences that unambiguously grouped as either Rad51 or Dmc1 were retained, while those that did not most likely represented other Rad51 paralogs such as Rad55 or Rad57 (Lin et al. 2006).
and were removed prior to subsequent analysis (Figure 3.2). Uncorrected pairwise protein sequence distances were calculated with ClustalX v2.0.12 (Thompson et al. 1997). Pairs of sequences with less than 0.10 protein sequence distance were identified. One member of this pair was removed on the basis of observed protein sequence-lengths or branch-lengths determined with phylogenetic analyses, usually reducing representation to one species per genus (Stiller and Harrell 2005). We removed the most divergent sequences during subsequent analyses, as necessary, to minimize the effects of long-branch attraction (Felsenstein 2004; Hampl et al. 2009).
Table 3.1: Support for eukaryotic supergroups and first order groups from phylogenetic analyses of Rad51, Dmc1, and concatenated protein sequence data.

<table>
<thead>
<tr>
<th></th>
<th>Opisthokonta</th>
<th>Amoebozoa</th>
<th>Excavata</th>
<th>Chromalveolata</th>
<th>Archaeplastida</th>
<th>Rhizaria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rad51</td>
<td></td>
<td>+++</td>
<td></td>
<td></td>
<td></td>
<td>+++</td>
</tr>
<tr>
<td>Dmc1</td>
<td>+++</td>
<td>+++</td>
<td>++</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Concat.</td>
<td>+++</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Rad51</td>
<td>++</td>
<td>+++</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>-</td>
<td>N/A</td>
<td>+</td>
<td>+++</td>
<td>++</td>
<td>-</td>
<td>+++</td>
<td>N/A</td>
</tr>
<tr>
<td>Dmc1</td>
<td>+++</td>
<td>+++</td>
<td>N/A</td>
<td>N/A</td>
<td>++</td>
<td>+</td>
<td>+++</td>
<td>+++</td>
<td>-</td>
<td>-</td>
<td>+++</td>
<td>-</td>
<td>N/A</td>
</tr>
<tr>
<td>Concat.</td>
<td>+++</td>
<td>+++</td>
<td>N/A</td>
<td>N/A</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+++</td>
<td>+++</td>
<td>-</td>
<td>+++</td>
<td>++</td>
<td>-</td>
</tr>
</tbody>
</table>

Note: Support for eukaryotic groups was assessed with PhyloBayes posterior probabilities from phylogenetic analyses performed on Rad51, Dmc1, and concatenated protein sequences (Figure 1). Pluses indicate that monophyletic groups were retrieved (ignoring the placement of Apusozoa) (+++= > 0.90, ++= 0.70-0.90, +< < 0.70) and minuses indicate the relationship was not retrieved. N/A indicates only one representative of the group was in the alignment.
<table>
<thead>
<tr>
<th>S.c. Rad51 amino acid position</th>
<th>R3</th>
<th>R5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Rad51</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>F6 R1</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>F7 R1</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>F8 R1</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>F6 R2</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>F7 R2</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>F8 R2</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>F6 R3</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>F7 R3</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>F8 R3</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>F6 R4</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>F7 R4</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>F8 R4</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>F6 R5</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>F7 R5</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>F8 R5</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 3.1:** Graphic representation of *Rad51* or *Dmc1* gene sequence fragments amplified with degenerate PCR from representatives of four eukaryotic supergroups and Apusozoa relative to *Saccharomyces cerevisiae* Rad51 protein sequence. Amoebozoa are labeled with blue, Rhizaria with eggplant, Chromalveolata with orange, Excavata with brown, and Apusozoa with black. Amino acid positions are *Saccharomyces cerevisiae* (S.c.) Rad51 protein sequence positions. Grey bars indicate regions encoded by fragments amplified with degenerate PCR. Letters and numbers on each side of grey bars indicate degenerate primers used (Table 3.2 and Figure 3.1).
Figure 3.2: **Unrooted phylogenetic tree of 47 Dmc1 homologs.** Trees were estimated with PhyML (LG+G) and PhyloBayes (LG+G) from 312 aligned amino acids. Opisthokonta are highlighted in purple, Amoebozoa in blue, Archaeplastida in green, Chromalveolata in orange, Rhizaria in violet, and Excavata in brown. Asterisks indicate data was obtained with degenerate PCR. The consensus topology of 2 PhyloBayes chains is shown.
**Figure 3.3: Unrooted phylogenetic tree of 47 Dmc1 homologs with accession numbers.** Trees were estimated with PhyML (LG+G) and PhyloBayes (LG+G) from 312 aligned amino acids. Opisthokonta are highlighted in purple, Amoebozoa in blue, Archaeplastida in green, Chromalveolata in orange, Rhizaria in violet, and Excavata in brown. Asterisks indicate data was obtained with degenerate PCR. All references are GenBank unless otherwise noted. The consensus topology of 2 PhyloBayes chains is shown.
Figure 3.4: Unrooted phylogenetic tree of 54 Dmc1 and RadA homologs with accession numbers. Trees were estimated with PhyML (LG+G) and PhyloBayes (LG+G) from 312 aligned amino acids. Opisthokonta are highlighted in purple, Amoebozoa in blue, Archaeplastida in green, Chromalveolata in orange, Rhizaria in violet, and Excavata in brown. Asterisks indicate data was obtained with degenerate PCR. All references are GenBank unless otherwise noted. The consensus topology of 2 PhyloBayes chains is shown.
Figure 3.5: Unrooted phylogenetic tree of 105 Rad51 and Dmc1 homologs. Trees were estimated with PhyML (LG+G) and PhyloBayes (LG+G) from 315 aligned amino acids. Opisthokonta are highlighted in purple, Amoebozoa in blue, Archaeplastida in green, Chromalveolata in orange, Rhizaria in violet, and Excavata in brown. Asterisks indicate data was obtained with degenerate PCR. The consensus topology of 2 PhyloBayes chains is shown.
Figure 3.6: Unrooted phylogenetic tree of 112 Rad51, Dmc1, and RadA homologs. Trees were estimated with PhyML (LG+G) and PhyloBayes (LG+G) from 315 aligned amino acids. Opisthokonta are highlighted in purple, Amoebozoa in blue, Archaeplastida in green, Chromalveolata in orange, Rhizaria in violet, and Excavata in brown. Asterisks indicate data was obtained with degenerate PCR. The consensus topology of 2 PhyloBayes chains is shown.
Figure 3.7: Unrooted phylogenetic tree of 157 Rad51, Dmc1 and RadA homologs. Trees were estimated with PhyloBayes (LG+G) from 314 aligned amino acids. Opisthokonta are highlighted in purple, Amoebozoa in blue, Archaeplastida in green, Chromalveolata in orange, Rhizaria in violet, and Excavata in brown. Asterisks indicate data was obtained with degenerate PCR. The consensus topology of 2 PhyloBayes chains is shown.
Figure 3.8: Unrooted phylogenetic tree of 52 Rad51 homologs. Trees were estimated with PhyML (LG+G) and PhyloBayes (LG+G) from 307 aligned amino acids. Opisthokonta are highlighted in purple, Amoebozoa in blue, Archaeplastida in green, Chromalveolata in orange, Rhizaria in violet, and Excavata in brown. Asterisks indicate data was obtained with degenerate PCR. The consensus topology of 2 PhyloBayes chains is shown.
Figure 3.9: Unrooted phylogenetic tree of 58 Rad51 homologs with accession numbers. Trees were estimated with PhyML (LG+G) and PhyloBayes (LG+G) from 307 aligned amino acids. Opisthokonta are highlighted in purple, Amoebozoa in blue, Archaeplastida in green, Chromalveolata in orange, Rhizaria in violet, and Excavata in brown. Asterisks indicate data was obtained with degenerate PCR. The consensus topology of 2 PhyloBayes chains is shown.
Figure 3.10: Unrooted phylogenetic tree of 65 Rad51 and RadA homologs with accession numbers. Trees were estimated with PhyML (LG+G) and PhyloBayes (LG+G) from 314 aligned amino acids. Opisthokonta are highlighted in purple, Amoebozoa in blue, Archaeplastida in green, Chromalveolata in orange, Rhizaria in violet, and Excavata in brown. Asterisks indicate data was obtained with degenerate PCR. All references are GenBank unless otherwise noted. The consensus topology of 2 PhyloBayes chains is shown.
Figure 3.11: Unrooted phylogenetic tree of 40 Concatenated Rad51 and Dmc1 homologs. Trees were estimated with PhyML (LG+G) and PhyloBayes (LG+G) from 603 aligned amino acids. Opisthokonta are highlighted in purple, Amoebozoa in blue, Archaeplastida in green, Chromalveolata in orange, Rhizaria in violet, and Excavata in brown. Asterisks indicate data was obtained with degenerate PCR. The consensus topology of 2 PhyloBayes chains is shown.
Figure 3.12: Unrooted phylogenetic tree of 40 Concatenated Rad51 and Dmc1 homologs with accession numbers (Dmc1/Rad51). Trees were estimated with PhyML (LG+G) and PhyloBayes (LG+G) from 603 aligned amino acids. Opisthokonta are highlighted in purple, Amoebozoa in blue, Archaeplastida in green, Chromalveolata in orange, Rhizaria in violet, and Excavata in brown. Asterisks indicate data was obtained with degenerate PCR. All references are GenBank unless otherwise noted. The consensus topology of 2 PhyloBayes chains is shown.
Figure 3.13: Protein sequence alignment of prokaryotic and eukaryotic RecA orthologs with amino acids conserved among 158 protein sequences indicated. Two eubacterial RecA, seven archaeobacterial RadA, 98 eukaryotic Rad51, and 51 eukaryotic Dmc1 protein sequences were aligned and analyzed for conserved amino acids. Seven exemplar Rad51 and Dmc1 protein sequences and two RadA protein sequences are presented. Amino acids that were present 100% among all domains, archaeobacteria and eukaryotes only, eukaryotes only, and Rad51 or Dmc1 only are highlighted with black, blue, green and yellow respectively. In addition, sites present ≥ 95% in one eukaryotic paralog but different or variable in the other paralog are highlighted in red. Dots mark residues identified during this study for which no function has been determined. Opisthokonta are labeled in purple, Amoebozoa in blue, Chromalveolata in orange, Excavata in brown, Rhizaria in eggplant, and Apusozoa in black. Arrows indicate positions of degenerate PCR primers. Numbers indicate amino acid positions of *Saccharomyces cerevisiae* Rad51. Supergroups were represented at each amino acid position except Apusozoa (126-356) and Rhizaria (188-397).
<table>
<thead>
<tr>
<th></th>
<th>recA</th>
<th>RadA</th>
<th>Rad51</th>
<th>Dmc1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dmc1</strong></td>
<td>0.77</td>
<td>0.55</td>
<td>0.46</td>
<td>0.34</td>
</tr>
<tr>
<td><strong>Rad51</strong></td>
<td>0.80</td>
<td>0.57</td>
<td>0.29</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Saccharomyces</th>
<th>Entamoeba</th>
<th>Oryza</th>
<th>Plasmodium</th>
<th>Trypanosoma</th>
<th>Cercomonas*</th>
<th>Ancyromonas*</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Rad51</strong></td>
<td>0.33</td>
<td>0.32</td>
<td>0.28</td>
<td>0.35</td>
<td>-</td>
<td>0.23</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td>0.22</td>
<td>0.30</td>
<td>-</td>
<td>0.34</td>
<td>-</td>
<td>0.26</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>0.37</td>
<td>0.35</td>
<td>0.34</td>
<td>-</td>
<td>0.28</td>
<td>-</td>
<td>0.39</td>
</tr>
<tr>
<td></td>
<td>0.00-0.19</td>
<td>0.20-0.29</td>
<td>0.30-0.39</td>
<td>0.40-0.49</td>
<td>0.50-0.59</td>
<td>0.60-0.69</td>
<td>0.70-0.79</td>
</tr>
</tbody>
</table>

**Figure 3.14: p-distance matrix of prokaryotic and eukaryotic RecA orthologs.** Uncorrected distances between eukaryotic Rad1 and Dmc1, archaeabacterial RadA, and eubacterial RecA protein sequences. All currently recognized eukaryotic supergroups are represented (purple=Opisthokonta, light blue=Amoebozoa, green=Archaeplastida, orange=Chromalveolata, brown=Excavata, blue=Rhizaria), and the currently unclassified Apusozoa (*Ancyromonas* and *T. trahens*). Calculations were performed with MEGA. Asterisks designate sequences obtained with degenerate PCR.
Table 3.2: Degenerate primers and their positions.

