

The Pennsylvania State University

The Graduate School

Department of Biology

**DISPARATE MODES OF EVOLUTION IN CHLOROPLAST AND NUCLEAR
GENOMES**

A Thesis in

Biology

by

Liyang Cui

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

May 2006

The thesis of Liying Cui was reviewed and approved* by the following:

Claude dePamphilis
Associate Professor of Biology
Thesis Advisor
Chair of Committee

Hiroshi Akashi
Assistant Professor of Biology

Hong Ma
Professor of Biology

Webb Miller
Professor of Biology and Computer Science and Engineering

Bruce Lindsay
Willaman Professor of Statistics

Douglas Caverner
Professor of Biology
Head of the Department of Biology

*Signatures are on file in the Graduate School

ABSTRACT

This thesis explores evolution of gene order in chloroplasts, genome duplications in flowering plants, and the relationship between phylogenetic position, polyploidy, and gene numbers. Comparisons of chloroplast gene orders in inferred ancestral and extant green algae and land plants, together with simulations under a neutral evolution model, indicate that gene order is under strong selection possibly resulting from an advantage for co-transcription of neighboring genes. We also show that genome duplications occurred in many lineages of flowering plants, and that basal angiosperm species express more genes in flowers than do derived eudicot and monocot species. The project illustrates how novel statistical approaches, applied to the rapidly growing database of expressed gene and genome sequences, can reveal mechanisms that underlie processes of genome evolution.

TABLE OF CONTENTS

LIST OF FIGURES	vi
LIST OF TABLES	viii
ACKNOWLEDGEMENTS	ix
Chapter 1 Evolution on the Genome Level.....	1
References	5
Chapter 2 ChloroplastDB: The chloroplast genome database	7
Preface	7
Introduction	10
Data Management and Organization	13
The Chloroplast Genome Database Interface.....	17
How to use the Chloroplast Genome Database	19
Future Prospects	20
Acknowledgements	21
References	21
Chapter 3 Adaptive evolution of chloroplast genome structure inferred using a parametric bootstrap approach	24
Preface	24
Abstract	26
Background	27
Results.....	30
Discussion	41
Conclusions	47
Methods.....	48
Acknowledgements	52
References	52
Additional Files	57
Chapter 4 Widespread genome duplications throughout the history of flowering plants.....	66
Preface	66
Abstract	67
Introduction	68
Results.....	72
Discussion	86

Methods.....	92
Acknowledgements	97
References	101
Chapter 5 Large number of genes expressed in flowers of basal angiosperms: inference from EST data	107
Preface	107
Abstract	107
Introduction	108
Methods.....	111
Results.....	114
Discussion	118
Figures	124
Tables.....	127
References	128
Concluding Remarks	131
References	133
Appendix Inferring ancestral chloroplast genomes with duplications	134
Preface	134
Abstract.....	136
Introduction	137
Methods.....	139
Results.....	148
Discussion	150
Conclusions	152
Acknowledgements	152
References	152

LIST OF FIGURES

Figure 2-1: ChloroplastDB overview.	14
Figure 2-2: Examples of analysis using the ChloroplastDB web interface.	18
Figure 3-1: Extensive rearrangement in <i>Chlamydomonas reinhardtii</i> and <i>Chlorella vulgaris</i> cpDNAs.	32
Figure 3-2: The phylogeny of cpDNAs.	34
Figure 3-3: Comparison of sidedness and functional cluster indices in <i>C. reinhardtii</i> cpDNA to those of simulated genomes.....	37
Figure 3-4: Selected functional clusters from <i>C. reinhardtii</i> cpDNA.....	40
Figure 4-1: Effect of gene death rate and time of genome duplication on the K_s distribution for paralogs.....	74
Figure 4-2: K_s distribution from a sample of <i>Arabidopsis</i> unigenes and the diagnostic test according to the constant birth-death model (null model).....	76
Figure 4-3: K_s distributions of paralogs in selected angiosperm species, with overlying fitted density from mixture model analysis, suggesting paleopolyploidy in eudicots and monocots.....	78
Figure 4-4: K_s distributions of paralogs and orthologs among magnoliids, suggesting independent duplications and possibly shared genome duplication events in Laurales (<i>Persea</i>) and Magnoliales (<i>Liriodendron</i>).	81
Figure 4-5: K_s distributions suggest possible genome duplications in basal angiosperms, no evidence for genome duplication events in some gymnosperm species.	85
Figure 4-6: Phylogenetic summary of paleopolyploidy events estimated by the mixture model approach and their distribution in angiosperm and gymnosperm lineages.	91
Figure 5-1: Flowchart of data processing. The ISO error correction and gene number estimation are implemented in the software ESTstat.....	124
Figure 5-2: The relationship of sample size and estimate bias in simulations based on three microarray experiments on <i>Arabidopsis</i> flower/inflorescence tissue. ...	125
Figure 5-3: Distribution of transcripts by expression levels in different floral libraries..	126

Figure A-1 : The reference phylogeny of chloroplast genomes from land plants and green algae.....	140
Figure A-2 : Estimated gene contents for IR (in black) and SSC regions (in red).....	143
Figure A-3 : The inferred gene content evolution process from <i>Int2</i> to <i>nt</i> (left) and <i>Int1</i> to <i>no</i> (right). Only IR (in square brackets) and SSC (in red) are shown.	144
Figure A-4 : The revised gene content for each region (only IR and SSC are shown).....	145
Figure A-5 : False negative rates for GRAPPA-IR (solid line) and GRAPPA (dashed line) as a function of evolutionary rate r on the simulated datasets.	148
Figure A-6 : The best tree returned by GRPPA-IR. The topology is the same as the reference tree.	149
Figure A-7 : The best tree obtained by GRAPPA without duplicate genes and SC/IR boundaries, which is different from the reference tree.	149

LIST OF TABLES

Table 3-1: The Kolmogorov-Smirnov test of gene clustering by the functional category in cpDNAs	31
Table 4-1: Genome sizes and base chromosome numbers for the angiosperm and gymnosperm species in this study..	98
Table 4-2: Summary of EST data sets and paralogous pairs identified in this study.....	99
Table 4-3: Mixture model estimates for K_s distributions in each species.....	100
Table 5-1: Relationship of EST clustering stringency (percent identity) and the gene number (N) estimated from a flower bud cDNA library of <i>Arabidopsis thaliana</i>	127
Table 5-2: Estimated total transcripts from multiple tissues of <i>Arabidopsis thaliana</i>	127
Table 5-3: Number of genes detected and expressed in flowers of eudicot and monocot species.....	127
Table 5-4: Estimated number of genes expressed in basal angiosperm flowers.....	128

ACKNOWLEDGEMENTS

The author thanks the thesis committee, Drs. Claude W. dePamphilis, Hong Ma, Webb Miller, Bruce G. Lindsay, and Hiroshi Akashi, for their guidance and generous support during the research progress. I was supported by a Graduate Fellowship from the graduate school of Penn State University in 2000-01, and the Braddock Fellowship from the Department of Biology in 2000-01. During 2001-05, the work has been supported by NSF grants DBI0115684 and DEB 0120705 to Claude dePamphilis, Teaching Assistantships from the Department of Biology, and travel grants from Deep Gene Research Coordination Network and the Floral Genome Project.

Thanks also go to my mentors, collaborators and colleagues: Jim Leebens-Mack, Victor Albert, John Carlson, Dawn Field, Mike Frohlich, Jeff Doyle, Pamela Soltis, Douglas Soltis, Laura Zahn, Dan Ilut, Naomi Altman, Changxuan Mao, Ji-Ping Wang, Hongzhi Kong, Jijun Tang, Li-San Wang, Bernard M. E. Moret, Tandy Warnow, Kerr Wall, Jill R. Duarte, Ali Barakat, Yan Zhang, Barbara Bliss, Joel McNeal, Anthony Carrol, Anthony DiSante, Josh Marion, Alexander Richter, Izabela Makalowska, Steve Schaeffer, Robert K. Jansen, Jeffrey Boore, David Stern, Jason Lilly. Thanks to Lena Scheaffer, Sheila Plock, Yi Hu and Donglan Tian for their efforts in the lab, to Paula Ralph, Kathryn McClintock, and Bronnie McLaughlin for their all-around help in the department. The TA coordinators, Carla Hass, Denise Woodward, Mitch Price, and Dianne Burpee taught me the art of lab instruction. I am grateful for friends made at Penn State, especially through the Chinese Friendship Association and the Yan Xin Qigong Club: He Huang, Jingfen Zhu, Sara Hua Zhong, Xinyi Li, Nelson Hayes, Yifan Ma, Joe

Gyekis, Elody Gyekis, Loanne Snavely, Chung-chen Shelby Kuo, Yi Fang, Denise O'hara, and Peng Qiu, my advisors, teachers and friends when I was enrolled in the Statistics Department, including Mosuk Chow, KB Boomer, Andrea Piccinin, Laura Simon, Steven Thompson, William Harkness, Jia Li and John Dziak.

The academic pursuit is made possible with strong financial support and unflappable love from a closely-knit family: my parents and brothers in China, and my husband, Jingzhi Zhu.

Chapter 1

Evolution on the Genome Level

A genome refers to a single set of chromosomes with all the genetic information. Haploid organisms, such as bacteria and organelles in eukaryote cells, usually contain one to many circular DNA molecules, which may be referred to as one chromosome (1). Most eukaryote cells contain two or more sets of chromosomes and thus they are from diploid or polyploid species. Changes of the genome between related species can be studied on multiple levels. On the DNA sequence level, a genome contains protein-coding genes, RNA genes and non-coding DNA (2). Some regions may be single-copy and others duplicated through tandem duplication, segmental duplication and insertions of transposable elements. The entire genome may also be duplicated. On the gene expression (mRNA) level, different genes in the genomes are expressed according to the tissue and developmental stages. One can construct a phylogenetic tree of related organisms (or genes) based on DNA or protein sequences. With the advancement of genome sequencing, it is possible to construct whole-genome phylogenies using gene content and gene order data. On the other hand, evolutionary changes could be mapped onto a phylogenetic tree, leading to the inference of ancestral states (DNA, protein or other characters) and paths of evolution. In the following chapters I will focus on three aspects of genome evolution in a phylogenetic context, including gene order, genome duplication and expressed genes, illustrated by examples in algal and land plants, especially flowering plants. For each study, I have developed new methods and applied

them to data generated from whole genome and EST sequencing projects. These principles of genome evolution are not limited to the species studied. It is expected that methods extension would be applicable to other genomes, but some of the biological interpretations may be different.

First, I utilized complete chloroplast genomes to study gene order evolution and its consequences in a unicellular green alga. The chloroplast is the photosynthetic organelle in green plants and algae, and it contains a single circular genome, encoding 50 ~ 250 genes. In 1986, whole genome sequencing became a reality for plant organelles (3), 47 chloroplast genomes have been fully sequenced and publicly deposited. Studies of individual genes or proteins provide rich sequences for target regions of the genome, but whole genome sequences offer another dimension of information that is only available with complete genome data. Comparisons of gene content changes in chloroplast genomes generally indicate parallel gene losses during evolution of algae and land plant lineages (4). Gene order changes, due to inversions and transpositions, have been observed in chloroplast genomes (5). Phylogeny reconstruction of ancestral gene orders have been conducted on chloroplast data (6). Chloroplast genomes are also unique in that most duplicate genes are located in the inverted repeats (IR), two identical regions in opposite orientations flanked by single copy regions (3). Chapter 2 describes a database of fully-sequenced chloroplast genomes, which provides tools to quickly identify and extract orthologous genes. I developed and used the prototype version of this database to extract orthologous protein sequences and constructed reference phylogenies for the chloroplast genomes studied in the next two chapters. Chapter 3 illustrates that the gene order changes in the evolution of an algal chloroplast genome may be under

selection for co-transcription of neighboring genes, resulting in clustered regions that were potential transcription units. The method used to reconstruct ancestral chloroplast genomes is detailed in the Appendix.

Next, the evolutionary changes in plant nuclear genomes are explored in the context of genome duplication. It is possible to observe chromosome pairing under a microscope, so that long before DNA sequences were available, cytogenetic studies were conducted on many plant species. Many plants are found to be polyploid, and polyploid species are distributed among all major lineages of flowering plants (angiosperms), from basal angiosperms to derived monocots and eudicots (7). However, it was not until the whole genome sequence of the model flowering plant, *Arabidopsis thaliana*, became available that large segmental duplications within this genome were discovered (8). Further comparative studies suggest that perhaps this genome has undergone two or three rounds of duplication in the last 350 million years (9). The dates for the ancient whole genome duplications are still under debate (10). Due to the large size of nuclear genomes of most plant species, it is not yet practical to sequence their whole genomes. Expressed Sequence Tag (EST) sequencing provides an economical alternative to study the expressed fraction of a genome (11). Chapter 4 reports studies of genome duplication based on EST data. We collected EST sequences from basal angiosperm lineages, basal eudicots and basal monocots, and a few eudicot crop species. We further developed methods to test the presence of large-scale duplications based on the distribution of silent substitutions (K_s) for duplicated genes in these genomes. This method addressed the variation of EST sequencing quality and the errors in K_s estimates and tested observed distributions against an expected age distribution of duplicate genes based on a birth and

death process. Results from basal angiosperm lineages suggest that genome duplications may have been recurrent in the history of flowering plants, and perhaps correlated with the radiation of basal angiosperms, monocots and eudicots.

Although EST sequences cover only a fraction of genes expressed, the gene expression profiles estimated from ESTs of floral cDNA libraries provided the first insight into the diverse sets of genes expressed in flowers of diverse lineages (12). Chapter 5 describes a comparison of the estimated numbers of expressed genes between flowers of some basal angiosperm species and selected eudicot and monocot species. The method has been specifically developed for EST sequences to correct for over-estimation of rare transcript due to sequence clustering errors (13), and statistically rigorous estimation of the total number of transcript species in the underlying library (14). The estimate of total number of genes (or unique transcript types) in the floral transcriptome is compared to empirical observations from whole-genome microarrays and simulations. Here the estimated total number of expressed genes is much larger in basal lineages, which could indicate a systematic and repeated reduction in gene coding content in more derived lineages even in the face of pervasive genome duplication. The difference in the inferred gene number is not explained by technical reasons such as EST clustering errors. This apparent reduction of expressed genes during flowering plant evolution has not been reported previously. An alternative mechanism that could also give rise to the observed pattern would be relaxed gene transcription regulation differences in basal angiosperms compared to derived angiosperm lineages.

Lastly, I discuss the extension of genome evolution studies in animal lineages. The study of genome rearrangements could be applied to chromosome evolution in *Drosophila* (fruit fly) species and in eutherian mammals. The challenges lie in implementation of different operations (tandem duplications, chromosome fission and fusions) and the size of data which requires far more computation time. Genome duplications are also found in some vertebrates, including fishes and amphibians, Because of large number of alternative splicing forms, regulation of transcription and post-transcription modification may be more important in those genomes. Comparative studies in these organisms are needed to address that how different modes of evolution interact to shape the genomes.

References

1. Casjens, S. (1998) The diverse and dynamic structure of bacterial genomes. *Annu Rev Genet*, **32**, 339-377.
2. Raven, P.H. and Johnson, G.B. (2002), *Biology*. 6th ed. McGraw-Hill, pp. 386-387.
3. Sugiura, M. (1989) Organization and expression of the Nicotiana chloroplast genome. *Biotechnology*, **12**, 295-315.
4. Martin, W., Stoebe, B., Goremykin, V., Hapsmann, S., Hasegawa, M. and Kowallik, K.V. (1998) Gene transfer to the nucleus and the evolution of chloroplasts. *Nature*, **393**, 162-165.
5. Doyle, J.J., Davis, J.I., Soreng, R.J., Garvin, D. and Anderson, M.J. (1992) Chloroplast DNA inversions and the origin of the grass family (Poaceae). *Proc Natl Acad Sci U S A*, **89**, 7722-7726.
6. Bafna, V. and Pevzner, P.A. (1995) Sorting by Reversals - Genome Rearrangements in Plant Organelles and Evolutionary History of X-Chromosome. *Mol Biol Evol*, **12**, 239-246.
7. Stebbins, G.L. (1950) *Variation and evolution in plants*. Columbia University Press, New York.
8. Blanc, G., Barakat, A., Guyot, R., Cooke, R. and Delseny, M. (2000) Extensive duplication and reshuffling in the Arabidopsis genome. *Plant Cell*, **12**, 1093-1101.

9. Bowers, J.E., Chapman, B.A., Rong, J. and Paterson, A.H. (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature*, **422**, 433-438.
10. Raes, J., Vandepoele, K., Simillion, C., Saeys, Y. and Van de Peer, Y. (2003) Investigating ancient duplication events in the Arabidopsis genome. *J Struct Funct Genomics*, **3**, 117-129.
11. Boguski, M.S., Tolstoshev, C.M. and Bassett, D.E., Jr. (1994) Gene discovery in dbEST. *Science*, **265**, 1993-1994.
12. Albert, V.A., Soltis, D.E., Carlson, J.E., Farmerie, W.G., Wall, P.K., Ilut, D.C., Solow, T.M., Mueller, L.A., Landherr, L.L., Hu, Y. *et al.* (2005) Floral gene resources from basal angiosperms for comparative genomics research. *BMC Plant Biol*, **5**, 5.
13. Wang, J.P., Lindsay, B.G., Leebens-Mack, J., Cui, L., Wall, K., Miller, W.C. and dePamphilis, C.W. (2004) EST clustering error evaluation and correction. *Bioinformatics*, **20**, 2973-2984.
14. Wang, J.-P. and Lindsay, B.G. (2005) A penalized nonparametric maximum likelihood approach to species richness estimation. *J Am Stat Assoc*, **100**, 942-959.

Chapter 2

ChloroplastDB: The chloroplast genome database

Preface

This manuscript has been published in *Nucleic Acids Research*. The authors are Liying Cui, Narayana Veeraraghavan, Alexander Richter, Kerr Wall, Robert K. Jansen, Jim Leebens-Mack, Izabela Makalowska and Claude W. dePamphilis. LC and KW designed the original database; NV, AR and IM continue to develop the public database, RKJ, JLM and CWD contributed to the annotation and corrections. LC wrote the manuscript and all authors suggested changes or edits, and approved the manuscript.

ChloroplastDB: the chloroplast genome database

Liying Cui, Narayanan Veeraraghavan¹, Alexander Richter¹, Kerr Wall, Robert K. Jansen², Jim Leebens-Mack, Izabela Makalowska¹, Claude W. dePamphilis*

Department of Biology and Institute of Molecular Evolutionary Genetics, ¹Center for Computational Genomics, Huck Institutes of Life Sciences, The Pennsylvania State University, University Park, PA 16802, USA

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

*To whom correspondence should be addressed. Tel: +1 814 863 6412; Fax: +1 814 865 9131; Email: cwd3@psu.edu

Abstract

The Chloroplast Genome Database (ChloroplastDB) is an interactive, web-based database for fully sequenced plastid genomes, containing genomic, protein, DNA, and RNA sequences, gene locations, RNA-editing sites, putative protein families and alignments (<http://chloroplast.cbio.psu.edu/>). With recent technical advances, the rate of generating new organelle genomes has increased dramatically. However, the established ontology for chloroplast genes and gene features has not been uniformly applied to all chloroplast genomes available in the sequence databases. For example, annotations for some published genome sequences have not evolved with gene naming conventions. ChloroplastDB provides unified annotations, gene name search, BLAST and download functions for chloroplast encoded genes and genomic sequences. A user can retrieve all orthologous sequences and alignments with one search regardless of gene names in GenBank. This feature alone greatly facilitates comparative research on sequence evolution including phylogeny, changes in gene content, codon usage, gene structure, and post-transcriptional modifications such as RNA editing. Orthologous protein sets are classified by TribeMCL and each set is assigned a standard gene name. Over the next few years, as the number of sequenced chloroplast genomes increase rapidly, the tools available in ChloroplastDB will allow researchers to easily identify and compile target data for comparative analysis of chloroplast genes and genomes.

Introduction

As the site in the eukaryotic cell where photosynthesis takes place, chloroplasts are responsible for much of the world's primary productivity, making chloroplasts essential to the lives of plants and animals alike. The oxygen in our atmosphere, all agricultural commodities, and fossil fuels such as coal and oil are "products" of photosynthesis (1). Other important activities that occur in chloroplasts (and several types of non-photosynthetic plastids) include the production of starch (2), certain amino acids and lipids (3,4), some of the colorful pigments in flowers (5), and key aspects of sulfur and nitrogen metabolism (6,7).

All plastids studied to date contain their own distinct genomes derived from a cyanobacterial ancestor that was captured early in the evolution of the eukaryotic cell (8). Although much smaller than the nuclear genome, chloroplast genomes typically contain around 110-120 unique genes including conserved open reading frames annotated as *ycf* genes (hypothetical chloroplast open reading frame) (9). Additional possible coding regions are designated as ORFs (open reading frames). These are typically annotated with the number of amino acids encoded (e.g., ORF1995)(10). Some algae have retained a large chloroplast genome with more than 200 genes, whereas the plastid genomes from non-photosynthetic organisms may retain only a few dozen genes.

Chloroplast gene sequences have been widely used as genetic markers for plant and algal phylogenetic studies for nearly two decades (11,12). Whereas one or a few genes have been the focus of study most of this time (*rbcL*, *atpB*, *matK*; but see studies by Graham and Olmstead (13,14)), rapid growth in the number of chloroplast genome

sequences is now making it possible for a wide range of phylogenetic issues to be addressed with genome scale data sets (15,16). For population-level studies, polymorphic regions for targeted sequencing can be identified through comparison of complete genome sequences for exemplar taxa (17). Chloroplast genome sequences are also being used to address a wide range of evolutionary questions about changes in gene content and gene order (18), the dynamics of insertion and deletion events (19), intergenomic gene transfer (20), and photosynthetic evolution (21). The development of genetic transformation of chloroplasts has been very exciting (22) and the list of target species will increase as the locations and flanking sequences for intergenic spacer regions are identified from an expanding number of chloroplast genome sequences (23). Genome-scale functional analyses, including investigations of plastid transcriptomes and proteomes are also progressing rapidly (24).

Several bioinformatic resources provide information on organelle genomes, and tools specific for these genomes have been developed (25). The standard repository for full genome sequences, the GenBank, EMBL and DDBJ nucleotide sequence databases, currently includes 44 complete plastid genomes sequenced since 1986. The NCBI GenBank genome section lists entire organelle genome sequences submitted to the database and reviewed by NCBI staff (26). GOBASE (27) also maintains a list of sequenced organelle genomes. A standardized nomenclature for plastid-encoded protein genes is available through the UniProt database (<http://www.expasy.org/txt/plastid.txt>). A web-based annotation tool, DOGMA, provides a graphical user interface to annotate draft and finished organelle genomes based on sequence similarity searches and RNA secondary structure prediction (28). The program GRAPPA has been used for

phylogenetic analysis of chloroplast gene order changes (29). A plastid gene order database was developed with uniform gene names for 32 plastid genomes (30). In addition, 500 primers are now available for targeted PCR amplification of sequences from chloroplast genomes (<http://bfw.ac.at/200/1859.html>).

A prerequisite of future research is the accessibility of well-annotated, easy-to-use sequence data. However, several major limiting factors exist including flat file presentation of annotated organelle genomes, lack of standard data structure for relational databases, and non-uniform annotation quality. Errors in the annotation typically persist in the standard databases (for example, the gene *rp12* is annotated as *rp12* in the *Oryza sativa* chloroplast DNA). As a heritage of early annotations, gene name variants, unidentified *ycfs* and ORFs, and unannotated genes are present in some genomes. Given the ubiquity of phylogenetic studies based on plastid gene sequences, the flat file format makes search and data retrieval cumbersome.

RNA editing, a post-transcriptional process that alters specific RNA bases prior to translation, is common in the chloroplast genomes of some land plants (31). RNA editing can result in the creation of start codons and removal of stop codons, as well as making radical amino acid substitutions that would not be predicted based on the DNA sequence alone. Accurate genome annotation and inference of protein sequences often cannot be accomplished without knowledge of RNA editing sites (e.g., in the chloroplast DNA of *Anthoceros formosae*, *Adiantum capillus-veneris*, and *Zea mays*). The pace of new data generation and large-scale analyses demand a better integration of resources for chloroplast genome research. ChloroplastDB is a relational database with a user-friendly interface and tools to aid the analysis of chloroplast genome sequences.

Data Management and Organization

ChloroplastDB was designed using a MySQL database structure. The tables in the relational database store data related to the genes, nucleotide sequences, and annotated protein sequences for coding sequences (CDS). The databases contain fully sequenced plastid genomes obtained from the NCBI RefSeq section (http://www.ncbi.nlm.nih.gov/genomes/static/euk_o.html). All genes, including protein-coding genes, tRNA, rRNA, hypothetical open reading frames (*ycf*, ORF) were parsed and incorporated into the database (Figure 2-1).

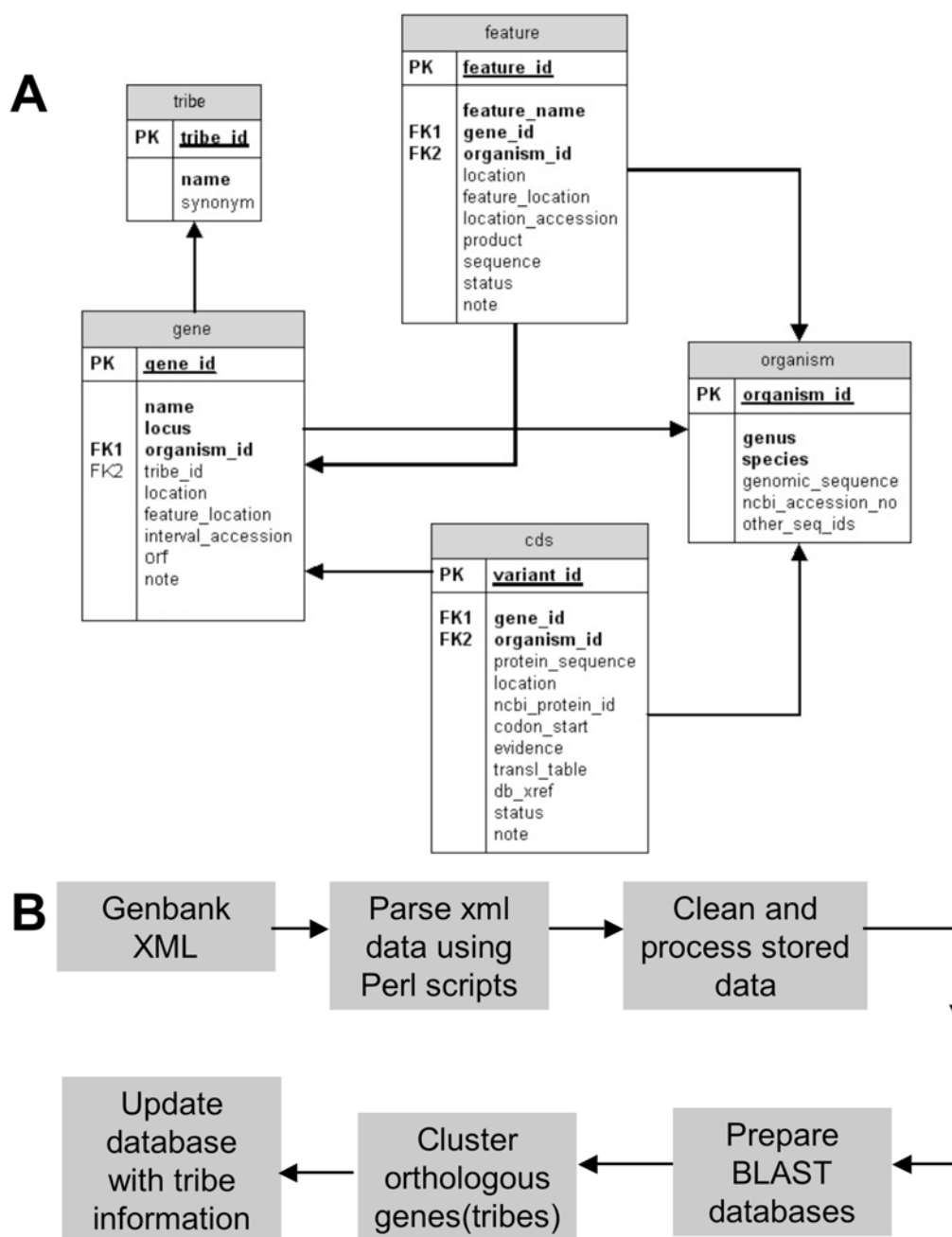


Figure 2-1: ChloroplastDB overview. (A) Database structure and relationship of data tables. PK: primary key. FK: foreign key. (B) Data flow and filtering steps to ensure the high quality of data stored in the database.

The standard process for extracting and storing data was carried out as follows:

1. A GenBank XML file containing a plastid genome sequence is downloaded. The XML format ensures better integrity of parsed data than GenBank flat files.
2. Using in-house XML parsers written in Perl, the XML data is extracted, filtered through quality control steps and formatted properly. The cleaned data are stored in the database in a form conducive to efficient data transactions.
3. Using the coordinates from the features (CDS, tRNA, rRNA, intron), the corresponding nucleotide sequence is extracted from the genome and stored in the database. The nucleotide coding sequences were used for sequence analysis using NCBI BLAST (v 2.2.10).
4. In a few instances when a parsed sequence lacks appropriate annotation, the GenBank records are updated with expert annotations after automatic processing of the XML file.
5. Three BLAST databases are created: one for the whole genome sequence from all organisms, a second for the annotated protein sequences of all organisms in the database, and a third for the generated nucleotide coding sequences from each organism.

When new sequences are added to the database, the proteins are sorted into potentially orthologous sets or "tribes" using tribeMCL (32). First, a sequence similarity profile is obtained by all-against-all BLAST on the protein sequences at a threshold of 1E-3. The BLAST output is fed to tribeMCL, which then generates a list of tribes representing protein families. This output is parsed and the tribes are updated in the database.

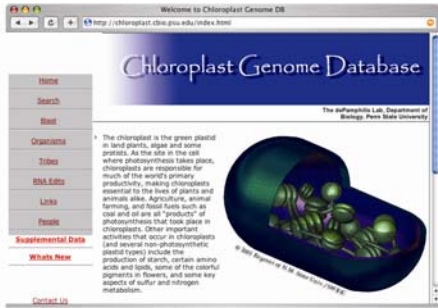
The quality control procedure is crucial in maintaining the integrity and accuracy of the extracted data. There appear to be some irregularities with the GenBank annotations. The genomic region spanning one gene and the gene features (CDS, tRNA, rRNA) share the same gene name. In case of overlapping and nested genes, the annotation for the second (or nested) gene could be attributed to the first gene, resulting in confusions. Also, there are instances where the gene names are not included in the feature description. We have avoided the problem by using the coordinates for each "gene" feature as the primary reference, and after that, the coordinates of other features are checked and assigned new gene names. For example, *rps12* is a trans-spliced gene containing three exons in the angiosperm chloroplast genome. The first exon is located about 30kb upstream of the second exon, and on the opposite coding strand. The initial parsed record for the gene contained many other genes nested in the intron region. After the filtering step, those nested genes were dissociated from the name "*rps12*" and assigned to their appropriate names. If no gene name was found, the feature was deemed to be an "orphan" and assigned a local name (starting with "lcl_anno").

When the GenBank annotation included RNA edited sites, both the location and type of edits were extracted from the record. The information was used to generate an edited pseudosequence that was stored with a list of edited sites. Just 38 genes with 541 annotated, experimental verified sites from *Anthoceros formosae* and *Physcomitrella patens* are included in the current GenBank annotations. RNA editing has been reported in other plants including tobacco, maize and *Adiantum* chloroplast DNA. Because the GenBank record does not contain a standard feature to store the RNA editing information, some edited sites were not reported while others were reported as exceptions

since the protein sequence did not match conceptual translation of the protein coding gene. To maintain quality and consistency of the data, we report annotated locations and the edited mRNA sequence.

The Chloroplast Genome Database Interface

Web user interfaces, developed using Perl CGI scripts, interact with the above mentioned data repository and provide users with basic sequence analysis tools (Figure 2-2). ChloroplastDB can be queried by gene name, and query results are returned in a table with links to individual genes. The BLAST similarity search was implemented for search against whole genome, extracted proteins, or extracted CDS. Sequences returned in BLAST searches can be exported to a fasta file. A user can also browse the list of organisms and all genes by specified subtypes (tRNA, rRNA, protein-coding) from each organism. The set of extracted genes vary from 56 in a non-photosynthetic parasitic plant, *Epifagus virginiana*, to 254 in the red alga *Porphyra purpurea*, including duplicate genes that are present in the genome. Tribes represent putatively orthologous genes across organisms, which can be downloaded to construct multiple sequence alignments. Together, these web interfaces provides a workbench for query, search, and sequence compilation and analysis. The various functions are seamlessly linked for a smooth user experience.

A 

B **Search Results**
- 43 Results
Click the gene name to view the feature and sequences

Gene Name	Organism
rbcL	<i>Acorus calamus</i>
rbcL	<i>Adiantum capillus-veneris</i>
rbcL	<i>Amborella trichopoda</i>
rbcL	<i>Anthoceros formosae</i>
rbcL	<i>Arabidopsis thaliana</i>
rbcL	<i>Atropa belladonna</i>
rbcL	<i>Calycanthus floridus var. glaucus</i>
rbcL	<i>Chaetosphaeridium globosum</i>
rbcL	<i>Chlamydomonas reinhardtii</i>
rbcL	<i>Chlorella vulgaris</i>

C

ID 326
Gene name rbcL
Locus AnfoCp036
Organism *Anthoceros formosae*
Location Start:End 72912:74339

Protein
HSPQETRAQVFRAGVDFKLTYYTTPDYETKTDILAAFRHTFPQVFPFEZAGAAVAESSTGTHTVW
TDCGLTGLDRTGCTDTEFVAGEEYQYIAYVATPLSEGGVYENHTSIVDQVFGELRALRLEEDLR
FPATERTQGFPIQVVERDKLNYGRFLDCTIKFKLGLSANNYGRVTECLRGLDFTKDDERNVNSQP

mRNA Edited
Edits at following locations:
70 C -> U
119 C -> U
121 U -> C
133 U -> C
302 C -> U

Edited sequence
AUGUCACCAAAAACGGAGCUGAAAGCAGGUGUGGAUUUAAAGCUGGUGUUAAAGAUUAGAUUAAACC
AUUUAUACCCUGAUACGAGACCAAGGAUACUGAUUUUUGGCAGCGUUCGCAUGACUCCUGAACACAGG

D
blastp 2.2.10 [Oct-19-2004]
Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller.
Query = 475 letters
Database = /var/apache/html/cv/chloroplast/chloroplast_publ [3661880451 letters]
Sequences producing significant alignments:

Hit	Definition	Bits	E-value
32	Adiantum capillus-veneris(rbcL_AdcaCp032). Genbank Acc:NC_0047266.1	945	0.0
3293	Psilotum nudum(rbcL_PsnCp032). Genbank Acc:NC_003386.1	914	0.0
208	Anthoceros formosae(rbcL_AnfoCp036). Genbank Acc:NC_004543.1	914	0.0
571	Chaetosphaeridium globosum(rbcL_CholCp045). Genbank Acc:NC_004115.1	912	0.0

E Select tribe member(s) to download its sequence in FASTA format

ID	Name	Organism
<input type="checkbox"/> 53	rbcL_AdcaCp032	<i>Adiantum capillus-veneris</i>
<input type="checkbox"/> 185	rbcL_AmtrCp030	<i>Amborella trichopoda</i>
<input type="checkbox"/> 326	rbcL_AnfoCp036	<i>Anthoceros formosae</i>
<input type="checkbox"/> 455	rbcL_ArthCp030	<i>Arabidopsis thaliana</i>
<input type="checkbox"/> 570	rbcL_AtbcCp030	<i>Atropa belladonna</i>
<input type="checkbox"/> 709	rbcL_CafCp031	<i>Calycanthus floridus var. glaucus</i>

F Select organism(s) to download its genomic sequence in FASTA format

	Organism Name	All Genes	Protein Coding Gene	tRNA	rRNA	Intron
1	Acorus calamus	135	85	38	8	25
2	Adiantum capillus-veneris	130	87	35	8	42
3	Amborella trichopoda	133	86	36	8	48
4	Anthoceros formosae	136	91	37	8	50
5	Arabidopsis thaliana	117	87	37	8	48

Figure 2-2: Examples of analysis using the ChloroplastDB web interface. (A)

Homepage of the database. (B) Search results for the gene "rbcL". (C) The gene view page linked to search result for each gene, including mRNA editing information. (D) BLAST results, with options to download sequences from the BLAST search. (E) Putative orthologous gene set listed as "Tribes". (F) The organism page presents a summary of genomes and extracted features in the database for batch download.

