The Pennsylvania State University

The Graduate School

Department of Biology

COMPUTATIONAL TOOLS AND THEIR APPLICATIONS IN PLANT COMPARATIVE GENOMICS

A Dissertation in

Biology

by

P. Kerr Wall

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2008

The thesis of P. Kerr Wall was reviewed and approved* by the following:

Claude W. dePamphilis
Professor of Biology
Dissertation Advisor
Co-Chair of Committee

Webb Miller
Professor of Biology and Computer Science and Engineering
Co-Chair of Committee

Hong Ma
Distinguished Professor of Biology

Naomi S. Altman
Associate Professor of Statistics

John E. Carlson
Professor of Molecular Genetics, School of Forest Resrouces
Co-director, Bioinformatics and Genomics

Douglas Caverner
Professor of Biology
Head of the Department of Biology

*Signatures are on file in the Graduate School

ABSTRACT

The integration and advancements of molecular biology, evolution, and computer science over the past few decades have led to the development of several new fields of study. Comparative genomics, the study of the similarities and differences between two or more genomes, continues to be fueled by the rapidly growing number of fully sequenced genomes in our public databases. As of 2008, the plant scientific community has sequenced ten plant genomes, with plans to sequence more than twenty genomes over the next few years. Therefore, there is a need for flexible, gene family focused databases that provide rich toolsets for comparative analyses of plants. The PlantTribes database is based on the results of a series of controlled protein clustering experiments performed at multiple stringencies, which produce sets of objectively defined plant gene families. Nearly a dozen published articles to date have relied on data extracted from PlantTribes including the recent *Populus* and *papaya* genome sequence papers, expression divergence following gene duplication, identification of gene families for intensive phylogenetic analysis, identification of microRNAs and their associated targets, and genome duplication history of basal angiosperms. Comparative genomics has also been aided in the rapid advancements in sequencing technologies over the last few decades. Next Generation (NG) sequencing technologies have become a great resource to the genomics community because of the extremely low 'per base' cost of sequencing. A simulation approach was developed to help determine the optimal mixture of sequencing methods for most complete and cost effective transcriptome sequencing. In terms of sequence coverage alone, the NG sequencing platforms are a dramatic advance over capillary-based sequencing. Sequencing and microarray outcomes from multiple experiments suggest that our simulator will be useful for guiding Next Generation transcriptome sequencing projects in a wide range of organisms.

TABLE OF CONTENTS

LIST OF FIGURES....vi

LIST OF TABLES....vii

ACKNOWLEDGEMENTS....viii

Chapter 1 Introduction....1

References....8

Chapter 2 PlantTribes: A gene family resource for comparative genomics in plants....12

Preface....12
Abstract....13
Introduction....14
Database production....15
Phylogenetic analysis pipeline....18
Data access and retrieval....20
The utility of PlantTribes for gene family analyses....22
Conclusions and future perspectives....25
Funding....26
Acknowledgements....27
References....27

Chapter 3 Gene Family Characteristics of *Populus trichocarpa*, a model tree, and comparisons with the model systems of *Arabidopsis* and rice....30

Preface....30
Abstract....31
Introduction....32
Results....34
Tribe size distributions....34
Many tribes are stable and strongly supported....39
The distribution of *Arabidopsis*, rice, and *Populus* genes among tribes elucidates the pattern of genome expansion....44
Tribe classification is not skewed by highly abundant and promiscuous domains....47
Discussion....51
Methods....55
Proteome clustering....55
Tribe stability and support....55
Phylogenetic analysis....56
PFAM domain analysis....56
Super tribes....57
Acknowledgements....57
References....58

Chapter 4 Comparison of next generation sequencing technologies for transcriptome characterization 63

Preface....63
Abstract....65
Background....65
Results....65
Conclusions....66
Background....67
Results....69
Next Generation Transcriptome sequencing of *Arabidopsis* floral tissue....69
Transcriptome sequencing of *Eschscholzia californica* using oligo-dT and random-primed libraries....74
Transcriptome sequencing in a normalized library of *Persea americana*....75
Correlation of observed *Arabidopsis* transcript frequencies with microarray data....77
Next Generation transcriptome simulation study....79
Combinations of traditional and NG sequencing....87
Discussion....90
NG transcriptome sequencing....90
NG sequencing simulation studies and comparisons of platforms....91
Analysis of gene expression by NG sequencing....95
NG sequencing can be scaled to suit different project goals....96
Conclusion....97
Methods....98
RNA preparation....98
mRNA purification and 454 library construction for *Arabidopsis thaliana* and *Eschscholzia californica*....99
Normalized cDNA library construction in *Persea americana*....99
Sequence analysis....100
Simulation studies....101
Acknowledgements....104
References....105

Chapter 5 Conclusions and Future Directions....109

References....111

LIST OF FIGURES

Figure **1-1**: Phylogenetic Tree of Viridiplantae.....3

Figure **2-1**: PlantTribes Database Production.....16

Figure **2-2**: Schematic Diagram describing navigation through the PlantTribes database.....21

Figure **2-3**: Tribes as Gene Family Approximations.....24

Figure **3-1**: Tribe e-value frequency distribution. 37

Figure **3-2**: Median tribe e-value as a function of tribe size. 38

Figure **3-3**: Tribe size in relation to clustering stringency for one, two, and three species tribe analyses. 40

Figure **3-4**: Cumulative frequencies of jackknife support values for one, two, and three-species tribe analyses. 42

Figure **3-5**: Jackknife support values in relation to tribe size. 43

Figure **3-6**: Number of *Populus* and *Arabidopsis* genes per tribe in all tribes with at least one gene from each species. 45