<table>
<thead>
<tr>
<th>Name</th>
<th>Nucleotide Sequence</th>
<th>Amino Acid Sequence</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>forward 1</td>
<td>ATC AAG GGC TTR AGY GA</td>
<td>I K G L S D/E</td>
<td>122</td>
</tr>
<tr>
<td>forward 2</td>
<td>ATC AAG GGA CTN TCN GA</td>
<td>I K G L S D/E</td>
<td>122</td>
</tr>
<tr>
<td>forward 3</td>
<td>ATC AAG GGA CTN AGY GA</td>
<td>I K G L S D/E</td>
<td>122</td>
</tr>
<tr>
<td>forward 4</td>
<td>ATC AAG GGC ATH TCN GA</td>
<td>I K G I S D/E</td>
<td>122</td>
</tr>
<tr>
<td>forward 5</td>
<td>ATC AAG GGC ATH AGY GA</td>
<td>I K G I S D/E</td>
<td>122</td>
</tr>
<tr>
<td>forward 6</td>
<td>GAG ATG TTC GGC GAR TTY MG</td>
<td>E M F G E F R/S</td>
<td>184</td>
</tr>
<tr>
<td>forward 7</td>
<td>GAC AGG GAA GGC ACN TTY MG</td>
<td>D R E G T F R/S</td>
<td>221</td>
</tr>
<tr>
<td>forward 8</td>
<td>ACT GAA GGC ACN TTY MGN CC</td>
<td>T E G T F R/S P</td>
<td>222</td>
</tr>
<tr>
<td>reverse 1</td>
<td>AGC GAC GAC YTG RTT NGT</td>
<td>T N Q V V A</td>
<td>331</td>
</tr>
<tr>
<td>reverse 2</td>
<td>ATG CGC GAC NAC YTG RTT</td>
<td>T N Q V V A H</td>
<td>332</td>
</tr>
<tr>
<td>reverse 3</td>
<td>TGC GAA DAT RTK NCC NCC</td>
<td>G G H/N I F A</td>
<td>355</td>
</tr>
<tr>
<td>reverse 4</td>
<td>TTC ACC YTT NCC YTT</td>
<td>K G K G E</td>
<td>370</td>
</tr>
<tr>
<td>reverse 5</td>
<td>AGT CTC ACC ACK NCC YTT</td>
<td>K G R G E T</td>
<td>371</td>
</tr>
<tr>
<td>Perkinsus forward</td>
<td>ATT GAC CAG GGC ATA GGT</td>
<td>I D Q G I G T</td>
<td>98</td>
</tr>
<tr>
<td>Perkinsus reverse</td>
<td>AAT TCT GAG CGC AAC AGG TT</td>
<td>N L L R S E F/L</td>
<td>290</td>
</tr>
</tbody>
</table>

Note: positions are relative to the *Saccharomyces cerevisiae* Rad51 protein
### Table 3.3: Proposed functions of residues identified during this study.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Residue</th>
<th>Function</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>RecA</td>
<td>G185</td>
<td>ATP binding</td>
<td>(Pellegrini et al. 2002)</td>
</tr>
<tr>
<td></td>
<td>D219</td>
<td></td>
<td>(Story, Weber, and Steitz 1992)</td>
</tr>
<tr>
<td>RadA</td>
<td>E221</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rad51</td>
<td>D280</td>
<td>DNA binding</td>
<td></td>
</tr>
<tr>
<td></td>
<td>S281</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dmc1</td>
<td>N325</td>
<td>DNA binding</td>
<td></td>
</tr>
<tr>
<td></td>
<td>G346</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>G347</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RadA</td>
<td>R287</td>
<td>DNA binding</td>
<td>(Okorokov et al. 2010)</td>
</tr>
<tr>
<td>Rad51</td>
<td>E144</td>
<td>Subunit binding</td>
<td>(Chen et al. 2007)</td>
</tr>
<tr>
<td>Dmc1</td>
<td>E176</td>
<td></td>
<td>(Conway et al. 2004)</td>
</tr>
<tr>
<td></td>
<td>F224</td>
<td></td>
<td>(Grigorescu et al. 2009)</td>
</tr>
<tr>
<td></td>
<td>R225</td>
<td></td>
<td>(Okorokov et al. 2010)</td>
</tr>
<tr>
<td></td>
<td>L285</td>
<td></td>
<td>(Pellegrini et al. 2002)</td>
</tr>
<tr>
<td>Rad51</td>
<td>L296</td>
<td>DNA binding</td>
<td>(Seong et al. 2009)</td>
</tr>
<tr>
<td></td>
<td>H356</td>
<td></td>
<td>(Shin et al. 2003)</td>
</tr>
<tr>
<td></td>
<td>E182</td>
<td></td>
<td>(Story, Weber, and Steitz 1992)</td>
</tr>
<tr>
<td></td>
<td>E295</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>R299</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>A351</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>H352</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>A320</td>
<td>Rad52 binding</td>
<td></td>
</tr>
<tr>
<td>Dmc1</td>
<td>K371</td>
<td>Subunit binding</td>
<td>(Okorokov et al. 2010)</td>
</tr>
<tr>
<td></td>
<td>E382</td>
<td></td>
<td>(Pellegrini et al. 2002)</td>
</tr>
<tr>
<td></td>
<td>E186</td>
<td>Rad54 binding</td>
<td>(Shin et al. 2003)</td>
</tr>
<tr>
<td></td>
<td>R293</td>
<td>DNA binding</td>
<td>(Story, Weber, and Steitz 1992)</td>
</tr>
<tr>
<td></td>
<td>G294</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>K343</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>K191</td>
<td>ATP binding</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Q193</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rad51</td>
<td>N254</td>
<td>Subunit binding</td>
<td>(Chen et al. 2007)</td>
</tr>
<tr>
<td>≠</td>
<td>A265</td>
<td></td>
<td>(Shin et al. 2003)</td>
</tr>
<tr>
<td>Dmc1</td>
<td>R308</td>
<td>Subunit/BRC4</td>
<td>(Story, Weber, and Steitz 1992)</td>
</tr>
<tr>
<td></td>
<td>D332</td>
<td>DNA binding</td>
<td></td>
</tr>
<tr>
<td></td>
<td>I349</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dmc1</td>
<td>Q301</td>
<td>Subunit binding</td>
<td>(Shin et al. 2003)</td>
</tr>
</tbody>
</table>

Note: Functions were determined by analysis of RecA, RadA, Rad51, and Dmc1 protein structures. Amino acids were either identified in 100% of sequences at that position or 95% among one eukaryotic paralog but different or variable in the other. Amino acid positions are relative to the *Saccharomyces cerevisiae* Rad51 protein.
MEIOSIS-SPECIFIC GENES AROSE BY DUPLICATION PRIOR TO THE LAST COMMON ANCESTOR OF EUKARYOTES

Abstract

Meiosis is a distinct and nearly universal feature of eukaryotes. However, the origins and evolutionary histories of genes that encode proteins that function during meiosis remain largely unknown. Whether the last eukaryotic common ancestor (LECA) was capable of meiosis is unknown. Also, whether meiosis in the LECA may have used the same machinery used by extant eukaryotes to complete important functions during meiosis, such as: 1) synaptonemal complex formation; 2) interhomolog DNA strand exchange; 3) Holliday junction resolution; and 4) sister chromatid cohesion, is unknown. We present our inventory of 20 genes whose products catalyze these important functions (Hop1, Rad21, Rec8, Spo11-1, Spo11-2, Spo11-3, Rad51, Dmc1, Hop2, Mnd1, Pms1, Mlh1, Mlh2, Mlh3, Msh2, Msh3, Msh4, Msh5, Msh6, and Mer3) among 46 diverse eukaryotes. For the first time, genomes of representatives from all eukaryotic supergroups and the Apusozoa (*Thecamonas trahens*) were tested for the presence of these meiotic components. We used alignments of phylogenetically verified protein sequence data to search nucleotide, EST, and protein sequence repositories. We determined that 10 of 20 genes are present in all eukaryotic supergroups and the unclassified Apusozoa, and 19 were likely present in the LECA. I also performed phylogenetic analyses on the protein sequence data obtained for all of the eukaryotes tested, revealing a pattern of gene duplications, most prior to the LECA. Many genes that encode proteins known to function only during meiosis in model organisms are paralogs of genes whose products also function during mitotic DNA damage repair or maintenance. In addition, these genes most likely arose by duplication of genes involved
in DNA damage repair. These data indicate that meiosis itself likely arose by gene duplication.

Introduction

The evolutionary forces that gave rise to meiosis are unknown (Cavalier-Smith 2002d; d'Erfurth et al. 2009; Wilkins and Holliday 2009; Bernstein and Bernstein 2010). Efforts to collect data on the origins of meiotic genes are ongoing (Villeneuve and Hillers 2001; Ramesh, Malik, and Logsdon 2005; Malik et al. 2008; Cavalier-Smith 2010; Wickstead, Gull, and Richards 2010). Meiosis is distinguished from mitosis by two nuclear divisions (reductional and equational) following a single genome-wide replication event (generally resulting in four genetically distinct haploid products), rather than one nuclear division (generally resulting in two genetically identical diploid products). Despite this dramatic difference, many events that occur during meiosis are analogous to mitosis, which itself depends upon functional components of somatic DNA mismatch and damage repair (Borts, Chambers, and Abdullah 2000; Marcon and Moens 2005; d'Erfurth et al. 2009; Wilkins and Holliday 2009; Bernstein and Bernstein 2010). Furthermore, some genes that encode products that process DNA damage in all domains of life are homologous to genes that encode proteins that function during meiosis in eukaryotes (Ramesh, Malik, and Logsdon 2005; Malik et al. 2007; Malik et al. 2008). It is from these observations that the generally held notion in which meiosis arose from mitosis early during eukaryotic evolution naturally emanates (Wilkins and Holliday 2009). Determining when meiotic genes appeared during the history of eukaryotes (especially those that function only during meiosis in model organisms) and what genetic mechanisms were responsible will help to clarify how meiosis arose (Ramesh, Malik, and Logsdon 2005).

Differences between meiotic and mitotic forms of nuclear division are apparent during meiotic prophase I, during which interactions between homologous chromosomes ensure their appropriate alignment and subsequent segregation (Dudas
and Chovanec 2004; Krogh and Symington 2004; Filippo, Sung, and Klein 2008; d'Erfurth et al. 2009; Wilkins and Holliday 2009). Events necessary for completion of meiotic prophase I in many eukaryotes include: 1) synaptonemal complex formation; 2) interhomolog DNA strand exchange; 3) sister chromatid cohesion; and 4) Holliday junction resolution (d'Erfurth et al. 2009; Wilkins and Holliday 2009). Several studies established that genes that encode products known to function during these events in model organisms arose very early during eukaryotic evolution (Paques and Haber 1999; Villeneuve and Hillers 2001; Dudas and Chovanec 2004; Krogh and Symington 2004; Ramesh, Malik, and Logsdon 2005; Filippo, Sung, and Klein 2008; Malik et al. 2008; Wickstead, Gull, and Richards 2010). Definitive evidence that genes involved in different stages of meiosis were present in the last common ancestor to all eukaryotes have not been produced because of several limiting conditions, including: a) failure to place a root on the eukaryotic phylogeny (Baldauf 2003; Simpson and Roger 2004; Roger and Simpson 2009); b) lack of genome-sequence data for all eukaryotic supergroups; and c) the existence of currently unclassified eukaryotes. The failure to completely resolve evolutionary relationships among eukaryotes makes it possible that any unsampled supergroups or unclassified eukaryotes may be the earliest-diverging eukaryotes and exclusion from analyses could result in misestimations of the presence of genes in the common ancestor of extant eukaryotes. Furthermore, biased taxonomic sampling of eukaryotes has been problematic for phylogenetic analyses (Dunn et al. 2008), clouding any conclusions regarding the evolution of meiotic genes.

We performed extensive searches of sequence databases for 20 genes that encode products involved in sister chromatid cohesion, pairing of homologous chromosomes, synaptonemal complex formation, and interhomolog DNA strand exchange (Hop1, Rad21, Rec8, Spo11-1, Spo11-2, Spo11-3, Rad51, Dmc1, Hop2, Mnd1, Pms1, Mlh1, Mlh2, Mlh3, Msh2, Msh3, Msh4, Msh5, Msh6, and Mer3) in the genomes of 46 diverse eukaryotes. Ten of these genes (Hop1, Rec8, Spo11-1, Spo11-2,
Dmc1, Hop2, Mnd1, Msh4, Msh5, and Mer3) are known to function only during meiosis in model organisms in model organisms (Malik et al. 2007). Access to newly sequenced genomes of eukaryotes from previously neglected groups greatly increased the breadth of sampling and provided our first glimpses of the suites of meiotic genes present in the supergroup Rhizaria (Bigelowiella natans), the first order group Haptophyta (Emiliania huxleyi), and currently unclassified Apusozoa (Thecamonas trahens). Distributions of 10 genes that encode proteins that function during four distinct stages of meiosis (eight that are meiosis-specific) indicate these genes were present in the last common ancestor to all eukaryotes. These analyses also provide data supporting the additional presence of Spo11-3 and Msh3 in the last common ancestor of eukaryotes; only Rec8 and Spo11-2 may have arisen later during eukaryotic evolution. Eukaryotic homologs that encode products that function during mitosis and DNA repair are frequently paralogs of meiosis-specific gene products. Several genes most likely arose from other genes that encode products that function during DNA repair, replication, or transformation (i.e. orthologs of archaebacterial Ski2, Top6A, RadA, MutL, and MutS genes). The bulk of meiotic genes tested arose once by duplication prior to the last common ancestor of eukaryotes (Malik et al. 2007; Bernstein and Bernstein 2010). The presence of so many genes in the last common ancestor that encode products necessary for a range of important steps during meiosis in extant eukaryotes strongly implies that meiosis in the last common ancestor was similar to meiosis observed in most eukaryotes today.

Results and discussion

Distributions of meiotic genes

We present the distribution of 20 genes that encode proteins that function during meiosis (Hop1, Rad21, Rec8, Spo11-1, Spo11-2, Spo11-3, Rad51, Dmc1, Hop2, Mnd1, Mlh1-3, Pms1, Msh2-6, and Mer3) among 46 diverse eukaryotes (Figure 4.1). These genes were selected on the basis that their products are important for four different stages
of meiosis: 1) synaptonemal complex formation; 2) interhomolog DNA strand exchange; 3) sister chromatid cohesion; and 4) Holliday junction resolution (Table 4.1) (Kleckner 1996; d'Erfurth et al. 2009). The taxonomic sampling includes representatives of every currently recognized eukaryotic supergroup and the Apusozoa (*Thecamonas trahens*).

Ten genes (*Rad51, Dmc1, Hop2, Mnd1, Mlh1, Mlh3, Msh2, Msh4, Msh5, and Msh6*) are present in representatives of every supergroup and *T. trahens*, implying that they were present in the last eukaryotic common ancestor (LECA) (Figure 4.2). An additional six genes (*Hop1, Rad21, Spo11-1, Pms1, and Mer3*) are missing from representatives of at least one eukaryotic supergroup and/or *T. trahens* (Figures 4.1 and 4.2) but are likely to have been present in the LECA, given their distributions and our current understanding of the evolutionary relationships among eukaryotes (Figure 1.2).

In addition, we interpreted the distribution of genes in the context of phylogenetic analyses performed on translated amino acid sequences of putative paralogs, with and without products of archaebacterial orthologs (Figures 4.3 - 4.15). Tree topologies retrieved with phylogenetic analyses of protein sequences translated from the 16 genes inferred to have been present in the LECA feature strongly supported monophyletic clades for many paralogs. Similarly, strongly supported topologies from analyses including Spo11-3, Msh3, Mlh2, and Rec8 protein sequences support the monophyly of their genes (Figures 4.5; 4.6; 4.14; and 4.15). Since these paralogs arose simultaneously during eukaryotic evolution, the distribution of one paralog can be inferred to be true for the other.