How to use the Chloroplast Genome Database

Gene Search

The basic query page allows a user to search for individual gene of interest. For example, search of "*rbcL*" returns all *rbcL* gene entries, in which two copies are from *Nephroselmis olivacea* since they are duplicated and located in the inverted repeats. Each gene is then linked to a gene view page. The gene view displays the gene name, organism, coordinates on the genome, exon boundaries, and DNA or protein sequences. Annotated RNA edits are highlighted with colors for easy identification.

BLAST

Customized BLAST searches against nucleotide coding sequences, proteins or genomic sequences allows a researcher to quickly identify novel sequences, to construct alignments and to annotate chloroplast genes. The returned entries are linked to respective gene view page or the whole genome record in NCBI. Selected list of entries can be exported as fasta sequences. A user can also run BLAST against the *Arabidopsis* or rice proteome to identify nuclear encoded homologs of chloroplast genes.

Tribes

An important feature of this database is pre-computed orthologous protein sets, which could be used for phylogenetic analysis. The tribes presented a uniform, automatic classification of chloroplast proteins using MCL clustering on all-by-all BLAST search results. With few exceptions, all other tribes represent orthologous gene sets, and a standard name is displayed for each tribe according to the UniProt list of plastid and cyanelle genes. The paralogous *psaA* and *psaB* are highly similar duplicate genes

(BLAST E-value < 1.0E-150) which are grouped together in a single tribe. In contrast, rapidly evolving *ycf1* genes are split into three tribes including seed plant, ferns plus bryophytes, and algal orthologs. Tribes also become a discovery tool for unannotated proteins. For example, ORF288 in hornwort, *Anthoceros formosae*, was sorted to the *cysT* tribe, together with an unannotated orthologous sequence from liverwort, *Marchantia polymorpha*. We also provide pre-computed protein and DNA alignments for each tribe.

Whole genome comparison and batch sequence retrieval

The plastid genomes from land plants, green algae, red algae and Apicomplexian represents a great range of diversity of the organelle genomes. The organism page presented direct link to the GenBank genome sequences, and ability to download genome sequences and genes by organisms. The user can use the downloaded sequence for organism specific analysis, or comparison for a specific type of sequences across organisms.

Future Prospects

Over the next few years, the growth of full organelle genome sequences will provide new opportunities for whole-genome comparative analyses. Cross-species investigations of genome-wide structural evolution, context-specific substitution processes (33), RNA editing, gene regulation and gene function will be more tractable for organelle genomes than much larger and more complex nuclear genomes. Organelle genomes may be an ideal proving ground for methods of analysis being developed to

understand genome and gene order evolution. The mission of ChloroplastDB is to promote comparative analyses of plastid genomes by addressing the community need for better, uniform annotation, quick sequence retrieval and homology search tools. The functionality of ChloroplastDB will grow as new genomes and alignments and other analyses are added, gene clustering techniques are improved, and visualization tools with gene order browsers are developed.

Acknowledgements

The authors thank Kevin Beckmann and Stacia Wyman for programming support, and Paul Wolf for providing RNA editing information. This work was supported through grants DBI 01-15684 to C.W.D. and DEB 01-20709 to R.K.J. and C.W.D., and Eberly College of Science, Pennsylvania State University.

References

1. Halliwell, B. (1978) The chloroplast at work. A review of modern developments in our understanding of chloroplast metabolism. *Prog Biophys Mol Biol*, **33**, 1-54.
2. Baroja-Fernandez, E., Munoz, F.J., Akazawa, T. and Pozueta-Romero, J. (2001) Reappraisal of the currently prevailing model of starch biosynthesis in photosynthetic tissues: a proposal involving the cytosolic production of ADP-glucose by sucrose synthase and occurrence of cyclic turnover of starch in the chloroplast. *Plant Cell Physiol*, **42**, 1311-1320.
3. Kirk, J.T. (1971) Chloroplast structure and biogenesis. *Annu Rev Biochem*, **40**, 161-196.
4. Vothknecht, U.C. and Westhoff, P. (2001) Biogenesis and origin of thylakoid membranes. *Biochim Biophys Acta*, **1541**, 91-101.
5. Reinbothe, S. and Reinbothe, C. (1996) The regulation of enzymes involved in chlorophyll biosynthesis. *Eur J Biochem*, **237**, 323-343.

6. Hatzfeld, Y., Lee, S., Lee, M., Leustek, T. and Saito, K. (2000) Functional characterization of a gene encoding a fourth ATP sulfurylase isoform from *Arabidopsis thaliana*. *Gene*, **248**, 51-58.
7. Schiltz, S., Gallardo, K., Huart, M., Negroni, L., Sommerer, N. and Burstin, J. (2004) Proteome reference maps of vegetative tissues in pea. An investigation of nitrogen mobilization from leaves during seed filling. *Plant Physiology*, **135**, 2241-2260.
8. Margulis, L. (1975) Symbiotic theory of the origin of eukaryotic organelles; criteria for proof. *Symp Soc Exp Biol*, 21-38.
9. Rochaix, J.D. (1997) Chloroplast reverse genetics: new insights into the function of plastid genes. *Trends in Plant Science*, **2**, 419-425.
10. Hallick, R.B. and Bairoch, A. (1994) Proposal for the naming of chloroplast genes. III. Nomenclature for open reading frames encoded in chloroplast genomes. *Plant Mol Biol Report*, **12**, S29-S30.
11. Cattolico, R.A. (1985) Chloroplast biosystematics: chloroplast DNA as a molecular probe. *Biosystems*, **18**, 299-306.
12. Clegg, M.T. (1993) Chloroplast gene sequences and the study of plant evolution. *Proc Natl Acad Sci U S A*, **90**, 363-367.
13. Graham, S.W. and Olmstead, R.G. (2000) Utility of 17 chloroplast genes for inferring the phylogeny of the basal angiosperms. *Am J Bot*, **87**, 1712-1730.
14. Graham, S.W. and Olmstead, R.G. (2000) Evolutionary significance of an unusual chloroplast DNA inversion found in two basal angiosperm lineages. *Curr Genet*, **37**, 183-188.
15. Leebens-Mack, J., Raubeson, L.A., Cui, L., Kuehl, J.V., Fourcade, M.H., Chumley, T.W., Boore, J.L., Jansen, R.K. and Depamphilis, C.W. (2005) Identifying the Basal Angiosperm Node in Chloroplast Genome Phylogenies: Sampling One's Way Out of the Felsenstein Zone. *Mol Biol Evol*, **22**, 1948-1963.
16. Goremykin, V.V., Holland, B., Hirsch-Ernst, K.I. and Hellwig, F.H. (2005) Analysis of *Acorus calamus* Chloroplast Genome and Its Phylogenetic Implications. *Mol Biol Evol*, **22**, 1813-1822.
17. Provan, J., Powell, W. and Hollingsworth, P.M. (2001) Chloroplast microsatellites: new tools for studies in plant ecology and evolution. *Trends Ecol Evol*, **16**, 142-147.
18. Gray, M.W. (1999) Evolution of organellar genomes. *Curr Opin Genet Dev*, **9**, 678-687.
19. Ingvarsson, P.K., Ribstein, S. and Taylor, D.R. (2003) Molecular evolution of insertions and deletion in the chloroplast genome of silene. *Mol Biol Evol*, **20**, 1737-1740.
20. Martin, W., Stoebe, B., Goremykin, V., Hapsmann, S., Hasegawa, M. and Kowallik, K.V. (1998) Gene transfer to the nucleus and the evolution of chloroplasts. *Nature*, **393**, 162-165.
21. Bungard, R.A. (2004) Photosynthetic evolution in parasitic plants: insight from the chloroplast genome. *Bioessays*, **26**, 235-247.
22. Daniell, H. and Chase, C.D. (eds.) (2004) *Molecular Biology and Biotechnology of Plant Organelles*. Kluwer Academic Publishers, Dordrecht.

23. Daniell, H. (1999) New tools for chloroplast genetic engineering. *Nat Biotechnol*, **17**, 855-856.
24. Rochaix, J.D. (2001) Posttranscriptional control of chloroplast gene expression. From RNA to photosynthetic complex. *Plant Physiol*, **125**, 142-144.
25. Jansen, R.K., Raubeson, L.A., Boore, J.L., dePamphilis, C.W., Chumley, T.W., Haberle, R.C., Wyman, S.K., Alverson, A.J., Peery, R., Herman, S.J. *et al.* (2005) Methods for obtaining and analyzing whole chloroplast genome sequences. *Methods Enzymol*, **395**, 348-384.
26. Pruitt, K.D., Tatusova, T. and Maglott, D.R. (2003) NCBI Reference Sequence project: update and current status. *Nucleic Acids Res*, **31**, 34-37.
27. O'Brien, E.A., Badidi, E., Barbasiewicz, A., deSousa, C., Lang, B.F. and Burger, G. (2003) GOBASE--a database of mitochondrial and chloroplast information. *Nucleic Acids Res*, **31**, 176-178.
28. Wyman, S.K., Jansen, R.K. and Boore, J.L. (2004) Automatic annotation of organellar genomes with DOGMA. *Bioinformatics*, **20**, 3252-3255.
29. Moret, B.M., Wang, L.S., Warnow, T. and Wyman, S.K. (2001) New approaches for reconstructing phylogenies from gene order data. *Bioinformatics*, **17 Suppl 1**, S165-173.
30. Kurihara, K. and Kunisawa, T. (2004) A gene order database of plastid genomes. *Data Science Journal*, **3**, 60-79.
31. Sugiura, M. (1995) The chloroplast genome. *Essays Biochem*, **30**, 49-57.
32. Enright, A.J., Kunin, V. and Ouzounis, C.A. (2003) Protein families and TRIBES in genome sequence space. *Nucleic Acids Res*, **31**, 4632-4638.
33. Morton, B.R. and Clegg, M.T. (1995) Neighboring base composition is strongly correlated with base substitution bias in a region of the chloroplast genome. *J Mol Evol*, **41**, 597-603.

Chapter 3

Adaptive evolution of chloroplast genome structure inferred using a parametric bootstrap approach

Preface

The manuscript has been published in *BMC Evolutionary Biology*, and was formatted for that journal. Authors for this original manuscript are Liying Cui, Jim Leebens-Mack, Li-San Wang, Jijun Tang, Linda Rymarquis, David B. Stern and Claude W. dePamphilis. LC conducted the analysis and drafted the manuscript . JLM and CWD conceived the study, helped with the analyses and contributed to the text. LSW contributed the code for the genome simulator. JT carried out the ancestral genome reconstruction. LR conducted the RNA analysis. DBS provided further experimental data review and revision of the draft.

Adaptive evolution of chloroplast genome structure inferred using a parametric bootstrap approach

Liyang Cui¹, Jim Leebens-Mack¹, Li-San Wang², Jijun Tang³, Linda Rymarquis⁴, David B. Stern⁴ and Claude W. dePamphilis^{1§}

¹Department of Biology, Institute of Molecular Evolutionary Genetics, and Huck Institutes of Life Sciences, The Pennsylvania State University, University Park, PA 16802, USA

²Department of Biology, University of Pennsylvania, Philadelphia, PA 19104, USA

³Department of Computer Science and Engineering, University of South Carolina, Columbia, SC 29208, USA

⁴Boyce Thompson Institute, Cornell University, Ithaca, NY 14853, USA

[§]Corresponding author

Email addresses:

LC: liying@psu.edu

JLM: jhl10@psu.edu

LSW: lswang@med.upenn.edu

JT: jtang@cse.sc.edu

LR: lar24@cornell.edu

DBS: ds28@cornell.edu

CWD: cwd3@psu.edu

Abstract

Background Genome rearrangements influence gene order and configuration of gene clusters in all genomes. Most chloroplast DNAs (cpDNAs) share a highly conserved gene content and with notable exceptions, a largely co-linear gene order. Conserved gene orders may reflect a slow intrinsic rate of neutral chromosomal rearrangements, or selective constraint. It is unknown to what extent observed changes in gene order are random or adaptive. We investigate the influence of natural selection on gene order in association with increased rate of chromosomal rearrangement. We use a novel parametric bootstrap approach to test if directional selection is responsible for the clustering of functionally related genes observed in the highly rearranged chloroplast genome of the unicellular green alga *Chlamydomonas reinhardtii* relative to ancestral chloroplast genomes.

Results Ancestral gene orders were inferred and then subjected to simulated rearrangement events under the random breakage model with varying ratios of inversions and transpositions. We found that adjacent chloroplast genes in *C. reinhardtii* were located on the same strand much more frequently than in simulated genomes that were generated under a random rearrangement processes (increased sidedness; $p < 0.0001$). In addition, functionally related genes were found to be more clustered than those evolved under random rearrangements ($p < 0.0001$). We reported evidence of co-transcription of neighboring genes, which may be responsible for the observed gene clusters in *C. reinhardtii* cpDNA.

Conclusions Simulations and experimental evidence suggest that both selective maintenance and directional selection for gene clusters are determinants of gene orders in chloroplasts.

Background

The influence of genotype on phenotype is not limited to the coding of peptides and functional RNAs by nucleotide sequences. An organism's phenotype is also affected by the chromosomal arrangement of genes and the interaction of gene products. Comparative genomics has revealed a number of gene clusters and chromosomal segments that have remained intact over hundreds of millions of years (1). Selection for clustering of co-transcribed genes has been hypothesized to influence gene order within bacterial and organelle genomes where gene clusters typically encode multiple components of a functional pathway (2). For example, the ribosomal proteins are encoded by similar operons in archaeobacteria, eubacteria and plastids (3). In eukaryotic genomes, co-expression of neighboring genes is significantly associated with the functional roles of the genes (such as housekeeping genes or genes in the same metabolic pathway) (4,5). One way that those genes become clustered is through tandem duplication, which usually results in functionally related genes being adjacent. On the other hand, unrelated genes may also be brought together through chromosome rearrangements (recombination, inversion and transposition). Unless selection is acting to maintain or promote gene clusters, gene orders in genomes subjected to rearrangements should become randomized with respect to function or co-expression profiles. Significant clustering has been inferred

using permutation tests that compare observed physical distances between pairs or blocks of co-expressed or functionally related genes to a null distribution constructed from randomized gene orders (4,5). However, this approach is limited since the evolutionary history of the genome was not considered. When comparing gene orders among related species, it is possible to estimate the ancestral genome and to simulate a null distribution for changes in gene order using a model. This evolutionary approach can be used to directly test the influence of selection on genome structure, that is, whether present-day genome structure has been influenced by directional selection for clustering of functionally related genes.

Small genomes, especially those of organelles and bacteria, are well suited to global comparisons of gene order. Like eukaryotic genomes, they are subject to structural changes such as inversion, transposition or translocation, as well as gene loss and (more rarely) gene gain. Chloroplast DNAs in most land plants share a highly conserved gene content and similar gene orders (6). Most cpDNAs include two identical regions in opposite orientations called the inverted repeat (IR), flanked by large single copy (LSC) and small single copy (SSC) regions. The IRs generally contain the bacterial-like rRNA gene clusters, and the genes involved in photosynthesis (photosystem I/II, cytochrome *b₆/f*, and ATP synthase) are arranged similarly in chloroplast and cyanobacterial genomes (2,3,7). Despite these well characterized patterns, it is unknown to what extent the conserved gene order reflects a slow intrinsic rate of neutral chromosomal rearrangements, rather than selection against alternative gene orders. A model of neutral rearrangement of gene order is required to formally test whether gene orders evolve under natural selection which prefers some gene arrangements over others.

Nadeau and Taylor first proposed a model for the neutral evolution of gene order in comparisons of mouse and human chromosomes (8). This “random breakage model” provides a null hypothesis for the evolution of gene order. It assumes a random distribution of break points and allows all possible gene orders without restrictions.

The random breakage model has been used to infer organismal phylogenies from gene order data (9). The gene order difference can be measured using the inversion distance, which is the minimal number of inversions necessary to transform one gene order to the other. Currently, the most accurate heuristic approach is implemented in the GRAPPA software (10), which is generally suitable for small taxon sets because the algorithm scores inversion medians for all nodes iteratively across all possible phylogenies. Algorithms for genomes with arbitrary rearrangements, a few deletions and duplications have been developed (11), and the capacity of GRAPPA can be scaled up with the disc-covering method (DCM) to potentially very large data sets (12).

The random breakage model does not account for recombination hotspots, which have been reported from human-mouse genome comparisons (13). However, at this time it may be difficult to model these hotspots, because the precise locations of reused breakpoints are unknown due to insufficient resolution of gene orders and potential errors in homology assessment given the scale of eukaryotic chromosomes (14). Thus, the fragile breakage model (13), as an alternative to the random breakage model, has not been well established.

Whereas gene order is generally conserved among land plant cpDNAs, very little synteny is observed between land plant cpDNAs and those of the chlorophytic green algae *Chlamydomonas reinhardtii* (15,16) and *Chlorella vulgaris* (17). The increased

rearrangement rate is associated with the invasion by a large number of short dispersed repeat elements, especially in *C. reinhardtii* (15) and *C. vulgaris* (16). The large number of rearrangements provides an excellent opportunity to test whether natural selection has preferred some changes in gene order more likely than others. Here we present novel statistics and parametric tests that lead us to reject the models of random rearrangement in favor of directional selection for clustering of functionally related genes in *C. reinhardtii* cpDNA. We also present experimental evidence that adaptive evolution of chloroplast genome structure could be driven by the advantage of concerted regulation conferred by polycistronic transcription.

Results

Functional clusters are not randomly distributed

We compared gene orders of representative cpDNAs from land plants, including tobacco (*Nicotiana tabacum*, [GenBank:NC_001879]) (18) and liverwort (*Marchantia polymorpha*, [GenBank:NC_001319]) (19), a charophytic green alga (*Chaetosphaeridium globosum* [GenBank:NC_004115]) (20), chlorophytic green algae (*Nephroselmis olivacea* [GenBank:NC_000927] (21), *C. vulgaris* [GenBank:NC_001865] (17), *C. reinhardtii* [GenBank:BK000554] (16)), a green flagellate alga with uncertain affinities (*Mesostigma viride* [GenBank:NC_002186]) (22), and the plastid of *Cyanophora paradoxa* [GenBank:NC_001675] (23) (**Figure 3-1**) (Additional file 1, Additional file 3). To measure the genome structure in terms of clustering by chromosome locations and by gene function, we defined "sided blocks" as contiguous genes coded on the same strand

of the plastid chromosome, and "functional clusters" as blocks of functionally related genes (see Methods). The randomness in the observed distribution of shared genes in chloroplast genomes with respect to gene function was assessed using a Kolmogorov-Smirnov test. The null hypothesis was rejected in all seven cpDNAs investigated (p-values in Table 3-1). While this test suggests some degree of functional clustering in all chloroplast genomes, it does not take into account the phylogenetic relationship of these organisms, so it is unclear whether functional clustering in chloroplast genomes is a legacy of genome organization in of an bacteria-like ancestor, or the product of selection on gene order in the face of genome rearrangements.

Table 3-1: The Kolmogorov-Smirnov test of gene clustering by the functional category in cpDNAs §

cpDNA	D_n (p-value)	Translation and transcription	Photosystem I and II	Electron Transport	ATP synthase
<i>Chlorella</i>	0.214(.6418)	0.488(.0066)	0.750(.0000)	0.833(.0000)	
<i>Chlamydomonas</i>	0.198(.6866)	0.473(.0060)	0.780(.0000)	0.769(.0000)	
<i>Nephroselmis</i>	0.209(.6207)	0.484(.0046)	0.703(.0000)	0.846(.0000)	
<i>Mesostigma</i>	0.275(.2786)	0.549(.0008)	0.769(.0000)	0.846(.0000)	
<i>Chaetosphaeridium</i>	0.242(.4388)	0.484(.0046)	0.714(.0000)	0.846(.0000)	
<i>Marchantia</i>	0.341(.0986)	0.473(.0060)	0.714(.0000)	0.846(.0000)	
<i>Nicotiana</i>	0.264(.3283)	0.473(.0060)	0.769(.0000)	0.846(.0000)	

§The test statistic D_n measures whether the distribution of functionally related genes is random in gene clusters. Total 85 shared genes between seven cpDNAs were included

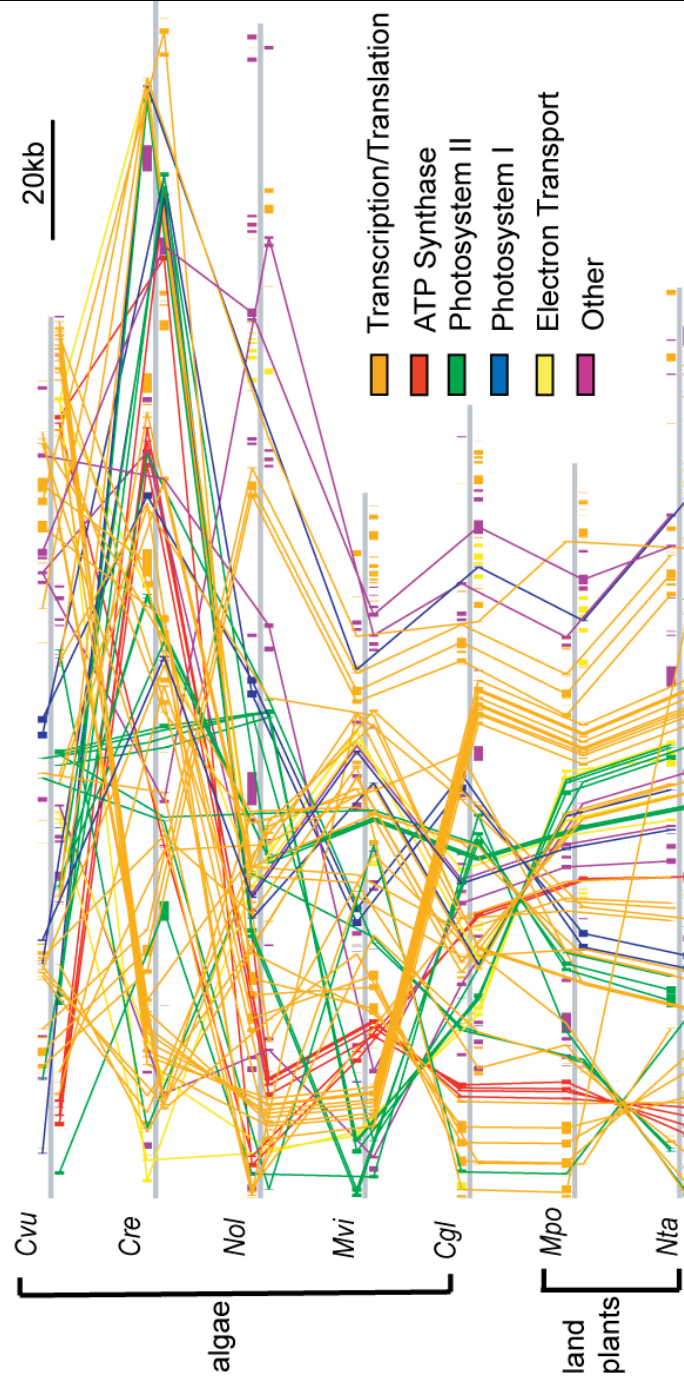


Figure 3-1: Extensive rearrangement in *Chlamydomonas reinhardtii* and *Chlorella vulgaris* cpDNAs.

(Continued) Representative cpDNAs from land plants and green algae are arranged to reflect their phylogenetic relationships. The scale bar indicates 20 kb. Each genome is

linearized and drawn as a grey bar. Genes are drawn as colored rectangles and with those encoded on the positive strand above the genome bar. Colored lines connect the homologs included in this study and the functional category is shown by specific colors. Abbreviations: Cre, *Chlamydomonas reinhardtii*, Cvu, *Chlorella vulgaris*, Nol, *Nephroselmis olivacea*, Mvi, *Mesostigma viride*, Cgl, *Chaetosphaeridium globosum*, Mpo, *Marchantia polymorpha*, Nta, *Nicotiana tabacum*.

Extensive rearrangements from the ancestral chloroplast genome to *C.*

reinhardtii

In order to investigate evolutionary changes of gene order, we constructed a phylogeny of seven representative cpDNAs and rooted with the sequence of *C. paradoxa* (23). Maximum parsimony, neighbor joining and maximum likelihood analyses of an alignment of 50 concatenated protein sequences including a total of 19,836 aligned sites (Additional file 2), all yielded identical fully resolved topologies with high bootstrap support (**Figure 3-2A**). *Mesostigma* was placed as a basal charophyte lineage in one previous analysis (24). The unrooted phylogeny of seven cpDNAs (**Figure 3-2B**) is congruent with the alternative placement of *Mesostigma* either as a basal charophyte (24) or basal to both charophyte and chlorophyte lineages (22). This tree was used as the reference phylogeny for gene order inference.

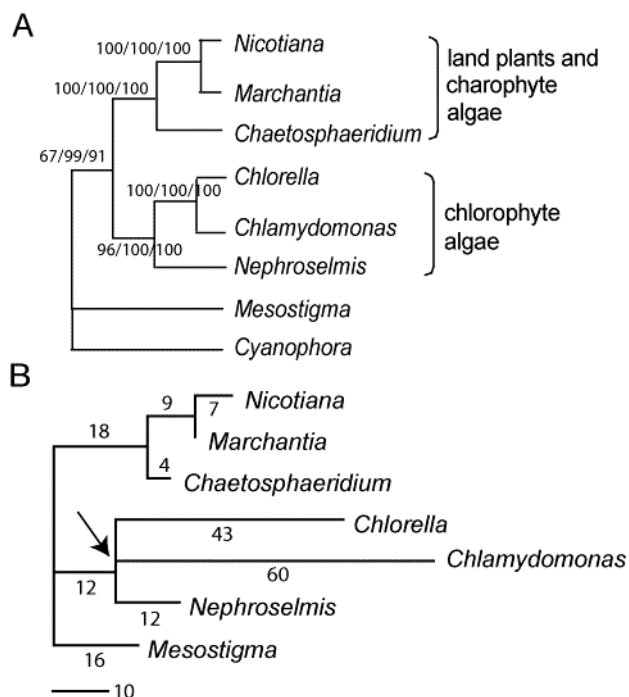


Figure 3-2: The phylogeny of cpDNAs. (A) The cpDNA phylogeny based on analysis of 50 concatenated proteins. The phylogeny includes major green plant and algal lineages and the outgroup Cyanophora. The bootstrap support values from maximum parsimony/neighbor joining/maximum likelihood analyses are labeled near each node. (B) Estimated inversion distances considering 85 common genes on the cpDNA phylogeny. There is an increase of rearrangements on branches leading to *C. reinhardtii* and *C. vulgaris*, from a common ancestor indicated by an arrow.

We scored the gene orders of 85 genes shared in the seven genomes (Gene orders in the additional file 3). Then we used modified versions of GRAPPA (11,25) to compute the inversion distance between ancestral nodes and each terminal node (Figure 3-2B; see Methods). The branches leading to two chlorophytic green algae, *C. reinhardtii* and *C. vulgaris*, are much longer than the branches leading to the other taxa. Many more steps were inferred on the *C. reinhardtii* lineage relative to the *C. vulgaris* lineage. Gene duplications or deletions were mapped before scoring the ancestral genomes with

inversions and not counted as rearrangements. IRs were present in all inferred ancestral nodes and one copy was lost in *C. vulgaris*. Ancestral gene orders were scored on all the phylogenies using a two-step approach (see Methods). Due to the computational time limit (the full search for ancestral gene orders may require months), we stopped scoring all possible ancestral gene orders with the data set after 25 days and took the best scored ancestral gene orders at that time (Additional file 4).

The cpDNAs of two land plants, *N. tabacum* and *M. polymorpha*, were separated by 7 inversions estimated based on the data set. One large inversion (~ 30kb) in the LSC region has long been recognized to separate the two genomes (26). Additional gene order rearrangements are directly observable through comparison of gene order files for the two species (see additional file 5 for the sequences of gene order rearrangements). Using GRAPPA, all rearrangements were inferred as inversions, but the total number of inversion events estimated by GRAPPA may be greater than the true (but unknown) mixture of inversions and transpositions because one transposition could result in the same change in gene order as two or three inversions.

Increased order in the genome structure after rearrangements

Two genomic structural characteristics were measured: the propensity of adjacent genes to be clustered on the same strand (using the sidedness index C_s) and the clustering of functionally related genes (using the functional cluster index, C_f) (see Methods). Both indices were calculated for the inferred ancestral gene orders and extant daughter lineages. Among land plants and charophytes, the inferred sidedness among ancestral genomes was similar to extant lineages, however, among the chlorophytes an opposite trend was observed, especially in the *C. reinhardtii* lineage (Additional file 3). The large

number of rearrangements in the *C. reinhardtii* cpDNA lineage resulted in dramatically increased sidedness relative to the inferred most recent common ancestor of *C.*

reinhardtii and *C. vulgaris* (C_s ancestor = 0.6966, C_s observed = 0.8710; **Figure 3-3A**).

A small increase of C_s was found in the *N. olivacea* lineage and there was almost no change in the lineage leading to *C. vulgaris*. A large increase was also observed in the functional clustering index, C_f , for *C. reinhardtii* (C_f ancestor = 0.01674, C_f observed = 0.03397; **Figure 3-3B**), whereas the trend was less profound in other lineages (Additional file 3). Thus, even if the ancestral genome already had a "sided" structure, sidedness increased with genome rearrangements as the *C. reinhardtii* chloroplast genome evolved. The inferred increase in sidedness and functional clustering in the face of the large number of rearrangements on the lineage leading to *C. reinhardtii* might be adaptive if such increases were not expected under random rearrangements.

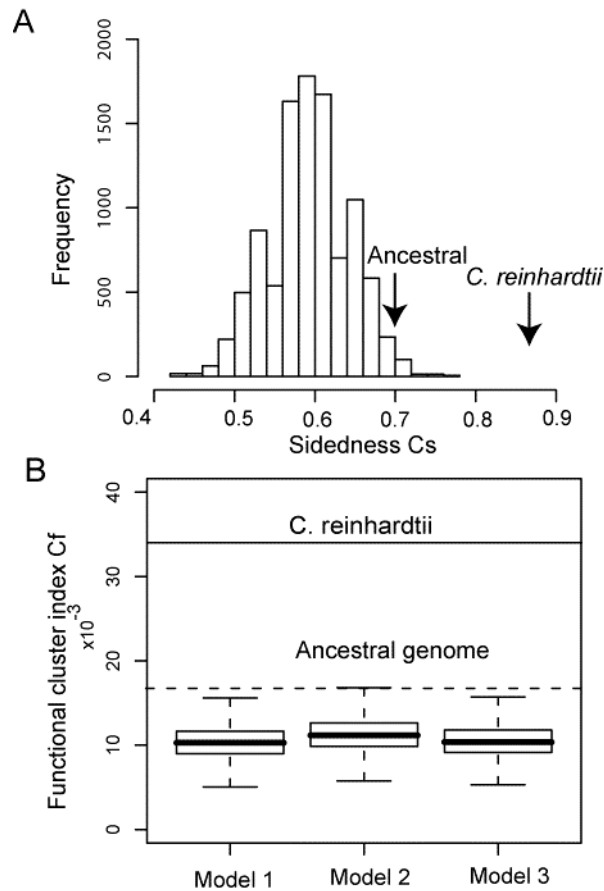


Figure 3-3: Comparison of sidedness and functional cluster indices in *C. reinhardtii* cpDNA to those of simulated genomes. (A) The sidedness index C_s observed in *C. reinhardtii* (indicated by an arrow) is significantly larger than C_s of gene orders simulated under the random breakage model (inversion only) and the estimated ancestral genome indicated in Figure 3-2B. (B) The functional cluster index C_f for *C. reinhardtii* (indicated by a solid horizontal line) is greater than that for the inferred ancestral genome (dashed line), in contrast to the decrease predicted by three sets of simulations under the random breakage model. Models 1, 2 and 3 specified the inversion/transposition ratios to be 1:0, 10:1 and 1:1, respectively, in simulations with 10,000 replicates. The box section of the box plot indicates the first quartile, median and the third quartile of the distribution.

To test the null hypothesis that the changes in C_s and C_f were consequence of random genome rearrangements rather than a consequence of directional selection (H_0 : random rearrangement; H_A : constraints in rearrangements), we simulated random

rearrangements starting with the inferred ancestral genome along the branch leading to *C. reinhardtii*. Although inversions are the most abundant type of rearrangement in cpDNAs (27), we also considered the contribution of transpositions and simulated genomes under three inversion to transposition ratios, while the total number of rearrangements was fixed according to the branch length inferred using GRAPPA (Figure 3-2B). Three simulations with 10,000 replicates were conducted with inversion to transposition ratios of 1:0, 10:1 and 1:1 under the random breakage model. The mean C_s values for the three sets were 0.5929, 0.6084 and 0.5948, respectively, and the 95% confidence intervals were (0.5056, 0.6742), (0.5281, 0.6854) and (0.5169, 0.6742), respectively. All datasets simulated under the random breakage model showed a significant decrease of sidedness from the ancestral level ($p < 0.0001$). In contrast, the C_s value calculated for *C. reinhardtii* increased significantly to 0.8710 (Figure 3-3A), greatly exceeding the sidedness that would be expected in a genome that had undergone this much evolutionary change relative to its ancestor. Simulations using inferred ancestral genomes for land plant lineages (e.g., *N. tabacum*) also strongly reject the null hypothesis of random rearrangements (results not shown).