Since the *Spo11-1* gene is inferred to have been present in the LECA and the *Spo11-3* gene arose at the same time, *Spo11-3* is also likely to have been present in the LECA. This inference is especially important if genes are apparently absent from particular groups (e.g. Discoba and/or Metamonada) that have been hypothesized as the earliest-diverging eukaryotes. As such, absences from such groups could indicate that a genes was not present in the LECA but arose early during eukaryotic evolution, after the
divergence of some eukaryotes. Thus, the \textit{Spo11-3}, \textit{Msh3}, \textit{Mlh2}, and \textit{Rec8} genes were likely present in the LECA (Figure 4.2). Only one gene (\textit{Spo11-2}) may have arisen later during eukaryotic evolution given its distribution (Figures 4.1; 4.5; and 4.6). The \textit{Spo11-2} gene is apparently absent from genomes of the Metamonada tested (\textit{Trichomonas vaginalis}, \textit{Giardia intestinalis}, and \textit{Spironucleus vortens}). In addition, the phylogenetic analyses of \textit{Spo11-2} protein sequence data (Figures 4.5 and 4.6) retrieve topologies in which the \textit{Spo11-2} clade is nested within the \textit{Spo11-1} clade. This indicates that the \textit{Spo11-2} gene may have arisen later during eukaryotic evolution, if the Metamonada are the earliest-diverging eukaryotes (Figure 1.2).

The phylogenetic analyses performed here indicate also that all of the meiotic genes tested arose by gene duplication (Figure 4.16) and many are orthologous to archaebacterial genes that encode proteins that function during DNA damage repair. Interestingly, several genes that encode proteins that function only during meiosis in model organisms (\textit{Hop1}, \textit{Rec8}, \textit{Spo11-1}, \textit{Spo11-2}, \textit{Dmc1}, \textit{Msh4}, \textit{Msh5}, and \textit{Mer3}) are paralogs of genes whose products function also during meiosis, mitosis, DNA damage repair, or maintenance (\textit{Rev7}, \textit{Rad21}, \textit{Spo11-3}, \textit{Rad51}, \textit{Msh2}, \textit{Msh3}, \textit{Msh6}, \textit{Brr2}, and \textit{Slh1}) (Table 4.2). Further, some genes are orthologs of archaebacterial genes whose products function during DNA damage repair or maintenance (\textit{Top6A}, \textit{Ski2}, \textit{RadA}, \textit{MutS}, and \textit{MutL}). Whether the duplications of meiosis-specific genes occurred simultaneously, due to large-scale genome duplication events is unknown. However, prior studies indicate that great numbers of gene duplications are likely to have occurred in the LECA (Zhou, Lin, and Ma 2010). Although we cannot be certain that the gene duplication events yielding meiosis-specific genes mark the origin of meiosis itself, these gene duplications most certainly resulted in the meiotic functions observed today.

Assessment of distributions

To determine the likelihood that observed gene absences indicate true losses of genes from genomes (Figures 4.1 and 4.17 and Table 4.2) the heuristic metric developed
in Chapter 2 was applied. Here, the proportions of observed absences explained by sequence detection failures (type II error) were estimated. Among observed absences from genomes of any of the *Dmc1*, *Pms1*, *Msh3*, *Msh4*, *Msh6*, and *Mer3* genes there is a \( \sim 1-10\% \) chance that the gene is present in the genome sequence but was not detected. If a given organism’s genome is well covered, then the gene has most likely been lost by the organism (*e.g.* *Ustilago maydis Rad51*). However, if there is a possibility that the genome sequencing is incomplete, the gene may be present in the genome but not in the genome assembly (*e.g.* *Bigelowiella natans Pms1*). The data analyzed here indicate that sporadic secondary gene losses occur frequently among diverse eukaryotes (Figures 4.1 and 4.17), a pattern first demonstrated among genes that encode DNA strand exchange proteins (Chapter 2).

**Case study: the *Spo11* genes**

Some apparent absences of meiotic genes are more ambiguous. For example, the absence of the *Spo11*-1 and *Spo11*-2 genes from the genome sequences of *D. purpureum* and *Polysphondylium pallidum* may be due to either true losses of genes from the genomes or false negatives caused by sequence detection failures. Assessment of *Spo11*-1 and *Spo11*-2 protein sequences (meiosis-specific transesterases that introduce dsDNA breaks necessary for homologous recombination (*Keeney, Giroux, and Kleckner 1997; Baudat and Keeney 2001; Lichten 2001; Szekvolgyi and Nicolas 2010*)) indicate that a high proportion of observed absences are likely due to false negatives caused by the inability to detect gene sequences (0.67 and 0.45, respectively) (Table 4.2). In addition, only the genome of *D. purpureum* has been completely sequenced. The authors of a previous study in which the distribution of *Spo11*-1, *Spo11*-2, and *Spo11*-3 genes hypothesized that the observed absences of *Spo11*-1 and *Spo11*-2 genes from the genome of *D. discoideum* may be due to incomplete genome sequence coverage or a result of mutagenesis during axenic cultivation (*Malik et al. 2007*).
However, the additional absence these genes from *D. purpureum* implies that the *Spo11-1* and *Spo11-2* genes were absent in the common ancestor of the two *Dictyostelium* species. Recent population genetic data indicate that *D. discoideum* populations display a rapid decay of linkage disequilibrium and recombinant genotypes, consistent with meiotic recombination (Flowers *et al.* 2010). In addition, formation of macrocysts (resulting from the fusion of two haploid cells) has been observed with *D. purpureum* (Mehdiabadi *et al.* 2009) and *D. giganteum* (Mehdiabadi *et al.* 2010) and synaptonemal complexes have been observed in *D. discoideum* (Okada *et al.* 1986). Therefore, it is likely that meiosis and homologous recombination occurs in *D. purpureum*. There are three possibilities that explain the apparent absences of the *Spo11-1* and *Spo11-2* genes in *D. purpureum*: i) the rate of dsDNA breaks is sufficiently high to stimulate interhomolog DNA strand exchange without *Spo11-1* or *Spo11-2*; ii) another nuclease is introducing dsDNA breaks, or iii) the sequences have diverged, making them difficult to detect. This study cannot distinguish among these possibilities.

**Conclusions**

We performed extensive search for homologs of 20 genes that encode products that are known to catalyze at least four important tasks during meiosis: 1) synaptonemal complex formation; 2) homologous recombination; 3) Holliday junction resolution; and 4) sister chromatid cohesion (Table 4.2). The distributions of ten genes (*Rad51, Dmc1, Hop2, Mnd1, Mlh1, Mlh3, Msh2, Msh4, Msh5, and Msh6*) indicate they are present their presence in the genomes of representatives from every currently recognized eukaryotic supergroups and the Apusozoa (*Thecamonas trahens*) (Figures 4.1 and 4.2). Some genes are absent from the genomes of one or more eukaryotic supergroups or *T. trahens* (*Hop1, Rad21, Spo11-1, Pms1*, and *Mer3*). However, based upon our current understanding of the evolutionary relationships of eukaryotes (Figure 1.2), we determined that these genes are likely to have been present in the last eukaryotic common ancestor (LECA). We also performed phylogenetic analyses on the proteins translated from all of the genes collected
(Figures 4.3 – 4.15). We used protein sequences of paralogs and archaeabacterial orthologs to root the phylogenies. On the basis of these analyses we determined that an additional four genes (Rec8, Spo11-3, Mlh2, and Msh3) are likely to have been present in the LECA, despite their apparent absences from representatives of multiple eukaryotic supergroups. Only one gene (Spo11-2) may have arisen later during eukaryotic evolution, based upon its distribution and phylogenetic analyses that retrieve topologies in which the Spo11-2 clade is nested within the Spo11-1 clade.

Frequently, we observed that genes arose by duplication, often in the LECA, of genes that are likely to have encoded proteins that functioned during DNA damage repair (Figure 4.16). In addition, we noticed that many homologs are paralogs in which at least one gene encodes products that function only during meiosis in model organisms and at least one other paralog that functions during both meiosis and mitosis. Nearly all of the genes here (except Hop2 and Mnd1 for which no other eukaryotic or archaeabacterial orthologs have been identified) likely arose by duplications of genes that encode DNA repair proteins, yielding multiple genes whose products are both meiosis-specific and generalist in nature, within the LECA. These data are most consistent with the possibility that meiosis arose from mitosis (Marcon and Moens 2005; d'Erfurth et al. 2009; Wilkins and Holliday 2009).

Methods

Database Searches

Keyword searches (e.g. Saccharomyces cerevisiae Rad51) of the National Center for Biotechnology Information (NCBI, www.ncbi.nlm.nih.gov/)protein sequence database retrieved protein sequences for representatives of animals, fungi, and plants. In addition, the Clusters of euKaryotic Orthologous Groups of proteins (KOGs) database for each protein was accessed (Tatusov et al. 2000). Sequence identities were initially verified using the tBLASTn (Altschul et al. 1997) option of the Basic Local Alignment Search Tool (BLAST), in which the translated nucleotide database is
searched using a protein query and evaluating the results (bi-directional BLAST).

These protein sequences were subsequently used as queries to search genome sequence databases at NCBI and other publicly available sites (Table 4.3) with BLASTp, tBLASTn, and BLASTn, as necessary, for all available *Hop1, Rev7, Rad21, Rec8, Spo11-1, Spo11-2, Spo11-3, Rad51, Dmc1, Hop2, Mnd1, Mlh1-3, Pms1, Msh2-6, Mer3, Slh1,* and *Brr2* sequences available for a set of 46 taxa from June through August 2010. Once additional protein sequence data were obtained, searches were also performed using protein sequence data from closely related organisms likely to share more recent common ancestors as queries. Identities of sequences were again confirmed with bi-directional BLAST (BLASTx and tBLASTn, as necessary). When necessary, phylogenetically verified (see below) protein sequences were aligned with MUSCLE v3.7 (Edgar 2004) and used to create position specific scoring matrices (PSSMs) with the tBLASTn module (available at http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=Download). Matrices were then used as queries with the PSI-BLAST module to search nucleotide genome sequence databases. When multiple sequences were found for a species, only the most complete was retained. If no previously annotated protein sequence was available in a database, then nucleotide sequences were annotated by hand, using Sequencher v4.5 (Genecodes, Ann Arbor, MI). Exons were identified on the basis of inferred translations using BLASTx pairwise comparisons to the NCBI protein sequence database and locations of putative intron splice donor and acceptor site sequences (*e.g.* G/GT to AG/G, although others may be observed among diverse eukaryotes). Additional comparisons of resulting amino acid sequences to other homologs were performed with alignments created using MUSCLE v3.7 (Edgar 2004) and observed with BioEdit v7.0.5.3 (Hall 1999).
Phylogenetic analyses

We aligned all potential eukaryotic protein sequences with archaeabacterial protein sequences using MUSCLE v3.7, manually edited them (removing ambiguously aligned columns and gaps) with BioEdit v7.0.5.3 (Hall 1999; Edgar 2004) and performed phylogenetic analyses on the set. Optimal protein substitution models and parameters were determined for each alignment independently with Modelgenerator v0.85 (Keane et al. 2006). Analyses were performed with RAxML v7.2.7 (Stamatakis, Hoover, and Rougemont 2008), for 1000 replicates at the CIPRES Science Gateway v3.0 (Miller et al. 2009).

Inventory Assembly

Genes were determined to be present in an organism when putative orthologs were discovered and identified with phylogenetic analyses. To determine the numbers of observed sequence absences attributable to failures of the sequence detection regimen, Smith-Waterman pairwise alignment scores (Homo sapiens versus Saccharomyces cerevisiae) were calculated with the PRSS/PRFX tool (http://fasta.bioch.virginia.edu/fasta_www2/fasta_www.cgi?rm=shuffle) (Table 2.2). The numbers of sequence detection failures expected for each protein, given its Smith-Waterman score, were determined with a Poisson regression analysis of protein sequence data previously collected (Chapter 2) for ten RNA Polymerase I and three Replication Protein A subunits among diverse eukaryotes with completed genome sequences.
Figure 4.1: Distribution of 20 homologs that function during meiosis among 46 eukaryotes representing all eukaryotic supergroups. Cells filled in with color (Opisthokonta in purple, Amoebozoa in blue, Archaeplastida in green, Chromalveolata in orange, Rhizaria in eggplant, Excavata in brown, and Apusozoa in black) indicate the homolog was found and phylogenetically verified. Labels of proteins known to function only during meiosis in model organisms are blue. Shades of grey indicate the proportion of observed absences attributed to sequence detection failures, estimated from Smith-Waterman pairwise alignment scores (Saccharomyces cerevisiae versus Homo sapiens) (see Methods). Darker greys indicate the gene is not present in the genome sequence sampled while lighter greys indicate the gene may be present but was not detected. Black protein labels identify sequences discovered in all eukaryotes sampled. Asterisks identify completed genome sequences (>8.0x WGS coverage or sequenced end-to-end).
Ratio of number of undetected sequences expected to observed

<table>
<thead>
<tr>
<th>TAXA/PROTEIN</th>
<th>Hop1</th>
<th>Rad51</th>
<th>Spo11-1</th>
<th>Spo11-2</th>
<th>Spo11-3</th>
<th>Rad5</th>
<th>Hop2</th>
<th>Msd1</th>
<th>Pms1</th>
<th>Mh1</th>
<th>Mh2</th>
<th>Mh3</th>
<th>Mh4</th>
<th>Mh5</th>
<th>Mh6</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>METAMONADA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trichomonas vaginalis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Giardia intestinalis*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spironucleus vortens</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>DISCOBA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Naegleria gruberi*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leishmania major/donovani</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trypanosoma cruzi*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>ARCHAEOPLASTIDA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arabidopsis thaliana*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oryza sativa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physcomitrella patens*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chlamydomonas reinhardtii*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chlorella sp.*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ostreococcus tauri*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Galdieria sulphuraria</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>HAPTOPHYTA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Emiliania huxleyi*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>STROMATOPILA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thalassiosira pseudonana*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phaeodactylum tricornutum*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fragilariopsis cylindrus</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phytophthora ramorum/sojae*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aureococcus anophagefferens</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blastocystis hominis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>ALVEOLATA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plasmodium vivax*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Toxoplasma gondii</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thelileria parva/annulata*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cryptosporidium muris*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perkinsus marinus</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paramaecium tetraurelia*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>RHIZARIA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bigelowiella natans</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>AMOEBOZOA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Entamoeba dispar</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dictyostelium purpureum*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Polysphondylium pallidum</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>HOLOZOA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Homo sapiens*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ciona*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nematostrasia vectensis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trichoplax adhaerens*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Monosiga brevisalis*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Salpingoeca rosetta*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Capaspora owczarzaki*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>FUNGI</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Saccharomyces cerevisiae*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aspergillus fumigatus*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ustilago maydis*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cryptococcus neoformans</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Laccaria bicolor*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coprinopsis cinerea*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mucor circinelloides*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Batrachochytrium dendrobatidis*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>APUSOZOA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thecamonas trahens*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>no failures observed</th>
<th>0.01-0.10</th>
<th>0.11-0.20</th>
<th>0.21-0.50</th>
<th>0.51-1.00</th>
<th>&gt; 1.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>no failures observed</td>
<td>0.01-0.10</td>
<td>0.11-0.20</td>
<td>0.21-0.50</td>
<td>0.51-1.00</td>
<td>&gt; 1.00</td>
</tr>
</tbody>
</table>
**Figure 4.2**: Presence of twenty homologs that function during meiosis in the last eukaryotic common ancestor (LECA) inferred by their distribution among eukaryotic supergroups. Cells filled in with color (Opisthokonta in purple, Amoebozoa in blue, Archaeplastida in green, Chromalveolata in orange, Rhizaria in eggplant, Excavata in brown, and Apusozoa in black) indicate the homolog was found and phylogenetically verified within that group. Labels of proteins known to function only during meiosis in model organisms are blue. Black plusses indicate the gene was most likely present in the LECA based upon its distribution among all eukaryotic supergroups, red plusses indicate the presence of the gene in the LECA on the basis of phylogenetic analyses.
Figure 4.3: Unrooted phylogenetic tree of 50 eukaryotic Hop1 and Rev7 homologs. Trees were estimated with maximum likelihood inference (LG+G; 1000 replicates) from 129 aligned amino acids. Opisthokonta are labeled with purple, Amoebozoa with blue, Archaeplastida with green, Chromalveolata with orange, Rhizaria with eggplant, Excavata with brown, and Apusozoa with black. Labels of proteins known to function only during meiosis in model organisms are blue. The best RAxML v7.2.7 tree is shown.
Figure 4.4: Unrooted phylogenetic tree of 49 eukaryotic Rad21 and Rec8 homologs. Trees were estimated with maximum likelihood inference (LG+G; 1000 replicates) from 171 aligned amino acids. Opisthokonta are labeled with purple, Amoebozoa with blue, Archaeplastida with green, Chromalveolata with orange, Rhizaria with eggplant, Excavata with brown, and Apusozoa with black. Labels of proteins known to function only during meiosis in model organisms are blue. The best RAxML v7.2.7 tree is shown.
Figure 4.5: Unrooted phylogenetic tree of 69 eukaryotic Spo11-1, Spo11-2, and Spo11-3 homologs with 6 archaebacterial Top6A homologs. Trees were estimated with maximum likelihood inference (LG+G+I; 1000 replicates) from 170 aligned amino acids. Opisthokonta are labeled with purple, Amoebozoa with blue, Archaeplastida with green, Chromalveolata with orange, Rhizaria with eggplant, Excavata with brown, and Apusozoa with black. Labels of proteins known to function only during meiosis in model organisms are blue. The best RAxML v7.2.7 tree is shown.
Plasmodium, Theileria, Perkinsus, Cryptosporidium, Toxoplasma, Bigelowiella, Emiliana, Galdieria, Trypanosoma, Leishmania, Aureococcus, Blastocystis, Phaeodactylum, Fragilaria, Pythophthora, Thalassiosira, Oryza, Arabidopsis, Physcomitrella, Chlorella, Chlamydomonas, Ostreococcus, Naegleria, Giardia, Spironucleus, Paramecium.
Figure 4.6: Unrooted phylogenetic tree of 69 eukaryotic Spo11-1, Spo11-2, and Spo11-3 homologs. Trees were estimated with maximum likelihood inference (LG+G+I; 1000 replicates) from 170 aligned amino acids. Opisthokonta are labeled with purple, Amoebozoa with blue, Archaeplastida with green, Chromalveolata with orange, Rhizaria with eggplant, Excavata with brown, and Apusozoa with black. Labels of proteins known to function only during meiosis in model organisms are blue. The best RAxML v7.2.7 tree is shown.
Figure 4.7: Unrooted phylogenetic tree of 81 eukaryotic Rad51 and Dmc1 homologs with 6 archaeobacterial RadA homologs. Trees were estimated with maximum likelihood inference (LG+G; 1000 replicates) from 305 aligned amino acids. Opisthokonta are labeled with purple, Amoebozoa with blue, Archaeplastida with green, Chromalveolata with orange, Rhizaria with eggplant, Excavata with brown, and Apusozoa with black. Labels of proteins known to function only during meiosis in model organisms are blue. The best RAxML v7.2.7 tree is shown.
Figure 4.8: Unrooted phylogenetic tree of 81 eukaryotic Rad51 and Dmc1 homologs with. Trees were estimated with maximum likelihood inference (LG+G; 1000 replicates) from 305 aligned amino acids. Opisthokonta are labeled with purple, Amoebozoa with blue, Archaeplastida with green, Chromalveolata with orange, Rhizaria with eggplant, Excavata with brown, and Apusozoa with black. Labels of proteins known to function only during meiosis in model organisms are blue. The best RAxML v7.2.7 tree is shown.
Figure 4.9: Unrooted phylogenetic tree of 82 eukaryotic Hop2 and Mnd1 homologs.
Trees were estimated with maximum likelihood inference (LG+G; 1000 replicates) from 98 aligned amino acids. Opisthokonta are labeled with purple, Amoebozoa with blue, Archaeplastida with green, Chromalveolata with orange, Rhizaria with eggplant, Excavata with brown, and Apusozoa with black. Labels of proteins known to function only during meiosis in model organisms are blue. The best RAxML v7.2.7 tree is shown.
Figure 4.10: Unrooted phylogenetic tree of 131 eukaryotic Mlh1, Mlh2, Mlh3, and Pms1 homologs with 4 archaeabacterial MutL homologs. Trees were estimated with maximum likelihood inference (LG+G+F; 1000 replicates) from 185 aligned amino acids. Opisthokonta are labeled with purple, Amoebozoa with blue, Archaeplastida with green, Chromalveolata with orange, Rhizaria with eggplant, Excavata with brown, and Apusozoa with black. The best RAxML v7.2.7 tree is shown.
Figure 4.11: Unrooted phylogenetic tree of 131 eukaryotic Mlh1, Mlh2, Mlh3, and Pms1 homologs. Trees were estimated with maximum likelihood inference (LG+G+F; 1000 replicates) from 185 aligned amino acids. Opisthokonta are labeled with purple, Amoebozoa with blue, Archaeplastida with green, Chromalveolata with orange, Rhizaria with eggplant, Excavata with brown, and Apusozoa with black. The best RAxML v7.2.7 tree is shown.
Figure 4.12: Unrooted phylogenetic tree of 113 eukaryotic Mer3, Brr2, and Slh1 homologs with 6 archaeobacterial Ski2 homologs. Trees were estimated with maximum likelihood inference (LG+G+I; 1000 replicates) from 338 aligned amino acids. Opisthokonta are labeled with purple, Amoebozoa with blue, Archaeplastida with green, Chromalveolata with orange, Rhizaria with eggplant, Excavata with brown, and Apusozoa with black. Labels of proteins known to function only during meiosis in model organisms are blue. The best RAxML v7.2.7 tree is shown.
Figure 4.13: Unrooted phylogenetic tree of 113 eukaryotic Mer3, Brr2, and Slh1 homologs. Trees were estimated with maximum likelihood inference (LG+G+I; 1000 replicates) from 338 aligned amino acids. Opisthokonta are labeled with purple, Amoebozoa with blue, Archaeplastida with green, Chromalveolata with orange, Rhizaria with eggplant, Excavata with brown, and Apusozoa with black. Labels of proteins known to function only during meiosis in model organisms are blue. The best RAxML v7.2.7 tree is shown.
Figure 4.14: Unrooted phylogenetic tree of 183 eukaryotic Msh2, Msh3, Msh4, Msh5, and Msh6 homologs with 5 archaeabacterial MutS homologs. Trees were estimated with maximum likelihood inference (LG+G+F; 1000 replicates) from 259 aligned amino acids. Opisthokonta are labeled with purple, Amoebozoa with blue, Archaeplastida with green, Chromalveolata with orange, Rhizaria with eggplant, Excavata with brown, and Apusozoa with black. Labels of proteins known to function only during meiosis in model organisms are blue. The best RAxML v7.2.7 tree is shown.
Figure 4.15: Unrooted phylogenetic tree of 183 eukaryotic Msh2, Msh3, Msh4, Msh5, and Msh6 homologs. Trees were estimated with maximum likelihood inference (LG+G+F; 1000 replicates) from 259 aligned amino acids. Opisthokonta are labeled with purple, Amoebozoa with blue, Archaeplastida with green, Chromalveolata with orange, Rhizaria with eggplant, Excavata with brown, and Apusozoa with black. Labels of proteins known to function only during meiosis in model organisms are blue. The best RAxML v7.2.7 tree is shown.
Figure 4.16: Radial tree topologies of archaebacterial and eukaryotic homologs. Phylogenetic analyses were performed using a maximum likelihood approach on protein sequences of eukaryotic homologs encoding products that function during meiosis with paralogs and archaebacterial homologs. Blue bubbles indicate proteins known to function only during meiosis in model organisms and pink bubbles indicate proteins that function during meiosis, mitosis, and/or DNA mismatch repair. Branch colors of eukaryotes correspond to supergroups (Opisthokonta in purple, Amoebozoa in blue, Archaeplastida in green, Chromalveolata in orange, Rhizaria in eggplant, Excavata in brown, and Apusozoa in black). Letters indicate that at least one of the proteins in the tree is important for synaptonemal complex formation (A), sister chromatid cohesion (B), double-strand breaks (C), DNA strand exchange (D), or Holliday junction resolution (E).
Table 4.1: Proteins involved in four general categories of meiosis and their functions.

<table>
<thead>
<tr>
<th>Category</th>
<th>Protein</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synaptonemal complex formation and pairing of homologous chromosomes</td>
<td>Hop1</td>
<td>Binds discrete sites on axial elements and promotes synapsis between homologous DNA duplexes during meiotic prophase I (Anuradha and Muniyappa 2004a; Anuradha and Muniyappa 2004b; Latypov et al.). <em>Hop1</em> is a paralog of Rev7, a gene encoding an accessory subunit of DNA polymerase zeta that is involved in translesion synthesis during post-replication repair and dsDNA break repair (Acharya <em>et al.</em> 2005; Kolas and Durocher 2006; Lee and Myung 2008).</td>
</tr>
<tr>
<td>Sister-chromatid cohesion</td>
<td>Rad21</td>
<td>Members of cohesin complexes that bind Smc1/Smc3 heterodimers, forming large rings around chromosomes during S-phase. Proteolytic cleavage by the separase triggers sister-chromatid disjunction during Mitotic Anaphase (Rad21) or Meiotic Anaphase II (Rec8) (Gruber, Haering, and Nasmyth 2003).</td>
</tr>
<tr>
<td>Double-strand DNA breaks</td>
<td><em>Spo11</em>-1 and -2 <em>Spo11</em>-3</td>
<td>Form dimers that cut dsDNA, generating 5'-nucleoprotein linkages on either side of the break that may become sites of recombination. Monomers are removed from the ends as oligonucleotide-bound complexes, leaving ssDNA tails. <em>Spo11</em>-1 and -2 function only during meiosis. In plants, <em>Spo11</em>-3 functions during vegetative growth (Lin and Smith 1994; Keeney, Giroux, and Kleckner 1997; Dernburg <em>et al.</em> 1998; Baudat and Keeney 2001; Hartung <em>et al.</em> 2002; Sugimoto-Shirasu <em>et al.</em> 2002; Szekvolgyi and Nicolas).</td>
</tr>
<tr>
<td>DNA strand exchange</td>
<td>Rad51 Dmc1</td>
<td>Both form helical filaments on ss- and ds-DNA, catalyze strand exchange, and cause ssDNA extension and dsDNA rotational transition during Mitotic prophase (Rad51) and meiotic prophase I (Rad51/Dmc1). Rad51 may recruit Dmc1 to the pre-synaptic filament (Nishinaka <em>et al.</em> 1998; Krogh and Symington 2004; Lopez-Casamichana <em>et al.</em> 2008).</td>
</tr>
<tr>
<td></td>
<td>Hop2 Mnd1</td>
<td>Together, they form heterodimers that stabilize Dmc1-ssDNA pre-synaptic filaments and stabilize dsDNA during DNA strand exchange (Chen <em>et al.</em> 2004; Henry <em>et al.</em> 2006).</td>
</tr>
</tbody>
</table>

Note: names of proteins known to function only during meiosis are bolded.
Table 4.1: Proteins involved in four general categories of meiosis and their functions. Names of proteins known to function only during meiosis are bolded. – continued