Given the large number of rearrangements observed in *C. reinhardtii*, C_f was also predicted to decrease significantly under the random breakage model, but C_f did not decrease in the cpDNA of *C. reinhardtii* (Figure 3-3B). The simulations with three models described above (all inversions, a small fraction of transpositions, and equal inversions and transpositions) all yielded a large decrease in clustering as expected (the observed C_f in *C. reinhardtii* was 0.03397, and the 95% confidence intervals for C_f in simulated genomes were 0.00744-0.01401, 0.00812-0.014299 and 0.0750-0.01418,

respectively). When transposition was included in simulations, decreases of C_f were on a similar scale to the inversion-only simulations, and no increase of C_f occurred in the simulated data sets. Taken together, these results indicate that the remarkable increase in sidedness and functional clustering observed in *C. reinhardtii* cpDNA has not been the outcome of solely chance events. Instead, the strong deviation from the range of outcomes expected under various random breakage models implies that the genome structure is the outcome of a directional selective process.

The increased level of organization in *C. reinhardtii* cpDNA was associated with both maintenance of ancestral clusters and growth of new clusters. There were six conserved blocks containing 19 of the 85 genes shared between the *C. reinhardtii* and the *C. vulgaris* cpDNAs. These blocks include concentrations of genes from a single functional category, such as ribosomal proteins (*rpl23-rpl12-rps19*, *rpl16-rpl14-rps8*), photosystem II (*psbL-psbF*, *psbB-psbT-psbN-psbH*), translation apparatus (*rrn16- trnI-GAU - trnA-UGC -rrn23-rrn5*), and ATP synthase subunits (*atpF-atpH*). Moreover, a number of small clusters of functionally related genes inferred in the ancestral genome were brought together in *C. reinhardtii* ("rearranged clusters" in **Figure 3-4B**). These include transcription/translation genes (*trnH-M-F*; *rpl/rps*; *rps3-rpoC2*), electron transport genes (*petA-petD*), and photosynthetic genes (*psbD-psaA exon 2-psbJ*) (**Figure 3-4B**). The new clusters contributed to the increase of C_f in the *C. reinhardtii* chloroplast genome.

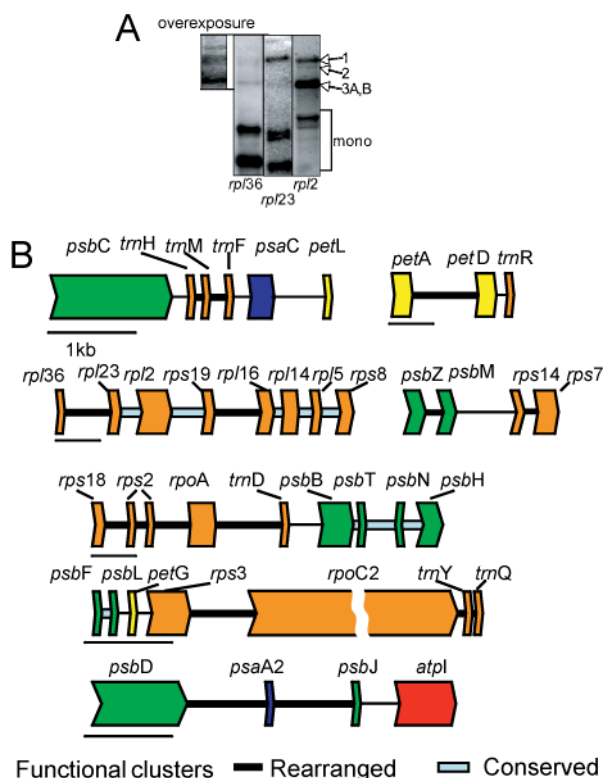


Figure 3-4: Selected functional clusters from *C. reinhardtii* cpDNA. (A) Evidence for co-transcription of the genes *rpl36-rpl23-rpl2-rps19*. The gel was loaded with total RNA from wild-type cells, and shows new evidence for co-transcription (see text). The top left lane is an over-exposure of the *rpl36* gel. Transcripts 1 and 2 (3.5 and 3.3 kb) are tricistronic *rpl36-rpl23-rpl2*, transcript 3A (2.5 kb) is *rpl36-rpl23*, and transcript 3B (2.5 kb) is probably *rpl2-rps19*. Single gene transcripts are labeled “mono”. (B) Rearranged functional clusters, which were absent from the inferred common ancestor of *C. reinhardtii* and *C. vulgaris*, were identified in *C. reinhardtii* (genes connected by bold black lines). Cyan lines connect conserved clusters retained from the ancestor cpDNA. The genes are displayed in the coding direction, and from top to bottom relative to their order in the genome. The exception is *psbN*, which is on the opposite strand relative to other genes shown (*psbT-B-N-H*). A scale bar of 1 kb is shown below and at the left of each gene cluster.

Coordinated expression of genes in functional clusters

Co-transcription of several clusters shown in Figure 3-4B has been previously documented, including *psbD-psaA* exon 2-*psbJ-atpI* (28), *psbF-psbL* (29), *petA-petD* (30), and *psbM-psbZ* (31). Co-transcription of *rpl* and *rps* genes have been found in land

plant chloroplasts (32). We documented co-transcription for an additional novel functional cluster, shown in Figure 3-4A. Using RNA gel blots, tricistronic transcripts of *rpl36-rpl23-rpl2* and possibly dicistronic *rpl2-rps19* species could be detected. Taken together, it appears that the clusters of functionally related genes observed in the *C. reinhardtii* cpDNA may be frequently co-transcribed.

Discussion

By reconstructing the ancestral gene order in chloroplast genomes and simulating genome rearrangements, we have been able to formally test and reject the null hypothesis that the chloroplast genome of *C. reinhardtii* has evolved through random rearrangements of an ancestral genome. The observed gene order of *C. reinhardtii* cpDNA deviates strongly from the degree of sidedness and clustering expected under a random breakage model. The cpDNA of *Euglena gracilis* also has a high degree of sidedness (33), however, the asymmetry of its coding strand is concentrated in one-half of the genome, and is associated with GC content, which could be influenced by asymmetrical replication of the chromosome (33). In the *C. reinhardtii* cpDNA, the sidedness is not associated with GC content and we hypothesize that it is driven by co-transcription of genes in a functional cluster. Whereas some clusters of co-transcribed genes such as the ribosomal proteins (*rpl23-rpl2-rps19*, *rpl16-rpl14-rps8*) were maintained in both *C. reinhardtii* and *C. vulgaris*, novel clusters clearly formed in the *C. reinhardtii* lineage (Figure 3-4B).

Co-transcription of neighboring genes in the *C. reinhardtii* chloroplast is a widely documented phenomenon (28-31). We demonstrated that in addition to the ribosomal protein clusters, global analyses support the elevated level of clustering of other functionally related genes. The aggregate of genes in clusters include most essential genes involved in translation and transcription, and some photosynthetic genes. Coordinated transcription may play a crucial role in the regulation of plastid gene expression in response to light or circadian rhythms (34,35). It is also possible that some clusters contain cis-elements, similar to the artificial polydeoxyadenosine sequences (36), which enhances transcription efficiency. Moreover, most of the putative co-transcription units are not conserved across chlorophytes. Therefore, the majority of functional clusters observed in *C. reinhardtii* represent new gene arrangements rather than ancestral conservation.

In the chloroplast gene order phylogeny (Figure 3-2B), the *C. reinhardtii* lineage resides on a long branch compared to the *C. vulgaris* lineage, and both genomes are more rearranged than that of *N. olivacea*, relative to the common ancestral genome of the three chlorophyte lineages. The elevated rate of chloroplast genome rearrangement in *C. reinhardtii* is associated with invasion of short repeat sequences, which heavily populate the non-coding regions, increasing the total length of the intergenic regions compared to *C. vulgaris* cpDNA by one-third (16). Although simple sequence repeats are common to microbial genomes (37), such elements are rare in most sequenced chloroplast genomes. Within the *Chlamydomonas* genus (Chlorophyceae), *C. reinhardtii* and *C. gelatinosa* cpDNAs exhibit a prevalence of repetitive DNA and a high degree of gene order variation compared to the *C. moewusii*/*C. pitschmannii* lineage (15,38,39). The sister

lineage to *C. reinhardtii* in our study, *C. vulgaris* (Trebouxiophyceae), contains numerous cpDNA repeat sequences. Besides chlorophyte algae, members of angiosperm families, including Campanulaceae (40), Fabaceae (41,42) and Geraniaceae (43), also contain repeat elements in rearranged cpDNAs, albeit of a much lower copy number (40-43). These repeat elements may act as molecular “grease” that facilitates non-homologous recombination and creates a pool of diverse genome structures subject to selective retention. Future investigations will test whether the increased rates of rearrangement in plastid genomes with dispersed repeats typically lead to increased sidedness and functional clustering as we infer for *C. reinhardtii*.

Gene order changes reflect relatively rare evolutionary events and are expected to result in much less homoplasy than substitution events in nucleotide or protein sequences over a deep time scale (44). Phylogeny reconstruction using GRAPPA is highly accurate even for divergent genomes (45), and thus the ancestral gene orders inferred in our study contained sufficient phylogenetic information. The only other software for genome rearrangement phylogeny, BADGER (46), performed much worse on this data set (results not shown). GRAPPA usually inferred unique ancestral gene orders on many data sets we tested. Furthermore, analyses on simulated data have shown that the inferred gene orders almost scored as good as true ancestral gene orders (47). In our simulation tests of three genomes with 85 genes each, and the branch lengths of 50, 20 and 20 (roughly corresponding to the branches leading to *C. reinhardtii*, *C. vulgaris* and *N. olivacea*; see Methods), the average score for ancestral gene orders computed by GRAPPA was only about 7% less than the true scores. In practice, we observed that the less optimal gene orders generally required more rearrangements. Therefore, it is quite likely that any error

in our estimation of ancestral gene order has resulted in a downward bias in the inferred number of rearrangements on the branch leading to *C. reinhardtii*. Increasing the number of rearrangements on this branch would only lead to a more certain rejection of the neutrality of rearrangements.

The accuracy of ancestral genome reconstruction also depends on the degree of divergence among extant taxa and taxon sampling. For example, accurate reconstruction of ancestral genomes at the mammalian CFTR locus was achieved at the DNA level (48). The high-quality reconstruction was attributed to a dense sampling of syntenic genome sequences from eutherian mammals, and the lack of gene order rearrangement at the locus. Because the *C. reinhardtii* cpDNA is one of the most rearranged chloroplast genomes sequenced to date, we included all available chlorophyte chloroplast genomes for evolutionary distance estimation and ancestral gene order reconstruction. The accuracy of our ancestral gene order estimation may improve with inclusion of additional chlorophyte plastid gene orders as they become available, but we do not foresee a substantial reduction in the inferred number of rearrangements separating *C. reinhardtii* and *C. vulgaris* from their common ancestor.

Inversions are thought to be much more common than transpositions in chloroplast genome evolution (27), and our estimation of ancestral genome order was made with the assumption that all rearrangements were inversions. However, we did consider the contribution of inversions and transpositions under different scenarios in the simulation from the ancestral genome. It should be noted that there is not a unique phylogeny distance measure using transposition only, because computationally one transposition is equivalent to two or three inversions (49). For this reason, we designed

our simulations to allow for various ratios of inversion and transposition events. The result of our simulation study does not vary significantly.

The GRAPPA-IR algorithm was developed to account for the inverted repeat (IR) region found in most plastid genomes (25). The IR region seems to evolve at a slower rate in both nucleotide sequence and gene order than the single copy regions, and frequent intra-molecular recombination homogenizes the two copies (50). The most conserved gene set in the IR region is the rRNA operon. In IR-containing green plastids, the order of rRNA genes are conserved, but the IR boundaries can vary greatly even within one genus (51). IR may restrict rearrangements that cross the single copy regions and thus concentrate gene order changes within single copy regions. However, this constraint of IR on genome rearrangements was lost in the *C. reinhardtii*/*C. vulgaris* lineage. We inferred that the loss of one IR occurred in the *C. vulgaris* lineage following divergence from the *C. reinhardtii* lineage. Notably, both lineages have undergone extensive rearrangements since their divergence from a common ancestor, and the only conserved clusters seem to be the translational apparatus (rRNA genes and the ribosomal protein cluster). In either genome, genes that typically reside together in the LSC region have often been scrambled and scattered. When comparing the ancestral genome to the *C. vulgaris* gene order, there was no distinction of LSC and SSC regions although many large clusters were still shared (Additional file 4). If there were constraints on the breakpoint locations, as experimentally identified in bacterial inversion mutants (52), it would limit the possible paths of evolution, and these constraints on the ancestral gene orders would increase the number of rearrangements relative to the estimations derived from GRAPPA. Therefore, as discussed above, our approach of detecting strong

deviation from expectation is conservative in that the number of rearrangements may be underestimated.

Recent studies of plant, animal and fungal genomes have shown that genes involved in the same pathways or genes sharing similar expression patterns are often spatially clustered (1,5,53). In eukaryotes, the operon structure has only been demonstrated in the nematode *Caenorhabditis* (54). Comparative analyses of yeast genomes indicate that rearrangements brought together duplicate genes forming the DAL cluster involved in allantoin metabolism (55). In this study, we demonstrated that positive selection for increased clustering has influenced gene order in the chloroplast genome. Gene clusters, as opposed to separated genes, permit polycistronic transcription and thus fewer transcriptional regulation units. Co-transcription may be facilitated by close spacing of genes in cpDNA because transcription termination is inefficient (56). Although post-transcriptional RNA processing often creates multiple single-gene transcripts, co-transcription foments an initial stoichiometric accumulation of RNA corresponding to each gene in a cluster. Thus, large clusters can be advantageous in coordinating gene expression on this level. Experimental approaches are necessary to understand whether these gene clusters function as operons. Because chloroplast primary transcripts are heavily processed – as just one example, the *psbB* cluster in maize accumulates as at least 15 distinct mRNA species with varying translational capacities (57) – direct analysis of the functional advantages of clustering in chloroplasts is challenging. Indeed, *Chlamydomonas* may be a special case, since unlike land plants it has a single rather than multiple RNA polymerases (58). This situation does not allow differential expression by promoter selectivity, and may therefore serve as a selective

force that favors physical grouping of genes rather than evolution of promoter sequences of dispersed genes.

Conclusions

In conclusion, we infer that gene order in the *C. reinhardtii* plastid genome evolved in a non-random fashion and hypothesize that genome structure has been influenced by directional selection acting on variation generated by an increased rate of genome rearrangement. Co-transcription of novel clusters of functionally related genes could convey advantages in response to environmental and developmental cues. Our results provide strong evidence that genetic responses to natural selection occur at the level of genome organization. By estimating the ancestral gene order and simulating rearrangements under a null model, we provide a formal demonstration that the chloroplast genome of *C. reinhardtii* has been shaped by natural selection. Although the model of natural selection on gene order is yet to be developed, application of our methods to sequences of additional chlorophyte plastid genomes would help to improve the accuracy of the ancestral genome reconstruction and inferred branch lengths. Experimental tests of strains with engineered gene orders may be possible in the future. The complex process of gene duplication and loss in bacterial and eukaryotic nuclear genomes is a challenge to reconstruct the ancestral gene order. Still, the development of new comparative tools (59) gives us hope that the type of analysis presented in this paper will soon be applicable to eukaryotic genomes.

Methods

Functional cluster of chloroplast genes

We defined a “functional cluster” as contiguous genes encoded on one strand from one of the following categories: transcription/translation, photosystem I and II, electron transport (cytochrome b6/f complex), and ATP synthase (See additional file 1).

Kolmogorov-Smirnov test of random clusters

A random cluster consists of genes from any functional category. The $n=85$ genes shared in the seven chloroplast genomes shown in Figure 3-1 were divided into 11 equal sized blocks of $r_j=7$ genes and one block of 8 genes so that the block sizes and number of blocks are equal. If m_{ij} genes were from the functional category i (total T_i genes) in the j th block, the observed cumulative frequency was $u_i = \sum_j m_{ij} / r_j$. The Kolmogorov-Smirnov test measures the deviation of the observed u_i from the expected from the random breakage model (13). The test statistic D_n was calculated for each functional category separately.

$$D_n = \max[\max(\frac{T_i}{n} - u_i), \max(u_i - \frac{T_i - 1}{n})]$$

Phylogeny of chloroplast genomes

Alignments of 50 proteins shared in the 8 chloroplast genomes shown in Figure 3-2A were concatenated into one data matrix (Additional file 2). 1,000 bootstrap replicates were conducted on the data set using PAUP* 4.0b10 with maximum parsimony and using MEGA with neighbor-joining methods and the Poisson-corrected distance. Maximum likelihood analysis with 100 bootstrap replicates was performed using

PHYLIP3.6 with JTT distance and gamma = 0.5. GRAPPA was not used to construct the reference phylogeny.

Inferring ancestral gene orders

The ancestral gene order was inferred from the gene orders of extant genomes on the best-scored tree following two steps. First, the gene contents for the LSC, SSC and IR regions of ancestral genomes of IR-containing cpDNAs were inferred based on parsimony. Changes in gene copy number due to IR expansion or contraction were considered the last step of gene order changes, and thus the gene contents of ancestral genomes were determined. The ancestral gene orders on the phylogeny for five genomes (excluding *C. vulgaris* and *C. reinhardtii*) were computed using GRAPPA-IR (25), which is a modified version of GRAPPA that scores rearrangements independently within LSC, SSC or IR. Second, the chlorophyte algal gene orders (the extant chloroplast gene orders of *N. olivacea*, *C. reinhardtii*, *C. vulgaris* and the inferred ancestral genome of *N. olivacea* from step one) and the gene order of *M. viride* were used for the inference of the common ancestral gene order of *C. vulgaris* and *C. reinhardtii*. The data set contains duplicated *trnV*-UAC and *trnG*-GCC in *C. vulgaris*, *trnE*-UUC and *psbA* in *C. reinhardtii* and three trans-splicing *psaA* exons in *C. reinhardtii*. The IR regions contained rRNA genes in the same order and orientation in each genome except that one copy was lost in the lineage leading to *C. vulgaris*. To score the genomes with gene duplications and deletions, multiple data sets were created each containing genomes with equalized gene contents by the following assignment rules: one copy of each duplicate genes outside the typical IR was chosen; the IR region lost in *C. vulgaris* was inserted to all possible locations in that genome. Preferably, we should test all these datasets (3,936

total) with inversion medians, however, such computation on one dataset alone will take more than a month. To overcome this limitation, these datasets were computed using breakpoint medians, and the assignment yielded the shortest tree was chosen for a full evaluation by GRAPPA. Because the gene contents of LSC and SSC in *C. reinhardtii* were different from other chloroplast genomes in the study, we allow free rearrangements such that genes in LSC or SSC could commute across the IR.

Ancestral gene order simulation

A set of simulation experiments were conducted to evaluate the accuracy of ancestral genome reconstruction with long branches. Three genomes with 85 genes each were generated from a defined ancestral gene order, and the branch lengths (inversion distances) were 50, 20 and 20, respectively. The true gene order score was 90 (equals the tree length). The scores were computed for inferred ancestral gene orders by GRAPPA and compared to the true score. The experiment was repeated on 30 data sets.

Random genome rearrangement simulation

Gene orders were simulated under the assumption that the rearrangements involve random breakpoints placed between genes. Initial gene orders were set based on the inferred ancestral gene orders estimated. Random rearrangement operations on the initial genomes were performed for the number of replicates according to the number of rearrangements inferred by GRAPPA. The parameters input to the model were the ratios of inversion and transposition (1:0, 10:1, 1:1) to test the sensitivity of the findings to the specific rearrangement model. The simulated genomes had identical gene content but scrambled gene orders relative to those observed in extant genomes, with the exception that inverted repeats were maintained. Test statistics (below) were calculated for each

simulated replicate of 10,000 total and the frequency distributions were used to test the null hypothesis of random rearrangement.

Sidedness index (C_s)

We designed the sidedness index (C_s) to measure the degree to which neighboring genes are clustered on the same strand (side) of the chromosome. A "sided block" includes only adjacent genes on one strand, and the number of sided blocks in a genome is designated as n_{SB} , while the total number of genes is n . C_s is defined as

$$C_s = (n - n_{SB}) / (n - 1).$$

When C_s reaches the maximum of 1, all genes are located on one side. If every gene resides on the strand opposite its neighbors, C_s approaches a minimum of zero.

Functional cluster index (C_f)

We divided a genome of total n genes to J sided blocks (r_1, r_2, \dots, r_J). In a block, we assigned genes to functional categories. Let the numbers of genes in the i th functional category and the j th block be m_{ij} , the functional cluster index C_f is

$$C_f = \frac{1}{J} \sum_{j=1}^J \frac{r_j}{n} \sum_{i=1}^4 \binom{m_{ij}}{2} / \binom{r_j}{2}.$$

A larger value of C_f indicates that functionally related genes are more clustered into blocks.

RNA analysis

Wild-type CC-124 cells were grown in Tris-Acetate-Phosphate medium [67] under continuous light to mid-log phase. RNA was isolated from 10 mL of cells as previously described (60). For filter hybridization, 5 μ g of total RNA was fractionated in 1.2% agarose and 6% formaldehyde gels, transferred to nylon membranes, and probed

with gene-specific PCR products labeled by random priming according to Church and Gilbert (61).

Acknowledgements

We thank A. Jarosz, H. Ma, J. Marden, W. Martin, W. Miller, and D. Schemske for valuable suggestions and comments. This work was supported by NSF awards DBI 0115684 and DEB 0120709 to C.W.D.. J. T. is supported by the University of South Carolina and part of the work was done while he was visiting the National Evolutionary Synthesis Center. *Chlamydomonas* genomics work at BTI was supported by NSF awards MCB 9975765 and MCB 0091020 to D.B.S.

References

1. Hurst, L.D., Pal, C. and Lercher, M.J. (2004) The evolutionary dynamics of eukaryotic gene order. *Nat Rev Genet*, **5**, 299-310.
2. Kallas, T., Spiller, S. and Malkin, R. (1988) Primary structure of cotranscribed genes encoding the Rieske Fe-S and cytochrome f proteins of the cyanobacterium *Nostoc* PCC 7906. *Proc Natl Acad Sci U S A*, **85**, 5794-5798.
3. Stoebe, B. and Kowallik, K.V. (1999) Gene-cluster analysis in chloroplast genomics. *Trends Genet*, **15**, 344-347.
4. Lercher, M.J., Urrutia, A.O. and Hurst, L.D. (2002) Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nat Genet*, **31**, 180-183.
5. Lee, J.M. and Sonnhammer, E.L. (2003) Genomic gene clustering analysis of pathways in eukaryotes. *Genome Res*, **13**, 875-882.
6. Palmer, J.D. (1985) In MacIntyre, R. J. (ed.), *Molecular Evolutionary Genetics*. Plenum Press, New York, pp. 131-240.
7. Pancic, P.G., Strotmann, H. and Kowallik, K.V. (1992) Chloroplast ATPase genes in the diatom *Odontella sinensis* reflect cyanobacterial characters in structure and arrangement. *J Mol Biol*, **224**, 529-536.

8. Nadeau, J. and Taylor, B.A. (1984) Length of chromosome segments conserved since divergence of man and mouse. *Proc. Natl. Acad. Sci. USA*, **81**, 814-818.
9. Cosner, M.E., Jansen, R.K., Moret, B.M., Raubeson, L.A., Wang, L.S., Warnow, T. and Wyman, S. (2000) A new fast heuristic for computing the breakpoint phylogeny and experimental phylogenetic analyses of real and synthetic data. *Proc Int Conf Intell Syst Mol Biol*, **8**, 104-115.
10. Moret, B.M., Wang, L.S., Warnow, T. and Wayman, S.K. (2001) New approaches for reconstructing phylogenies from gene order data. *Bioinformatics*, **17**, S165-S173.
11. Tang, J.J. and Moret, B.M.E. (2003), *Lecture Notes in Computer Science*. Springer-Verlag Berlin, Berlin, Vol. 2748, pp. 37-46.
12. Tang, J. and Moret, B.M. (2003) Scaling up accurate phylogenetic reconstruction from gene-order data. *Bioinformatics*, **19 Suppl 1**, i305-312.
13. Pevzner, P. and Tesler, G. (2003) Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proc. Natl. Acad. Sci. USA*, **100**, 7672-7677.
14. Trinh, P., McLysaght, A. and Sankoff, D. (2004) Genomic features in the breakpoint regions between syntenic blocks. *Bioinformatics*, **20 Suppl 1**, I318-I325.
15. Boudreau, E. and Turmel, M. (1996) Extensive gene rearrangements in the chloroplast DNAs of *Chlamydomonas* species featuring multiple dispersed repeats. *Mol Biol Evol*, **13**, 233-243.
16. Maul, J.E., Lilly, J.W., Cui, L., dePamphilis, C.W., Miller, W., Harris, E.H. and Stern, D.B. (2002) *Chlamydomonas* chloroplast chromosome: islands of genes in a sea of repeats. *Plant Cell*, **14**, 2659-2679.
17. Wakasugi, T., Nagai, T., Kapoor, M., Sugita, M., Ito, M., Ito, S., Tsudzuki, J., Nakashima, K., Tsudzuki, T., Suzuki, Y. *et al.* (1997) Complete nucleotide sequence of the chloroplast genome from the green alga *Chlorella vulgaris*: the existence of genes possibly involved in chloroplast division. *Proc Natl Acad Sci U S A*, **94**, 5967-5972.
18. Shinozaki, K., Ohme, M., Tanaka, M., Wakasugi, T., Hayashida, N., Matsubayashi, T., Zaita, N., Chunwongse, J., Obokata, J., Yamaguchishinozaki, K. *et al.* (1986) The Complete Nucleotide-Sequence of the Tobacco Chloroplast Genome - Its Gene Organization and Expression. *Embo Journal*, **5**, 2043-2049.
19. Ohyama, K., Fukuzawa, H., Kohchi, T., Sano, T., Sano, S., Shirai, H., Umesono, K., Shiki, Y., Takeuchi, M., Chang, Z. *et al.* (1988) Structure and organization of *Marchantia polymorpha* chloroplast genome. I. Cloning and gene identification. *J Mol Biol*, **203**, 281-298.
20. Turmel, M., Otis, C. and Lemieux, C. (2002) The chloroplast and mitochondrial genome sequences of the charophyte *Chaetosphaeridium globosum*: insights into the timing of the events that restructured organelle DNAs within the green algal lineage that led to land plants. *Proc Natl Acad Sci U S A*, **99**, 11275-11280.
21. Turmel, M., Otis, C. and Lemieux, C. (1999) The complete chloroplast DNA sequence of the green alga *Nephroselmis olivacea*: insights into the architecture of ancestral chloroplast genomes. *Proc Natl Acad Sci U S A*, **96**, 10248-10253.

22. Lemieux, C., Otis, C. and Turmel, M. (2000) Ancestral chloroplast genome in *Mesostigma viride* reveals an early branch of green plant evolution. *Nature*, **403**, 649-652.
23. Loffelhardt, W., Bohnert, H.J. and Bryant, D.A. (1997) The cyanelles of *Cyanophora paradoxa*. *Crit Rev Plant Sci*, **16**, 393-413.
24. Karol, K.G., McCourt, R.M., Cimino, M.T. and Delwiche, C.F. (2001) The closest living relatives of land plants. *Science*, **294**, 2351-2353.
25. Cui, L., Tang, J., Moret, B.M.E. and dePamphilis, C.W. (2005), *TR-CS-2005-08*. University of New Mexico.
26. Palmer, J.D. (1990) Contrasting modes and tempos of genome evolution in land plant organelles. *Trends Genet*, **6**, 115-120.
27. Boudreau, E. and Turmel, M. (1995) Gene rearrangements in *Chlamydomonas* chloroplast DNAs are accounted for by inversions and by the expansion/contraction of the inverted repeat. *Plant Mol Biol*, **27**, 351-364.
28. Choquet, Y., Goldschmidt-Clermont, M., Girard-Bascou, J., Kuck, U., Bennoun, P. and Rochaix, J.D. (1988) Mutant phenotypes support a trans-splicing mechanism for the expression of the tripartite *psaA* gene in the *C. reinhardtii* chloroplast. *Cell*, **52**, 903-914.
29. Mor, T.S., Ohad, I., Hirschberg, J. and Pakrasi, H.B. (1995) An unusual organization of the genes encoding cytochrome b559 in *Chlamydomonas reinhardtii*: *psbE* and *psbF* genes are separately transcribed from different regions of the plastid chromosome. *Mol Gen Genet*, **246**, 600-604.
30. Sturm, N.R., Kuras, R., Buschlen, S., Sakamoto, W., Kindle, K.L., Stern, D.B. and Wollman, F.A. (1994) The *petD* gene is transcribed by functionally redundant promoters in *Chlamydomonas reinhardtii* chloroplasts. *Mol. Cell. Biol.*, **14**, 6171-6179.
31. Higgs, D.C., Kuras, R., Kindle, K.L., Wollman, F.A. and Stern, D.B. (1998) Inversions in the *Chlamydomonas* chloroplast genome suppress a *petD* 5' untranslated region deletion by creating functional chimeric mRNAs. *Plant J*, **14**, 663-671.
32. Kugita, M., Yamamoto, Y., Fujikawa, T., Matsumoto, T. and Yoshinaga, K. (2003) RNA editing in hornwort chloroplasts makes more than half the genes functional. *Nucleic Acids Res*, **31**, 2417-2423.
33. Morton, B.R. (1999) Strand asymmetry and codon usage bias in the chloroplast genome of *Euglena gracilis*. *Proc. Natl. Acad. Sci. USA*, **96**, 5123-5128.
34. Thompson, R.J. and Mosig, G. (1990) Light affects the structure of *Chlamydomonas* chloroplast chromosomes. *Nucl. Acids Res.*, **18**, 2625-2631.
35. Eberhard, S., Drapier, D. and Wollman, F.A. (2002) Searching limiting steps in the expression of chloroplast-encoded proteins: relations between gene copy number, transcription, transcript abundance and translation rate in the chloroplast of *Chlamydomonas reinhardtii*. *Plant J*, **31**, 149-160.
36. Lisitsky, I., Rott, R. and Schuster, G. (2001) Insertion of polydeoxyadenosine-rich sequences into an intergenic region increases transcription in *Chlamydomonas reinhardtii* chloroplasts. *Planta*, **212**, 851-857.

37. Saunders, N.J., Peden, J.F., Hood, D.W. and Moxon, E.R. (1998) Simple sequence repeats in the *Helicobacter pylori* genome. *Mol Microbiol*, **27**, 1091-1098.
38. Lemieux, B., Turmel, M. and Lemieux, C. (1985) Chloroplast DNA variation in *Chlamydomonas* and its potential application to the systematics of this genus. *Biosystems*, **18**, 293-298.
39. Boudreau, E., Otis, C. and Turmel, M. (1994) Conserved gene clusters in the highly rearranged chloroplast genomes of *Chlamydomonas moewusii* and *Chlamydomonas reinhardtii*. *Plant Mol Biol*, **24**, 585-602.
40. Cosner, M.E., Jansen, R.K., Palmer, J.D. and Downie, S.R. (1997) The highly rearranged chloroplast genome of *Trachelium caeruleum* (Campanulaceae): multiple inversions, inverted repeat expansion and contraction, transposition, insertions/deletions, and several repeat families. *Curr Genet*, **31**, 419-429.
41. Perry, A.S., Brennan, S., Murphy, D.J., Kavanagh, T.A. and Wolfe, K.H. (2002) Evolutionary re-organisation of a large operon in adzuki bean chloroplast DNA caused by inverted repeat movement. *DNA Res*, **9**, 157-162.
42. Milligan, B.G., Hampton, J.N. and Palmer, J.D. (1989) Dispersed repeats and structural reorganization in subclover chloroplast DNA. *Mol Biol Evol*, **6**, 355-368.
43. Price, R.A., Calie, P.J., Downie, S.R., Logsdon, J., J.M. and Palmer, J.D. (1990) In Vorster, P. (ed.), *Proc. Int. Geraniaceae Symp.* University of Stellenbosch, Monvillia, South Africa, pp. 235-244.
44. Rokas, A. and Holland, P.W. (2000) Rare genomic changes as a tool for phylogenetics. *Trends Ecol Evol*, **15**, 454-459.
45. Wang, L.S., Jansen, R. K., Moret, B. M., Raubeson, L. A., Warnow, T. (2002), *Pac. Symp. Biocomput.* World Scientific Pub., River Edge, pp. 524-535.
46. Simon, D. and Larget, B. (2004).
47. Siepel, A.C. and Moret, B.M.E. (2001), *Lecture Notes in Computer Science.* Springer-Verlag, Vol. 2149, pp. 189-203.
48. Blanchette, M., Green, E.D., Miller, W. and Haussler, D. (2004) Reconstructing large regions of an ancestral mammalian genome in silico. *Genome Res*, **14**, 2412-2423.
49. Wang, L.-S. (2001), *Lecture Notes in Computer Science.* Springer-Verlag, Vol. 2149, pp. 175-188.
50. Lemieux, B., Turmel, M. and Lemieux, C. (1990) Recombination of *Chlamydomonas* Chloroplast DNA Occurs More Frequently in the Large Inverted Repeat Sequence Than in the Single-Copy Regions. *Theor Appl Genet*, **79**, 17-27.
51. Goulding, S.E., Olmstead, R.G., Morden, C.W. and Wolfe, K.H. (1996) Ebb and flow of the chloroplast inverted repeat. *Mol Gen Genet*, **252**, 195-206.
52. Segall, A.M. and Roth, J.R. (1989) Recombination between homologies in direct and inverse orientation in the chromosome of *Salmonella*: intervals which are nonpermissive for inversion formation. *Genetics*, **122**, 737-747.
53. Williams, E.J. and Bowles, D.J. (2004) Coexpression of neighboring genes in the genome of *Arabidopsis thaliana*. *Genome Res*, **14**, 1060-1067.

54. Blumenthal, T., Evans, D., Link, C.D., Guffanti, A., Lawson, D., Thierry-Mieg, J., Thierry-Mieg, D., Chiu, W.L., Duke, K., Kiraly, M. *et al.* (2002) A global analysis of *Caenorhabditis elegans* operons. *Nature*, **417**, 851-854.
55. Wong, S. and Wolfe, K.H. (2005) Birth of a metabolic gene cluster in yeast by adaptive gene relocation. *Nat Genet*, **37**, 777-782.
56. Monde, R.A., Schuster, G. and Stern, D.B. (2000) Processing and degradation of chloroplast mRNA. *Biochimie*, **82**, 573-582.
57. Barkan, A. (1988) Proteins encoded by a complex chloroplast transcription unit are each translated from both monocistronic and polycistronic RNAs. *EMBO J*, 2637-2644.
58. Cahoon, A.B. and Stern, D.B. (2001) Plastid transcription: A ménage à trois? *Trends Plant Sci*, **6**, 45-46.
59. Kent, W.J., Baertsch, R., Hinrichs, A., Miller, W. and Haussler, D. (2003) Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc Natl Acad Sci U S A*, **100**, 11484-11489.
60. Drager, R.G., Higgs, D.C., Kindle, K.L. and Stern, D.B. (1999) 5' to 3' exoribonucleolytic activity is a normal component of chloroplast mRNA decay pathways. *Plant J*, **19**, 521-531.
61. Church, G.M. and Gilbert, W. (1984) Genomic sequencing. *Proc Natl Acad Sci U S A*, **81**, 1991-1995.

Additional Files

Additional file 1 – Gene coding and functional categories.

The gene rps12 was coded as 55 and 74 in the cases when the first exon and the second exon are separated by other genes.

Gene	Code	Functional Category
atpA	9	ATP synthase
atpB	39	ATP synthase
atpE	38	ATP synthase
atpF	10	ATP synthase
atpF	11	ATP synthase
atpH	12	ATP synthase
atpI	13	ATP synthase
petA	43	electron transport
petB	61	electron transport
petD	62	electron transport
petG	49	electron transport
petL	48	electron transport
ccsA	85	Other
cemA	42	Other
clpP	56	Other
rbcL	40	Other
ycf1	82	Other
psaA	31	photosystem I and photosystem II
psaB	30	photosystem I and photosystem II
psaC	86	photosystem I and photosystem II
psaJ	52	photosystem I and photosystem II
psbA	2	photosystem I and photosystem II
psbB	57	photosystem I and photosystem II
psbC	24	photosystem I and photosystem II
psbD	23	photosystem I and photosystem II
psbE	47	photosystem I and photosystem II
psbF	46	photosystem I and photosystem II
psbH	60	photosystem I and photosystem II
psbI	6	photosystem I and photosystem II
psbJ	44	photosystem I and photosystem II
psbK	5	photosystem I and photosystem II
psbL	45	photosystem I and photosystem II
psbM	19	photosystem I and photosystem II
psbN	59	photosystem I and photosystem II
psbT	58	photosystem I and photosystem II

psbZ	26	photosystem I and photosystem II
ycf4	41	photosystem I and photosystem II
rpl14	67	transcription/translation
rpl16	68	transcription/translation
rpl2	71	transcription/translation
rpl20	54	transcription/translation
rpl23	72	transcription/translation
rpl32	83	transcription/translation
rpl36	65	transcription/translation
rpoA	63	transcription/translation
rpoB	17	transcription/translation
rpoC1	16	transcription/translation
rpoC2	15	transcription/translation
rps11	64	transcription/translation
rps12	55	transcription/translation
"rps12, exon2"	74	transcription/translation
rps14	29	transcription/translation
rps18	53	transcription/translation
rps19	70	transcription/translation
rps2	14	transcription/translation
rps3	69	transcription/translation
rps4	32	transcription/translation
rps7	73	transcription/translation
rps8	66	transcription/translation
rrn16	75	transcription/translation
rrn23	78	transcription/translation
rrn5	79	transcription/translation
trnA-UGC	77	transcription/translation
trnC-GCA	18	transcription/translation
trnD-GUC	20	transcription/translation
trnE-UUC	22	transcription/translation
trnF-GAA	35	transcription/translation
trnFM	28	transcription/translation
trnG-GCC	27	transcription/translation
trnH-GUG	1	transcription/translation
trnI-GAU	76	transcription/translation
trnK-UUU	3	transcription/translation
trnL-UAA	34	transcription/translation
trnL-UAG	84	transcription/translation
trnM-CAT	37	transcription/translation
trnN-GUU	81	transcription/translation
trnP-UGG	51	transcription/translation
trnQ-UUG	4	transcription/translation
trnR-ACG	80	transcription/translation
trnR-UCU	8	transcription/translation
trnS-GCU	7	transcription/translation
trnS-UGA	25	transcription/translation

trnT-UGU	33	transcription/translation
trnV-UAC	36	transcription/translation
trnW-CCA	50	transcription/translation
trnY-GUA	21	transcription/translation

Additional file 2 – Protein alignment matrix.

Text file, with a NEXUS format data matrix of concatenated proteins from seven chloroplast genomes and the outgroup, *Cyanophora paradoxa*. Data file omitted.

Additional file 3 – The gene order data set.

The file contains gene orders of seven chloroplast genomes, computed C_s and C_f indices, and the inferred rearrangement phylogeny.

Gene orders

>cv *Chlorella*

```
-42  -2  86  -10  -11  -12  -13  -14  -33  6  -7  75
76  77  78  79  -24  -23  25  28  22  54  53  50  -51
52  55  73  41  -3  -26  -1  -8  -36  -44  -45  -46  -
47  -36  84  -62  -61  -56  -27  19  -21  34  -60  59
-58  -57  31  30  -35  -4  -5  80  81  85  83  82  -18
17  16  15  40  29  27  37  -20  -32  -38  -39  -49  -
48  -43  9  -63  -64  -65  -66  -67  -68  -69  -70  -
71  -72
```

>cr *Chlamydomonas*

```
43  62  80  51  5  22  -18  -33  -9  -54  -25  -50  -
56  -84  -61  65  72  71  70  68  67  66  31  27  32
75  76  77  78  79  -2  -7  -38  -73  -29  -19  -26  -
85  -34  -31  -22  -60  59  -58  -57  -20  -63  -14  -
53  -41  -47  -37  -17  -17  46  45  49  69  15  21  4
30  -8  -40  10  6  42  12  11  64  3  2  -79  -78  -
77  -76  -75  -39  -82  55  -52  -13  -44  -31  -23
24  1  28  35  86  48  81  -16  -36
```

>no *Nephroselmis*

```
29  34  4  84  -19  37  2  39  38  -62  -61  -63  -64
-65  -66  -67  -68  -69  -70  -71  -72  -10  -11  -12
-13  -56  -1  -36  8  -18  17  16  15  14  28  23  24
-50  -51  52  55  73  41  42  43  48  49  3  5  35  27
-44  -45  -46  -47  -53  -32  -9  -54  20  -22  -26
25  -60  59  -58  -57  -33  -7  -6  -21  31  30  -40
75  76  77  78  79  80  85  -81  86  83  82  81  -85
-80  -79  -78  -77  -76  -75  40
```

>mv *Mesostigma*

```
23  24  -2  -56  57  58  -59  60  61  62  -63  -64  -
65  -66  -67  -68  -69  -70  -71  -72  -40  39  38  37
-8  -9  -36  -10  -11  -12  -13  -14  -15  -16  -17
18  7  -6  -5  -4  31  30  29  28  3  35  -1  -19  -44
-45  -46  -47  -27  -26  25  -20  54  -53  -52  51  50
```

```

55 73 41 42 43 48 49 -22 -21 34 -32 33 75
76 77 78 79 80 86 81 -85 -82 -83 84 -80 -
79 -78 -77 -76 -75 -33
>cg Chaetosphaeridium
74 73 19 -18 17 16 15 14 13 12 11 10 -9 -
8 -3 -4 5 6 -7 -62 -61 -60 59 -58 -57 56
55 54 -53 -52 51 50 -49 -48 29 28 -21 -22
1 36 35 37 -38 -39 40 41 42 43 -44 -45 -46
-47 -27 -26 25 -24 -23 -2 -32 -33 34 31 30
20 -63 -64 -65 -66 -67 -68 -69 -70 -71 -72
75 76 77 78 79 80 -81 -83 84 85 -86 -82 81
-80 -79 -78 -77 -76 -75
>mp Marchantia
55 73 19 -18 17 16 15 14 13 12 11 10 -9 -
8 7 -6 -5 4 3 2 1 -20 -21 -22 23 24 -25
26 27 -28 -29 -30 -31 -32 -33 34 35 -36 37
-38 -39 40 41 42 43 -44 -45 -46 -47 48 49
-50 -51 52 53 -54 -56 57 58 -59 60 61 62 -
63 -64 -65 -66 -67 -68 -69 -70 -71 -72 75
76 77 78 79 80 -81 83 84 85 -86 -82 81 -80
-79 -78 -77 -76 -75
>nt Nicotiana
-1 -2 -3 -4 5 6 -7 8 9 -10 -11 -12 -13 -
14 -15 -16 -17 18 -19 -20 -21 -22 23 24 -25
26 27 -28 -29 -30 -31 -32 -33 34 35 -36 37
-38 -39 40 41 42 43 -44 -45 -46 -47 48 49
-50 -51 52 53 -54 -55 -56 57 58 -59 60 61
62 -63 -64 -65 -66 -67 -68 -69 -70 -71 -72
-73 -74 75 76 77 78 79 80 -81 82 83 84 85
-86 81 -80 -79 -78 -77 -76 -75 74 73 72 71

```

Cs for the data set

```

>cv Chlorella 0.7093
>cr Chlamydomonas 0.8710
>no Nephroselmis 0.7742
>mv Mesostigma 0.7582
>cg Chaetosphaeridium 0.7174
>mp Marchantia 0.6703
>nt Nicotiana 0.6875

```

Cf for the data set (x1E-3)

```

>cv Chlorella 19.51
>cr Chlamydomonas 33.97

```

```

>no Nephroselmis      19.64
>mv Mesostigma    22.58
>cg Chaetosphaeridium 23.16
>mp Marchantia    19.43
>nt Nicotiana     19.93

```

Inferred rearrangement phylogeny

1. Step 1, 5 IR-containing genomes

Tree length=78

```
(no:24,(mv:16,(cg:4,(mp:0,nt:7):9):18):0);
```

2. Step 2, loss of IR occurred in *C. vulgaris*. nop, the inferred parent of *N. olivacea* in Step 1.

Tree length = 144

```
(cv:43,((cr:60,(nop:1,mv:15):13):3,no:9):0);
```

3. Resolved with the reference tree

```
(mv:16,(no:12,(cv:43,cr:60):0):12,(cg:4,(mp:0,nt:7):9):18));
```

Additional file 4 – Comparison of gene clusters.

This file shows gene clusters shared (in brackets) between the inferred ancestral genome of *C. reinhardtii* and *C. vulgaris* to the cpDNA of *C. vulgaris* and *N. olivacea*.

```

>ancestral gene order of C. reinhardtii and C. vulgaris (p-cr-cv)
1   -3   -6  -21 [ 31   30] -40 [75  76  77   78
79]   80  -86   81  -85  -84 [82   83][ -79 -78 -77 -
76 -75]   29   34  -56   5   35   22  -20   54   9
32   2   [39   38] [ -62  -61][ -60   59  -58  -57 ]
-33  -7  -27  -26   25 [-63  -64  -65  -66  -67  -68
-69  -70  -71  -72 ][ -10  -11  -12  -13 ] -14 [ -15
-16  -17   18] -19   37 [ -24  -23 ] -28   4   -8
36  -53  -50  -51 [52   55   73   41]   42 [43   48
49][ -44  -45  -46  -47]
>C. vulgaris (cv)
-42  -2   86 [-10  -11  -12  -13] -14  -33   6  -7
[75  76  77  78  79] [-24  -23] 25  28  22  54  53
50  -51 [52  55  73  41] -3  -26  -1  -8  -36 [-44
-45  -46  -47] -36  84 [-62  -61] -56  -27  19  -21
34 [-60  59  -58  -57] [31  30] -35  -4  -5  80  81
85 [83  82][-18  17  16  15] 40  29  27  37  -20  -

```



```

32 [ -38 -39][-49 -48 -43] 9 [-63 -64 -65 -66
-67 -68 -69 -70 -71 -72]
>N. olivacea (no)
29 34 4 84 -19 37 2 [39 38][-62 -61][-63 -64
-65 -66 -67 -68 -69 -70 -71 -72][-10 -11 -12
-13] -56 -1 -36 8 [-18 17 16 15] 14 28 [ 23
24] -50 -51 [52 55 73 41] 42 [43 48 49] 3 5
35 27 [-44 -45 -46 -47] -53 -32 -9 -54 20 -
22 -26 25 [-60 59 -58 -57] -33 -7 -6 -21 [31
30] -40 [75 76 77 78 79] 80 85 -81 86 [83 82]
81 -85 -80 [-79 -78 -77 -76 -75] 40

```

Additional file 5 – Inversions separating *N. tabacum* and *M. polymorpha* cpDNA.

The inversion distance between the chloroplast DNAs of *Marchantia polymorpha* (mp) and *Nicotiana tabacum* (nt) is 7. To illustrate the sequence of gene order changes, the genomes are represented in a condensed form as following:

```
gene 19-->1 is renamed as gene 1
gene -20-->-54 is renamed as gene 2
gene 55 is renamed gene 3
gene -56-->-72 is renamed gene 4
gene -73-->-74 is renamed gene 5
gene 75-->-81 is renamed gene 6
gene 82 is renamed gene 7
gene 83-->-86 is renamed gene 8.
```

So the dataset becomes:

mp

```
3 -5 -1 2 4 6 8 -7
```

nt

```
1 2 -3 4 5 6 7 8
```

These two genomes can be divided into two parts, i.e., gene 1-5 corresponding to LSC, gene 6-8 corresponding to IR-SSC. They two parts can be viewed as independent to each other.

1. To translate (6 8 -7) into (6 7 8), there are two ways:

flip 8 in mp, then flip (-8 -7), i.e.

```
(6 8 -7)
-->(6 -8 -7)
-->(6 7 8)
```

Alternatively, flip (8 -7) in mp, then flip -8.

```
(6 8 -7)
-->(6 7 -8)
-->(6 7 8)
```

2. There are many choices to translate (3 -5 -1 2 4) into (1 2 -3 4 5),

The total number of possible ways are around 1000. One of the possible scenarios is:

```
(3 -5 -1 2 4)
--> (3 -5 1 2 4)
--> (3 -5 -2 -1 4)
--> (3 -5 -4 1 2)
--> (3 -2 -1 4 5)
--> (1 2 3 4 5)
```

The total sequences are any combination of the above two parts.

Chapter 4

Widespread genome duplications throughout the history of flowering plants

Preface

This manuscript has been accepted for publication in *Genome Research*. Authors for the original manuscript are Liying Cui, P. Kerr Wall, James H. Leebens-Mack, Bruce G. Lindsay, Douglas E. Soltis, Jeff J. Doyle, Pamela S. Soltis, John E. Carlson, K. Arumuganathan, Abdelali Barakat, Victor A. Albert, Hong Ma and Claude W. dePamphilis. LC conducted the data collection, data analysis, discussion and wrote the draft based in part on some introductory paragraphs from DS and PS. KW wrote the computing pipeline. BGL helped especially in statistical Methods and tests. JLM conducted the rate comparison. DS, JJD, PS, JC, AB, VA, HM and CWD all provided thoughtful comments and helped writing the manuscript. AA conducted the genome size measurements. CWD designed and supervised the study.

Abstract

Genomic comparisons provide evidence for ancient genome-wide duplications in a diverse array of animals and plants. We developed a birth-death model to identify evidence for genome duplication in EST data, and applied a mixture model to estimate the age distribution of paralogous pairs identified in EST sets for species representing the basal-most extant flowering plant lineages. We found evidence for episodes of ancient genome-wide duplications in the basal angiosperm lineages including *Nuphar advena* (Nymphaeaceae), and the magnoliids, *Persea americana* (Lauraceae), *Liriodendron tulipifera* (Magnoliaceae) and *Saruma henryi* (Aristolochiaceae). In addition, we detected independent genome duplications in a basal eudicot *Eschscholzia californica* (Papaveraceae) and the basal monocot *Acorus americanus* (Acoraceae), both of which were distinct from duplications documented for ancestral Poaceae and core eudicot lineages. In gymnosperms, we found equivocal evidence for ancient polyploidy in *Welwitschia mirabilis* (Welwitschiaceae) and no evidence for ancient polyploidy in *Pinus* (Pinaceae), although gymnosperms generally have much larger genomes than the angiosperms investigated. Cross-species sequence divergence estimates suggest that synonymous substitution rates in the basal angiosperms are less than half those previously reported for core eudicots and members of the Poaceae. The lower substitution rate allows inference of older duplication events. We hypothesize that evidence of an ancient duplication observed in the *Nuphar* data may represent a genome duplication in the common ancestor of all or most extant angiosperms (except *Amborella*).

There are 2 supplemental tables for the manuscript. Teri Solow and Lukas Muller provided the EST sequence assembly for eight species (*Acorus americanus*, *Amborella trichopoda*, *Eschscholzia californica*, *Liriodendron tulipifera*, *Nuphar advena*, *Persea americana*, *Saruma henryi*, and *Welwitschia mirabilis*), now available through the Plant Genome Network (<http://pgn.cornell.edu/>).

Introduction

Gene duplication has long been recognized to be a major force in evolution (1). Genome doubling as a consequence of polyploidy has had a profound influence on the evolutionary history of extant lineages. Ohno proposed that whole-genome duplications occurred in the early history of all vertebrates (1). While the hypothesis of whole-genome duplication in the earliest vertebrates has been somewhat controversial (2-5), ancient polyploidy is supported by genetic and genomic investigations of individual gene families as well as large syntenic chromosomal segments (6-9). The importance of genome duplication in the evolution of amphibians (10) and the yeast *Saccharomyces cerevisiae* has been more widely accepted (3,11,12).

Polyploidy is known to be common in many plant lineages (13-15). The angiosperms in particular have been the subject of considerable speculation regarding the frequency of polyploidy. Classic studies estimated that 30% to 50% of angiosperms are polyploids (13,16,17), and more recently most if not all extant angiosperms have been implicated as ancient polyploids (18-20). These inferences were based on comparisons of nuclear DNA content (C-value) or genome size, across a broad spectrum of species.

However, the rapid reduction of duplicate genes after polyploidization can drastically shrink genome size and gene content following genome duplication (1,21,22). Despite the small size of the *Arabidopsis thaliana* genome (157 Mb)(23), recent investigations have revealed two or more rounds of genome duplications (24-26). Analysis of the rice genome also suggested ancient polyploidy in the early history of the grass family (Poaceae) (27,28) It now appears that perhaps all major lineages of eukaryotic genomes possess considerable numbers of duplicate genes that may have resulted from genome duplications (1,29).

Whole-genome duplication, tandem gene duplication and segmental duplication all generate paralogous gene pairs. For species with complete genome sequences, such as *Arabidopsis*, rice and now *Populus*, it is possible to differentiate whole genome duplications from segmental and tandem gene duplications by mapping chromosomal locations of duplicate genes or blocks of genes (25-27,30,31). Lynch and Conery (29) proposed a genomic scale approach to estimate the age of gene duplication events and the fate of resulting paralogous gene pairs by evaluating the frequency distribution of per site synonymous divergence levels (K_s) for pairs of duplicate genes. After gene duplication, some paralogs will be silenced, and eventually be eliminated, while many of the preserved paralogs may be subject to changes in DNA sequence or gene expression leading to sub- or neofunctionalization (32-34).

Synonymous substitutions are largely immune to the strong selective pressures that greatly impact the rate of protein divergence (29,35)., and when corrected for multiple substitutions that occur in highly diverged sequences, these nearly neutral substitutions in protein-coding regions can be used as a proxy for the amount of time that

has passed since gene duplication. A genome-wide duplication event results in a sudden increase in the frequency of paralogous pairs. Evidence of past genome duplications can be seen as peaks in the distribution of K_s values for sampled paralogous pairs (29,36,37). This method does not depend on genomic positional information, and can be applied to any species for which there are moderately large EST sets. Identification of duplicated blocks of genes in genome sequences, however, provides much stronger evidence of ancient polyploidy, and average K_s values [or K_a (24)] can be used to date the origin of duplicated blocks. Using the large amount of DNA sequences generated by EST and genome sequencing projects, Blanc and Wolfe (36) investigated 14 model plant species (mostly crop species with known recent polyploid history), finding spikes in the distribution of older paralogous pairs (with higher K_s values) in 9 species. Schlueter et al. (37) advanced the analysis of K_s distributions by applying a finite mixture model (38) to sets of paralogous pairs identified in large EST databases for 8 major crop species, including soybean, *Medicago*, tomato, potato, maize, *Sorghum*, rice and barley, and inferred multiple independent genome duplications in Fabaceae, Solanaceae, and Poaceae over the last 14-60 million years. In general this method is only suitable for duplicated genes with similar codon usage, because K_s is affected by the codon usage bias (39,40).

All of the plants previously investigated using K_s distributions (36,37) belonged to either derived monocot (a single family, the Poaceae) or eudicot lineages. Most of the species examined were either crop species or close relatives, where a predisposition to polyploidy might have increased the chances of having traits important for domestication and agriculture [but see (41)]. Until recently, there has been very little sequence data for phylogenetically pivotal taxa representing the basal lineages of the eudicots, monocots or

all angiosperms, and the genome histories of these lineages are poorly understood. Here and throughout this paper we use the term “basal” as shorthand when referring to a lineage that is sister to a larger clade containing all other members of a particular group. An understanding of ancient genome duplication in the basal-most angiosperm lineages is especially important in understanding the role of polyploidy in the origin and early diversification of flowering plants (42-44). We utilize sets of 9000 to 10,000 ESTs generated for a number of species representing these basal lineages (45) to assess the frequency of ancient genome duplications across all major extant angiosperm lineages (Table 1) and evaluate whether these data can elucidate the timing of ancient genome duplication events in early angiosperm history.

To facilitate the interpretation of K_s distributions, we have modeled the gene birth-and-death process with and without genome-wide duplication events. Our model provides a predicted age distribution for any sample of duplicate genes while accounting for empirical estimation errors in K_s . The model was used to generate predicted K_s distributions for sets of paralogous pairs under the null hypothesis that the gene births and deaths occurred at constant rates. Null distributions were modeled using parameter values and error corrections estimated for each data set (see METHODS). When the null hypothesis of a constant birth-and-death process was rejected, the log-transformed K_s distribution for each taxon was analyzed using a mixture model to identify subpopulations of paralogous pairs generated through one or more large scale duplication events (37,38). Our results provide evidence of ancient polyploidy throughout the major angiosperm lineages, and support the possibility that a genome-scale duplication event occurred prior to the rapid diversification of flowering plants (46).

Results

Model parameters and their influence on the observed age distribution of paralogs

To add statistical rigor to the interpretation of K_s distributions for paralogous pairs, we modeled the expected age and K_s distributions under a constant rate birth-death model (see METHODS). Whereas recent studies have shown that evidence of paleopolyploidy is often (but not always) discernible in K_s plots for paralogous pairs (36,37,47), the accumulation of single gene duplications, variation in the rates of gene death, and error in K_s estimates have not been studied quantitatively. We take a modeling approach to account for the rate of gene death, the time since gene (or genome) duplication, and the error in K_s estimates in analyses of paralogous pairs. Our null model assumes gene birth and death are independent events, each with a constant rate over time. Under this model, the expected age distribution for paralogous pairs is a declining exponential with a decay parameter corresponding to the rate of gene death. K_s distributions derived from simulations under this model are influenced by the random nature of nucleotide substitution and error in K_s estimation. In order to formally test for deviation from a constant rates model using empirical data, we generate a null distribution for the frequency of K_s values using parameters estimated from the data for the rate of gene death and the error in K_s estimation.

Our model was also used to simulate K_s distributions for paralogous pairs arising from a mixture of single gene duplications and ancient polyploidy events. Empirical estimates of variation in K_s were based on analyses of *Arabidopsis thaliana* paralogous pairs. Figure 4-1 shows K_s distributions for data simulated with different rates of gene

death and different times since the genome duplication event. These K_s distributions contain two components; the first one is always a declining exponential distribution corresponding to "background" single gene duplications, and the second component represents paralogous pairs arising from a polyploidy event. Very recent genome duplications may be obscured by background gene duplications when the modal K_s values do not appear as distinct peaks. Conversely, increases in the number of gene deaths and variance in K_s with time render older genome duplications less detectable than younger events, and we were not able to detect a significant duplication signal for events with an expected K_s of 1.5 (Figure 4-1 C,F,I). High gene death rates also eroded the impact of genome duplications on K_s distributions (Figure 4-1 G-I). These results corroborate previous evidence that ancient genome duplication events are not always detectable in analyses of K_s distributions (36,48).

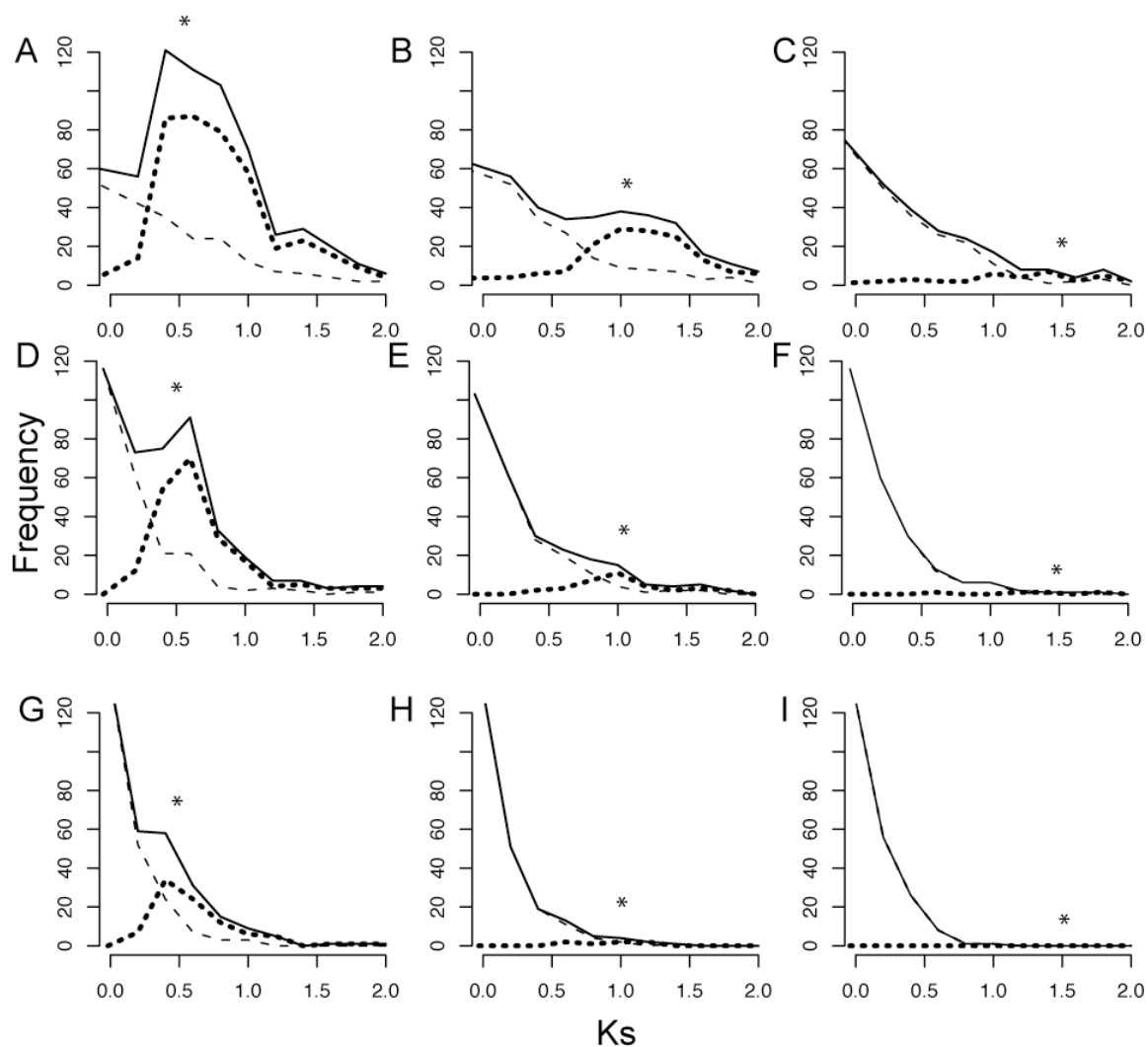


Figure 4-1: Effect of gene death rate and time of genome duplication on the K_s distribution for paralogs. A single genome duplication was simulated, where the time since duplication (corresponding to $K_s = 0.5$ in A, D, and G, 1.0 in B, E, and H, or 1.5 in C, F, and I) was indicated by a star. The death rate of duplicate pairs (δ) increases from the top row to the bottom row ($\delta = 0.67$ for A, D, G, as estimated from *Arabidopsis* data, 1.34 for B, E, H, and 2.68 for C, F, I). In each plate, the observed frequency of paralogs from the background gene duplication was plotted with a dashed line, while the distribution deriving from the genome duplication was plotted with a dotted line. The K_s distribution of all paralogs was drawn with a solid line.

Evidence of genome duplications in diverse lineages of flowering plants

Model validation: duplications detected in eudicots. EST sets from *Arabidopsis thaliana*, *Glycine max* (soybean) and *Solanum lycopersicum* (tomato) were used to validate our test of the constant-birth-death-rate model. The genome duplication histories for these species have been elucidated in several previous analyses (24,26,49-52). To make these analyses comparable to analyses of the other EST sets in this study, we randomly sampled sets of 6000 unigenes, or about 10,000 ESTs, from a much larger set of available ESTs for each of these taxa (see METHODS).

Model validation: duplications detected in eudicots. EST sets from *Arabidopsis thaliana*, *Glycine max* (soybean) and *Solanum lycopersicum* (tomato) were used to validate our test of the constant-birth-death-rate model. The genome duplication histories for these species have been elucidated in several previous analyses (24,26,49-52). To make these analyses comparable to analyses of the other EST sets in this study, we randomly sampled sets of 6000 unigenes, or about 10,000 ESTs, from a much larger set of available ESTs for each of these taxa (see METHODS).

To determine whether inference of genome-wide duplication events depends on the method of synonymous substitution estimation, we compared four methods of K_s estimation, including the original Nei-Gojobori (NG) method (53), the modified Nei-Gojobori (modified NG) method (54), the Goldman and Yang maximum likelihood (ML) method (55), and the YN00 (YN) method (56). Results were similar across all K_s estimation procedures in analyses of the *Arabidopsis* data set (Figure 4-2A). Analyses of replicate subsamples from the *Arabidopsis* unigenes gave very similar results to analyses of all paralogous pairs (Figure 4-2B)(29,30,47), suggesting that 6000 unigenes are sufficient for estimating K_s distributions for the other species in this study (Table 4-2).

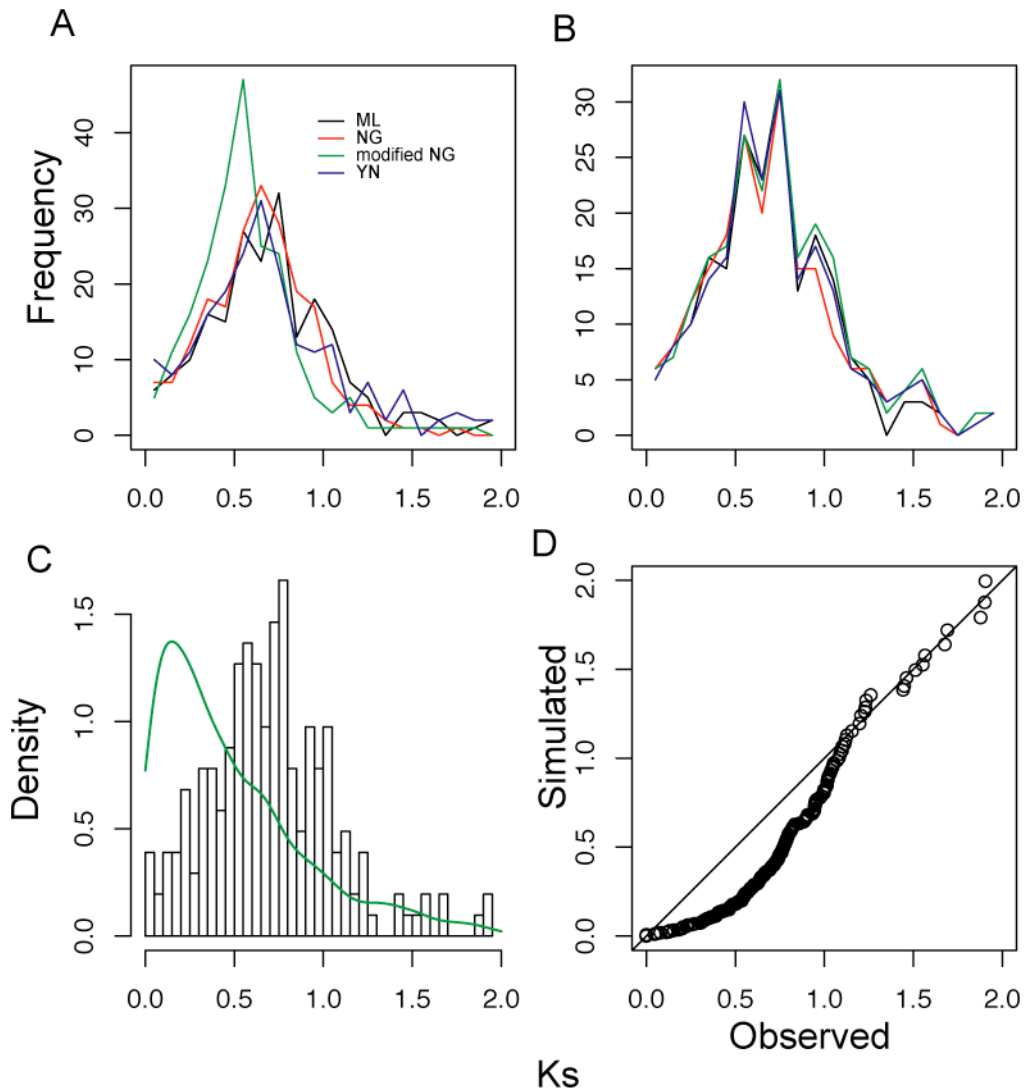


Figure 4-2: K_s distribution from a sample of *Arabidopsis* unigenes and the diagnostic test according to the constant birth-death model (null model). A. K_s estimates from four methods show strong agreement. ML: maximum likelihood method by Goldman and Yang. NG: Nei-Gojobori method. mNG: modified Nei-Gojobori method. YN: Yang and Nielsen method. These sample sizes are comparable to the unigenes available for the species sequenced in this study. B. K_s distributions for paralogs from four replicate unigene samples of 6000 sequences each. C. The density plot of observed K_s distribution and simulated data based on the null model with parameter $\delta = 0.67$. D. The Quantile-Quantile plot of observed and expected K_s values shows the poor fit of the null hypothesis that gene birth and death rates are constant ($p < 0.0001$).

We estimated the rate parameter for *Arabidopsis* data ($\delta = 0.67$) assuming a constant-birth-death model (the null model) and tested the expected distribution against the observed distribution using a chi-squared test (Figure 4-2C). The null model was rejected ($p \ll 0.0001$), and the Quantile-Quantile plot showed obvious deviation from the expected distribution of K_s values (bootstrap Kolmogorov-Smirnov test, $p \ll 0.0001$) (Figure 4-2D). Next, we applied the mixture model to estimate the median age (in K_s equivalent unit) of duplicate genes from recent or older duplication events (Table 4-3). This analysis, using ML distances, identifies two significant components, a background component with median $K_s = 0.2889$, and a prominent second component including 79% of the paralogous pairs with a median $K_s = 0.7510$ that corresponds to the polyploidy peak detected by Blanc and Wolfe (36). Similar results were obtained when the YN, NG, and modified NG K_s estimates were used, so only ML distance estimates are reported for all other analyses since they are typically less biased with lower error, especially for more divergent sequences (56). We obtained similar results to those reported in previous studies (36,37), with much smaller subsamples of ESTs (Figure 4-2B).