<table>
<thead>
<tr>
<th>Holliday junction resolution</th>
<th>Mlh1-3 Pms1</th>
<th>Msh2,3,6 4,5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mlh1 forms heterodimers with Pms1, Mlh2, and Mlh3. Mlh1-Pms1 functions during mismatch repair, interacting with Msh2/3 and Msh2/6 heterodimers. During meiosis Mlh1/2 and Mlh1/3 function to resolve heteroduplexes (Hunter and Borts 1997; Borts, Chambers, and Abdullah 2000; Hoffmann et al. 2003).</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Form heterodimer sliding clamps that may diffuse along duplex DNA adjacent to mismatches (Msh2/3 or Msh2/6), marking the location of the lesion and signaling downstream machinery. During meiosis Msh4/5 form clamps that bind Holliday junctions and Msh2/6 (Snowden et al. 2004).</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A helicase with roles in synaptonemal complex formation, crossover interference, and unwinding of Holliday junctions during meiosis (Bishop and Zickler 2004; Borner, Kleckner, and Hunter 2004; Sugawara et al. 2009; Wang et al. 2009). Mer3 is paralogous to Shh1 that encodes a putative RNA helicase involved in translation inhibition of non-poly (A) mRNAs and is required suppressing dsRNA viruses and Brr2 that encodes an RNA helicase required for activation of spliceosomal catalysis (Noble and Guthrie 1996; de la Cruz, Kressler, and Linder 1999; Seafoss, Dever, and Wickner 2001).</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Mer3
Figure 4.17: Number of detection failures as predicted by Poisson regression analysis of RNA Polymerase I and Replication Protein A subunits with observed numbers of detection failures for 18 meiotic genes. The numbers of observed meiotic gene detection failures (indicated with open circles) are plotted against the natural logarithm of Smith-Waterman pairwise alignment scores of *Homo sapiens* and *Saccharomyces cerevisiae*. Poisson regression analyses were performed on the observed numbers of failures to detect RNA Polymerase I subunits (A190, A135, AC40, AC19, AC12.2, Rpb5, Rpb6, Rpb8, Rpb10, and Rpb12) and Replication Protein A subunits (RPA1-3) among 34 taxa with at least 8.0X whole-genome shotgun sequencing coverage relative to Smith-Waterman scores (Methods). The predicted numbers of failures relative to Smith-Waterman scores (black dots) were plotted with Wald 90% confidence limits (green dots). Shades of grey indicate the proportion of observed absences attributed to sequence detection failures, estimated from Smith-Waterman pairwise alignment scores (*S. cerevisiae versus H. sapiens*) (see Methods). Darker greys indicate the gene is not present in the genome sequence sampled while lighter greys indicate the gene may be present but was not detected. Black labels identify sequences discovered in all eukaryotes sampled.
<table>
<thead>
<tr>
<th>Protein</th>
<th># Observed absences</th>
<th>Smith-Waterman alignment score</th>
<th># Predicted absences</th>
<th>Ratio (Exp./Obs.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hop1</td>
<td>14</td>
<td>261</td>
<td>9.24</td>
<td>0.66</td>
</tr>
<tr>
<td>Rad21</td>
<td>8</td>
<td>225</td>
<td>10.26</td>
<td>1.28</td>
</tr>
<tr>
<td>Rec8</td>
<td>34</td>
<td>102</td>
<td>14.66</td>
<td>0.43</td>
</tr>
<tr>
<td>Spo11-1</td>
<td>15</td>
<td>233</td>
<td>10.03</td>
<td>0.67</td>
</tr>
<tr>
<td>Spo11-2</td>
<td>23</td>
<td>220</td>
<td>10.41</td>
<td>0.45</td>
</tr>
<tr>
<td>Spo11-3</td>
<td>31</td>
<td>480</td>
<td>4.90</td>
<td>0.16</td>
</tr>
<tr>
<td>Rad51</td>
<td>2</td>
<td>1538</td>
<td>0.23</td>
<td>0.11</td>
</tr>
<tr>
<td>Dmc1</td>
<td>8</td>
<td>1178</td>
<td>0.65</td>
<td>0.08</td>
</tr>
<tr>
<td>Hop2</td>
<td>7</td>
<td>161</td>
<td>12.35</td>
<td>1.76</td>
</tr>
<tr>
<td>Mnd1</td>
<td>3</td>
<td>328</td>
<td>7.61</td>
<td>2.54</td>
</tr>
<tr>
<td>Pms1</td>
<td>3</td>
<td>1544</td>
<td>0.22</td>
<td>0.07</td>
</tr>
<tr>
<td>Mlh1</td>
<td>0</td>
<td>1664</td>
<td>0.16</td>
<td>0.00</td>
</tr>
<tr>
<td>Mlh2</td>
<td>35</td>
<td>405</td>
<td>6.09</td>
<td>0.17</td>
</tr>
<tr>
<td>Mlh3</td>
<td>15</td>
<td>427</td>
<td>5.71</td>
<td>0.38</td>
</tr>
<tr>
<td>Msh2</td>
<td>0</td>
<td>2285</td>
<td>0.03</td>
<td>0.00</td>
</tr>
<tr>
<td>Msh3</td>
<td>22</td>
<td>1534</td>
<td>0.23</td>
<td>0.01</td>
</tr>
<tr>
<td>Msh4</td>
<td>12</td>
<td>1175</td>
<td>0.65</td>
<td>0.05</td>
</tr>
<tr>
<td>Msh5</td>
<td>11</td>
<td>868</td>
<td>0.59</td>
<td>0.14</td>
</tr>
<tr>
<td>Msh6</td>
<td>3</td>
<td>1780</td>
<td>0.11</td>
<td>0.04</td>
</tr>
<tr>
<td>Mer3</td>
<td>16</td>
<td>1572</td>
<td>0.21</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Note: Smith-Waterman alignment scores were calculated from pairwise alignments of *Saccharomyces cerevisiae* and *Homo sapiens* protein sequences and used to determine the numbers of absences predicted due to detection failures. The ratio of expected and observed values indicates the proportion of observed absences that are likely due to sequence detection failures. Protein names in bold indicate meiosis-specific function.
<table>
<thead>
<tr>
<th>Organism</th>
<th>Database</th>
<th>Web Address</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Trichomonas vaginalis</em></td>
<td>TrichDB</td>
<td><a href="http://trichdb.org/trichdb/">http://trichdb.org/trichdb/</a></td>
<td>(Aurrecoechea et al. 2009a)</td>
</tr>
<tr>
<td><em>Spironucleus vortens</em></td>
<td>JGI</td>
<td><a href="http://genome.jgi-psf.org/Spivo0/Spivo0.home.html">http://genome.jgi-psf.org/Spivo0/Spivo0.home.html</a></td>
<td></td>
</tr>
<tr>
<td><em>Naegleria gruberi</em></td>
<td>JGI</td>
<td><a href="http://genome.jgi-psf.org/Naegr1/Naegr1.home.html">http://genome.jgi-psf.org/Naegr1/Naegr1.home.html</a></td>
<td></td>
</tr>
<tr>
<td><em>Leishmania</em> and <em>Trypanosoma</em></td>
<td>TriTrypDB</td>
<td><a href="http://tritrypdb.org/tritrypdb/">http://tritrypdb.org/tritrypdb/</a></td>
<td>(Aslett et al. 2010)</td>
</tr>
<tr>
<td><em>Physcomitrella patens</em></td>
<td>JGI</td>
<td><a href="http://genome.jgi-psf.org/Physcomitrella/Physcomitrella.home.html">http://genome.jgi-psf.org/Physcomitrella/Physcomitrella.home.html</a></td>
<td>(Rensing et al. 2008)</td>
</tr>
<tr>
<td><em>Chlamydomonas reinhardtii</em></td>
<td>JGI</td>
<td><a href="http://genome.jgi-psf.org/chlamy/chlamy.home.html">http://genome.jgi-psf.org/chlamy/chlamy.home.html</a></td>
<td>(Merchant et al. 2007)</td>
</tr>
<tr>
<td><em>Chlorella</em></td>
<td>JGI</td>
<td><a href="http://genome.jgi-psf.org/ChlNC64A_1/ChlNC64A_1.home.html">http://genome.jgi-psf.org/ChlNC64A_1/ChlNC64A_1.home.html</a></td>
<td></td>
</tr>
<tr>
<td><em>Ostreococcus tauri</em></td>
<td>JGI</td>
<td><a href="http://genome.jgi-psf.org/Ostta4/Ostta4.home.html">http://genome.jgi-psf.org/Ostta4/Ostta4.home.html</a></td>
<td>(Palenik et al. 2007)</td>
</tr>
<tr>
<td><em>Galdieria sulphuraria</em></td>
<td>Galdieria Genome Project</td>
<td><a href="http://genomics.msu.edu/cgi-bin/galdieria/blast.cgi">http://genomics.msu.edu/cgi-bin/galdieria/blast.cgi</a></td>
<td>(Barbier et al. 2005)</td>
</tr>
<tr>
<td><em>Emiliania huxleyi</em></td>
<td>JGI</td>
<td><a href="http://genome.jgi-psf.org/Emihu1/Emihu1.home.html">http://genome.jgi-psf.org/Emihu1/Emihu1.home.html</a></td>
<td></td>
</tr>
<tr>
<td><em>Phaeodactylum tricornutum</em></td>
<td>JGI</td>
<td><a href="http://genome.jgi-psf.org/Phatr2/Phatr2.home.html">http://genome.jgi-psf.org/Phatr2/Phatr2.home.html</a></td>
<td>(Bowler et al. 2008)</td>
</tr>
<tr>
<td><em>Fragilariopsis cylindrus</em></td>
<td>JGI</td>
<td><a href="http://genome.jgi-psf.org/Fracy1/Fracy1.home.html">http://genome.jgi-psf.org/Fracy1/Fracy1.home.html</a></td>
<td></td>
</tr>
<tr>
<td><em>Phytophthora ramorum</em></td>
<td>JGI</td>
<td><a href="http://genome.jgi-psf.org/Phyra1_1/Phyra1_1.home.html">http://genome.jgi-psf.org/Phyra1_1/Phyra1_1.home.html</a></td>
<td>(Tyler et al. 2006)</td>
</tr>
<tr>
<td><em>P. sojae</em></td>
<td>JGI</td>
<td><a href="http://genome.jgi-psf.org/Physol_1/Physol_1.home.html">http://genome.jgi-psf.org/Physol_1/Physol_1.home.html</a></td>
<td>(Tyler et al. 2006)</td>
</tr>
<tr>
<td><em>Aureococcus anophagefferens</em></td>
<td>JGI</td>
<td><a href="http://genome.jgi-psf.org/Auran1/Auran1.home.html">http://genome.jgi-psf.org/Auran1/Auran1.home.html</a></td>
<td></td>
</tr>
</tbody>
</table>
Table 4.3: Genome sequence databases searched with web address and references. - Continued

<table>
<thead>
<tr>
<th>Organism</th>
<th>Database</th>
<th>Web Address</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Plasmodium vivax</em></td>
<td>PlasmoDB</td>
<td><a href="http://plasmodb.org/plasmo/">http://plasmodb.org/plasmo/</a></td>
<td>(Aurrecoechea et al. 2009b)</td>
</tr>
<tr>
<td><em>Toxoplasma gondii</em></td>
<td>ToxoDB</td>
<td><a href="http://toxodb.org/toxo/">http://toxodb.org/toxo/</a></td>
<td>(Aurrecoechea et al. 2007)</td>
</tr>
<tr>
<td><em>Cryptosporidium muris</em></td>
<td>CryptoDB</td>
<td><a href="http://cryptodb.org/cryptodb/">http://cryptodb.org/cryptodb/</a></td>
<td>(Heiges et al. 2006)</td>
</tr>
<tr>
<td><em>Paramecium tetraurelia</em></td>
<td>Paramecium DB</td>
<td><a href="http://paramecium.cgm.cnrs-gif.fr/">http://paramecium.cgm.cnrs-gif.fr/</a></td>
<td>(Arnaiz et al. 2007)</td>
</tr>
<tr>
<td><em>Dictyostelium purpureum</em></td>
<td>JGI</td>
<td><a href="http://genome.jgi-psf.org/Dicpu1/Dicpu1.home.html">http://genome.jgi-psf.org/Dicpu1/Dicpu1.home.html</a></td>
<td></td>
</tr>
<tr>
<td><em>Ciona intestinalis</em></td>
<td>JGI</td>
<td><a href="http://genome.jgi-psf.org/Cioin2/Cioin2.home.html">http://genome.jgi-psf.org/Cioin2/Cioin2.home.html</a></td>
<td></td>
</tr>
<tr>
<td><em>Nematostella vectensis</em></td>
<td>JGI</td>
<td><a href="http://genome.jgi-psf.org/Nemve1/Nemve1.home.html">http://genome.jgi-psf.org/Nemve1/Nemve1.home.html</a></td>
<td>(Putnam et al. 2007)</td>
</tr>
<tr>
<td><em>Trichoplax adhaerens</em></td>
<td>JGI</td>
<td><a href="http://genome.jgi-psf.org/Triad1/Triad1.home.html">http://genome.jgi-psf.org/Triad1/Triad1.home.html</a></td>
<td>(Srivastava et al. 2008)</td>
</tr>
<tr>
<td><em>Monosiga brevicollis</em></td>
<td>JGI</td>
<td><a href="http://genome.jgi-psf.org/Monbr1/Monbr1.home.html">http://genome.jgi-psf.org/Monbr1/Monbr1.home.html</a></td>
<td>(King et al. 2008)</td>
</tr>
<tr>
<td><em>Salpingoeca rosetta</em></td>
<td>BROAD - Origins of Multicellularity Database</td>
<td><a href="http://www.broadinstitute.org/annotation/genome/multicellularity_project/GenomeDescripti">http://www.broadinstitute.org/annotation/genome/multicellularity_project/GenomeDescripti</a> ons.html#%3Ci%3ESalpingoeca_rosetta%3C/i%3EFormerly_know n_as_%3Ci%3EProterospongia_sp.%3C/i%3E_AT_ CC_50818</td>
<td></td>
</tr>
<tr>
<td><em>Capsaspora owcarzaki</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Thecamonas trahens</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Laccaria bicolor</em></td>
<td>JGI</td>
<td><a href="http://genome.jgi-psf.org/Lacbi1/Lacbi1.home.html">http://genome.jgi-psf.org/Lacbi1/Lacbi1.home.html</a></td>
<td>(Martin et al. 2008)</td>
</tr>
<tr>
<td><em>Coprinus cinereus</em></td>
<td>JGI</td>
<td><a href="http://genome.jgi-psf.org/Copci1/Copci1.home.html">http://genome.jgi-psf.org/Copci1/Copci1.home.html</a></td>
<td></td>
</tr>
<tr>
<td><em>Mucor circinelloides</em></td>
<td>JGI</td>
<td><a href="http://genome.jgi-psf.org/Mucci2/Mucci2.home.html">http://genome.jgi-psf.org/Mucci2/Mucci2.home.html</a></td>
<td></td>
</tr>
<tr>
<td><em>Batrachochytrium dendrobatidis</em></td>
<td>JGI</td>
<td><a href="http://genome.jgi-psf.org/Batde5/Batde5.home.html">http://genome.jgi-psf.org/Batde5/Batde5.home.html</a></td>
<td></td>
</tr>
</tbody>
</table>
CHAPTER 5

CONCLUDING REMARKS

The studies presented in this thesis were designed to provide insight into the evolutionary history of meiosis. Although different hypotheses for the prevalence and maintenance of meiosis at the population level are well developed, scant data elucidating the origin and subsequent evolution of meiotic genes are available. In the following text, I will bring the research presented in this thesis into context by outlining our current understanding of meiosis at the levels of populations, individuals, and genes. I will also present a unifying hypothesis for the origin of meiosis. Finally, I will suggest experiments that will further elucidate the origin and evolution of meiosis.

Why meiosis?

Meiosis is necessary for sexual reproduction in eukaryotes (Weismann, Parker, and Ronnfeldt 1893). Two (usually haploid) products of meiosis (e.g. spores and gametes) are combined (cell fusion), yielding offspring with the parental numbers of homologous chromosomes (usually diploid) (Figure 1.3 – B) (Weismann, Parker, and Ronnfeldt 1893). Thus, the halving of organisms’ genomes during meiosis ensures the maintenance of ploidy in offspring; sexual reproduction results in the alternation of haploid and diploid phases in eukaryotic life cycles (Maynard Smith and Szathmary 1995). There are several costs associated with sexual reproduction: 1) the time and energy to switch from mitotic to meiotic cell divisions; 2) the search for appropriate mating partners; 3) the risks of failing to find appropriate mates; 4) the risk of contracting sexually transmitted diseases; 5) the disruption of genomes that are well adapted to their environments; and 6) the transmission of only half of genetic material to offspring (the twofold cost of sex) (Nei 1967; Lewontin 1971; Feldman 1972; Maynard Smith 1978; Michod and Levin 1988; Kondrashov 1993; Barton and Charlesworth 1998; West, Lively, and Read 1999; Otto and Lenormand 2002). These costs of sexual reproduction
and meiosis would seem to be prohibitively expensive, giving a fitness advantage to asexually reproducing populations.