We next analyzed public EST sets from selected libraries of soybean and tomato. Soybean ESTs were sampled from flower, young seedling, root and other vegetative tissue libraries. Mixture model analysis suggests that 71% of the paralogs were likely to arise from a large-scale duplication (Table 4-3), which appears as a significant peak in the K_s distribution with estimated median $K_s = 0.6705$ (Figure 4-3A). This species is a relatively recent tetraploid (36,37,51,57). Thus many of the duplicate pairs assigned to the first component in the mixture model are likely derived from polyploidization rather than background single gene duplications.

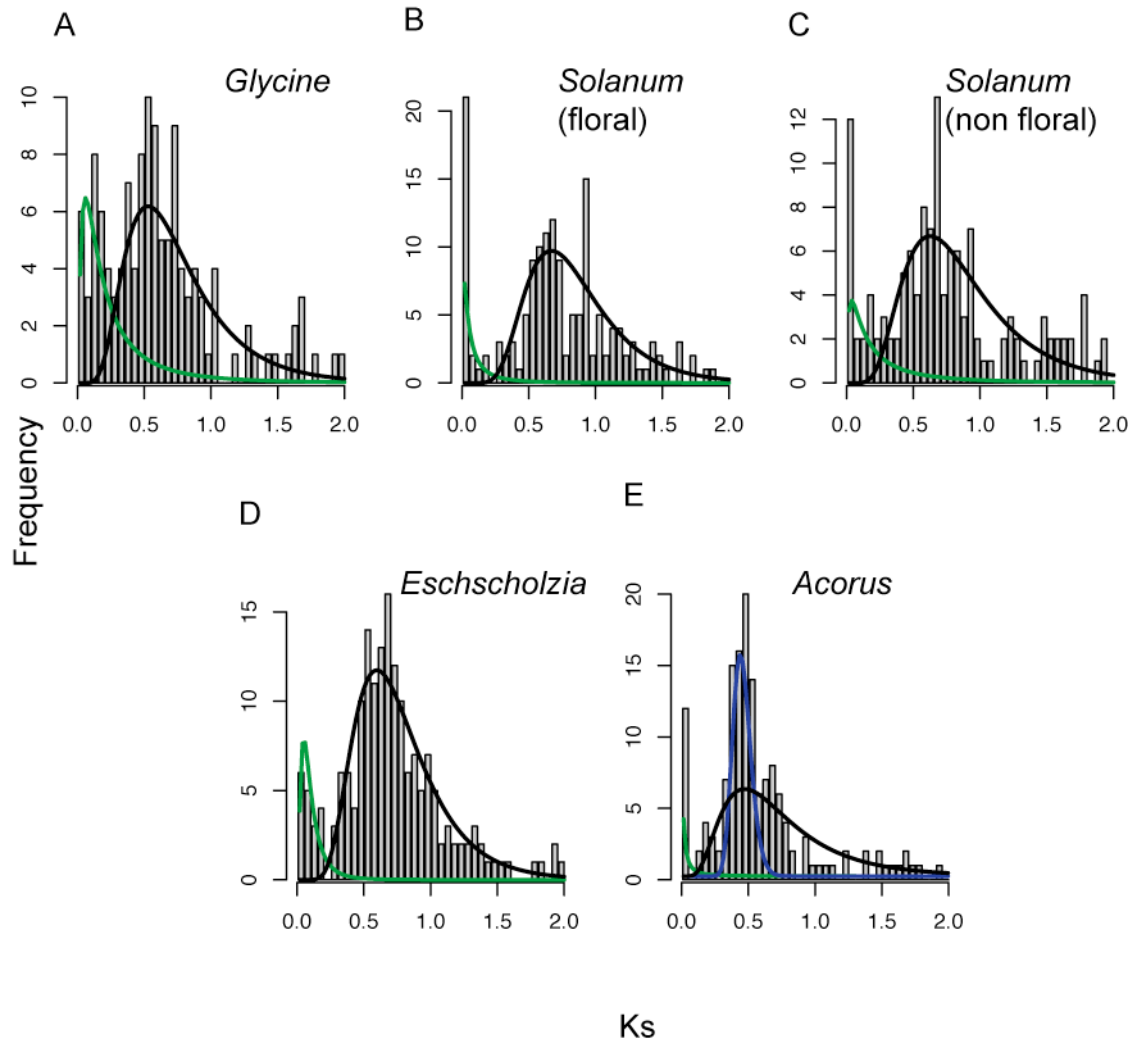


Figure 4-3: K_s distributions of paralogs in selected angiosperm species, with overlaying fitted density from mixture model analysis, suggesting paleopolyploidy in eudicots and monocots. Each fitted line indicates a subpopulation in the mixture. Green, the first component, corresponds to the paralogs from background gene duplications; other colors indicate the estimated median K_s for ancient duplications. Blue, $K_s < 0.5$; black, $0.5 < K_s < 1.0$; yellow, $K_s > 1.0$. A. *Glycine max* (soybean). B-C, *Solanum lycopersicum* (tomato), data from floral tissue (B) and non-floral tissue (C). D. a basal eudicot, *Eschscholzia californica* (California poppy). E. a basal monocot, *Acorus americanus*.

Results for tomato also suggest large-scale duplications which account for over 90% of paralogs. Moreover, the distributions for paralogous gene pairs sampled from two tissue sources (floral vs. non-floral organs) were similar (Figure 4-3B,C), and in

agreement with previous analyses based on all duplicate gene pairs in this species (median $K_s = 0.277$ and 0.632) (37). Together, our tests found strong signals of deviation from the null model, and as expected, mixture model analyses suggest ancient polyploidy events in *Arabidopsis*, *Glycine* and *Solanum*. Further, our results suggest that unbiased K_s distributions can be obtained from as few as 6000 unigenes sampled from complex cDNA libraries derived from developing floral organs.

Ancient polyploidy in a basal eudicot. *Eschscholzia californica* (California poppy, Papaveraceae) is a member of Ranunculales, the sister lineage to all other eudicots (58-61). Analysis of the K_s distribution of 149 pairs of *Eschscholzia* paralogs rejected the constant birth-and-death model ($P \ll 0.0001$) and two components in the distribution were identified by the mixture model. The second component dominated the distribution, with 89% of the duplicate pairs (Figure 4-3C), providing the first strong evidence of probable ancient genome duplication in a basal eudicot. Phylogenetic analyses of duplicated *AGAMOUS* and *AP3* homologs (44,62,63) suggest that this duplication event occurred after the split between Ranunculales and core eudicots. Thus, the genome-wide duplication event evident in the *Eschscholzia* paralogous pairs was probably independent of the genome duplications that have been inferred from analyses of the *Arabidopsis* genome (24,26,47).

Basal monocot. *Acorus americanus* (Acoraceae, Acorales) represents the sister lineage to all other monocots (58-61,64,65). Three components were identified in the paralogous pairs by the mixture model approach. The second component, accounting for 33% of all duplicates, was shown as a sharp peak in the K_s distribution, while the third component, containing 65% of the duplicates, appeared as a broader peak (Figure 4-3E).

Based on the distinct modes observed in raw K_s distribution, we hypothesize that the second and third components estimated in the mixture model represent two distinct large-scale duplication events. This hypothesis will be tested in future phylogenetic analyses of well-sampled gene families.

Magnoliids. Both shared and lineage-specific genome duplications were inferred from analyses of unigenes from three magnoliid species: *Liriodendron tulipifera* (Magnoliaceae, Magnoliales), *Persea americana* (Lauraceae, Laurales) and *Saruma henryi* (Aristolochiaceae, Piperales). A total of 92 paralogous pairs were detected in the *Liriodendron* unigene set. The constant birth-death model was rejected ($p < 0.001$), and a mixture of two components was identified in the K_s distribution, with the second component being dominant (Figure 4-4A). The null birth-death model was also rejected in the *Persea americana* (avocado) analysis ($p \ll 0.0001$) with 196 paralogous gene pairs. The optimal mixture model also included two components very similar to those seen for *Liriodendron* (Figure 4-4B; Table 4-3).

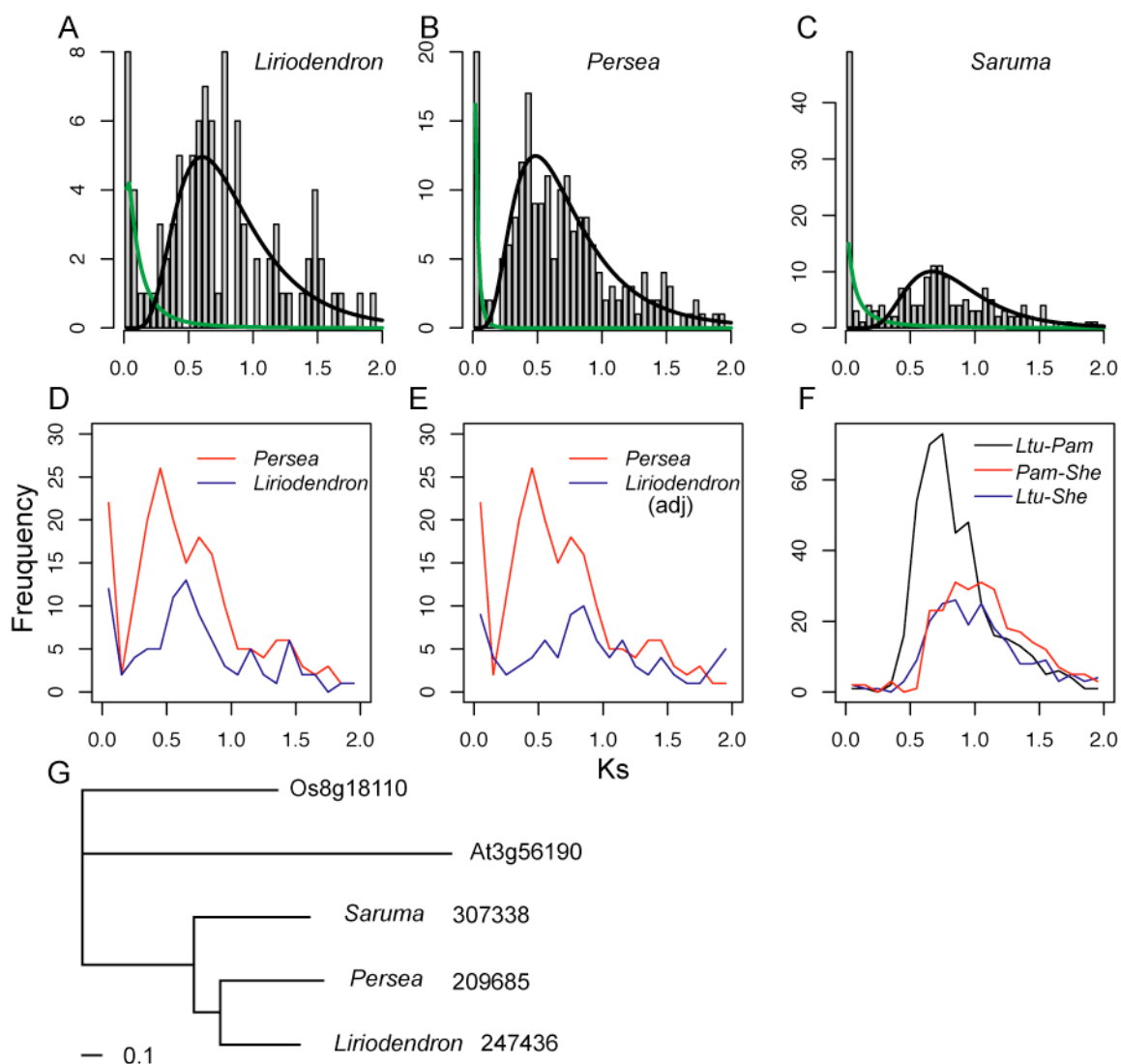


Figure 4-4: K_s distributions of paralogs and orthologs among magnoliids, suggesting independent duplications and possibly shared genome duplication events in Laurales (*Persea*) and Magnoliales (*Liriodendron*). A-C. The distribution in (A) *Liriodendron*, (B) *Persea* and (C) *Saruma*, with fitted lines based on the mixture model analysis. D. The K_s distribution for *Liriodendron* and *Persea*, without scaling for rate differences between lineages. E. K_s distribution for paralogs in *Liriodendron* after rate calibration (= adj.), compared with that of *Persea*, suggesting recent independent duplication and older shared genome scale duplications. F. K_s distribution for orthologs of two magnoliid species. Ltu, *Liriodendron*; Pam, *Persea*; She, *Saruma*. G. Phylogeny of one representative orthologous gene set used for relative rate estimate. The branch length shows the estimated relative rate of synonymous evolution in respective species.

To determine whether the duplication events inferred from the K_s distributions of *Liriodendron* and *Persea* represented events in a common ancestor, we first computed the median K_s of putatively orthologous gene pairs (total 408 pairs identified as reciprocal best hits in BLAST searches) and compared the median K_s for orthologs with K_s values for paralogous pairs within each species. The K_s distribution of putative ortholog pairs showed a single major component (median = 0.8057, variance = 0.0858; Figure 4-4F), which was slightly older than the probable genome duplication observed in *Persea* (median = 0.6464, variance = 0.1197) ($p < 0.0001$, Wilcoxon test). The timing of the duplication event inferred from the *Liriodendron* K_s distribution (median = 0.7616, variance = 0.1328) relative to the divergence of the *Persea* and *Liriodendron* lineages was ambiguous ($p = 0.35$), and direct comparison of the *Persea* and *Liriodendron* K_s distribution may be confounded by unequal substitution rates.

To account for variation in synonymous substitution rates between the *Persea* and *Liriodendron* lineages, we aligned putatively orthologous genes from *Liriodendron*, *Persea* and *Saruma* and estimated K_s values for each lineage on a phylogeny. We examined 19 putative orthologous gene sets in the three species with alignments of at least 400 base pairs for all taxa (see METHODS) and found that the synonymous substitution rate on the lineage leading to *Liriodendron* was slower on average than the rate on the lineage leading to *Persea*. For example, in the tree for the orthologous set shown in Figure 4-4G, the branch length (in K_s units) for the branch to *Persea* is 1.31 times the branch to *Liriodendron*. The ratio of synonymous substitutions on the *Persea* branch relative to the *Liriodendron* branch ranged from 0.86 to 2.68, and the ratio was greater than one in 16 of 19 cases. When *Liriodendron* paralog K_s values were multiplied

by the median branch-length ratio, 1.29, the peak in the scaled *Liriodendron* K_s distributions matched an older, but non-significant peak in the *Persea* K_s distribution (Figure 4-4E). Taken together, these analyses suggest that the prominent peak in the *Liriodendron* K_s distribution (median = 0.82) represents a duplication event in the common ancestral genome of Magnoliales and Laurales that was not identified as a distinct component in the mixture model for the *Persea* K_s distribution. In line with the comparison of K_s values for *Persea* paralogs and putative *Liriodendron-Persea* orthologs, we interpret the dominant peak in the *Persea* K_s distribution to represent a genome-scale duplication event that occurred after the divergence of Magnoliales and Laurales. This hypothesis needs to be tested with additional data.

Saruma henryi is a member of Piperales, which is sister to the Magnoliales and Laurales clade (58-61). The K_s distribution of *Saruma* paralogs showed a distinct peak with median $K_s = 0.7927$ (Figure 4-4C; Table 4-3). This is lower than the median K_s for 202 *Saruma – Liriodendron* ortholog pairs (0.9555, $p=0.0001$) and the median K_s for 254 putative *Saruma – Persea* ortholog pairs (1.0121, $p<0.0001$; Figure 4-4F). We therefore surmise that the peak in the K_s distribution of *Saruma* paralogous pairs represents a large-scale duplication in Piperales after divergence from the Magnoliales and Laurales lineages.

Basal-most angiosperms. *Amborella trichopoda* (Amborellaceae) and the water lilies (Nymphaeales) are either successive sister lineages to all other extant angiosperms or together form a clade that is sister to the rest of the angiosperms (60,66,67). The K_s distribution for a total of 69 *Amborella* paralogous pairs appeared to follow an exponential distribution, but the uniform birth-death process was rejected ($p < 0.01$;

Figure 4-5A). However, the mixture model analysis only identified one component containing all of the gene pairs (Table 4-3). The Nymphaeales are represented in this study by *Nuphar advena*. A total of 138 paralogous pairs were identified and the resulting K_s distribution did not fit the constant birth-death model ($p < 0.01$). Three mixture components were estimated from the K_s distribution. The second component, accounting for 56% of the paralogous pairs, provided strong evidence for ancient polyploidy in the history of the *Nuphar* genome (Figure 4-5B). The third component, with a median K_s of 1.3273, may represent the oldest genome duplication to be detected in analyses of K_s distributions. The median K_s for the third component was not distinguishable from the median K_s value for putative *Amborella* – *Nuphar* orthologs (Figure 4-5C)(median K_s [orthologs] = 1.24, variance 0.1918, based on 113 putatively orthologous sequence pairs; $p = 0.05$, two-sample t test on the $\log K_s$ [orthologs] and $\log K_s$ [third component of *Nuphar* paralogs]). Therefore, the third component in the *Nuphar* K_s distribution may correspond to a polyploidy event that occurred at approximately the time of the divergence between the *Amborella* and *Nuphar* lineages (see DISCUSSION below).

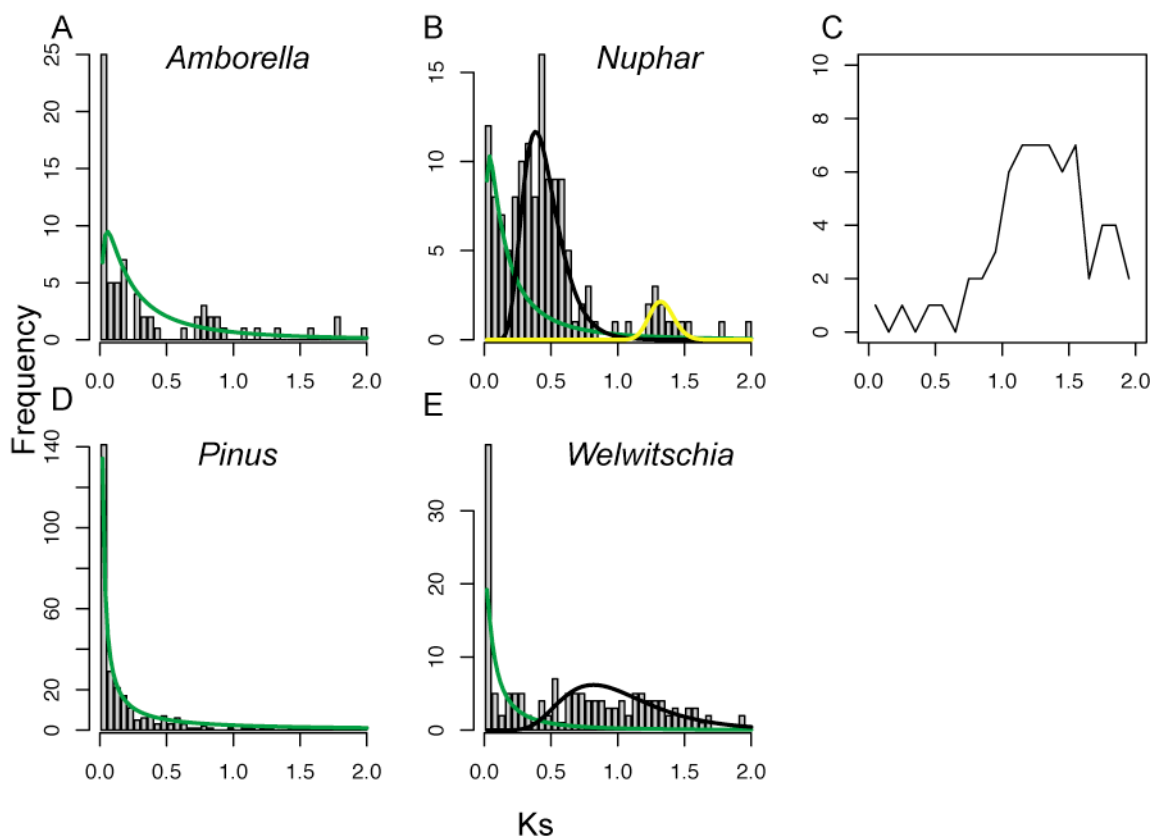


Figure 4-5: K_s distributions suggest possible genome duplications in basal angiosperms, no evidence for genome duplication events in some gymnosperm species. A. K_s distribution in *Amborella*, a basal-most angiosperm. No significant large-scale duplication is detected. B. Three distinct components in the K_s distribution for *Nuphar*, also a basal-most angiosperm, suggest at least two large-scale duplications in the genome. C. K_s distribution for putative orthologs between *Amborella* and *Nuphar*. D. *Pinus taeda* (loblolly pine) paralogous pairs follow the null model (METHODS). E. K_s distribution for paralogs in a gymnosperm *Welwitschia*. The second component based on the mixture model analysis suggests possible continuous duplications in the genome history.

Gymnosperms. We obtained 52,527 unigenes for *Pinus taeda* (loblolly pine) from PlantGDB (68), and a random sample of 6000 unigenes was drawn to match the sample size for other species we investigated. The K_s distribution showed a clear monotonous decay of paralogs with increasing age and no detectable sign of genome duplication in the recent history ($p = 0.16$; Figure 4-5D). The frequency distribution for

all paralogous pairs was essentially identical. The analysis of *Pinus pinaster* yielded a similar exponential distribution (Table 4-3).

The constant-birth-death model was rejected for *Welwitschia* ($p < 0.01$), and a mixture analysis of the K_s distribution identified two components (Figure 4-5E). The second component, corresponding to the heavy right hand tail of the distribution, may represent one or more ancient duplication events, or a reduced rate of gene death for older duplicates.

Discussion

While previous studies using K_s distributions provided significant insights into genome duplications (29,32,36,37), we introduced a model-based statistical test of deviation from a constant rate of gene birth and death. The test accounted for estimation error in K_s values. The birth-death model developed here for duplicated genes is a natural extension of stochastic birth and death models that have been widely used in population genetics, population dynamics and phylogenetic analysis of gene families (69). Simulations based on this model have allowed us to investigate how specified death rates and duplication time result in K_s distributions with (or without) secondary peaks or heavy tails (e.g., Figure 4-1). This model can be extended to incorporate variable rates of gene birth or death over time, and to the extreme, an instant burst of gene birth corresponding to a whole-genome duplication. Although we could not exclude partial and segmental duplications, the model has been validated with genomes with known duplication history and such whole-genome events were most likely to be detected.

We found that three major factors influence the frequency and observed divergence of paralogous pairs arising from genome-wide duplications. Time since the duplication event, the rate of gene death, and the background rate of gene birth all influence the observed K_s distribution at present time. Very recent genome duplication events are associated with K_s values for the resulting paralogous pairs that are indistinguishable from those of background single-gene duplications using EST data. For example, polyploidy is not clearly evident in the K_s distribution for hexaploid wheat because there has been little divergence among the parental or homeologous gene copies, and the range of divergence for allelic variants were not distinct from paralogs arising from recent gene duplications (36). At the same time, evidence of very ancient genome duplications is eroded as synonymous substitutions reach saturation and variance in K_s increases. This may be the case in K_s plots for wheat, maize, rice, and barley, in which evidence for a genome duplication event some 50-60 million years ago (mya) in the common ancestor of all major grain lineages has been obscured (36,48). Detection of very old duplication events in K_s distributions is especially difficult in species with high synonymous substitution rates. Conversely, evidence for the oldest detectable genome-wide duplications will be found in K_s distributions for species with the slowest substitution rates (see below).

Concurrent expansion of a few gene families could lead to moderate deviations from the null model. This is especially true if ancient duplication events are overrepresented in the set of sampled paralogous pairs, or if major adaptive radiations of individual gene families preceded or accompanied the radiation of members of the lineage under study. In this study, we avoided over counting of ancient gene duplications

by constraining genes to be included in only one paralogous pair. Our analysis of duplicated *Arabidopsis* genes verified that this approach produced K_s distributions similar to those of previous studies that implemented more elaborate corrections for gene family expansions (47). Moreover, sampled paralogous genes were not particularly biased towards large gene families. Whereas most sampled duplicate genes belonged to the housekeeping functional categories, such as protein synthesis, proteolysis and energy metabolism, none of the duplicate gene sets were dominated by a single gene family. Several transcription factor families were also identified in our paralog pairs, but again no family accounted for more than a few percent of the duplicate gene pairs.

Our results for *Persea* (Lauraceae) and *Liriodendron* (Magnoliaceae) corroborate previous evidence of ancient polyploidy from isozyme studies (70). Soltis and Soltis (1990) found that 25-29% of the loci investigated were duplicated in both families, and hence could have arisen via polyploidy. All members of Magnoliaceae examined shared the very same isozyme duplications (PGI, TPI, 6PGD) while the Lauraceae species shared a different suite of isozyme duplications (PGM, TPI, 6PGD, GDH). These were interpreted as evidence for independent paleopolyploid events occurring very early in the evolutionary history of Magnoliaceae and Lauraceae. The *Persea* and *Liriodendron* paralogous genes suggest polyploidy in a common ancestor at least 100 mya (71) followed by a second round of polyploidy in the *Persea* lineage (Figure 4-4E), but this hypothesis must be tested with analyses of additional gene family phylogenies. If this scenario is correct, the duplicated isozyme loci observed in the Magnoliaceae and Lauraceae may have arisen from a polyploidy event that predated the separation of the two families (72).

Over time, the accumulation of nucleotide substitutions can become saturated, and therefore lineages with slow synonymous substitutions rates will allow a longer view into genome history relative to lineages with faster substitution rates. It is estimated that the synonymous substitution in palm (2.61×10^{-9} synonymous substitutions/per year) (73) is only about half of the rate reported for grasses, eudicots (29) and grass-eudicot comparisons (74). We infer a similarly slow substitution rate for other basal angiosperms based on the Magnoliales – Laurales divergence as a calibration point. We estimated a synonymous site divergence of $K_s = 0.7$ for *Liriodendron* and *Persea* ortholog pairs (Figure 4-4F). Using a divergence date estimate of ca. 116 mya for the Magnoliales - Laurales split (71), we estimate an average synonymous substitution rate of 3.02×10^{-9} synonymous substitutions/per year. The low substitution rate in *Liriodendron* and *Persea* may be explained in part by their longer generation times relative to model eudicot and grass species.

We found that the median for the oldest component in the *Nuphar* K_s distribution is close to the median K_s for putative *Amborella-Nuphar* orthologs (median $K_s = 1.24$; Figure 4-5C). This level of divergence is compatible with the synonymous divergence for the very early duplication in *Arabidopsis* (i.e., γ duplication) (26,42,47). Direct dating of the early *Nuphar* peak based on the K_s data is challenging because of uncertainty in the branching relationships between *Amborella*, *Nuphar*, and the rest of the angiosperms, and the possibility of additional rate variation as were seen for magnoliids. We adopted two approaches to date the earliest event in *Nuphar*. First, using an *Amborella-Nuphar* ortholog divergence of 1.24 and a calibration range of 134–165 mya (67) gives a rate of 4.66 to 3.79×10^{-9} substitutions per silent site per year. Therefore, $K_s = 1.33$ would predict

an age range between 143-173 mya for the split between these two lineages. An alternative calculation, using the magnoliid calibration of 3.02×10^{-9} substitutions per silent site per year, leads to an estimate of 220 mya for the divergence of lineages leading to *Amborella* and *Nuphar*.

This range of age estimates supports two alternative interpretations of the *Nuphar* and *Amborella* paralog K_s distributions. The third component in the *Nuphar* K_s distribution may represent polyploidy in a common ancestor of all angiosperms (Figure 6) in agreement with recent analyses of MADS-box gene families (44,75). This scenario would require that evidence of ancient polyploidy has been sufficiently eroded as to be undetected in analyses of EST samples from *Amborella* and other angiosperm species due to gene death and/or saturation of synonymous substitutions as discussed above. For example, non-significant peaks around $K_s = 1.5$ in the *Liriodendron* and *Persea* K_s distributions (Figure 4-4A and 4-4B) may provide weak evidence of polyploidy early in angiosperm history. Alternatively, the earliest duplication peak detected in the *Nuphar* analysis may trace back to a genome duplication in the common ancestor of *Nuphar* and all extant angiosperm lineages other than *Amborella* (Figure 4-6). Such a scenario would be consistent with the hypothesis that *Amborella* is sister to all other extant angiosperms (e.g., solid line on Figure 4-6), and the extremely low proportion of duplicate genes found in the *Amborella* unigene set. This scenario also would narrow the timing of a genome duplication to about 10 million years separating the branch points for *Amborella* and all other extant angiosperm lineages (67). As discussed above, however, there have been instances where known genome duplication events have not been detected in K_s distributions (Figure 4-1)(27,36), so lack of evidence for ancient polyploidy in the

Amborella K_s distribution does not exclude the possibility of polyploidy in an ancestral genome.

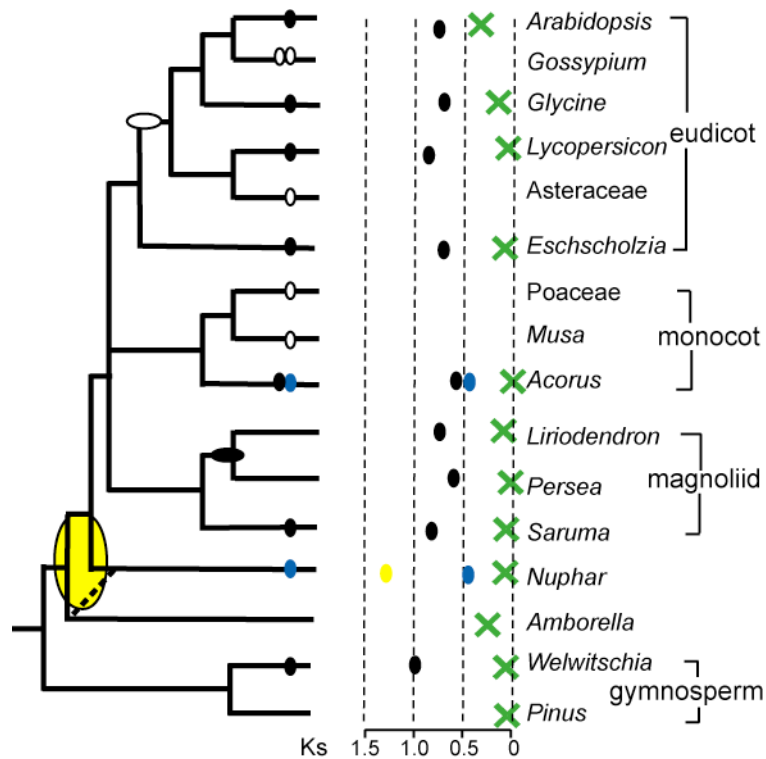


Figure 4-6: Phylogenetic summary of paleopolyploidy events estimated by the mixture model approach and their distribution in angiosperm and gymnosperm lineages. Scaled graph in center with green Xs corresponding to median K_s of pairs from background gene duplications, while blue and black ovals indicate the median K_s of possible concentrated duplications in the history of the lineage. Phylogenetic tree at left shows likely placement of detected genome scale duplications. Uncertainty in phylogenetic timing of the duplication event is indicated with a wide oval that covers possible branch points compatible with the current K_s evidence. Hollow ovals indicate duplications identified in previous studies using paralogous genes or genomic data from those lineages.

While the genomic sequences have revealed evidence of polyploidy in the Poaceae and core eudicots, the secondary peaks found in paralog K_s distributions for representatives of virtually all major angiosperm lineages, support the notion that genome duplications are common in angiosperm history and gene birth and death are important

processes in plant evolution (29). The evidence now supports the hypothesis proposed initially decades ago by Stebbins (13) that angiosperms have experienced repeated rounds of polyploidization throughout their evolutionary history. Many questions follow: How many polyploidy events separate different plant lineages? What is the typical fate of genes generated through these duplication events? And perhaps most intriguingly, have polyploidy events been important engines of angiosperm diversification. Genome scale sequencing of phylogenetically crucial angiosperm species would provide the data necessary to directly test whether the rapid diversification of flowering plants following the origin of the angiosperms (46) was associated with one or more polyploidy events.

Methods

EST sequencing and assembly EST sequences from floral cDNA libraries of seven species (*Amborella trichopoda*, spatterdock water lily [*Nuphar advena*], avocado [*Persea americana*], yellow poplar [*Liriodendron tulipifera*], wild ginger [*Saruma henryi*]), sweet flag [*Acorus americanus*], and California poppy [*Eschscholzia californica*]) are available through the Plant Genome Network (www.pgn.cornell.edu). cDNA library construction, EST sequencing and assembly were described previously (45).

Public EST sets from selected libraries for *Arabidopsis thaliana*, soybean (*Glycine max*, Williams 82) and tomato (*Solanum lycopersicum*, cultivar TA496) were downloaded from GenBank dbEST section, trimmed using `seqclean`, and assembled using CAP3 with the percent identity parameter $p = 90$ and overlap length 40bp.

Arabidopsis thaliana ESTs were from four libraries (root, flower, green silique and two to six week above-ground organs). To minimize the allelic variations in the EST sequence collection, the unigenes were mapped to the *Arabidopsis* genome, and redundant unigenes matching the same genomic locus were discarded. Only the sequences that matched the protein coding regions were retained. From this screened unigene set, we drew replicate samples with 6000 unigenes in each sample. The sample size of 6000 *Arabidopsis* unigenes approximates the number of unigenes from new EST data sets we analyzed. To compare if library sources influence the estimates, we analyzed two samples of tomato ESTs, one from floral cDNA libraries and one from vegetative cDNA libraries. The soybean ESTs were sampled from cDNA libraries of flower, young seedling, root and other vegetative organs. Unigenes for gymnosperms *Pinus taeda* and *Pinus pinaster* were downloaded from PlantGDB (68), which were built with public ESTs from all libraries. For each species, we sampled 6000 unigenes for K_s analysis.

K_s calculation for paralogs and orthologs: Paralogous pairs of sequences were identified from best reciprocal matches in all-by-all BLASTN searches. For datasets with trace files, we discarded bases with Phred (76,77) quality values lower than 20. Only sequence pairs with alignment lengths over 300 bp were used for K_s calculations. Translated sequences of unigenes generated by ESTScan (78) were aligned using MUSCLE v3.3 (79). Nucleotide sequences were then forced to fit the amino acid alignments. The K_s value for each sequence pair was calculated using the Goldman and Yang maximum likelihood method (55) implemented in codeml with the F3x4 model (80). In order to assess whether the shape of K_s distributions was dependent of the estimation procedure, the Nei-Gojobori method, the modified Nei-Gojobori method and

the YN00 method (56) were also applied on the *Arabidopsis* set. The K_s frequency in each interval size of 0.05 within the range [0, 2.0] was plotted.