Despite the costs associated with sexual reproduction, it is pervasive among eukaryotes (Bell 1982). Obligate asexual lineages are uncommon among eukaryotes and persist for relatively short periods of evolutionary time (White 1978; Bell 1982; Richards 1986). These observations beg the following question: Why should so many eukaryotes take these risks? In essence, this is the “paradox of sex” (Michod and Levin 1988; Kondrashov 1993; Barton and Charlesworth 1998; West, Lively, and Read 1999; Otto and Lenormand 2002). Some organisms reduce the costs of sex by alternating between sexual and asexual modes of reproduction (facultative sex) (Dacks and Roger 1999). However, the questions of why facultatively sexual organisms should bother with meiosis at all and why a great number of organisms rely exclusively upon sexual reproduction remain.

The question of why eukaryotes undergo the costly process of sexual reproduction is often answered with the benefits of genetic recombination to produce variable offspring, upon which natural selection can act (Fisher 1930; Muller 1932; Hill and Robertson 1966), especially in response to changing environments (Van Valen 1973). In fact, that recombination increases the efficacy of natural selection has been demonstrated convincingly in the laboratory with the fruit fly *Drosophila melanogaster* (Rice and Chippindale 2001) and the green alga *Chlamydomonas reinhardtii* (Kaltz and Bell 2002). In these organisms, populations that underwent multiple generations of sexual reproduction were more fit than asexually reproducing populations. Populations with genetic recombination may also be able to purge deleterious mutations more rapidly than exclusively asexual populations (Muller 1964; Kondrashov 1988). Although various hypotheses have been offered, revealing differences of opinion regarding the importance of selection for positive mutations and the elimination of deleterious mutations (*i.e.* the roles of natural selection and random genetic drift, respectively), there is little doubt that
the population level effects of genetic recombination provide sufficient selection for the long-term maintenance of meiosis (Otto and Gerstein 2006). More contentious, is the notion that there are short-term benefits of genetic recombination at the level of the individual.

Prior to the origin of meiosis, eukaryotes must have relied only upon asexual modes of reproduction (probably mitosis) (Szathmary and Smith 1995). However, once meiosis arose, these organisms (like many extant eukaryotes) were probably facultative sexual reproducers; sometimes they reproduced sexually and sometimes they reproduced asexually. Therefore, to study the selective advantages of meiosis that led to its origin, we can observe the conditions in which extant facultatively sexual organisms undergo meiosis (Michod, Bernstein, and Nedelcu 2008). It is well known that several unicellular eukaryotes that normally divide by mitosis will occasionally, when exposed to environmental stressors, divide by meiosis (Michod, Bernstein, and Nedelcu 2008). For example, both *Saccharomyces cerevisiae* (Herskowitz 1988) and *C. reinhardtii* (Sager and Granick 1954) will switch from mitotic to meiotic divisions in nutrient-poor media and *Volvox carteri* will undergo meiosis during heat shock (Kirk and Kirk 1986). Thus it is tempting to conclude that these organisms undergo meiosis to introduce variability to their offspring that may deal better with these stressful environments than their parents (Otto and Lenormand 2002; Otto and Gerstein 2006; Otto 2008). However, in the green alga *C. reinhardtii*, researchers found that calculated fitnesses of sexually reproducing populations were lower than those of asexually reproducing populations during the first generation (Colegrave, Kaltz, and Bell 2002). Only after subsequent episodes of sexual reproduction did fitnesses of the sexually reproducing populations exceed those of asexually reproducing populations (Colegrave, Kaltz, and Bell 2002). This negative, early effect of genetic recombination was also shown separately in *D. melanogaster* (Charlesworth and Barton 1996).
The observations that sexually reproducing organisms are initially less fit than asexually reproducing populations may be explained by a concept called recombination load (Charlesworth and Barton 1996; Colegrave, Kaltz, and Bell 2002). Simply put, the variation in the fitness of a population increases during the first sexually reproducing generation due to genetic recombination (Otto and Lenormand 2002). Previously linked genes become shuffled by genetic recombination, producing novel combinations of genes (Agrawal 2006). The first, most obvious, problem is that combinations of genes that have been selected for in a given environment are broken apart (Charlesworth and Barton 1996). However, many organisms benefit from inheriting combinations of genes that increase their fitnesses, while other organisms inherit deleterious combinations (Kouyos, Otto, and Bonhoeffer 2006). This is why genetic recombination increases the efficacy of natural selection; beneficial combinations of genes are selected for, increasing in populations, and deleterious combinations are purged from populations (Feldman, Christiansen, and Brooks 1980; Kondrashov 1984; Kondrashov 1988). The problem is that, initially, the increased fitness provided by the beneficial combinations of genes is outweighed by the decrease in fitness caused by the deleterious combinations of genes and the breaking apart of previously fit genomes (Otto and Lenormand 2002).

Theoretically, there are conditions (weak and negative epistasis) in which short-term advantages of genetic recombination could be realized and positively selected (Otto and Lenormand 2002). However, evidence that these conditions exist in nature is weak (Elena and Lenski 1997; Rice 2002; Bonhoeffer et al. 2004). For these reasons, it is unlikely that the production of variable offspring upon which natural selection acts could have provided the immediate selective benefits required for the origin of meiosis.

Some environmental conditions, such as exposure to metabolically or environmentally produced oxygen-containing compounds, can result in double-strand DNA breaks (i.e. oxidative stress) (Nedelcu and Michod 2003; Nedelcu, Marcu, and Michod 2004). While single-strand DNA damage can be repaired using the
complementary strand of DNA in a helix, double-strand DNA damage requires recombination with homologous chromosomes (Michod, Bernstein, and Nedelcu 2008). Thus meiosis may have arisen as an adaptation for damage DNA repair (Bernstein et al. 1984). This possibility is evidenced by S. pombe and V. carteri, both of which undergo meiosis in response to oxidative stress (Bernstein and Johns 1989; Nedelcu and Michod 2003). Furthermore, the connection between DNA damage and recombination is supported by the observations that mutations in recombination genes make cells sensitive to UV damage and exposure of cells to mutagens increases recombination rates (Bernstein and Bernstein 1991).

Of course, the hypothesis in which double-strand DNA damage repair supplies the primary benefit of meiosis relies upon the presence of at least a diploid number of homologous chromosomes. Indeed, diploid yeast cells are more resistant to DNA damage than haploid cells (Herskowitz 1988). However, unlike many diploid eukaryotes (e.g. metazoans (Schrader and Hughes-Schrader 1931)) whose haploid states are transient and associated exclusively with sexual reproduction (e.g. gametes), many eukaryotes, especially unicellular, experience longer haploid stages (Lewis 1985). The question, then, is why would organisms risk the integrity of their genomes by having extended haploid lifecycle stages? If DNA damage occurs during the haploid state and no other appropriate haploid cells are available for cell fusion and replication of chromosomes is not possible, due to the damage, the cells would seem to be in danger. The benefits of having haploid stages during the lifecycles of unicellular eukaryotes must outweigh the risks of their diminished capacities to repair double-strand DNA damage.

Another hypothesis posits that environmental stressors (e.g. oxidative stress and starvation) may provide an advantage to organisms with haploid-diploid ploidy cycles (Maynard Smith and Szathmary 1995). That is, during times of oxidative stress, diploidy may be beneficial for repair of double-strand DNA damage, while, during starvation, haploids may benefit from faster growth relative to diploids (Cleveland 1947; Szathmary
et al. 1990; Hurst and Nurse 1991). Haploid populations of S. cerevisiae do, indeed, have higher fitnesses than diploid populations in nutrient-limiting environments (Adams and Hansche 1974). It has been argued, however, that, since the ancestral eukaryote in which meiosis arose was probably phagotrophic, diploidy should have been favored during periods of starvation as diploids are larger and should, therefore, be able to engulf larger prey (Lewis 1985; Maynard Smith and Szathmary 1995). This logic leads us back to the conclusion that diploidy should be beneficial and haploidy should be rare, begging, once again, the question of why eukaryotes should exist as haploids at all. Furthermore, these hypotheses suffer from lack of supporting data (Kondrashov 1994). There is no evidence that ancestral eukaryotes would have been subjected to such alternating environments.

The fact that the majority of genetic mutations are deleterious (Lewontin 1974) would seem to support the notion that diploidy should be selectively advantageous, due to the presence of two copies of every gene (Otto and Goldstein 1992). If an allele carrying a deleterious mutation can be masked by the presence of a wildtype allele in a homologous chromosome (i.e. low dominance) then fitnesses of organisms may not be affected (Crow and Kimura 1965; Maynard Smith 1978; Charlesworth 1991; Kondrashov and Crow 1991; Perrot, Richerd, and Valero 1991). Mathematical models confirm this prediction, but only in cases of large genomes with relatively high rates of genetic recombination (Perrot, Richerd, and Valero 1991; Otto and Goldstein 1992). This is because increasing numbers of heterozygous loci result in greater ability to mask deleterious alleles (Otto and Goldstein 1992). However, the increased fitness of heterozygotes (heterosis) has a price: the maintenance of deleterious alleles in the population (mutation load) (Kondrashov 1994). In haploid cells with small genomes and low levels of genetic recombination, deleterious mutations are unlikely to persist as selection should efficiently purge them from the population (Scudo 1967). The ancestral eukaryotes in which meiosis arose most likely had few chromosomes (maybe only one) and relatively low rates of genetic recombination (Michod and Levin 1988). Therefore,
they would have benefited from maintaining a haploid number of chromosomes (Otto and Goldstein 1992).

In ameiotic haploid organisms, diploidization might occur either endogenously, due to errors that occur during mitosis (Cleveland 1947; Hurst and Nurse 1991), or exogenously, due to the fusion of two haploid cells (Cavalier-Smith 1975). This diploidization is likely to have reduced the fitnesses of the ameiotic ancestral eukaryotes in which meiosis arose (Otto and Goldstein 1992). Therefore, I argue that meiosis arose due to the selective benefits of retrieving haploid numbers of chromosomes after spontaneous diploidization. Furthermore, I propose also that meiosis could only have arisen in the presence of a strong constant selective pressure. Therefore, the cause of these diploidization events is important to consider. If mutations occurred that resulted in endogenous diploidization then they would simply have been selected against and purged from the population. Thus, cytological sources of diploidization would not have provided the constant selective pressures necessary for the origin of meiosis. However, diploidization that occurred because of repeated fusions of two haploid eukaryotes could have provided a constant selective force. Such fusions could be explained as an artifact of the ancestral eukaryotes’ phagocytic lifestyle. That is, when one cell attempted to engulf another eukaryotic cell (either by accident or cannibalism), their membranes could have occasionally fused. Unlike endogenous sources of diploidization, such an exogenous source could not easily be purged from populations. A method of identification could have evolved in order to avoid fusions but, if cannibalization was common, then chemical signaling may not have been selectively advantageous. Of course, the problem could have been remedied by abandoning their phagotrophic lifestyles but this solution would have required a switch to another food source, an endeavor that would most certainly have required many changes to the cells. Simply put, meiosis was a less costly way, evolutionarily speaking, to cope with constant and spontaneous diploidization than eliminating the cause (phagocytosis) altogether. This
scenario provides the constant, immediate selective benefits to individuals that would have been necessary for meiosis to arise.

**Meiosis arose from mitosis**

There are two main theories for the origin of meiotic genes: 1) meiotic genes arose directly from prokaryotic genes encoding products that were involved primarily in transformation (reviewed in (Bernstein and Bernstein 2010)); and 2) meiotic genes arose from genes encoding products that were involved primarily in mitosis (reviewed in (Wilkins and Holliday 2009)). Distinguishing between these possibilities is important to our understanding of the origin and evolution of meiosis.

Prokaryotic organisms are able to exchange genetic material via parasexual processes (*i.e.* conjugation (Lederberg and Tatum 1946), transformation (Griffith 1928; Avery, Macleod, and McCarty 1944), and transduction (Lederberg *et al.* 1951)), utilizing recombination enzymes that are also important for DNA damage repair (Maynard Smith and Szathmary 1995). In this regard, prokaryotic parasexual processes are analogous to sexual reproduction in eukaryotes. More specifically, recombination of prokaryotic genomes during transformation appears similar to meiotic recombination in eukaryotes (Bernstein and Bernstein 2010). Indeed, many genes necessary for bacterial transformation are orthologs of genes necessary for recombination of nonsister homologous chromosomes during meiosis (Marcon and Moens 2005). In addition, bacterial and eukaryotic orthologs may have similar functions during transformation and meiosis, respectively. For example, bacterial RecA, which stimulates DNA strand exchange during transformation, is orthologous to the eukaryotic gene encoding Dmc1, which stimulates interhomolog DNA strand exchange during meiosis in most eukaryotes. In addition, both transformation and meiosis can be induced by similar types of stress. Following these observations, it has been proposed that meiosis in eukaryotes arose immediately from eubacterial transformation (Bernstein and Bernstein 2010). This
hypothesis explains the evolution of sex as a continuous evolutionary process from bacteria to eukaryotes (Bernstein and Bernstein 2010).

Central to the argument that meiotic recombination in eukaryotes arose directly from eubacterial transformation is the observation that many genes were horizontally transferred from mitochondria (likely the result of the engulfment of eubacteria by early eukaryotes (Margulis 1970)) to the nuclear genome of eukaryotes (Gabaldon and Huynen 2003). Eubacterial recA homologs and eukaryotic recA homologs (Rad51 and Dmc1) share a high level of sequence similarity (0.20 and 0.23, respectively; Figure 3.14). Therefore, eukaryotic Rad51 and Dmc1 homologs may have arisen from recA orthologs that were transferred from eubacteria after their engulfment by eukaryotes (Lin et al. 2006). However, this model also predicts that Rad51 and Dmc1 should be more closely related to eubacterial recA genes than to archaebacterial RadA genes and distance analyses indicate that Rad51 and Dmc1 are most similar to archaebacterial RadA genes (0.43 and 0.45, respectively; Figure 3.14). Also, phylogenetic analyses indicate that Rad51 and Dmc1 share a more recent ancestor with archaebacterial RadA genes than with eubacterial recA genes (Stassen et al. 1997; Lin et al. 2006). In sum, these data indicate that eukaryotes inherited a recA homolog (RadA) vertically from archaebacteria and not horizontally from a eubacteria.

Since the first eukaryotes were certain to have been capable of nuclear divisions (i.e. mitosis), it is most likely that mitosis arose very early during eukaryotic evolution. The protoeukaryotes could have been mitotically dividing organisms that were also capable of bacteria-like transformation. Then meiosis could have arisen from transformation in the presence of mitosis. The crux of this argument is that meiotic recombination originating from bacterial transformation would have been a continuous evolutionary process (Bernstein and Bernstein 2010). That is, if mitosis arose first and there was neither bacteria-like transformation nor meiosis then a gap exists, during which eukaryotes did not undergo genetic recombination or sex (Bernstein and Bernstein 2010).
This argument assumes that nonsister homologous recombination did not occur during mitosis. However, crossing over has been shown in animal and fungal vegetative cells (mitotic crossing-over), albeit at much lower frequencies than meiotic crossing over (Cardoso et al.; Xu and Rubin 1993). Genetic recombination could have occurred if protoeukaryotes were capable of mitosis but neither transformation nor meiosis; there need not have been a “sex gap” during eukaryotic evolution if mitosis arose first.

Phylogenetic and distance analyses of the translated protein sequences of eukaryotic and prokaryotic recA homologs indicate that the eukaryotic Rad51 and Dmc1 genes are paralogs (Figure 3.14). That is, the genes encoding Rad51, which functions during both mitotic and meiotic DNA strand exchange reactions in model organisms, and Dmc1, which functions only during meiotic DNA strand exchange reactions in model organisms, arose by a single gene duplication event that occurred during eukaryotic evolution. There are three possible outcomes of gene duplication events: 1) the “extra” gene copy quickly degrades and its products (if any) do not function (nonfunctionalization) (Ohno 1970); 2) a division of labor occurs such that the two gene copies encode products that perform distinct complementary functions previously accomplished by the products a single gene (subfunctionalization) (Force et al. 1999); or, 3) one gene copy, free from the constraints of purifying selection, is free to mutate and its products then perform novel functions (neofunctionalization) (Ohno 1970). Since both the eukaryotic Rad51 and Dmc1 genes have been retained, either subfunctionalization or neofunctionalization of the genes occurred after they arose. Either the ancestral gene encoded products that functioned during both mitotic and meiotic DNA strand exchange and the duplication event yielded genes whose products divided these functions or the ancestral gene encoded products that functioned during only one reaction (mitotic or meiotic DNA strand exchange) and the gene duplication event resulted in the origin of a novel function. Put another way, either both mitotic and meiotic DNA strand exchange reactions were present at the time of the gene duplication or one arose from the other.
In addition to *Dmc1* and *Rad51*, there are several genes whose products are known to function only during meiosis in model organisms that are paralogs of genes encoding products that function during both mitosis and meiosis (Chapter 4). These genes encode products that are involved in several important (if not critical) events that occur during meiosis, including sister chromatid cohesion, dsDNA cutting, DNA strand exchange, and Holliday junction resolution. Thus the phenomenon is not restricted to genes encoding products involved in DNA strand exchange reactions but include many other genes necessary for successful completion of meiosis. Additionally, the distributions of these genes among diverse eukaryotes and phylogenetic analyses indicate that many of these duplications occurred prior to the common ancestor of all extant eukaryotes, making it possible that they all occurred at the same time or during a very small window during eukaryotic evolution. It is likely that either both mitosis and meiosis were present at the time of the gene duplication event(s) or one arose from the other.

Since the earliest eukaryotes were also most likely haploid, it seems unlikely that meiosis could have been the primary means of reproduction as it would have required two rounds of DNA synthesis or some combination of cell fusions and DNA synthesis to obtain the appropriate numbers of chromosomes. Although some single-celled organisms (*e.g.* *Saccharomyces cerevisiae*) have haploid stages of their lifecycles, during which they fuse with other cells to form diploid cells that may ultimately undergo meiosis, most of their nuclear divisions are mitotic (Herskowitz 1988). These observations and the greater cytological and genetical complexity of meiosis (Chapter 1) indicate that mitosis most likely arose first and meiosis is a derived process that arose later during eukaryotic evolution (Cavalier-Smith 1981b; Simchen and Hugerat 1993; Wilkins and Holliday 2009). In total, these data indicate that meiosis may have arisen from mitosis *de novo* as a result of one or more largescale gene duplication events. Below, I propose an
evolutionary model that includes a preadaptation that could have provided the selective benefits necessary for such a profound event to occur.

**A model for the evolution of meiotic DNA strand exchange genes**

The results obtained and the observations made during the studies performed in Chapters 2 through 4 revealed three major points regarding the evolution of meiotic DNA strand exchange genes: 1) While meiotic DNA strand exchange genes are often lost, *Rad51* appears to be present in all but one eukaryotic genome studied; 2) In *Saccharomyces cerevisiae*, *rad51* functional mutations or *Rad51* overexpression rescue(s) the null mutant phenotypes of other DNA strand exchange genes studied in Chapter 2 (Table 2.5); and, 3) *Rad51* and *Dmc1* may have overlapping functions in some organisms, such that one paralog may perform the activities of the other. These points have culminated in a model of meiotic DNA strand exchange gene evolution that explains the various complements of genes observed in different eukaryotes (Figure 5.1).

The presence of ten DNA strand exchange genes (*Rad52*, *Rad59*, *Rad51*, *Rad55*, *Rad57*, *Dmc1*, *Hop2*, *Mnd1*, *Rad54*, and *Rdh54*) in representative genomes of all the eukaryotic supergroups studied in Chapter 2 (Opisthokonta, Amoebozoa, Archaeplastida, Chromalveolata, and Excavata) indicate that they were likely to have been present in the last eukaryotic common ancestor. It is, therefore, feasible that the ancestor of eukaryotes was capable of meiotic DNA strand exchange and meiosis (Figures 2.1 and 5.1 – A). Also, given their distributions, two genes (*Rad59* and *Rad54*) may have arisen later during eukaryotic evolution (Figures 2.1 and 5.1 – B). So, although eukaryotes began with a core set of meiotic DNA strand exchange machinery, additional genes have since been added during the evolution of different eukaryotic lineages. However, the distributions of meiotic DNA strand exchange genes also indicate that frequent independent losses of important genes have occurred. These apparently contradictory observations beg the question: How can eukaryotes lose genes so important for meiotic
DNA strand exchange in model organisms and, by inference, in the last common ancestor of all extant eukaryotes?

Only one organism (*Giardia intestinalis*) is confirmed to be without a *Rad51* gene, while other genes have often been lost (Figures 2.1, 3.5, and 4.1). Hence, I hypothesized that a connection exists between the nearly ubiquitous presence of *Rad51* among eukaryotes and the frequent loss of other meiotic DNA strand exchange genes in independent eukaryotic lineages. As it happens, these observations can be explained by the following: 1) There are no known suppressors of *rad51* animal or fungal null mutant phenotypes; and 2) Overexpression or functional mutations of the *Rad51* gene suppresses *rad52, dmc1, rad55, rad57, hop2, mnd1, rad54, and rdh54* *Saccharomyces cerevisiae* null mutants (Milne and Weaver 1993; Klein 1997; Bishop *et al.* 1999; Krejci *et al.* 2002; Tsubouchi and Roeder 2003; Henry *et al.* 2006; Schild and Wiese 2009). I hypothesized that changes in *Rad51* expression or changes in its coding sequence may result in the relaxation of purifying selection on meiotic DNA strand exchange genes (such as the *Dmc1* gene) (Figure 5.1 – C). That is, when overexpressed or mutated *Rad51* products may ‘fill-in’ for missing components, performing their functions, or rendering the functions of other gene products altogether unnecessary. Such a dynamic would theoretically result in relaxation of the normally purifying selection that serves to preserve genes in populations of organisms. This relaxation of selection may then result in the loss of DNA strand exchange genes (Figure 5.1 – D). In addition, some meiotic DNA strand exchange components are known to interact only with a limited set of proteins (*e.g.* Hop2 and Mnd1 proteins only interact with Dmc1) (Chen *et al.* 2004; Henry *et al.* 2006). Therefore, the loss of the *Dmc1* gene may leave the *Hop2* and *Mnd1* genes vulnerable to loss (Figure 5.1 – E).

Finally, the complements of meiotic DNA strand exchange genes and the interactions of their products may provide a feedback loop in which subsequent mutations changing the expression of *Rad51* genes or creating beneficial functional mutants further
alter gene combinations (Figure 5.1 – F). Although lineage-specific genes and protein interactions are almost certainly affecting the complements of DNA strand exchange genes observed in different eukaryotes, this general model provides a eukaryote-wide hypothesis for understanding their evolution.

There is some preliminary evidence which suggests that one meiotic DNA strand exchange protein may perform the functions of another. As stated previously, *G. intestinalis* is the only organism known to lack a *Rad51* gene. However, during the study presented in Chapter 3, a search was conducted in the genome of a closely related diplomonad (*Spironucleus vortens*) with database mining and degenerate PCR, and no *Rad51* gene was found (data not shown). Interestingly, *G. intestinalis* does contain two copies of the *Dmc1* gene, both appearing to encode proteins that function during nuclear divisions in cysts (Poxleitner et al. 2008). It is possible that one copy of *Dmc1* may encode products that perform the functions normally completed by Rad51.

The Dmc1 proteins of *G. intestinalis* and *S. vortens* appear to have residues that are highly conserved among Rad51 protein sequences, with *G. intestinalis* Dmc1-A appearing slightly more ‘Rad51-like’ than Dmc1-B, especially at amino acid positions 331 and 332 (Figure 5.2). Residue D332 has been determined in eubacterial RecA and archaebacterial RadA proteins to bind DNA (Story, Weber, and Steitz 1992; Shin et al. 2003; Chen et al. 2007). The functions of these amino acids in Rad51 and Dmc1 proteins are unknown and it is possible that residues 331 and 332 are responsible for Rad51- or Dmc1-specific functions. Further studies will be needed to determine if these residues confer Rad51- or Dmc1-specific functions, but the possibility that one paralog may perform the functions of another in *G. intestinalis* is intriguing. Whether these sites are useful as diagnostic characters for Rad51- or Dm1-function (regardless of the paralog being observed) and if there is any functional significance of variations at these sites are questions worthy of scientific investigation. The point here is that the functions of
meiotic DNA strand exchange genes and the interactions between them may be more
dynamic than previously supposed.

**A model for the origin of meiosis**

The results of the scientific studies presented in this thesis have culminated in a
cohesive model for the origin of meiosis that I will now present. There are four main
events that distinguish meiosis from mitosis: 1) the pairing of homologous chromosomes
during meiosis I; 2) DNA strand exchange (recombination) between non-sister
homologous chromosomes; 3) sister-chromatid cohesion that persists through the first
meiotic division; and 4) the absence of DNA replication (S-phase) upon entering the
second meiotic division (Wilkins and Holliday 2009). Although there are other
differences between meiosis and mitosis (described in Chapter 1 and summarized in
Figure 1.3), understanding the possible origins of these four novel steps is considered
by many to be necessary for surmising the origin of meiosis itself (Kleckner 1996;
Villeneuve and Hillers 2001; Wilkins and Holliday 2009). That all (or almost all) of
these steps, each requiring its own set of specialized machinery, are necessary in
eukaryotes for successful completion of meiosis would seem to exclude any gradualist
explanations for the origin of meiosis. However, to suggest that these complex processes
could have arisen simultaneously seems to defy logic. For these reasons, the origin of
meiosis is considered one of the most formidable problems in evolutionary studies
(Maynard Smith 1978; Hamilton 1999; Wilkins and Holliday 2009). The following
evolutionary model, including mechanisms for the origins of the novel steps described
above, explains the origin of meiosis in a manner that is both feasible and testable.
Although I find much agreement with models presented by other researchers (especially
(Wilkins and Holliday 2009) and (Cavalier-Smith 2002d)), the timing of important events
and the mechanisms proposed here, responsible for the origins of the pairing of
homologous chromosomes, prolonged sister-chromatid cohesion, and meiotic
recombination, are, I think, unique.
I have shown that several genes whose products are known to function only during meiosis in model organisms must have been present in the common ancestor to all known extant eukaryotes (Chapters 3 and 4). Therefore, meiosis must have arisen in eukaryotes that existed prior to the last eukaryotic common ancestor. Such eukaryotes were probably phagotrophic, single-celled organisms with haploid numbers of chromosomes (possibly one chromosome) contained within nuclei (Figure 5.3 – A) (Cavalier-Smith 1975; Hurst and Nurse 1991; Cavalier-Smith 2002a; Wilkins and Holliday 2009). Like many extant eukaryotes, ancestral eukaryotes may have frequently engulfed prokaryotic organisms and, occasionally, other eukaryotes (Figure 5.3 – B) (Adl et al. 2005). It is possible that during phagocytosis, rather than one eukaryotic cell engulfing and digesting another eukaryotic cell, the cell membranes became fused, especially if the cells were genetically identical (Figure 5.3 – C). That the fusion of haploid eukaryotic cells may have been the precursor to meiosis is not a new concept, having been proposed on numerous occasions, probably due to its similarity to syngamy during the haploid-diploid cycles of many extant eukaryotes (Maynard Smith and Szathmary 1995). Following eukaryotic cell fusions, nuclear envelopes could have been followed quickly by nuclear fusions (Figure 5.3 – D). Again, such fusions may have been reminiscent of nuclear fusions observed during the sexual haploid-diploid lifecycles of extant eukaryotes (Wilkins and Holliday 2009).

The fusion of two haploid eukaryotic cells would, of course, have yielded a single diploid eukaryotic cell (Figure 5.3 – E). I believe that life would have proceeded somewhat normally for such newly formed diploid eukaryotes, until, that is, they attempted to undergo mitosis. The cells may have entered pre-mitotic S phase (DNA synthesis), copying each of the chromosomes present in the newly diploid nuclei. However, due to changes in gene expression levels and/or stoichiometry of protein and DNA molecules caused by the presence of diploid numbers of chromosomes, mitosis may not have proceeded normally. Recall, that during mitosis, Rad51 proteins function
during DNA strand exchange between sister chromatids (and, rarely, between non-sister homologous chromosomes) (Nishinaka et al. 1998; Krogh and Symington 2004; Lopez-Casamichana et al. 2008), while, during meiosis, Dmc1 proteins are necessary for DNA strand exchange between non-sister homologous chromosomes (Bishop et al. 1992; Bishop 1994; Bishop et al. 1999; Sehorn et al. 2004; Sauvageau et al. 2005) (Figure 1.4 and Table 2.5). Interhomolog DNA strand exchange during meiosis in Saccharomyces cerevisiae dmc1 null mutants is greatly reduced (Table 2.5) (Bishop 1994). However, overexpression of Rad51 significantly diminishes this phenotype, stimulating interhomolog DNA strand exchange (Bishop et al. 1999; Tsubouchi and Roeder 2003).