The age distribution of paralogs under a constant birth-death model (the null model). We modeled the birth and death of paralogs formed by gene duplications under a constant rate birth-death model in order to test whether an observed frequency distribution of K_s values indicates deviation from this process. The duplicate genes are generated by a Poisson process at rate β , and the number of duplicate pairs decreases by age at an exponential rate δ . We can estimate the age distribution of surviving paralogs (*survivors*), total N , by considering the process as sampling gene birth over time $[0, t]$, and decide if each birth was a survivor.

The distribution for the number of survivors of age t is

$$N(t) \sim Po\left(\gamma \int_0^t \delta \exp(\delta s) ds\right) = Po(\gamma \cdot F(t)),$$

where $\gamma = \beta/\delta$, and $F(t) = 1 - \exp(-\delta t)$, the cumulative density function of exponential(δ). From this we deduce that the population size $N(\infty) = Po(\gamma)$. Furthermore, the survivor's age distribution is an empirical distribution of a sample of exponentially distributed random variables, generated with the parameter δ .

To obtain the estimate of the true age, we must consider the error of K_s with respect to the true age of paralogs. If the true age is T , then we can calculate K_s (with error) as:

$$K_s = T + (s|t) z,$$

where $s|t$ is the standard error for K_s at $T = t$, and z is a standard normal random variable. The error can be estimated from the empirical standard error given by the

PAML software.

The mean of s is expected to correlate with the time t , as older K_s estimates have larger variances. The conditional distribution of s can be approximated by exponential($2/t$). The maximum likelihood estimate of the parameter δ from the data was obtained using a grid-based method and a simulated sample under the null model were compared to the observed using a chi-squared test. A Quantile-Quantile plot (Q-Q plot) is used to visualize the difference of observed data and a simulated data set according to the null model. A strong deviation from the 45° line in the Q-Q plot suggests that the two distributions differ, and a bootstrap Kolmogorov-Smirnov test (<http://sekhon.polisci.berkeley.edu/matching/ks.boot.html>) was applied to compare the observed and expected K_s distributions. The modeling and simulation scripts are available from the author upon request.

Finite mixture model of genome duplications. In order to explore further how genome-wide duplication events influence the age distribution of paralogs and K_s distributions, we defined the *background duplication* as gene duplication under the constant rate birth-death process, and a genome duplication as an instant spike of gene birth overlaid on top of the background. We modeled changes in the K_s distribution with increasing time since the duplication event, while assuming a constant rate of gene loss (death rate) and a constant background gene duplication rate (birth rate). Each simulation included a genome duplication (which led to new duplicates n) at time t . About 5% of duplicates were allowed to escape the death process.

In all instances when we rejected the constant rates hypothesis, we surmised that the observed K_s distributions actually reflect a compound distribution generated by

variable birth and/or death rates from the time of duplication. For example a genome duplication event would generate an immediate spike in the birth of paralogs. Mixture models treat the distribution of interest as a mixture of a number of component distributions in various proportions. The EMMIX software is suitable for mixed populations where each component can be described by a Gaussian density (38)(see <http://www.maths.uq.edu.au/~gjm/emmix/emmix.html> for the Users' Guide). Following Schlueter et al. (37), we model the log-transformed K_s distribution of paralogs (The actual distribution is a mixture of log transformed exponentials and normals). Observations with $K_s < 0.005$ were excluded to avoid fitting a component to infinity (37). This truncation could also reduce the proportion of gene pairs attributed to the background duplication. We modeled the mixed populations with one to four components and repeated the EM algorithm 100 times with random starting values, as well as 10 times with k-mean start values. One restriction imposed on the variance structure of K_s is that variance increases with the mean according to the empirical estimates. The observed data could often be fitted to more than one component, with different means, variances and mixture proportions. The mixture model with the best fit was identified using the Bayesian Information Criterion (81). The mean and variance for each component (subpopulation of $\log K_s$ values) for the selected model were back-transformed to the original scale for plotting and interpretation.

Calibrating rate of synonymous substitution across lineages. When comparing K_s distributions among taxa, variation in the substitution rates among lineages must be taken into account. We used a phylogenetic approach to estimate lineage-specific synonymous substitution rates on branches leading to the magnoliids *Liriodendron*

tulipifera, *Persea americana*, and *Saruma henryi*. Orthologous genes from *Arabidopsis thaliana*, rice and the three magnoliid species were classified by InParanoid (82). Protein alignments of *Arabidopsis* and rice gene models (the TIGR *Arabidopsis thaliana* database, the TIGR rice database) were first constructed, then DNA alignments were forced to protein alignments by codon positions. A maximum likelihood tree was estimated using the HKY model in PHYML v.2.4.3 (83) for each putative ortholog set including at least 400 aligned nucleotide positions. A per site estimate of synonymous substitution (K_s) was then made for each magnoliid branch in gene phylogenies consistent with the organismal relationships ((*Liriodendron*, *Persea*) *Saruma*) using codeml in the PAML package (80). The ratio of K_s values on the *Persea* branch relative to the *Liriodendron* branch was then estimated for each gene.

Acknowledgements

The authors thank Dr. Jongmin Nam for providing code for K_s computation, Lena Scheaffer, Yi Hu, Shelia Plock for technical support on cDNA library construction and sequencing, Lukas Mueller, Dan Ilut, Teri Solow, and Steve Tanksley for the PGN Database, and additional reviewers for critical comments on the manuscript. The work was supported by NSF Plant Genome award DBI-0115684.

Table 4-1: Genome sizes and base chromosome numbers for the angiosperm and gymnosperm species in this study. Relationships among the organisms and the major lineages are indicated in Figure 4-6. Source: KBG, Kew Botanical Garden Plant C-value databases. This study: DNA content determined by flow cytometry as described in (84).

Scientific name	Common name	Family	Group	Genome size(Mbp)	chromosome number (2n)	Source
<i>Arabidopsis thaliana</i>	thale cress	Brassicaceae	rosid	157	10	KBG
<i>Glycine max</i>	soybean	Fabaceae	rosid	2205	40	KBG
<i>Solanum lycopersicum</i>	tomato	Solanaceae	asterid	1005	24	KBG
<i>Eschscholzia californica</i>	California poppy	Papaveraceae	Ranunculales	502	12	This study
<i>Acorus americanus</i>	sweet flag	Acoraceae	monocot	392	24	This study
<i>Liriodendron tulipifera</i>	yellow poplar	Magnoliaceae	magnoliid	1710	38	This study
<i>Persea americana</i>	avocado	Lauraceae	magnoliid	907	24	KBG
<i>Saruma henryi</i>		Aristolochiaceae	magnoliid	3014	52	This study
<i>Nuphar advena</i>	spatterdock water lily	Nymphaeaceae	basalmost angiosperm	2772	34	This study
<i>Amborella trichopoda</i>		Amborellaceae	basalmost angiosperm	870	26	KBG
<i>Pinus taeda</i>	loblolly pine	Pinaceae	gymnosperm	21658	24	KBG
<i>Pinus pinaster</i>	pine	Pinaceae	gymnosperm	23863	24	KBG
<i>Welwitschia mirabilis</i>		Welwitschiaceae	gymnosperm	7056	42	KBG

 Table 4-2: Summary of EST data sets and paralogous pairs identified in this study.

Scientific name	ESTs	Unigenes	Pairs with $K_s < 2$	Source
<i>Arabidopsis thaliana</i>		6000 ¹	205	dbEST
<i>Glycine max</i>	10046	6240	125	dbEST
<i>Solanum lycopersicum</i>	10028	5303	143	dbEST
<i>Eschscholzia californica</i>	9079	5713	178	PGN
<i>Acorus americanus</i>	7484	4663	149	PGN
<i>Liriodendron tulipifera</i>	9531	6520	92	PGN
<i>Persea americana</i>	8735	6183	196	PGN
<i>Saruma henryi</i>	10273	6293	184	PGN
<i>Nuphar advena</i>	8442	6205	138	PGN
<i>Amborella trichopoda</i>	8629	6099	69	PGN
<i>Pinus taeda</i>		6000 ²	276	PlantGDB
<i>Pinus pinaster</i>		6000 ³	259	PlantGDB
<i>Welwitschia mirabilis</i>	9776	6048	157	PGN

¹ Sampled from 6369 unigenes

² Sampled from 52527 unigenes

³ Sampled from 8076 unigenes

Table 4-3: Mixture model estimates for K_s distributions in each species. Initial tests against the null model (no genome duplication) were conducted, then a mixture analysis was applied to each species. The final mixture model was selected according to the Bayesian Information Criterion (BIC) and the restriction on the mean/variance structure for K_s (see METHODS). n, sample size; p, number of mixture components, -lnL, log likelihood for the mixture model. For each mixture model, the proportions (prop.) for each component (subpopulation) sum to 1.

Scientific name	n	p	lnL	BIC	median	variance	prop.
<i>Arabidopsis thaliana</i>	202	2	-162.498	351.54	0.2889	0.0473	0.21
					0.751	0.0777	0.79
<i>Glycine max</i>	123	2	-147.358	318.78	0.1873	0.0398	0.29
					0.6705	0.1066	0.71
<i>Solanum lycopersicum</i> (floral)	139	2	-118.607	261.89	0.0643	0.0066	0.09
					0.7894	0.1021	0.91
<i>Solanum lycopersicum</i> (non floral)	119	2	-122.933	269.76	0.1857	0.0547	0.15
					0.7885	0.1425	0.85
<i>Eschscholzia californica</i>	178	2	-161.652	349.21	0.0871	0.0043	0.11
					0.7098	0.087	0.89
<i>Acorus americanus</i>	139	3	-103.568	246.61	0.0118	0.001	0.01
					0.455	0.0046	0.33
					0.5813	0.1309	0.65
<i>Liriodendron tulipifera</i>	87	2	-94.046	210.42	0.1005	0.0121	0.14
					0.7616	0.1328	0.86
<i>Persea americanus</i>	186	2	-196.998	420.12	0.0234	0.0004	0.07
					0.6464	0.1197	0.93
<i>Saruma henryi</i>	146	2	-162.789	350.5	0.0913	0.0168	0.2
					0.7927	0.1066	0.8
<i>Nuphar advena</i>	134	3	-159.416	358.02	0.1746	0.0461	0.37
					0.4291	0.0202	0.56
					1.3273	0.0084	0.07
<i>Amborella trichopoda</i>	49	1	-80.676	169.14	0.2698	0.1147	1
<i>Pinus taeda</i>	227	1	-405.77	822.39	0.0839	0.0147	1
<i>Pinus pinaster</i>	240	1	-373.135	757.23	0.2499	0.0819	1
<i>Welwitschia mirabilis</i>	132	2	-181.128	386.67	0.1139	0.0271	0.35
					0.9519	0.1374	0.65

References

1. Ohno, S. (1970) *Evolution by gene duplication*. Springer-Verlag, New York.
2. Hughes, A.L. (1999) Phylogenies of developmentally important proteins do not support the hypothesis of two rounds of genome duplication early in vertebrate history. *J Mol Evol*, **48**, 565-576.
3. Friedman, R. and Hughes, A.L. (2001) Gene duplication and the structure of eukaryotic genomes. *Genome Res*, **11**, 373-381.
4. Makalowski, W. (2001) Are We Polyploids? A Brief History of One Hypothesis. *Genome Res*, **11**, 667-670.
5. Hughes, A.L. and Friedman, R. (2003) 2R or not 2R: testing hypotheses of genome duplication in early vertebrates. *J Struct Funct Genomics*, **3**, 85-93.
6. Abi-Rached, L., Gilles, A., Shiina, T., Pontarotti, P. and Inoko, H. (2002) Evidence of en bloc duplication in vertebrate genomes. *Nat Genet*, **31**, 100-105.
7. Gu, X., Wang, Y. and Gu, J. (2002) Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. *Nat Genet*, **31**, 205-209.
8. McLysaght, A., Hokamp, K. and Wolfe, K.H. (2002) Extensive genomic duplication during early chordate evolution. *Nat Genet*, **31**, 200-204.
9. Dehal, P. and Boore, J.L. (2005) Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biology*, **3**, e314.
10. Bogart, J.P. (1979) Evolutionary implications of polyploidy in amphibians and reptiles. *Basic Life Sci*, **13**, 341-378.
11. Wolfe, K.H. and Shields, D.C. (1997) Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, **387**, 708-713.
12. Kellis, M., Birren, B.W. and Lander, E.S. (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature*, **428**, 617-624.
13. Stebbins, G.L. (1950) *Variation and evolution in plants*. Columbia University Press, New York.
14. Grant, V. (1981) *Plant Speciation*. Columbia University Press, New York.
15. Soltis, D.E. and Soltis, P.S. (1999) Polyploidy: recurrent formation and genome evolution. *Trends Ecol Evol*, **14**, 348-352.
16. Müntzing, A. (1936) The evolutionary significance of autopolyploidy. *Hereditas*, **21**, 263-378.
17. Darlington, C.D. (1937) *Recent advances in cytology*. P. Blakiston's Son & Co., Philadelphia.
18. Grant, V. (1963) *The origin of adaptations*. Columbia University Press, New York.
19. Masterson, J. (1994) Stomatal size in fossil plants: evidence for polyploidy in majority of angiosperms. *Science*, **264**, 421-424.

20. Otto, S.P. and Whitton, J. (2000) Polyploid incidence and evolution. *Ann Rev Genet*, **34**, 401-437.
21. deWet, J.M. (1979) Origins of polyploids. *Basic Life Sci*, **13**, 3-15.
22. Liu, B. and Wendel, J.F. (2003) Epigenetic phenomena and the evolution of plant allopolyploids. *Mol Phylogenet Evol*, **29**, 365-379.
23. Bennett, M.D., Leitch, I.J., Price, H.J. and Johnston, J.S. (2003) Comparisons with *Caenorhabditis* (approximately 100 Mb) and *Drosophila* (approximately 175 Mb) using flow cytometry show genome size in *Arabidopsis* to be approximately 157 Mb and thus approximately 25% larger than the *Arabidopsis* genome initiative estimate of approximately 125 Mb. *Ann Bot (Lond)*, **91**, 547-557.
24. Vision, T.J., Brown, D.G. and Tanksley, S.D. (2000) The origins of genomic duplications in *Arabidopsis*. *Science*, **290**, 2114-2117.
25. Simillion, C., Vandepoele, K., Van Montagu, M.C., Zabeau, M. and Van de Peer, Y. (2002) The hidden duplication past of *Arabidopsis thaliana*. *Proc Natl Acad Sci U S A*, **99**, 13627-13632.
26. Bowers, J.E., Chapman, B.A., Rong, J. and Paterson, A.H. (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature*, **422**, 433-438.
27. Paterson, A.H., Bowers, J.E., Chapman, B.A., Peterson, D.G., Rong, J. and Wicker, T.M. (2004) Comparative genome analysis of monocots and dicots, toward characterization of angiosperm diversity. *Curr Opin Biotechnol*, **15**, 120-125.
28. Yu, J., Wang, J., Lin, W., Li, S., Li, H., Zhou, J., Ni, P., Dong, W., Hu, S., Zeng, C. *et al.* (2005) The Genomes of *Oryza sativa*: a history of duplications. *PLoS Biol*, **3**, e38.
29. Lynch, M. and Conery, J.S. (2000) The evolutionary fate and consequences of duplicate genes. *Science*, **290**, 1151-1155.
30. Blanc, G., Hokamp, K. and Wolfe, K.H. (2003) A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. *Genome Res*, **13**, 137-144.
31. Cannon, S.B., Mitra, A., Baumgarten, A., Young, N.D. and May, G. (2004) The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*. *BMC Plant Biol*, **4**, 10.
32. Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.L. and Postlethwait, J. (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, **151**, 1531-1545.
33. Adams, K.L., Cronn, R., Percifield, R. and Wendel, J.F. (2003) Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proc Natl Acad Sci U S A*, **100**, 4649-4654.
34. Wang, J.P., Lindsay, B.G., Leebens-Mack, J., Cui, L., Wall, K., Miller, W.C. and dePamphilis, C.W. (2004) EST clustering error evaluation and correction. *Bioinformatics*, **20**, 2973-2984.
35. Li, W.H. and Grauer, D. (1991) *Fundamentals of Molecular Evolution*. Sinauer Associates, Sanderland, MA.

36. Blanc, G. and Wolfe, K.H. (2004) Widespread paleopolyploidy in model plant species inferred from age distribution of duplicate genes. *Plant Cell*, **16**, 1667-1678.
37. Schlueter, J.A., Dixon, P., Granger, C., Grant, D., Clark, L., Doyle, J.J. and Shoemaker, R.C. (2004) Mining EST databases to resolve evolutionary events in major crop species. *Genome*, **47**, 868-876.
38. McLachlan, G.J., Peel, D., Basford, K.E. and Adams, P. (1999) The EMMIX software for the fitting of mixtures of normal and t-components. *J Stat Softw*, **4**, 2.
39. Wang, H.C., Singer, G.A. and Hickey, D.A. (2004) Mutational bias affects protein evolution in flowering plants. *Mol Biol Evol*, **21**, 90-96.
40. Bierne, N. and Eyre-Walker, A. (2003) The problem of counting sites in the estimation of the synonymous and nonsynonymous substitution rates: implications for the correlation between the synonymous substitution rate and codon usage bias. *Genetics*, **165**, 1587-1597.
41. Hilu, K.W. (1993) Polyploidy and the evolution of domesticated plants. *Am. J. Bot.*, **80**, 2521-2528.
42. De Bodt, S., Maere, S. and Van de Peer, Y. (2005) Genome duplication and the origin of angiosperms. *Trends Ecol Evol*, **In press**.
43. Zahn, L.M., Leebens-Mack, J., DePamphilis, C.W., Ma, H. and Theissen, G. (2005) To B or Not to B a flower: the role of DEFICIENS and GLOBOSA orthologs in the evolution of the angiosperms. *J Hered*, **96**, 225-240.
44. Zahn, L.M., Kong, H., Leebens-Mack, J.H., Kim, S., Soltis, P.S., Landherr, L.L., Soltis, D.E., Depamphilis, C.W. and Ma, H. (2005) The evolution of the SEPALLATA subfamily of MADS-box genes: a preangiosperm origin with multiple duplications throughout angiosperm history. *Genetics*, **169**, 2209-2223.
45. Albert, V.A., Soltis, D.E., Carlson, J.E., Farmerie, W.G., Wall, P.K., Ilut, D.C., Solow, T.M., Mueller, L.A., Landherr, L.L., Hu, Y. *et al.* (2005) Floral gene resources from basal angiosperms for comparative genomics research. *BMC Plant Biol*, **5**, 5.
46. Darwin, C.D. (1903) *More letters of Charles Darwin*. John Murray, London.
47. Maere, S., De Bodt, S., Raes, J., Casneuf, T., Van Montagu, M., Kuiper, M. and Van de Peer, Y. (2005) Modeling gene and genome duplications in eukaryotes. *Proc Natl Acad Sci U S A*, **102**, 5454-5459.
48. Paterson, A.H., Bowers, J.E. and Chapman, B.A. (2004) Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc Natl Acad Sci U S A*, **101**, 9903-9908.
49. Blanc, G., Barakat, A., Guyot, R., Cooke, R. and Delseny, M. (2000) Extensive duplication and reshuffling in the Arabidopsis genome. *Plant Cell*, **12**, 1093-1101.
50. Grant, D., Cregan, P. and Shoemaker, R.C. (2000) Genome organization in dicots: genome duplication in Arabidopsis and synteny between soybean and Arabidopsis. *Proc Natl Acad Sci U S A*, **97**, 4168-4173.
51. Shoemaker, R.C., Polzin, K., Labate, J., Specht, J., Brummer, E.C., Olson, T., Young, N., Concibido, V., Wilcox, J., Tamulonis, J.P. *et al.* (1996) Genome duplication in soybean (*Glycine* subgenus soja). *Genetics*, **144**, 329-338.

52. Ku, H.M., Vision, T., Liu, J. and Tanksley, S.D. (2000) Comparing sequenced segments of the tomato and Arabidopsis genomes: large-scale duplication followed by selective gene loss creates a network of synteny. *Proc Natl Acad Sci U S A*, **97**, 9121-9126.
53. Nei, M. and Gojobori, T. (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol*, **3**, 418-426.
54. Zhang, J., Rosenberg, H.F. and Nei, M. (1998) Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proc Natl Acad Sci U S A*, **95**, 3708-3713.
55. Goldman, N. and Yang, Z. (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol*, **11**, 725-736.
56. Yang, Z. and Nielsen, R. (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol*, **17**, 32-43.
57. Lavin, M., Herendenn, P. and Wojciechowski, M.F. (2005) Evolutionary rates analysis of Leguminosae implicates a rapid diversification of lineages during the Tertiary. *Syst Biol*, **54**, 575-594.
58. Soltis, P.S., Soltis, D.E., Zanis, M.J. and Kim, S. (2000) Basal lineages of angiosperms: relationships and implications for floral evolution. *Int J Plant Sci*, **161**, S97-S107.
59. Soltis, D.E., Soltis, P.S. and Zanis, M.J. (2002) Phylogeny of seed plants based on evidence from eight genes. *Am J Bot*, **89**, 1670-1681.
60. Zanis, M.J., Soltis, D.E., Soltis, P.S., Mathews, S. and Donoghue, M.J. (2002) The root of the angiosperms revisited. *Proc Natl Acad Sci U S A*, **99**, 6848-6853.
61. Borsch, T., Hilu, K.W., Quandt, D., Wilde, V., Neinhuis, C. and Barthlott, W. (2003) Noncoding plastid trnT-trnF sequences reveal a well resolved phylogeny of basal angiosperms. *J Evol Biol*, **16**, 558-576.
62. Kramer, E.M., Dorit, R.L. and Irish, V.F. (1998) Molecular evolution of genes controlling petal and stamen development: duplication and divergence within the APETALA3 and PISTILLATA MADS-box gene lineages. *Genetics*, **149**, 765-783.
63. Kramer, E.M. and Irish, V.F. (1999) Evolution of genetic mechanisms controlling petal development. *Nature*, **399**, 144-148.
64. Duvall, M.R., Learn, G.H., Jr., Eguiarte, L.E. and Clegg, M.T. (1993) Phylogenetic analysis of rbcL sequences identifies *Acorus calamus* as the primal extant monocotyledon. *Proc Natl Acad Sci U S A*, **90**, 4641-4644.
65. Hilu, K.W., Borsch, T., Mueller, K., Soltis, D.E., Soltis, P.S., Savolainen, V., Chase, M.W., Powell, M., Alice, L.A., Evans, R. *et al.* (2003) Angiosperm phylogeny based on matK sequence information. *Am J Bot*, **90**, 1758-1776.
66. Stefanovic, S., Rice, D.W. and Palmer, J.D. (2004) Long branch attraction, taxon sampling, and the earliest angiosperms: Amborella or monocots? *BMC Evol Biol*, **4**, 35.
67. Leebens-Mack, J., Raubeson, L.A., Cui, L., Kuehl, J.V., Fourcade, M.H., Chumley, T.W., Boore, J.L., Jansen, R.K. and Depamphilis, C.W. (2005)

- Identifying the Basal Angiosperm Node in Chloroplast Genome Phylogenies: Sampling One's Way Out of the Felsenstein Zone. *Mol Biol Evol*, **22**, 1948-1963.
68. Dong, Q., Schlueter, S.D. and Brendel, V. (2004) PlantGDB, plant genome database and analysis tools. *Nucleic Acids Res*, **32**, D354-359.
 69. Karev, G.P., Wolf, Y.I., Berezovskaya, F.S. and Koonin, E.V. (2004) Gene family evolution: an in-depth theoretical and simulation analysis of non-linear birth-death-innovation models. *BMC Evol Biol*, **4**, 32.
 70. Soltis, D.E. and Soltis, P.S. (1990) Evidence of ancient polyploidy in primitive angiosperms. *Syst Bot*, **15**, 328-328.
 71. Bell, C.D., Soltis, D.E. and Soltis, P.S. (2005) The age of the angiosperms: a molecular timescale without a clock. *Evolution*, **59**, 1245-1258.
 72. Brysting, A.K. and Borgen, L. (2000) Isozyme analysis of the *Cerastium alpinum* C-arcticum complex (Caryophyllaceae) supports a splitting of C-arcticum Lange. *Plant Syst Evol*, **220**, 199-221.
 73. Gaut, B.S., Morton, B.R., McCaig, B.C. and Clegg, M.T. (1996) Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene *Adh* parallel rate differences at the plastid gene *rbcL*. *Proc Natl Acad Sci U S A*, **93**, 10274-10279.
 74. Wolfe, K.H., Li, W.H. and Sharp, P.M. (1987) Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proc Natl Acad Sci U S A*, **84**, 9054-9058.
 75. Kim, S., Yoo, M.-J., Albert, V.A., Farris, J.S., Soltis, P.S. and Soltis, D.E. (2004) Phylogeny and diversification of B-function MADS-box genes in angiosperms: evolutionary and functional implications of a 260-million-year-old duplication. *Am J Bot*, **91**, 2102-2118.
 76. Ewing, B., Hillier, L., Wendl, M.C. and Green, P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res*, **8**, 175-185.
 77. Ewing, B. and Green, P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res*, **8**, 186-194.
 78. Iseli, C., Jongeneel, C.V. and Bucher, P. (1999) ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Proc Int Conf Intell Syst Mol Biol*, 138-148.
 79. Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.
 80. Yang, Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci*, **13**, 555-556.
 81. Schwarz, G. (1978) Estimating the Dimension of a Model. *Annals of Statistics*, **6**, 461-464.
 82. Remm, M., Storm, C.E. and Sonnhammer, E.L. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol*, **314**, 1041-1052.
 83. Guindon, S. and Gascuel, O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*, **52**, 696-704.

84. Wang, W., Tanurdzic, M., Luo, M., Sisneros, N., Kim, H.R., Weng, J.K., Kudrna, D., Mueller, C., Arumuganathan, K., Carlson, J. *et al.* (2005) Construction of a bacterial artificial chromosome library from the spikemoss *Selaginella moellendorffii*: a new resource for plant comparative genomics. *BMC Plant Biol*, **5**, 10.

Chapter 5

Large number of genes expressed in flowers of basal angiosperms: inference from EST data

Preface

This is a manuscript in preparation. Major contributors are Liying Cui, Ji-Ping Wang, Bruce G. Lindsay and Claude W. dePamphilis. LC conducted the data analysis, participated in the development of the ESTstat software and wrote the manuscript. JPW designed the NPMLE method and developed software. BGL provided suggestions on the simulation. CWD conceived the study and supervised the draft writing. Part of the content was presented as a poster for Plant and Animal Genomes XIII, January 2005.

Abstract

Gene expression in a tissue or organ is developmentally regulated. The reproductive structure in angiosperms, the flower, has been a focus of genetic and molecular biology studies which uncovered a regulatory network of transcription factors. However, the expression of all genes in flower development is not well defined. Early studies on tobacco implied perhaps several thousand RNA species were present in floral organs. ESTs from a broad range of flowering plants, especially recently generated from basal angiosperm flowers, provided rich information on gene expression distribution in flowers among phylogenetically early-diverging lineages. We develop a robust estimator

for the number of expressed genes that is implemented in the software ESTstat. The method corrects for EST clustering errors that usually lead to overestimation of the number of unique transcripts (unigenes) sequenced. The model plant *Arabidopsis thaliana* is estimated to express 8,000~11,000 genes during early flower development. We expand the estimate of total number of genes expressed in flowers to other eudicots, monocots and basal angiosperms. Basal angiosperms especially show a large number of distinct transcripts, and differential expression is found between male and female flowers of *Amborella trichopoda*. Compared with the genome duplication history, the number of gene expressed in flowers may be largely downsized in *Arabidopsis* while maintained in basal lineages. Also, libraries of early stage flowers are shown to contain more transcripts expressed at a low level than mature flowers, consistent with microarray experiments on different floral developmental stages. From the early stage developing flowers, we expect a large number of unique genes be discovered in future EST or cDNA sequencing.

Key words: Gene number, EST, flower development

Introduction

Gene expression is highly regulated during development. In plants, gene expression regulation at the transcript level is shown to lead to tissue differentiation (1). As a complex structure, flowers include multiple organs; most commonly shared are sepals, petals, stamens (male reproductive organ) and carpels (female reproductive organ). Therefore, genes expressed in flowers should encompass broad functional categories including both basic metabolism and development. For example, several plant

MADS-box genes are shown to be specifically expressed in flowers and determine the floral organ identity in model organisms such as *Arabidopsis thaliana*, *Antirrhinum majus* (snapdragon), and petunia (2-4). Molecular and genetic studies of these model organisms identified a conserved genetic regulatory network involving transcription factors (such as MADS box, bZIP, zinc finger, homeobox, myb and AP2/EREBP domain proteins) (5-8), protein kinases (9,10), and transcription factor regulators (11,12). Expansion of these gene families have been linked to genome-wide duplication in some angiosperm lineages. Duplication of floral development genes, *APETALA3* and *APETALA1* was believed to accompany the origin of core eudicots (13). At least one round of genome duplication occurred in the *Arabidopsis* genomes since the core eudicots split from monocots, resulting in duplicate blocks in the genome (14,15). Genome duplication has likely expanded the set of genes expressed in flowers. The genes expressed in flowers are a result of regulation of a specific subset of genes in the genome. This study aims at estimation of the genes expressed in flowers, which we refer to as “gene number”. Under this premise, different transcript types from the same gene are included in our estimates of gene number.

The earliest efforts to estimate the number of genes expressed in a tissue were based on solution hybridization experiments. Kamalay and Goldberg demonstrated that tobacco nuclear RNAs are under developmental control, and tissue specific mRNAs are found in both unique and shared RNA subsets between vegetative and reproductive tissues (16). They estimated that the mRNA complexity in floral organs (petal, anther and ovary) was about 3.2×10^4 kb in each organ. If the average length of an mRNA is 1.6kb

(Average length of known cDNA in *Arabidopsis*), the result suggested about 20,000 genes expressed in each organ. With little genetic information, the data implied a high complexity of expressed genes in floral organs. Multiple high-throughput methods for measuring gene expression have become available. ESTs (17) and microarrays (18) provide rich data on the diversity and expression levels of different transcripts in specific tissues. Notably the full genome sequence of a model eudicot plant, *Arabidopsis thaliana*, yielded a surprise in that the number of genes is larger than many of the animals (19). Full genome sequences for rice suggested even more genes present in the compact genome of that monocot species (20,21).

However, the number of genes present in a genome does not necessarily determine the number of genes expressed in an organ. It has been shown by whole-genome microarray experiments that a large fraction of the *Arabidopsis* genome is differentially expressed between organs (18,22). Meristem tissues and reproductive tissues such as young flower buds have more genes expressed at a lower level than mature, differentiated tissues (22). While the total gene number remains unknown until the genome is finished and fully annotated, compiled tissue cDNA libraries would allow estimates of the transcript pool for the genome. Estimates of unique transcripts based on any individual tissue may provide more specific knowledge of the toolkit used in the developmental process.

We infer the extent of total genes expressed in various tissues or organs based on EST data. ESTs accumulated from cDNA libraries provide a suitable source of data to address the gene expression distribution in a wide range of organisms (23). We further compare the gene number estimates to that identified by microarray and by the genome

annotation using the *Arabidopsis* genome data. Because EST sequencing for most cDNA libraries is not exhaustive, a great majority of transcripts are not represented in the sequenced population. Simulations show that the downward bias may be mitigated by bootstrap confidence interval estimates. EST sequence cleaning and clustering also influences the accuracy of true gene expression level as measured by EST counts.

We compare the expressed gene numbers in flower libraries among derived and basal angiosperm lineages (24). Results indicate that genomes of basal lineages may express more genes than known in model organisms, however, the estimates do not correlate with genome size or genome duplication history of flowering plant lineages (25). Multiple biological interpretations, including heterozygosity, alternative splicing, non-genic transcript, and broader expression of gene family members, may have contributed to the number of genes expressed in basal angiosperm flowers.

Methods

EST sequence assembly

EST sequences from floral cDNA libraries were obtained from GenBank nucleotide section for *Arabidopsis thaliana* (flower buds), tomato (four libraries, including 0-3mm flower buds, 3-8mm buds, 8mm buds to preanthesis, and open flowers), soybean (flower buds, less than 3mm), grape (pre-bloom to nectary stage flower), barley (male and female inflorescence), rice (mixed stage panicle). ESTs from other species (California poppy, asparagus, *Liriodendron tulipifera*, *Persea americana*, *Saruma henryi*, *Nuphar advena* and *Amborella trichopoda*) were generated by the Floral Genome Project

(24). Only directional cloned and sequenced libraries were selected so that a data set contains all 5' ESTs or all 3' ESTs. Sequences were trimmed using seqclean to remove polyA tracks, vector sequences and contaminate sequences. Subsequently they were separated by the sequencing directions and assembled using CAP3 (26) at percent identity $P = 90$ and length overlap $O = 40$. The assembly criterion was chosen based on comparisons of assembled *Arabidopsis thaliana* unigenes under various assembly rules to the genome annotation (27). To evaluate the influence of clustering stringency on the gene number estimates, we also constructed EST clusters for one *Arabidopsis* data set at $P = 85, 90, 95$ and 97.5 .

EST clustering and gene number estimation

A unigene is a consensus sequence derived from EST clusters, which may not always correspond to a unique gene in the genome due to assembly errors. If only one EST is captured, the unigene is called a singleton. Non-overlapping ESTs from the same cDNA could be assembled into different unigenes. This is a major source of error for assembling 5' ESTs, and we call it the ISO error (for insufficient overlap error). To obtain a more accurate representation of gene expression levels captured by ESTs, we developed a correction method and implemented in ESTstat (27,28).

We are seeking to estimate the total number of genes expressed, which we call the gene number N . The EST count X follows a zero-truncated Poisson mixture distribution. We obtained the penalized maximum likelihood solution for the mixture distribution (details in (29)). A bootstrap sampling of gene clusters based on \mathbf{c} provides the confidence intervals (29).

Simulation based on microarray data

Data from three experiments using Affymetrix ATH1 GeneChip on *Arabidopsis thaliana* flowers (ATGenExpress_29, Col-0, inflorescence stage 1-6 flower; ATGenExpress_46, shoot apex, inflorescence; and X. Zhang experiment, wild type inflorescence(22)) were obtained and normalized with the Robust Multi-array Average (RMA) and Microarray Suite 5.0 (MAS5.0) methods. After normalization, the average intensity values on the chip were treated as the true expression level of each gene. The distribution of intensity values was used to generate sample EST counts with the total sample size S fixed. That is, for each gene, the number of ESTs sampled is

$$X_i \sim \text{Poisson}(\alpha\lambda) \text{ and the tuning parameter } \alpha = \frac{S}{N} \sum \text{Intensity}_i.$$

Two sets of simulations were conducted. First, repeated sampling of 10,000 ESTs from the same library was conducted to compare the variation of the gene number estimates. Next, variable sample sizes were compared. Five samples were generated from the expression profile represented by a microarray experiment and the sample sizes ranged from 5000 to 50,000.