In Chapters 3 and 4 I demonstrated that both Rad51 and Dmc1 genes are present in representatives of all known eukaryotic supergroups and, so, are likely to have been present in the last common ancestor of all extant eukaryotes (Figures 3.2 – 3.12 and 4.1). However, mitosis most likely arose prior to the origin of meiosis (Cavalier-Smith 1981b; Simchen and Hugerat 1993). It is well known that Rad51 and Dmc1 are paralogs (genes arising from a common ancestral gene by duplication) (Stassen et al. 1997; Ramesh, Malik, and Logsdon 2005; Lin et al. 2006; Malik et al. 2008). Therefore, it is also likely that the ancestor of Rad51 and Dmc1 was most similar to Rad51 (I will call this ancestral gene Rad51'), encoding products that functioned during mitotic DNA strand exchange. Thus I propose that the change in the numbers of chromosomes from a haploid to a diploid number resulted in ‘overexpression’ of Rad51’ genes, increasing the numbers of Rad51’ proteins relative to the numbers of DNA molecules. In addition to DNA strand exchange between sister chromatids, this overexpression could have stimulated DNA strand exchange between non-sister homologs (Figure 5.3 – F).

Because pairing of non-sister homologous chromosomes is important for successful completion of the reductive division of meiosis, its origin is considered key to the origin of meiosis itself (Wilkins and Holliday 2009). However, without sustained pairing of sister-chromatids through the first division and monopolar attachment of
spindles to each chromosome, equational divisions are equally likely to occur (Watanabe and Nurse 1999; Toth et al. 2000; Yokobayashi, Yamamoto, and Watanabe 2003; Hauf and Watanabe 2004). Although meiosis could have evolved with the equational divisions occurring first and the reductive division occurring second, rather than the other way around, we can imagine a mechanism by which sister-chromatids could have stayed bound until the second division. The Rad21 gene encodes products that bind sister-chromatids during mitosis (Table 4.1) (Gruber, Haering, and Nasmyth 2003). During meiosis, Rec8, a paralog of Rad21 (Parisi et al. 1999), performs a similar function (Parisi et al. 1999; Watanabe and Nurse 1999; Toth et al. 2000; Gruber, Haering, and Nasmyth 2003; Yokobayashi, Yamamoto, and Watanabe 2003). Like Dmc1, Rec8 proteins are known to function only during meiosis in model organisms (Parisi et al. 1999; Watanabe and Nurse 1999; Toth et al. 2000; Yokobayashi, Yamamoto, and Watanabe 2003).

Again, because we expect that eukaryotes were capable of mitotic nuclear divisions prior to the origin of meiotic divisions, we also expect that the ancestor of Rad21 and Rec8 was most similar to Rad21; encoding products that functioned in a manner similar to Rad21 proteins in extant organisms (I will call this ancestral gene Rad21'). In S. cerevisiae, expression of Rad21 by a Rec8 promoter in null rec8 mutants results in meiosis-like monopolar attachment of microtubules to chromosomes, rather than the mitosis-like bipolar attachment normally seen in null rec8 mutants (Figure 1.3) (Toth et al. 2000). In Schizosaccharomyces pombe null rec8 mutants, Rad21 will relocate to centromeres (Yokobayashi, Yamamoto, and Watanabe 2003). Both experiments resulted in equational, rather than reductive, divisions during meiosis I (Figure 1.3) (Toth et al. 2000; Yokobayashi, Yamamoto, and Watanabe 2003). That is, although Rad21 attaches to centromeres and monopolar attachment of microtubules are rescued, the reductive divisions are not. However, I suggest that Rad21 overexpression in addition to Rad51 overexpression may result in retrieving the reductive division during meiosis I in yeast rec8/dmc1 double null mutants. Similarly, changes in the numbers of Rad21' proteins
relative to the numbers of DNA molecules in primitive eukaryotes (in the presence of increased numbers of Rad51’ proteins) could have resulted in the monopolar attachment of mitotic spindles to and extended sister-chromatid cohesion. Essentially, a meiosis I-like reductional division may have resulted.

Although I cannot find other examples in the current data, it is possible that additional genes acted similarly when overexpressed to achieve pairing of homologous chromosomes or suppression of DNA synthesis upon entering the second round of meiosis in ancestral eukaryotes. However, these steps may be otherwise explained (Wilkins and Holliday 2009). In *S. cerevisiae*, pairing of homologous chromosomes occurs during the $G_1$ lifecycle stage (prior to pre-meiotic DNA synthesis) (Weiner and Kleckner 1994). Pairing is interrupted during the pre-meiotic S-phase and restored during meiotic prophase I. Like the pre-meiotic pairing that occurs during $G_1$, meiotic pairing initially occurs in the absence of meiotic recombination and synaptonemal complex formation (Burgess, Kleckner, and Weiner 1999). Similarly, pairing of homologous chromosomes occurs in mitotically dividing cells *S. cerevisiae* during $G_1$, pairing is interrupted during pre-mitotic S-phase, and pairing is restored during $G_2$ (Burgess, Kleckner, and Weiner 1999). Pairing of non-sister homologous chromosomes in somatic cells has also been observed in Diptera and a variety of plants (Stack and Brown 1969). Therefore, a mechanism for homologous pairing may have existed in ancestral eukaryotes, prior to the origin of meiosis. In addition, the changes the interactions of mitotic spindles with homolog kinetochores could have contributed to prolonged sister-chromatid cohesion, through the first division (Wilkins and Holliday 2009).

The suppression of DNA synthesis after one (reductional) division, as cells enter into a second (equational) division, distinguishes meiosis from mitosis. In *Xenopus laevis, S. cerevisiae*, and *S. pombe*, pre-mitotic DNA synthesis is stimulated by a licensing reaction, in which a complex (composed of Mcm2-7) is loaded onto chromatin
by Origin of Replication Complexes (ORCs), Cdc6, and Cdt1 (Blow and Dutta 2005). In budding and fission yeasts, the activities of these Mcm complex ‘loaders’ are down-regulated by Cycling Dependent Kinases (CDKs) during S-phase and early mitosis, preventing DNA synthesis (Broek et al. 1991; Hayles et al. 1994; Dahmann, Diffley, and Nasmyth 1995; Diffley 1996; Piatti et al. 1996). In animals, upregulation of CDKs and inhibition of Cdt1 by geminin act together to suppress DNA synthesis during mitosis (Wohlschlegel et al. 2000; Tada et al. 2001; Lee et al. 2004). In animals and fission yeast, overexpression of Cdt1 and Cdc6 results in extensive re-replication during mitosis (Nishitani et al. 2000; Vaziri et al. 2003; Thomer et al. 2004; Arias and Walter 2005). However, in *S. cerevisiae* and *X. laevis*, significant re-replication of DNA occurs only when CDKs or geminin are inactivated (Nguyen, Co, and Li 2001; Li and Blow 2005). Overexpression of CDKs and/or geminin should, then, suppress DNA synthesis. It is possible that, in ancestral eukaryotes, changes in the numbers and/or stoichiometry of CDKs caused by the presence of diploid numbers of chromosomes in otherwise haploid cells resulted in suppression of DNA synthesis after the first division. At that point, the cells could have simply entered into a normal mitotic division, yielding haploid cells (Figure 5.3 – G).

Assuming the presence of small genomes, selection likely favored haploid cells over diploid cells early during eukaryotic evolution. Therefore, diploid eukaryotes arising from the fusion of two haploid eukaryotes should have been at a selective disadvantage. This selective force may have resulted in further refinement of the process described here. Eventually large-scale gene duplication events, possibly due to frequent unequal pairing of non-sister homologous chromosomes, yielded the many paralogous gene groups seen today (Figure 4.16). The presence of gene paralogs allowed for divisions of labor to occur (Ohno 1970; Ridley 2004) such that some genes would encode products that functioned predominantly during mitosis, but, on the occasions in which cell fusions occurred, the other genes could have functioned to reduce the numbers of
chromosomes. As genomes became recombined, the longer-term benefits of genetic recombination may have been realized and meiosis would have become even more refined, including enhanced mechanisms for cell fusions, dsDNA cuts, crossing-over, cross-over interference, and synaptonemal complex formation.

In summary, the data presented in this thesis supports the idea that meiosis arose from mitosis by large-scale gene duplication following a preadaptation that served to reduce increased numbers of chromosomes (from diploid to haploid) caused by erroneous eukaryotic cell-cell fusions.

**Future directions**

The model for the origin of meiosis presented in this thesis makes two major predictions: 1) During mitosis, overexpression of *Rad51* and *Rad21* genes should promote reductional divisions; and 2) Meiosis should be possible in the absence of meiosis-specific machinery. Both of these hypotheses can be tested using modern genetic techniques. However, it should be noted that only positive results would be informative, while negative results could arise from behaviors that evolved since meiosis arose (*e.g.* cell cycle checkpoints). Below, I suggest experiments designed to explore the models presented in this thesis and provide further insight into the origin and evolution of meiosis.

Whether reductional divisions can be produced during mitosis and whether the reductional division in *dmc1/rec8* null mutants can be rescued, by overexpression of *Rad51* and *Rad21* could be tested with *Saccharomyces cerevisiae*. As described previously, increasing *Rad51* copy number in *S. cerevisiae* null *dmc1* mutants rescues the null mutant phenotype (Bishop *et al.* 1999), characterized by defective recombination, accumulation of double-strand break recombination intermediates, failure to form normal synaptonemal complexes, and arrest late in meiotic prophase I (Bishop 1994) (Table 2.5). Recall, that there is no null mutant phenotype during mitosis (Bishop *et al.* 1992). Interestingly, overexpression of another gene involved in meiosis (*Rad54*) also rescues
null *dmc1* phenotypes (Tsubouchi and Roeder 2003), but it probably does so by removing double-strand break recombination intermediates (Petukhova, Stratton, and Sung 1998; Petukhova *et al.* 1999; Kiianitsa, Solinger, and Heyer 2002), rather than by performing the job of Dmc1, as the increased numbers of Rad51 are likely to do.

Although overexpressing *Rad51* rescues the null *dmc1* mutant phenotype in *S. cerevisiae*, it is unlikely to rescue a null *dmc1/rec8* mutant alone. The *S. cerevisiae* null *rec8* mutant phenotype includes the loss of monopolar attachment of microtubules and equational divisions during meiosis I (Parisi *et al.* 1999; Watanabe and Nurse 1999). The increased expression of *Rad21* in *rec8* mutants rescues monopolar attachment of microtubules but does not retrieve reductional divisions (Toth *et al.* 2000). I propose that *Rad51* and *Rad21* overexpression in UV radiated *S. cerevisiae* cells may rescue the reductional divisions in null *rec8* mutants by increasing the numbers of DNA strand exchange events and extending sister-chromatid cohesion through meiosis I. In addition, I propose that *Rad51* and *Rad21* overexpression should rescue reductional divisions in *dmc1/rec8* double mutants. These experiments would test the idea that overexpression of paralogs whose products function during both mitosis and meiosis (*e.g.* *Rad51* and *Rad21*) results in the completion of functions normally fulfilled by products that function only during meiosis (*e.g.* *Dmc1* and *Rec8*). To test whether overexpression of *Rad51* and *Rad21* is sufficient to explain the origin of meiosis, the same experiments could be performed during mitosis. If reductional divisions are observed, then the overexpression of genes is sufficient to explain the origin of meiosis.
Figure 5.1: General model for the evolution of DNA strand exchange genes. A. many DNA strand exchange genes arose very early during eukaryotic evolution, B. additional components may have arisen later by gene duplication, C. *Rad51* gene overexpression or mutation results in relaxed selection for retention of other components, D. some components may be lost, E. other components known to function only in complexes may be lost (*i.e.* Hop2/Mnd1 heterodimers are known only to function with Dmc1 proteins), and F. suites of genes result in further selection for *rad51* mutations. Components in bold indicate they are known to function only during meiosis in model organisms.
### Figure 5.2: Alignment of conserved Rad51 and Dmc1 residues

Percent identities determined from the alignments of 98 Rad51 and 51 Dmc1 protein sequences (Chapter 3) are indicated. Amino acid residues that are at least 75% conserved in Rad51 are highlighted in yellow, Dmc1 in green. The actual percent identities are provided for each paralog below. The *Saccharomyces cerevisiae* Rad51 amino acid residue is indicated above for reference. Representatives are provided here for each eukaryotic supergroup (Opisthokonta are labeled purple, Amoebozoa blue, Archeplastida green, Chromalveolata orange, Excavata brown, and Rhizaria eggplant). *Giardia intestinalis* and *Spironucleus vortens* Dmc1 protein sequence data are also provided for comparison.
A) Two genetically similar/identical nucleated cells (possibly sisters) with haploid numbers of linear chromosomes

B) One cell attempts to engulf the other

C) Fusion of cell membranes

D) Fusion of nuclear envelopes

E) Single nucleus with diploid number of chromosomes

F) Entry into "mitosis", DNA synthesis, Rad51 and Rad21 overexpression, DNA strand exchange, and pairing of homologous chromosomes

G) CDK overexpression, suppression of DNA synthesis, and entry into normal haploid mitosis

Figure 5.3: Model for mitotic ploidy reduction in ancestral eukaryotes.
REFERENCES


Huxley, J. 1942. Evolution, the modern synthesis. G. Allen & Unwin ltd, London,


Martin, W. 1999. A briefly argued case that mitochondria and plastids are descendants of endosymbionts, but that the nuclear compartment is not. Proceedings of the Royal Society of London Series B-Biological Sciences 266:1387-1395.


Octobre, G., A. Lorenz, J. Loidl, and J. Kohli. 2008. The Rad52 homologs Rad22 and rtil of Schizosaccharomyces pombe are not essential for meiotic interhomolog DNA strand exchange, but are required for meiotic intrachromosomal recombination and mating type-related DNA repair. Genetics 178:2399-2412.


Searfoss, A., T. E. Dever, and R. Wickner. 2001. Linking the 3’ poly(A) tail to the subunit joining step of translation initiation: Relations of Pab1p, eukaryotic translation initiation factor 5B (Fun12p), and Ski2p-Slh1p. Molecular and Cellular Biology 21:4900-4908.