Classification of unigenes to PlantTribes

The PlantTribes database (<http://www.floralgenome.org/planttribes/>) contains putative gene families for *Arabidopsis* and rice clustered using TribeMCL (30). Unigenes from five basal angiosperm species were assigned to the tribes according to the top BLAST hit of *Arabidopsis* or rice proteins with the E value cutoff at 1E-10.

Results

EST clustering stringency affects gene number estimates

For the same EST set, the gene number estimates are robust within a range of EST clustering stringencies, although the observed gene cluster profiles vary (**Table 5-1**). As the stringency increases, ESTs from similar genes or genes with common domains are less likely to be clustered together, and long contigs may split into smaller clusters, which could lead to a significant increase in observed singletons. For the *Arabidopsis* flower ESTs, the number of unigenes identified increases by 8% and the observed singletons by 12% when the clustering stringency P increases from 85 to 97.5. Although the point estimate of gene number increases, the 95% confidence intervals still largely overlap within the range (8000, 11000), suggesting that the gene number estimate is robust. When the clustering stringency is so high that it leads to extreme false separation of ESTs from the same gene, the estimates become less reliable.

On the other hand, the most influential error is not dependent on the clustering stringency alone, but the overlap of ESTs in a cluster. The ISO error results in an overestimate of rare unigenes, especially the number of singletons. After mapping the EST clusters to the *Arabidopsis* genome, we estimated that the ISO error was around 5% for 3' ESTs at clustering stringency $P=85$ and 90, and increased to 13% at $P=95$. This problem is most serious in clustering 5' ESTs, where partially degraded transcript may constitute a large fraction of the cDNA pool. We estimated that most data sets of 5' ESTs contained about 10% singletons that should be clustered into larger contigs, but the error could be as high as 25%. ESTstat attempts to correct the error before estimation of the

gene number (see **Figure 5-1**). Without this correction step, the number of rare transcripts would be over estimated, and the subsequent estimate of gene number would be inflated.

Adequate sample size required for robust estimates

To evaluate the sample size effect on the gene number estimates, we simulated cDNA libraries with transcript abundance levels following those from one microarray experiment on *Arabidopsis* wild type flowers (stage 1-6) (22). From 100 repeated samples, the median estimated number of expressed genes is 14452 and the 95% confidence interval (C.I.) is (12249, 17061). This estimate agrees with the number of genes labeled “present” according to the MAS 5.0 normalization method (14748), and roughly corresponds to the number of genes with expression value > 50 (14127) out of 22,746 gene elements represented on the chip. It suggests that a sample size of 10,000 yields reliable estimates for the expressed genes in the floral tissue. At this sample size, the interval estimates agree with independent whole-genome array experiments.

We also compared the estimates based on different EST sample sizes using simulation data (**Figure 5-2**). The gene expression levels are derived based on three independent microarray experiments containing flower or inflorescence tissue. EST clusters are randomly drawn to reach the sample sizes of 5000 to 30000. Most bootstrap gene number estimates overlap within the range of 10,000 to 22,000. Simulations based on inflorescence and shoot apex microarray data also yield similar results (not shown). The estimator is designed to reach high accuracy under various distribution assumptions. Different from unbiased maximum likelihood estimator, the NPMLE result in a downward bias in order to reduce the variance. In other words, an unbiased ML estimator

may under some situations lead to much larger estimates (in the millions), which is unrealistic for the model flowering plants in this study. By controlling for sample size and EST clustering stringency, the results for diverse flower EST libraries become comparable.

Estimates of genes expressed in flowers and other tissues

We can estimate the gene number for multiple tissues and even whole plants by combining the EST data from well-differentiated tissues. When pooling flowers and non-flower tissues, such as roots and young shoots, we obtained that in each two-source combination, there are about 20,000 genes expressed (**Table 5-2**). The florally expressed genes contribute to 50%~60% of the genes expressed in the tissues combined. It suggests that a large fraction of the genes in the pooled tissues are only preferentially expressed in one source, which agrees with previously reported results from microarray experiments (22). When all three types of tissues were pooled (flower, root and shoot), we estimated that the total number of expressed genes is about 29,000, close to the predicted protein-coding genes in the whole genome (TIGR *Arabidopsis* genome release 5.0).

Theoretically, the proportions of ESTs contributed by each library do not bias the result, because the pooled set represents a different mixture distribution from each subset.

However, the estimates increased to much more than 30,000 when we pooled over 150,691 ESTs from five sources (flower, root, shoot, leaf and green silique). This is likely in part to be due to the presence of alternatively spliced transcripts that do not cluster together in unigene construction (27), and due to the presence of a low frequency of genomic fragments that have passed through the cDNA cloning process. We

considered that the sample was not representative of the true expression levels for genes in the genome. The increase of clustering errors resulted in much higher fraction of rare transcripts such that the correction method based on a single library is not sufficient.

Expressed gene numbers among angiosperm flowers

The number of expressed genes from flowers of eudicot and monocot species lies in a consistent range similar to the *Arabidopsis* estimate, with exceptions of California poppy, a basal eudicot, and rice (**Table 5-3**). We especially collected the EST data from young flowers when possible, which include mostly the early stage buds. These include cDNA libraries from *Arabidopsis*, soybean, California poppy, barley and asparagus. Tomato, grape and rice ESTs were derived from mixed stage flowers. Based on the mixture distribution, we estimated the fractions of genes with high, intermediate, and low expression levels to be around 75%, 18% and 7% in young *Arabidopsis* flowers. There appears to be some difference of the proportions and estimated levels of each population of transcripts among species. For example, in tomato and grape flower libraries, most genes are expressed at medium levels and few at very low levels (**Figure 5-3**).

For most basal angiosperms we found that the numbers of expressed genes are much larger (**Table 5-4**), about 50% to one time more than those of derived eudicots, when controlling for the sample sizes. In the basalmost angiosperm *Amborella*, it was estimated that more genes are expressed in male flowers than in female flowers, suggest possible sex difference in gene expression. These libraries all represent the early stage of flower development. Results also show that these tissues express most genes at a low level (**Figure 5-3**), so that the total numbers of unigenes after EST clustering are relatively high.

Discussion

This study analyzes expressed gene numbers in flowers of a broad range of angiosperms. With shallow sampling of total transcript pools by EST sequencing, the rare transcripts are most influential in the gene number estimate. Although the small sample size leads to negative bias of the gene number estimate (25), we showed that when the sample size of 10,000 ESTs is used across libraries, the results are robust and comparable. Also, results based on independent samples of ESTs from different species do not suggest strong correlation of EST sample sizes and the gene number estimates. Therefore, those estimated with a large number of transcripts are not due to difference in sample sizes.

EST sequencing and clustering quality are important factors in determining the accurate level of genes expressed. The clustering error correction model uses both EST read length and estimated cDNA length (inferred from EST contigs). We use the quality score model based sequence assembly (generated by CAP3) and average read lengths do not appear to correlate with gene number estimates. The estimates based on public data sets do not contain quality information so that clustering errors in data from sources other than data from the Floral Genome Project (24) may be higher.

Library preparation techniques also may influence the total pool of transcripts represented. The most serious source of bias that could result in significant increase of “false” transcripts is genomic sequence contaminants. Although we do not have direct measure of genomic DNA contamination in the cDNA library, chloroplast and mitochondrial DNAs can be used to estimate the degree of foreign DNA in the mRNA

preparation. In the libraries surveyed, less than 1% of clones yield significant matches to chloroplast or mitochondrial DNA. These clones are not over represented in singletons. The majority of sequences do not come from genomic sources and the EST sequences represent true transcripts.

Alternative splicing may lead to several variants of the same genes expressed, which could result in an increase of estimated gene numbers. The degree of alternative splicing is low in plants. It is estimated to be 6% in *Arabidopsis* and 16% in rice in a survey based on full length cDNAs mapped to the *Arabidopsis* and rice genomes (31), but for most plant species, the frequency will not be known until long genomic sequences become available. By our definition, splicing variants that fail to cluster will be counted as different genes, and included in our estimates of expressed gene number.

The sampling scheme for different plants may contribute to different representation of transcript pools in a cDNA library. For highly inbred lines, such as *Arabidopsis*, ESTs from the same gene are generally clustered together. For other outcross species sampled from the wild, if the plants are highly heterozygous and allelic divergence exceed 10%, some alleles will appear as separate clusters, and total gene number estimate may be inflated. It is unknown how common it is for allelic divergence to be so high that ESTs from different alleles do not cluster.

It is surprising that with similar sample depth and clustering approach, the basal angiosperms consistently recorded a large number of transcripts expressed in the floral tissue. The genome sizes of the species studied range from 125 Mb (*Arabidopsis thaliana*) to over 5000 Mb (barley). Except for soybean, a diploidized tetraploid (32),

most species are diploid, and some have had possible whole-genome duplications since the origin of major lineages of angiosperms (25,33). There is an apparent lack of correlation between the genome size measurements and the gene number expressed in flowers. In cereals, the large genome sizes is mainly due to proliferation of repeat elements, including LINES, SINEs, and retrotransposons (34). Expressed transposable elements may have contributed to the pool of transcripts in basal angiosperms and rice. Their transcription activities may lead to regulatory changes during development (35). While shared polyploidy events may be responsible for genome size increases in some lineages, the reduction of genome sizes is significant in the model organisms, *Arabidopsis* and rice, compared to closely related lineages (36). Thus, the decrease of expressed genes in eudicots is more likely a product of differential gene regulation than simple reduction of genomes since past genome duplication.

Previous studies suggest that perhaps most angiosperm species have had polyploidy history (33). Compared to the basal angiosperm lineages (*Amborella*, *Nuphar* and magnoliids: *Liriodendron*, *Persea*), eudicot and monocot species probably had genome duplications that were independent from most of those detected in basal angiosperm lineages (25). However, duplicate genes may have been eliminated at a lower rate due to the low rate of substitutions estimated from basal angiosperms compared to eudicots and grasses (37). The dynamic gene birth and death in the genome, combined with the genome duplication history, shapes the difference in potential gene space of basal angiosperms and derived lineages. In addition, duplicated floral development genes originated in eudicots (13,38,39) suggest that functional differentiation exists between these paralogs, which may lead to finer temporal regulation of the transcripts, instead of

increase of total transcripts. The expression pattern of MADS box genes in different flowering plant lineage support that the same gene set are expressed in a broader range of organs in basal lineages than in the derived eudicots (12).

The distribution of putative protein families for the unigenes from basal angiosperm species was compared to 29142 putative protein families in *Arabidopsis* and rice genome, classified in the PlantTribes database. We identified tribes which contain at least one member from the basal angiosperm species, while 261 and 231 lack any *Arabidopsis* or rice member, respectively. Most genes in these families have not been functionally characterized. Domain searches identified several F-box proteins, zinc finger proteins, and proline-rich extensins. One tribe is found with similarity to *Arabidopsis* extensin, and the *Arabidopsis* protein is expressed in flower buds and bolted flowers (40). Several groups of LTR-retrotransposon proteins which are absent from *Arabidopsis* are found in ESTs, and similar findings were reported in cereal species (41). It is possible that the basal angiosperm species contain a large gene set and many genes are activated in flower development, while derived lineages either have lost some gene sets, or have evolved more control over timing and location of gene expression, including exclusion of transposable elements.

We report that over 10,000 genes are expressed in most species, which may represent common developmental programs in flowers. The flower is a complex reproductive structure regulated by internal and external signals (42,43). Diverse classes of genes are required to complement the structural and functional roles involved in flower development in addition to basic metabolism. In the floral meristem, genes in floral induction and organogenesis, cell division and expansion are highly active. After floral

organs are determined, genes in constructing specific structure such as nectary and pigment synthesis are expressed in those organs. During the late stages, genes involved in meiosis and post meiotic development of male and female gametophytes are expressed more abundantly (22). The cDNA libraries constructed from floral tissues often consists of different stages of flowers and the cutoff stages may not be equivalent in developmental stages. For species with single flowers, such as *Nuphar*, we sampled single flower buds, which would contain more large buds than the source for *Arabidopsis* libraries made from inflorescences of very young buds. Thus, the *Nuphar* library may contain more genes required for late stages of flower development.

In addition, transcription activity in non-genic or inter-genic regions could contribute to the estimated gene numbers. Yamada *et al.* (44) reported active transcription in previously defined intergenic regions for *Arabidopsis*, and some active anti-sense transcripts for known genes using whole-genome tiling arrays. Most of the transcription activity was detected in annotated gene regions or verified to be ORFs. If the frequency of non-genic transcript in other flowering plants is higher than that of *Arabidopsis*, it may be due to difference in transcription regulation, or an increased precision of transcription control in derived eudicot and monocot species.

From estimates of gene numbers for individual tissues and pooled sets, we may detect differentiation among gene expression profiles between related structures. It is possible that the large number of genes estimated genes in the male flowers of *Amborella* compared to in the female flowers include sex-specific regulators, and other genes that are specific to the male structure. Due to the sample size for these two libraries, the gene

number estimates are quite conservative. Still, the difference in total transcript diversity is clearly supported based on the similar sampling depth in either library.

From flower cDNA libraries of basal angiosperm, eudicot and monocots, we uncovered rich diversity of transcripts especially in basal angiosperms. Our method provides guidance for sampling size and gene discovery in EST sequencing projects. According to our estimate, double sequencing efforts for these targeted tissue libraries that have potential large number of genes could lead to more gene discovery than comparatively sequencing similar number of ESTs from another tissue.

Figures

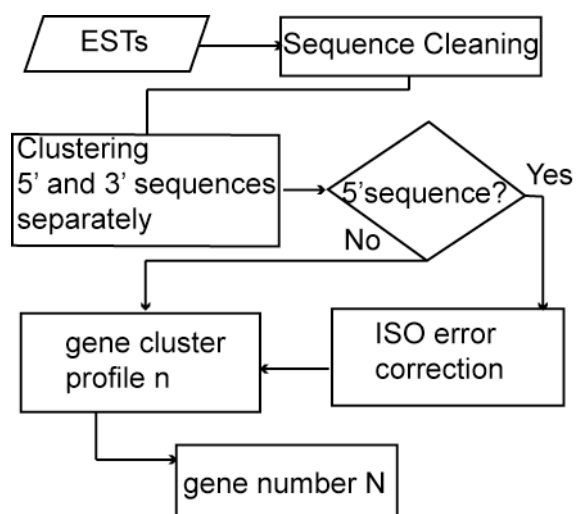


Figure 5-1: Flowchart of data processing. The ISO error correction and gene number estimation are implemented in the software ESTstat.

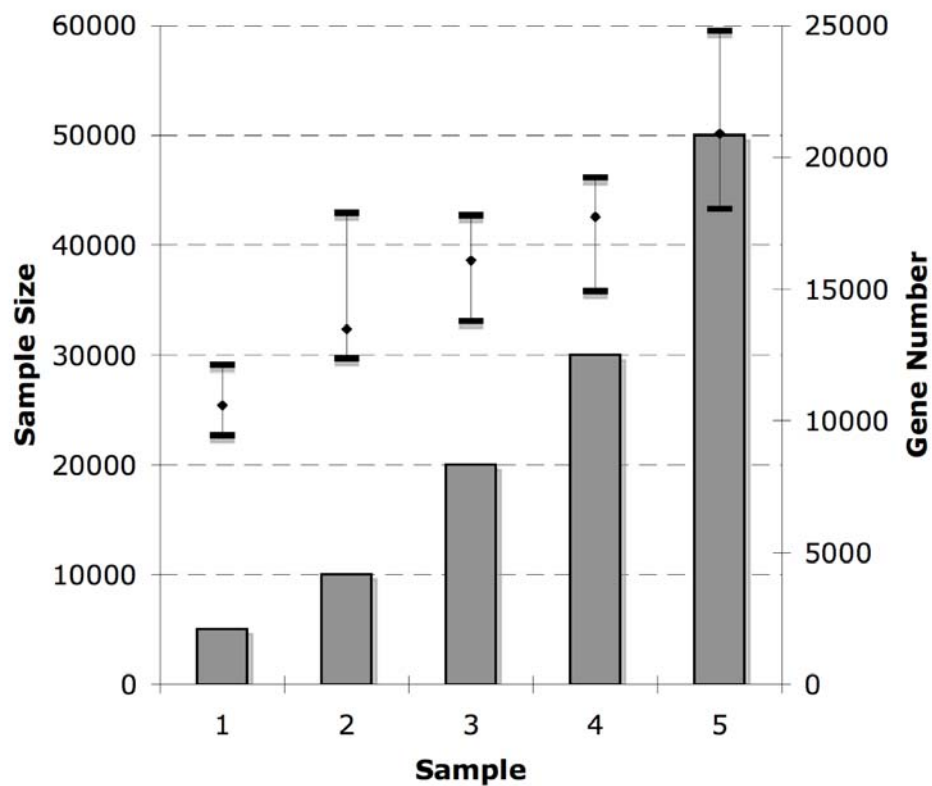


Figure 5-2: The relationship of sample size and estimate bias in simulations based on three microarray experiments on *Arabidopsis* flower/inflorescence tissue. Source, stage 1-6 flower. The sample size is labeled on the left Y axis and the gene number estimates (point estimate and the bootstrap 95% confidence interval) are labeled for each sample size. It shows that the gene number estimates stabilize when the total sample size reaches 10000.

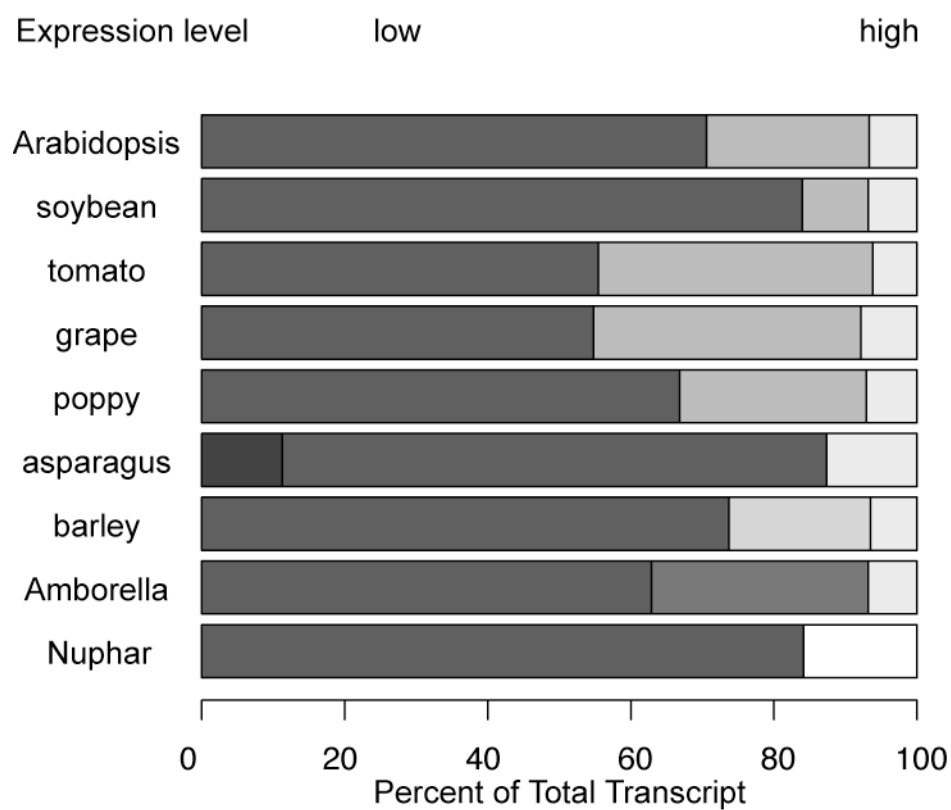


Figure 5-3: Distribution of transcripts by expression levels in different floral libraries. The expression levels are estimated by the Poisson parameter in the mixture distribution. Shades of gray indicate the expression levels from low (dark) to high (bright).

Tables

Table 5-1: Relationship of EST clustering stringency (percent identity) and the gene number (N) estimated from a flower bud cDNA library^a of *Arabidopsis thaliana*.

Percent Identity	Total Unigenes	Singletons	N	95% C.I.
85	2461	1743	8797	(7182, 11240)
90	2494	1778	9069	(7629, 11061)
95	2567	1859	10119	(8724, 11918)
97.5	2653	1957	11899	(9664, 14130)

a. Total ESTs n=5710.

Table 5-2: Estimated total transcripts from multiple tissues of *Arabidopsis thaliana*.

Tissue	Total ESTs	Total Unigenes	N	95% C. I.
flower + root	18788	7371	16262	(16136, 21951)
root + shoot	23695	9067	18864	(18510, 23293)
shoot + flower	16018	6683	18064	(16276, 23539)
whole plant	40597	12281	29625	(23359, 31621)

Table 5-3: Number of genes detected and expressed in flowers of eudicot and monocot species.

Species	Total ESTs	Total Unigenes	N	95% C. I.
tomato	12740	5732	12701	(11760, 15965)
soybean	9036	4690	12366	(11246, 15246)
grape	6495	3744	10041	(8656, 12724)
California poppy	9079	5164	15272	(12688, 18907)
barley	8604	4240	11041	(9952, 13965)
rice (mixed stage)	16205	7065	15931	(15134, 18871)
asparagus	7362	3709	10124	(7882, 11410)

Table 5-4: Estimated number of genes expressed in basal angiosperm flowers.

Library	Total ESTs	Total Unigenes	<i>N</i>	95% C. I.
<i>Amborella</i> male flower	4279	3331	18220	(13754,20313)
<i>Amborella</i> female flower	4427	3062	10782	(8947,11972)
<i>Amborella</i> (combined)	8706	5832	19340	(16798, 23809)
<i>Nuphar</i>	10208	6680	22045	(19683, 28451)
<i>Liriodendron</i>	9531	6195	28330	(22093,31450)
<i>Persea</i>	8735	5867	22819	(18927,25275)
<i>Saruma</i>	10213	5262	12243	(10629,15545)

References

1. Fernandes, J., Brendel, V., Gai, X., Lal, S., Chandler, V.L., Elumalai, R.P., Galbraith, D.W., Pierson, E.A. and Walbot, V. (2002) Comparison of RNA expression profiles based on maize expressed sequence tag frequency analysis and micro-array hybridization. *Plant Physiol*, **128**, 896-910.
2. Ma, H. (1994) The unfolding drama of flower development: recent results from genetic and molecular analyses. *Genes Dev*, **8**, 745-756.
3. Saedler, H. and Huijser, P. (1993) Molecular biology of flower development in *Antirrhinum majus* (snapdragon). *Gene*, **135**, 239-243.
4. Tsuchimoto, S., van der Krol, A.R. and Chua, N.H. (1993) Ectopic expression of pMADS3 in transgenic petunia phenocopies the petunia blind mutant. *Plant Cell*, **5**, 843-853.
5. Theissen, G., Becker, A., Di Rosa, A., Kanno, A., Kim, J.T., Munster, T., Winter, K.U. and Saedler, H. (2000) A short history of MADS-box genes in plants. *Plant Mol Biol*, **42**, 115-149.
6. Jakoby, M., Weisshaar, B., Droge-Laser, W., Vicente-Carbajosa, J., Tiedemann, J., Kroj, T. and Parcy, F. (2002) bZIP transcription factors in Arabidopsis. *Trends Plant Sci*, **7**, 106-111.
7. Takatsuji, H. (1998) Zinc-finger transcription factors in plants. *Cell Mol Life Sci*, **54**, 582-596.
8. Weigel, D. (1998) From floral induction to floral shape. *Curr Opin Plant Biol*, **1**, 55-59.
9. Jonak, C. and Hirt, H. (2002) Glycogen synthase kinase 3/SHAGGY-like kinases in plants: an emerging family with novel functions. *Trends Plant Sci*, **7**, 457-461.

10. Roe, J.L., Rivin, C.J., Sessions, R.A., Feldmann, K.A. and Zambryski, P.C. (1993) The *Tousled* gene in *A. thaliana* encodes a protein kinase homolog that is required for leaf and flower development. *Cell*, **75**, 939-950.
11. Weigel, D. (1997) Flower development: repressing reproduction. *Curr Biol*, **7**, R373-375.
12. Kim, S., Koh, J., Yoo, M.J., Kong, H., Hu, Y., Ma, H., Soltis, P.S. and Soltis, D.E. (2005) Expression of floral MADS-box genes in basal angiosperms: implications for the evolution of floral regulators. *Plant J*, **43**, 724-744.
13. Irish, V.F. and Litt, A. (2005) Flower development and evolution: gene duplication, diversification and redeployment. *Curr Opin Genet Dev*, **15**, 454-460.
14. Blanc, G., Hokamp, K. and Wolfe, K.H. (2003) A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. *Genome Res*, **13**, 137-144.
15. Ermolaeva, M.D., Wu, M., Eisen, J.A. and Salzberg, S.L. (2003) The age of the *Arabidopsis thaliana* genome duplication. *Plant Mol Biol*, **51**, 859-866.
16. Kamalay, J.C. and Goldberg, R.B. (1984) Organ-specific nuclear RNAs in tobacco. *Proc Natl Acad Sci U S A*, **81**, 2801-2805.
17. Rounsley, S.D., Glodek, A., Sutton, G., Adams, M.D., Somerville, C.R., Venter, J.C. and Kerlavage, A.R. (1996) The construction of *Arabidopsis* expressed sequence tag assemblies. A new resource to facilitate gene identification. *Plant Physiol*, **112**, 1177-1183.
18. Wellmer, F., Riechmann, J.L., Alves-Ferreira, M. and Meyerowitz, E.M. (2004) Genome-wide analysis of spatial gene expression in *Arabidopsis* flowers. *Plant Cell*, **16**, 1314-1326.
19. Poethig, R.S. (2001) Life with 25,000 genes. *Genome Res*, **11**, 313-316.
20. Goff, S.A., Ricke, D., Lan, T.H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H. *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science*, **296**, 92-100.
21. Yu, J., Hu, S., Wang, J., Wong, G.K., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X. *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science*, **296**, 79-92.
22. Zhang, X., Feng, B., Zhang, Q., Zhang, D., Altman, N. and Ma, H. (2005) Genome-wide expression profiling and identification of gene activities during early flower development in *Arabidopsis*. *Plant Mol Biol*, **58**, 401-419.
23. Boguski, M.S., Tolstoshev, C.M. and Bassett, D.E., Jr. (1994) Gene discovery in dbEST. *Science*, **265**, 1993-1994.
24. Albert, V.A., Soltis, D.E., Carlson, J.E., Farmerie, W.G., Wall, P.K., Ilut, D.C., Solow, T.M., Mueller, L.A., Landherr, L.L., Hu, Y. *et al.* (2005) Floral gene resources from basal angiosperms for comparative genomics research. *BMC Plant Biol*, **5**, 5.
25. Cui, L., Wall, K., Lindsay, B.G., Leebens-Mack, J.H., Doyle, J.J., Soltis, D.E., Soltis, P.S. and dePamphilis, C.W. (2006) Widespread genome duplications in flowering plants. *Genome Res*, In Press.
26. Huang, X. and Madan, A. (1999) CAP3: A DNA sequence assembly program. *Genome Res*, **9**, 868-877.

27. Wang, J.P., Lindsay, B.G., Leebens-Mack, J., Cui, L., Wall, K., Miller, W.C. and dePamphilis, C.W. (2004) EST clustering error evaluation and correction. *Bioinformatics*, **20**, 2973-2984.
28. Wang, J.P., Lindsay, B.G., Cui, L., Wall, P.K., Marion, J., Zhang, J. and dePamphilis, C.W. (2005) Gene capture prediction and overlap estimation in EST sequencing from one or multiple libraries. *BMC Bioinformatics*, **6**, 300.
29. Wang, J.-P. and Lindsay, B.G. (2005) A penalized nonparametric maximum likelihood approach to species richness estimation. *J Am Stat Assoc*, **100**, 942-959.
30. Enright, A.J., Kunin, V. and Ouzounis, C.A. (2003) Protein families and TRIBES in genome sequence space. *Nucleic Acids Res*, **31**, 4632-4638.
31. Itoh, H., Washio, T. and Tomita, M. (2004) Computational comparative analyses of alternative splicing regulation using full-length cDNA of various eukaryotes. *Rna*, **10**, 1005-1018.
32. Zhu, T., Schupp, J.M., Oliphant, A. and Keim, P. (1994) Hypomethylated sequences: characterization of the duplicate soybean genome. *Mol Gen Genet*, **244**, 638-645.
33. Blanc, G. and Wolfe, K.H. (2004) Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell*, **16**, 1667-1678.
34. Sandhu, D. and Gill, K.S. (2002) Gene-containing regions of wheat and the other grass genomes. *Plant Physiol*, **128**, 803-811.
35. Girard, L. and Freeling, M. (1999) Regulatory changes as a consequence of transposon insertion. *Dev Genet*, **25**, 291-296.
36. Bennetzen, J.L., Ma, J. and Devos, K.M. (2005) Mechanisms of recent genome size variation in flowering plants. *Ann Bot (Lond)*, **95**, 127-132.
37. Gaut, B.S., Morton, B.R., McCaig, B.C. and Clegg, M.T. (1996) Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene *Adh* parallel rate differences at the plastid gene *rbcL*. *Proc Natl Acad Sci U S A*, **93**, 10274-10279.
38. Zahn, L.M., Leebens-Mack, J., DePamphilis, C.W., Ma, H. and Theissen, G. (2005) To B or Not to B a flower: the role of DEFICIENS and GLOBOSA orthologs in the evolution of the angiosperms. *J Hered*, **96**, 225-240.
39. Kim, S., Soltis, P.S., Wall, K. and Soltis, D.E. (2005) Phylogeny and Domain Evolution in the APETALA2-Like Gene Family. *Mol Biol Evol*.
40. Van den Heuvel, K.J., Van Lipzig, R.H., Barendse, G.W. and Wullems, G.J. (2002) Regulation of expression of two novel flower-specific genes from tomato (*Solanum lycopersicum*) by gibberellin. *J Exp Bot*, **53**, 51-59.
41. Rudenko, G.N. and Walbot, V. (2001) Expression and post-transcriptional regulation of maize transposable element MuDR and its derivatives. *Plant Cell*, **13**, 553-570.
42. Okamoto, J.K., den Boer, B.G., Lotys-Prass, C., Szeto, W. and Jofuku, K.D. (1996) Flowers into shoots: photo and hormonal control of a meristem identity switch in *Arabidopsis*. *Proc Natl Acad Sci U S A*, **93**, 13831-13836.

43. Komeda, Y. (2004) Genetic regulation of time to flower in *Arabidopsis thaliana*. *Annu Rev Plant Biol*, **55**, 521-535.
44. Yamada, K., Lim, J., Dale, J.M., Chen, H., Shinn, P., Palm, C.J., Southwick, A.M., Wu, H.C., Kim, C., Nguyen, M. *et al.* (2003) Empirical analysis of transcriptional activity in the *Arabidopsis* genome. *Science*, **302**, 842-846.

Concluding Remarks

Here I would like to summarize some future directions or extensions of the methods/results from chapters 3-5. In general, the study of genome evolution on two major levels: gene order and genome duplication can be applied to almost all organisms with several related genome sequences available.

The methods I have developed for ancestral genome reconstruction was used on chloroplast genomes only. It may be extended to eukaryotic genomes. To do so, the genome rearrangement algorithm needs to address following challenges:

1. Multi-chromosome data which involve chromosome fusion, es. This is implemented in GRIMM/MGR.
2. Segmental duplication and overlapping genes probably should be considered as a unique unit in rearrangements. This occurrence of overlapping genes is frequent for *Drosophila pseudoobscura* and *Drosophila melanogaster* homologous chromosome I tested (Jijun Tang, Stephen Schaeffer, personal communication). It requires an automatic processing of “gene” or orthologous segments from draft sequences.
3. The operations between eukaryote chromosomes could be very large. Some simplifications are necessary to remove trivial problems and to speed up the computation.

Jijun Tang proposed adding “artificial genomes” to the data set to reduce branch lengths (personal communication). There still needs more theory development and tests in this area. Also, since eukaryote genomes are organized in chromosome domains, gene expression regulation mechanisms are different from chloroplast genomes (1). The clustering of genes in the eukaryote genomes is more likely related to shared *cis* elements.

Cases of whole genome duplications have been reported in teleost fishes, amphibian and yeast (2-4). The frequency of polyploidy species in vertebrates is much lower than in flowering plants, but the availability of sequenced genomes that are rich in orthologous sequences will enable comparison of fine scale duplications (perhaps most tandem duplications) and will help define the fate of duplicate genes. Currently, assumptions on the background birth-death process have not been directly linked to the biological evidence.

On the study of gene expression, the estimation method could be applied to much richer sources of tag sequencing and gene expression data other than ESTs. It will be necessary to redesign the clustering error correction because the sources of error may be different. Correlation between different technologies needs to be established so that the expression level detected by one approach can be used to compare to the other.

The expression and regulation of floral genes in basal angiosperms is still largely unknown. The phylogenetic inference is promising for retained duplicate genes in basal lineage vs. derived lineages based on gene number estimates. Furthermore, a thorough survey of the expression of paralogous genes (not necessarily flower development regulators) would help to identify what contributes to maintaining duplicated genes and

what leads to reduction of genes after duplication. Ideally, the gene space discovery from a few basal angiosperm species would complement the largely well studied *Arabidopsis* genome and enable functional studies in other species.

The chloroplast genomics project and the Floral Genome Project have brought together opportunities for cross-disciplinary training when I am enrolled in the Biology Program at Penn State. The bioinformatics tool kit developed through these projects would also benefit other research and perhaps lead to new research opportunities.

References

1. Levine, M. and Tjian, R. (2003) Transcription regulation and animal diversity. *Nature*, **424**, 147-151.
2. Jaillon, O., Aury, J.M., Brunet, F., Petit, J.L., Stange-Thomann, N., Mauceli, E., Bouneau, L., Fischer, C., Ozouf-Costaz, C., Bernot, A. *et al.* (2004) Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature*, **431**, 946-957.
3. Ohno, S. (1999) Gene duplication and the uniqueness of vertebrate genomes circa 1970-1999. *Semin Cell Dev Biol*, **10**, 517-522.
4. Wolfe, K.H. and Shields, D.C. (1997) Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, **387**, 708-713.

Appendix

Inferring ancestral chloroplast genomes with duplications

Preface

This manuscript appears as Tech Report TR-CS-2005-08 at Department of Computer Sciences, University of New Mexico, Albuquerque. Original authors are Liying Cui, Jijun Tang, Bernard M. E. Moret and Claude W. dePamphilis. LC collected the data, designed the test and wrote the manuscript. JT developed the source code, simulations and wrote relevant part of the manuscript. BMEM contributed to the algorithm development. CWD contributed to writing the manuscript and discussions.

Inferring ancestral chloroplast genomes with duplications

Liyang Cui¹, Jijun Tang², Bernard M.E. Moret³ and Claude W. dePamphilis¹

¹ Department of Biology, The Pennsylvania State University, University Park, PA 16802,
USA

² Department of Computer Science and Engineering, University of South Carolina,
Columbia, SC 29280, USA

³ Department of Computer Science, University of New Mexico, Albuquerque, NM
87131, USA

Abstract

Motivation: Genome structural evolution is shaped by gene rearrangements and gene content changes, including gene duplications, insertions and deletions. Ancestral genome reconstruction based on whole-genome alignments has been limited to cases where few deletions or duplications can be assumed. Since conserved large duplicated regions are present in many chloroplast genomes, the inference of such duplication event is needed in reconstruction of ancestral chloroplast genomes.

Results: We apply GRAPPA-IR, a modified GRAPPA algorithm, to reconstruct ancestral chloroplast genomes from divergent land plants and green algae. The reconstructed ancestral genomes contain inverted repeats (IRs), which supports that conservation of the feature in chloroplast evolution. IR expansion has contributed primarily to gene content changes in these genomes, opposing to gene loss or transfer to nuclear genomes in land plants. On the contrary, single gene duplications outside IRs are inferred to be independent and do not constrain the genome rearrangements.

Availability: The C source code for GRAPPA-IR is available upon request.

Contact: jtang@cse.sc.edu

Introduction

Mutations in a genome consist of not only base pair level changes but also events that alter the chromosome structure, such as inversions, duplications and deletions (1,2). Gene order phylogeny was first proposed by David Sankoff (3); this algorithm using breakpoint distance was implemented in `BPAnalysis` software, and applied to animal mitochondrial genomes (4). It was not able to uniquely map the gene order changes and usually produced several tie trees, i.e., equally parsimonious trees regarding the optimization criterion. The inversion distance and inversion median were introduced to improve the phylogenetic accuracy, and the algorithm has been implemented in the software `GRAPPA` (5). Extensive simulations showed that inversion medians were superior to breakpoint medians and the trees returned were more accurate using either distance-based or parsimony methods (6,7). Currently, `GRAPPA` (version 2.0) is able to estimate the phylogeny and true inversion medians using genomes with equal gene content (5,8). A scaled-up version, `DCM-GRAPPA`, is able to estimate the gene-order phylogeny with apparently high accuracy for thousands of genomes, thus greatly increasing the power of genome phylogeny using large datasets (9). Part of the code for `GRAPPA` has been integrated into `GRIMM` to apply to multi-chromosomal genomes, such as human and mouse (10).

Biologists are interested in simultaneous inference of ancestral genomes and the genome phylogeny from a set of known genomes. Ancestral genome reconstruction has advanced significantly since whole genome sequences became available. Comparisons of orthologous chromosomal segments showed heterogeneous rates of evolution of the X

chromosome in human, mouse and rat (11). However, on the genome level the evolutionary change of genome structure is less well understood. The reconstruction was most successful in regions where few rearrangements happened for the assemblage of species that radiated within a short evolutionary period (12,13). Tandem duplications appear to have occurred frequently in mammalian genomes, which altered the gene copy number but did not change the gene order of an orthologous segment (e.g., the alpha globin cluster) (14). Organelle genomes, on the other hand, exhibit high diversity of genome rearrangements including inversions, transpositions and non-tandem duplications. The inversion model is close to the biological process of genome rearrangements, and inversion medians can be regarded as ancestral gene orders. Still, reconstruction of ancestral organelle genome presents two challenges.

1. The algorithm needs to handle large segment duplications and single gene duplication or deletions.
2. The algorithm needs to compute phylogeny that includes heterogeneous branch lengths with high accuracy since the rate of genome rearrangements could vary significantly among lineages.

The gene order data of fully sequenced chloroplast genomes provide excellent opportunity for developing and testing new algorithms. Chloroplast genomes have undergone significant downsizing from a free-living cyanobacteria-like ancestor (15) while the genome structure has been maintained. Typical land plant and green algal chloroplast genomes are circular single chromosomes consisting of 60 -150 genes, which encode proteins, tRNAs, rRNAs and hypothetical open reading frames. Most chloroplast genomes consist of four distinct parts: two duplicated regions (inverted repeats, IRs)

separated by a large single copy (LSC) and a small single copy (SSC) region. One common characteristic of the chloroplast IR is the presence of three rRNA genes (*rrn5*, *rrn16* and *rrn23*), which are homologous to the cyanobacterial *rrn* operon. The chloroplast gene order of land plants is mostly conserved, except for elevated level of rearrangements in specific lineages (16-19). The gene content of these chloroplast genomes vary greatly, largely due to the expansion and contraction of the IR at the IR-SC boundaries; this “ebb and flow” of the IR boundary has been observed even within a genus (20,21). Chloroplast genomes of green algae (charophyte and chlorophyte algae) also contain more variations of gene order and some are highly rearranged (22).

Previously we reported an GRAPPA algorithm to infer gene order phylogeny using data sets with a limited number of deletions, but no duplication is allowed (23). Here we develop a new algorithm that allows for the large duplication resulting in a quadripartite structure (e.g., LSC-IR-SSC-IR) in chloroplast and other IR containing genomes. The assumption of the new approach is that inversions do not occur across inverted repeats, because the genome structure will be disrupted by such inversions that “flip” the repeats from inverted to the same orientation. We also test the performance of the new algorithm compared to the original algorithm when duplicate genes are excluded.

Methods

The Dataset

Chloroplast genomes representing major lineages of land plants and green algae were selected and gene orders were extracted, all of which share the quadripartite

structure. The organisms include *Nicotiana tabacum* (tobacco, *nt*), *Psilotum nudum* (whisk fern, *pn*), *Marchantia polymorpha* (liverwort, *mp*), *Chaetosphaeridium globosum* (a charophyte green alga, *cg*), *Nephroselmis olivacea* (a chlorophyte green alga, *no*) and *Mesostigma viride* (a photosynthetic protist, *mv*). A reference phylogenetic tree was constructed using the maximum parsimony method with 50 concatenated proteins (Figure A-1). The reference tree is the same as the phylogeny by Lemieux *et al.* (24) in which *Mesostigma* basal to other green plants.

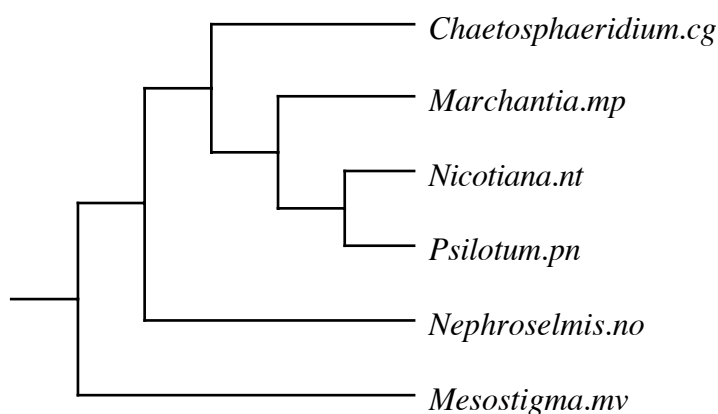


Figure A-1: The reference phylogeny of chloroplast genomes from land plants and green algae.

We extracted 70 unique genes from the six genomes. Actual number of genes included in each genome ranges from 73 to 80 due to duplicated genes in the IR. The gene set includes 62 characterized protein-coding genes, 10 tRNAs (identified by amino acid anticodons) and a hypothetical conserved open reading frame (*ycf1*). The encoding reflects the order and orientation of genes in the genome. The location of multi-exon genes was determined by the first exon. In one case (*psbD-psbC*), the order of

overlapping genes was determined by the position of the start codon. The data set was then applied to a two-stage analysis to estimate ancestral gene orders.

Mapping Gene Contents in the Ancestral Genomes

We first consider the case when the gene content for each region of the genome (LSC, SSC and IR) is relatively conserved. When the genome is on a leaf (i.e., it is an extant taxon and its gene order is known), the gene content for the LSC, SSC and IR regions can be determined through direct observation. However, the gene contents for the same region among the genomes may not be identical. We can only estimate the gene content for each region of the ancestral genomes based on the assumption that all evolutionary events that alter the gene order are rare and that concurrent (i.e., parallel) changes in two children are less likely than a change in the parent. Thus, at each internal node, for a given region, when the regional gene contents for the two children are known, we face three possibilities of assigning a gene to the region:

1. If both children have gene g in the same region, then the parent had g in that region; otherwise, both children need to expand (or shrink) IRs and include g in that region, with a very low probability.
2. If neither child has g , then g is most likely absent in the parent. However, the parent may have g in that region, with a very low probability.
3. If g is located in different regions between the children, then it could be in either region of the parent. The two choices are equally likely without further information

from the phylogeny. If the tree is rooted, we use the gene content in the evolutionary path to break the tie; otherwise, we are left with an undetermined outcome for g .

If a gene is undetermined in some internal node, it may become resolved using an iterative improvement algorithm similar to the core algorithm in GRAPPA itself. The same method was used (23) for data sets with unequal gene content:

1. For each sibling pair of *leaves*, if a gene appears in the same region at both children, we place it in the same region at the parent (an internal node); if the gene appears in different region at the leaves, we mark its status as undetermined in the parent.
2. Starting from an arbitrary root, we carry out a depth-first search of the tree to propagate resolutions according to our standard rule — if two neighbors have the gene presented in the same region, the node will have it in that region too — and thus to resolve undetermined states through look-ahead and cost propagation.

Using the method above, we were able to determine the most likely gene contents for each possible tree (105 trees in this case). The estimated gene content for the internal nodes of the reference tree is presented in Figure **A-2**.

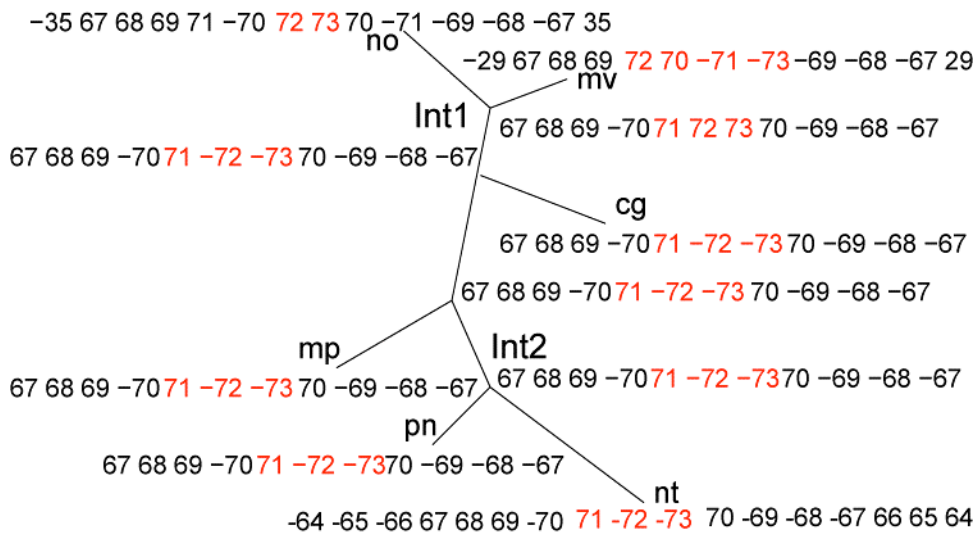


Figure A-2: Estimated gene contents for IR (in black) and SSC regions (in red).

Reconstruct Ancestral Gene Orders

From the observation of gene content mapping, the IR appears to be mostly conserved among land plants. We hypothesize the evolution of chloroplast genome structure as the following two steps:

1. The circular genome was divided into regions and inversions occurred in each region independently. No inversion spanning IR is allowed.
2. Once a segment from single copy regions was copied twice and joined to existing IRs, the new genomes with longer IRs propagated. Alternatively, a segment was spliced out from IR and joined the single copy region, and the new genome with smaller IRs propagated.

One should notice that the above two steps could happen several times along each edge. IR expansion is responsible for most gene duplications. Based on this assumption, we could infer the possible evolutionary process from the internal node of *Int2* to *nt* and *Int1* to *no*, shown in Figure A-3 . This is a case of IR expansion.

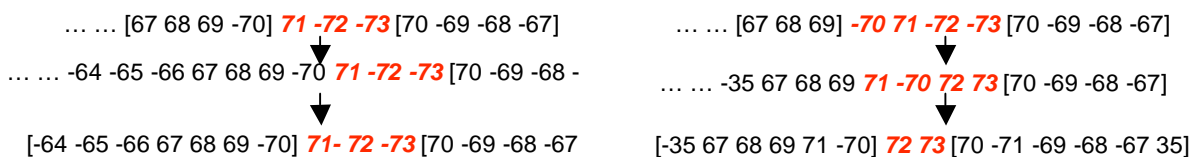


Figure A-3: The inferred gene content evolution process from *Int2* to *nt* (left) and *Int1* to *no* (right). Only IR (in square brackets) and SSC (in red) are shown.

For example, on the path that *Int2* was transformed to *nt*, the segment (-64 -65 -66) annealed with the original IR (67 68 69 -70) to form a new IR. If we remove the duplicates from the resulting IR, the gene contents of *Int2* and *nt* would be identical. From the observation above, we can further simplify the gene content of IR and SSC so that in the evolutionary path IR regions for all genomes (leaves and internal) contain gene (67 68 69), and the SSC regions contain gene (70 71 72 73). The simplified gene content map is shown in Figure A-4 .

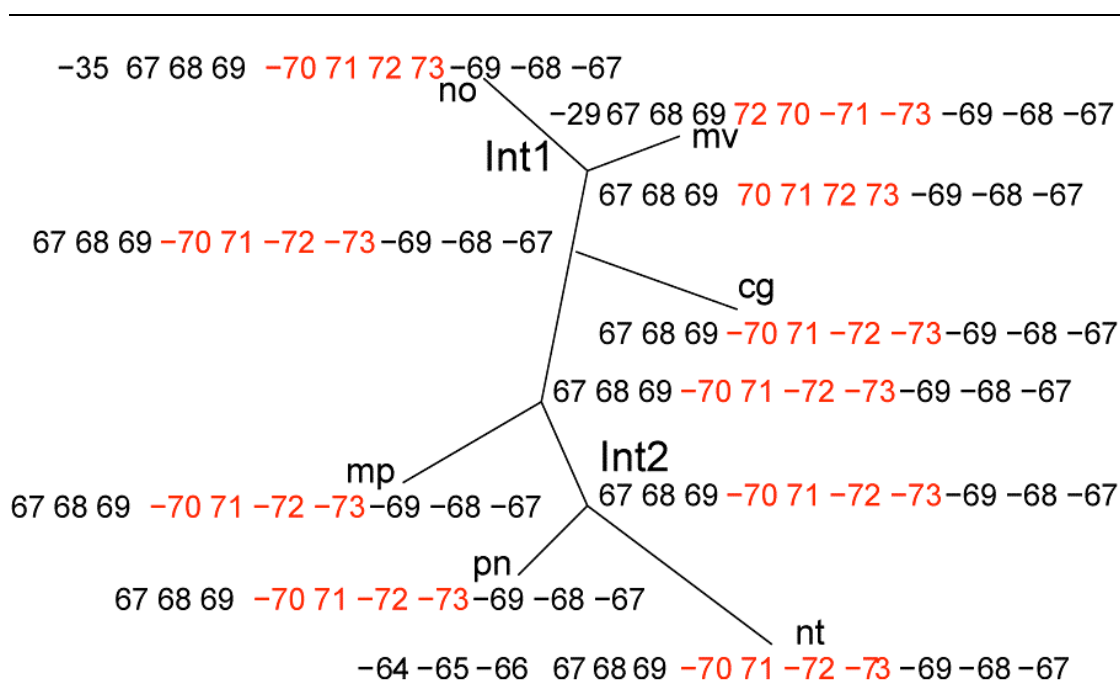


Figure A-4: The revised gene content for each region (only IR and SSC are shown).

By our observation, the movement of SC genes to IR is common, so that the additional rule is used to uniquely determine the gene order when duplicates exist in one child. If one gene in the IR of one child has been determined to belong to SC in the parent node, then the duplication of that gene was due to IR expansion in the child. Two gene orders are created, each containing one of the duplicate copies. On the other hand, if one IR gene in the parent is moved to SC region in the child, the “ancestral” IR gene is inserted to all possible locations in IR of the child to create multiple data sets. This operation treats duplicate genes at the boundaries of IR and SC as the last step towards the observed gene orders in the evolutionary path. Then it is possible to reduce the problem to all leaf genomes of equal gene content (or with deletions).

We then reconstructed the phylogeny after the gene contents of ancestral genomes were determined. Since the gene contents are reduced to equal after the simplification step, it is feasible to use GRAPPA to infer an inversion phylogeny after removing duplicate genes. However, each region may reflect a conflict history, although unlikely, which could lead to unresolved phylogeny. Thus, we develop a new method, called GRAPPA-IR, which estimates inversions bounded by the boundary of IRs.

The new method still uses the exhaustive approach: to score a tree, it needs to solve the median problems of three genomes iteratively until no improvement can be found. However, this method differs from GRAPPA in the way it solves the median problems.

For three given genomes G_1 , G_2 and G_3 , solving the median problem is to find a genome G_0 that can minimize the sum of distances from itself to three given genomes. Since inversions do not cross IR boundaries, thus inversions in each region (LSC, SSC or IR) occur independently from other regions. In other words, the median problem can be divided into three sub-median problems, each of which is constructed from genes in the same region of the genome G_1 , G_2 and G_3 . The sub-median problems can be solved separately using available inversion median solvers.

Simulations

We set out to test the accuracy of GRAPPA-IR by simulations. For this purpose, we generated datasets of 6 and 10 genomes and chose genomes of 78 genes (70 genes in the LSC, 5 in the SSC and 3 in the IR), roughly in the range of our dataset described in

the paper. We chose a large range of evolutionary rates r , the expected number of evolutionary events along an edge. We used r in the range of 4 — 10, which means that the actual number of inversions along each edge is sampled from a uniform distribution on the set $\{1, 2, \dots, 2r\}$. Given the model tree, we assigned the identity gene order to the root, and randomly generated gene order for each node based on the edge length and the gene order of its parent, with the assumption that inversions can not cross the IR boundaries. For each combination of parameter settings, we simulated 10 datasets and averaged the results.

Given an inferred tree (reconstructed phylogeny), we can assess the topological accuracy in terms of *false positives* and *false negatives* (25) with respect to the true tree. If an edge in the true tree is missing in the inferred tree, this edge is then called a *false negative* (FN). Similarly, a *false positive* edge (FP) appears in the inferred tree, but not in the true tree. The FP and FN rates are the number of false positives and false negatives divided by the number of edges (of non-zero length) in the true tree.

We compared the GRAPPA-IR to the original GRAPPA. We considered all trees with the minimum score given by both methods and took their strict consensus. Therefore, the trees returned by both methods need not to be fully resolved and they tend to have somewhat better rates for false positives than for false negatives. Thus we report FN rates rather than FP rates or a single Robinson-Foulds score (25). Figure A-5 shows simulation results. It indicates that when the evolutionary rate $r < 10$, GRAPPA-IR is more accurate than GRAPPA.

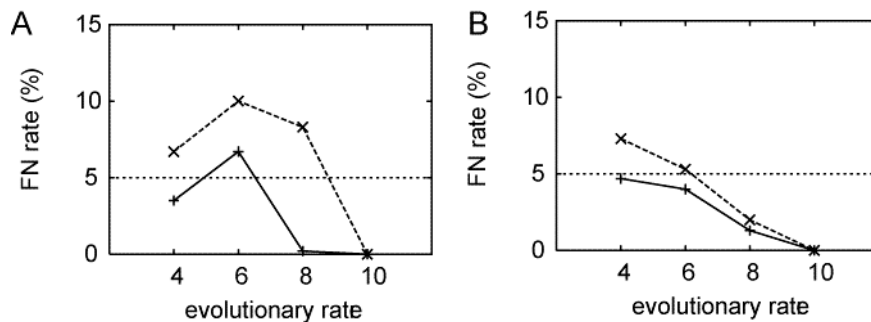


Figure **A-5**: False negative rates for GRAPPA-IR (solid line) and GRAPPA (dashed line) as a function of evolutionary rate r on the simulated datasets. A. 6 genomes. B. 10 genomes. The horizontal line indicates 5% error, a typical threshold of acceptability for accuracy in phylogenetic reconstruction (26)

Results

We evaluate all trees for the six genomes using the new method. The best score returned is 76 after 100 min of computation on a PIV 3.4GHz workstation. The best tree agrees with the reference tree (Figure **A-6**). All the other trees are clearly worse, with scores no less than 78.

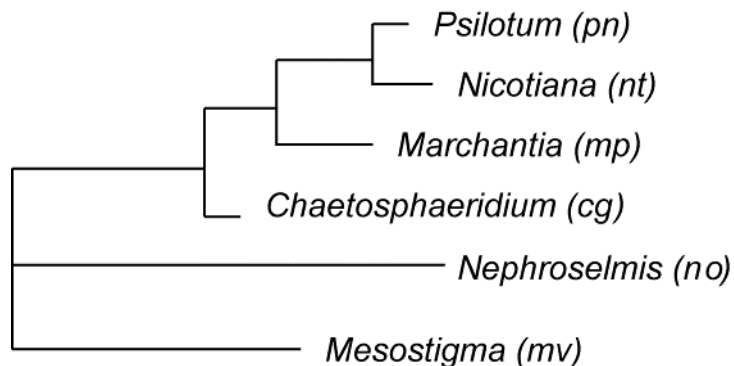


Figure A-6: The best tree returned by GRPPA-IR. The topology is the same as the reference tree.

We also test the data set with original GRAPPA, ignoring the region boundaries. The inference allows inversions to occur across IR and single copy regions. The best obtained has the same score of 76, yet the topology (Figure A-7) is very different from the result of the previous test and is in conflict with the reference tree.

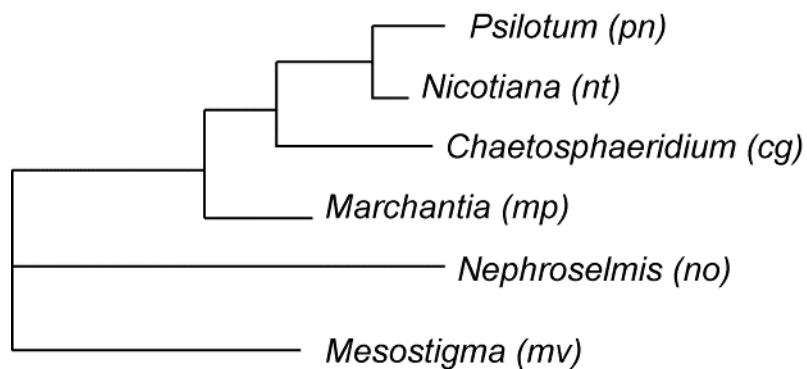


Figure A-7: The best tree obtained by GRAPPA without duplicate genes and SC/IR boundaries, which is different from the reference tree.

Discussion

Ancestral Gene Cluster

We are able to reconstruct ancestral gene orders from chloroplast genomes of land plants, green algae and a flagellate protist, which were separated by at least 450 million years of evolution (27). The ancestral chloroplast genomes of land plants and algae contain IRs, which is consistent with the hypothesis that the IR is a feature derived early in the chloroplast endosymbiosis (28). In addition, the ancestral IR contains the same gene content as that of *Mesostigma*, supporting that *Mesostigma* chloroplast genome encodes several ancestral gene clusters (24). By comparison of ancestral gene orders to the extant genomes, it is possible to test formally the evolutionary force of gene order changes. For example, maintenance of ancestral gene clusters may be related to function or transcriptional advantage, and thus these gene clusters are under constraints in the face of genome rearrangements.

IR and Genome Stability

The gene content of the IR varies across land plants, even in a single genus or family (20). It is known that homologous recombination is frequent between the two copies of IR. In a single chloroplast, hundreds of copies of chloroplast DNA co-exist as circular monomer, dimer and linear chromosomes (29). In the cellular endosymbiosis environment, the selection on accuracy of replication may have been relaxed to the degree that unequal recombination and replication slippage contribute to the expansion or shrinkage of IRs. On the other hand, the intra-molecular recombination process should

homogenize the sequences of the two IRs and thus the particular IR size and the gene content are maintained. The two counteracting phenomena may have played important roles in shaping the current diversity of chloroplast genome gene orders.

We found that incorrect gene order phylogenies were recovered without consideration of the IR boundary information. This strongly suggests that maintenance of IRs is necessary in the evolution of chloroplast genomes in most of the cases. We propose that IR provides an insulation mechanism that stabilizes the genome structure, and the genes in single copy regions do not commute across the IR. This agrees with the observation that gene rearrangements are more frequent in chloroplast genomes without IR (30). However, some genomes with residual IRs but infrequent gene movements between single copy regions compared to related lineages do not conform to the hypothesis (17). Future experimental studies on highly rearranged chloroplast genomes, for example, in the green alga *Chlamydomonas* lineage, may shed light on the maintenance of IR and genome rearrangements.

Comparison to Other Methods

Extensive tests show that trees returned by GRAPPA are superior to those returned by other gene-order phylogeny methods. The closely related package of Pevzner's group, MGR (31), is the only one that approaches its accuracy. GRAPPA-IR is mostly suitable for small data sets with insulated inversions, while for eukaryote genomes more efficient algorithms need to be developed to estimate much more rearrangements. For example, duplications and deletions are considered frequent as shown by the reconstruction of one 1.1 Mb region in the eutherian mammal ancestor (12). A combination of disc-covering

and other approaches may scale up the capability to infer ancestral gene order for large genomes (9,32).

Conclusions

We implement a new method to infer ancestral gene orders with duplications. Tests on a real data set show accurate recovery of the genome phylogeny as well as fast inference of ancestral gene orders. It provides new insight into the chloroplast genome evolutionary process.

Acknowledgements

The work was supported by the National Science Foundation grants DBI 0115684, DEB 0120709 to C.W.D and Department of Computer Science and Engineering, University of South Carolina to J.T.

References

1. Hurst, L.D., Pal, C. and Lercher, M.J. (2004) The evolutionary dynamics of eukaryotic gene order. *Nat Rev Genet*, **5**, 299-310.
2. Kent, W.J., Baertsch, R., Hinrichs, A., Miller, W. and Haussler, D. (2003) Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc Natl Acad Sci U S A*, **100**, 11484-11489.
3. Sankoff, D., Leduc, G., Antoine, N., Paquin, B., Lang, B.F. and Cedergren, R. (1992) Gene order comparisons for phylogenetic inference: evolution of the mitochondrial genome. *Proc Natl Acad Sci U S A*, **89**, 6575-6579.
4. Blanchette, M., Kunisawa, T. and Sankoff, D. (1999) Gene order breakpoint evidence in animal mitochondrial phylogeny. *J Mol Evol*, **49**, 193-203.

5. Moret, B.M., Wang, L.S., Warnow, T. and Wyman, S.K. (2001) New approaches for reconstructing phylogenies from gene order data. *Bioinformatics*, **17 Suppl 1**, S165-173.
6. Moret, B.M., Wyman, S., Bader, D.A., Warnow, T. and Yan, M. (2001) A new implementation and detailed study of breakpoint analysis. *Pac Symp Biocomput*, 583-594.
7. Bader, D.A., Moret, B.M. and Yan, M. (2001) A linear-time algorithm for computing inversion distance between signed permutations with an experimental study. *J Comput Biol*, **8**, 483-491.
8. Moret, B.M.E., Siepel, A.C., Tang, J. and Liu, T. (2002) Inversion medians outperform breakpoint medians in phylogeny reconstruction from gene-order data. *Lecture Notes in Computer Science*, **2452**.
9. Tang, J. and Moret, B.M. (2003) Scaling up accurate phylogenetic reconstruction from gene-order data. *Bioinformatics*, **19 Suppl 1**, i305-312.
10. Tesler, G. (2002) GRIMM: genome rearrangements web server. *Bioinformatics*, **18**, 492-493.
11. Gibbs, R.A., Weinstock, G.M., Metzker, M.L., Muzny, D.M., Sodergren, E.J., Scherer, S., Scott, G., Steffen, D., Worley, K.C., Burch, P.E. *et al.* (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*, **428**, 493-521.
12. Blanchette, M., Green, E.D., Miller, W. and Haussler, D. (2004) Reconstructing large regions of an ancestral mammalian genome in silico. *Genome Res*, **14**, 2412-2423.
13. Murphy, W.J., Larkin, D.M., Everts-van der Wind, A., Bourque, G., Tesler, G., Auvil, L., Beever, J.E., Chowdhary, B.P., Galibert, F., Gatzke, L. *et al.* (2005) Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science*, **309**, 613-617.
14. Hughes, J.R., Cheng, J.F., Ventress, N., Prabhakar, S., Clark, K., Anguita, E., De Gobbi, M., de Jong, P., Rubin, E. and Higgs, D.R. (2005) Annotation of cis-regulatory elements by identification, subclassification, and functional assessment of multispecies conserved sequences. *Proc Natl Acad Sci U S A*, **102**, 9830-9835.
15. Raven, J.A. and Allen, J.F. (2003) Genomics and chloroplast evolution: what did cyanobacteria do for plants? *Genome Biol*, **4**, 209.
16. Cosner, M.E., Jansen, R.K., Palmer, J.D. and Downie, S.R. (1997) The highly rearranged chloroplast genome of *Trachelium caeruleum* (Campanulaceae): multiple inversions, inverted repeat expansion and contraction, transposition, insertions/deletions, and several repeat families. *Curr Genet*, **31**, 419-429.
17. Cosner, M.E., Raubeson, L.A. and Jansen, R.K. (2004) Chloroplast DNA rearrangements in Campanulaceae: phylogenetic utility of highly rearranged genomes. *BMC Evol Biol*, **4**, 27.
18. Price, R.A., Calie, P.J., Downie, S.R., Logsdon, J., J.M. and Palmer, J.D. (1990) In Vorster, P. (ed.), *Proc. int. Geraniaceae symp.* University of Stellenbosch, Monvillia, South Africa, pp. 235-244.

19. Perry, A.S., Brennan, S., Murphy, D.J., Kavanagh, T.A. and Wolfe, K.H. (2002) Evolutionary re-organisation of a large operon in adzuki bean chloroplast DNA caused by inverted repeat movement. *DNA Res*, **9**, 157-162.
20. Goulding, S.E., Olmstead, R.G., Morden, C.W. and Wolfe, K.H. (1996) Ebb and flow of the chloroplast inverted repeat. *Mol Gen Genet*, **252**, 195-206.
21. Plunkett, G.M. and Downie, S.R. (2000) Expansion and Contraction of the Chloroplast Inverted Repeat in Apiaceae Subfamily Apioideae. *Syst Bot*, **25**, 648-667.
22. Maul, J.E., Lilly, J.W., Cui, L., dePamphilis, C.W., Miller, W., Harris, E.H. and Stern, D.B. (2002) The *Chlamydomonas reinhardtii* plastid chromosome: islands of genes in a sea of repeats. *Plant Cell*, **14**, 2659-2679.
23. Tang, J., Moret, B.M., Cui, L. and dePamphilis, C.W. (2004) Phylogenetic Reconstruction from Arbitrary Gene-Order Data. *Proc IEEE Symp Bioinform Bioeng (BIBE'04)*, 592-599.
24. Lemieux, C., Otis, C. and Turmel, M. (2000) Ancestral chloroplast genome in *Mesostigma viride* reveals an early branch of green plant evolution. *Nature*, **403**, 649-652.
25. Robinson, D.R. and Foulds, L.R. (1981) Comparison of phylogenetic trees. *Math Biosci*, **53**, 131-147.
26. Swofford, D.L., Olson, G., Waddell, P. and Hills, D.M. (1996) In Hills, D. M., Moritz, M. and Mable, B. (eds.), *Molecular Systematics*. 2nd ed. Sinauer Associates, Sunderland, pp. 407-514.
27. Herrmann, R.G., Maier, R.M. and Schmitz-Linneweber, C. (2003) Eukaryotic genome evolution: rearrangement and coevolution of compartmentalized genetic information. *Philos Trans R Soc Lond B Biol Sci*, **358**, 87-97; discussion 97.
28. Palmer, J.D. (1985) In MacIntyre, R. J. (ed.), *Molecular Evolutionary Genetics*. Plenum Press, New York, pp. 131-240.
29. Bendich, A.J. and Smith, S.B. (1990) Structure of chloroplast and mitochondrial DNAs. *Curr Genet*, **17**, 421-425.
30. Palmer, J.D. and Thompson, W.F. (1982) Chloroplast DNA rearrangements are more frequent when a large inverted repeat sequence is lost. *Cell*, **29**, 537-550.
31. Bourque, G. and Pevzner, P.A. (2002) Genome-Scale Evolution: Reconstructing Gene Orders in the Ancestral Species. *Genome Res.*, **12**, 26-36.
32. Hannenhalli, S. and Pevzner, P.A. (1999) Transforming cabbage into turnip: Polynomial algorithm for sorting signed permutations by reversals. *J ACM*, **46**, 1-27.

VITA

Liyong Cui

EDUCATION

Ph.D., Biology, expected in 2006

Pennsylvania State University, University Park PA

Focus on bioinformatics and molecular evolutionary biology

Master of Applied Statistics, May 2005

Pennsylvania State University, University Park, PA

B.S., Genetics, July 2000.

Fudan University, Shanghai, China

SELECTED PUBLICATIONS

Cui L, Leebens-Mack J, Wang LS, Tang J, Rymarquis L, Stern DB, Depamphilis CW. (2006) Adaptive Evolution of Chloroplast Genome Structure Inferred Using a Parametric Bootstrap Approach. *BMC Evol Biol.* 6(1):13

Cui L, Veeraraghavan N, Richter A, Wall K, Jansen RK, Leebens-Mack J, Makalowska I, dePamphilis CW. (2006) ChloroplastDB: the Chloroplast Genome Database. *Nucleic Acids Res.* 34(Database issue):D692-6.

Leebens-Mack J, Raubeson LA, Cui L, Kuehl JV, Fourcade MH, Chumley TW, Boore JL, Jansen RK, depamphilis CW. (2005) Identifying the basal angiosperm node in chloroplast genome phylogenies: sampling one's way out of the Felsenstein zone. *Mol Biol Evol.* 22(10):1948-63

Tang J, Moret BME., Cui L, dePamphilis CW. (2004) Phylogenetic reconstruction from arbitrary gene-order data. *Proceedings of the fourth IEEE conference on Bioinformatics and Bioengineering (BIBE'04)* 592-599.

Maul J, Lilly JW, Cui L, dePamphilis CW, Miller W, Harris EH, Stern DB. (2002) The *Chlamydomonas reinhardtii* plastid chromosome: islands of genes in a sea of repeats. *Plant Cell* 14(11): 2659-2679

ACTIVITIES

Executive Panel, Chinese Friendship Association, Penn State University, 2001-02

Biology Dept Graduate Student Association, Penn State University 2003-04

Student Red Cross Club, Penn State University, 2003-04

Penn State Yan Xin Qigong Club 2002-05