Figure **3-7**: Promiscuous versus non-promiscuous PFAM domains defined by InterDom. 49

Figure **4-1**: Distributions of relative start sites and number of reads per gene.....76

Figure **4-2**: Correlation of gene expression with number of transcripts. 78

Figure **4-3**: Simulation results for different Next Generation sequencing technologies. 81

Figure **4-4**: Simulation results for combinations of Next Generation sequencing technologies. 89

Figure **4-5**: Probe expression distributions and relative uniqueness of varying x-mer sizes. 93

LIST OF TABLES

Table **1-1**: Genome Sequencing Progress of Viridiplantae 4

Table **3-1**: The total number of tribes, number of tribes with two or more genes, and number of singletons for each clustering at each stringency.....36

Table **3-2**: Percentages of genes within a genome that have best BLASTp values against each genome. 45

Table **3-3**: The PFAM domains present in the *Arabidopsis* only tribes. 50

Table **4-1**: Sequencing Statistics of analyzed libraries. 71

Table **4-2**: *Arabidopsis* 454 Reads Mapped to the annotated genome.....71

Table **4-3**: Top 10 Most Frequently Detected unigenes in 454 cDNA libraries of *Arabidopsis*, *Eschscholzia*, and *Persea*. 72

ACKNOWLEDGEMENTS

Although I had been trained separately in computers and biology when I came to Penn State, I had never integrated them until I started working for the Floral Genome Project. Therefore, I am grateful to my advisor, Claude dePamphilis, for guiding me in my early development and struggles in genomics and bioinformatics. Claude taught me how to think scientifically, independently, and outside the box. He also helped me develop a passion for science, especially evolutionary genomics. I would also like to thank Jim Leebens-Mack, who was just as influential in my early training in bioinformatics. Although I knew how to debug normal computer programs, debugging computational biology programs was more challenging at first. Jim helped me find confidence in my work and helped me realize the similarities in the two different types of programming. I would like to thank my committee members – Hong Ma, John Carlson, Naomi Altman, and Webb Miller. I have had the wonderful privilege to work with all of you on different projects, and each has helped in my professional development. I would also like to thank all of the members of the dePamphilis lab that I have had the fortunate chance to work with over the years while at Penn State: Jill Duarte, Barbara Bliss, Lena Landherr, Sheila Plock, Ali Barakat, Liying Cui, Yan Zhang, Yuannian Jiao, Joel McNeal, Josh Marion, Erik Wolcott, and Tony Orenga. I would like to thank Jonas Price for guiding my spiritual development over the last 4 years. I would like to thank my family back home in Louisiana. My parents, Ron and Judi, gave me all of the chances in life that any person would wish for. They taught me the meaning of hard work and discipline, which helped me become the hard-working son that I am today. I am grateful to my brother and sister, Kevin and Krystle, for helping me realize there is a world outside of science. Finally, I am indebted to my wife of ten years, Courtney, and our two children, Jordan and Taylor. They have given me strength when they did not realize it, but especially when I needed it the most.

Chapter 1

Introduction

The modern biologist works in an unprecedented time in the history of scientific knowledge. There are several factors that have been converging over the last few decades that have increased scientific productivity in a profound way. First, molecular biology has continued to be advanced through the tools and techniques that help researchers study genetics at the molecular level. Next, the decreased price of generating biological data has led to an enormous volume of deposited data in our national databases (NCBI, TIGR, JGI, etc.). Finally, the large volume of sequence data combined with the continued advances in the desktop computer, have led to the integration of molecular biology, evolution, and computer science. This integration has led to the development of several new fields of study including bioinformatics, genomics, and comparative genomics. Bioinformatics is the use of computer science, mathematics, and information theory to model and analyze biological systems. While genomics is the study of an organism's entire genome, comparative genomics is the study of similarities and differences between two or more genomes.

Over the last few decades, evolutionary biologists have made significant advancements in our understanding of the evolutionary history of Viridiplantae (Figure 1-1). This group of species, commonly referred to as green plants, diverged from the lineage leading to animals and fungi more than 1,500 million years ago (MYA) (Yoon et al. 2004; Merchant et al. 2007). Plants emerged from the oceans to colonize land approximately 450 MYA (Rensing et al. 2008). The group of non-vascular land plants, commonly referred to as the bryophytes, include liverworts, hornworts, and mosses. The next major step in the evolution of plants occurred approximately 350 MYA and involved the development of vascular tissue (Bowman et al. 2007). A primary

advantage of vascular tissue was that it gave plants the ability to grow taller in order to reach sunlight. A group of non-seed vascular plants associated with this evolutionary breakthrough and still in existence (extant) includes the ferns and allies. Plants developed seeds approximately 300 MYA, thus increasing the chance of survival during harsh conditions (Troitsky et al. 1991; Chaw et al. 1997). The group of non-flowering seed-plants are commonly referred to as the gymnosperms, which include pines, cycads, and ginkos,. Finally, plants evolved flowers approximately 150 MYA (Bremer et al. 2004; Friis et al. 2004; Bell et al. 2005; Leebens-Mack et al. 2005; Leebens-Mack et al. 2006; Rydin et al. 2006; Jansen et al. 2007). Flowering plants, also known as angiosperms, are the most numerous of modern plant taxa. The angiosperms are broken up into three major subdivisions, basal angiosperms, monocots, and dicots.

A large fraction of genes in plant genomes are the product of duplication and novel gene creation processes that have occurred within plants over their 500 million year history (Blanc and Wolfe 2004; Blanc and Wolfe 2004; Cui et al. 2006). Therefore, gene classifications that attempt to capture all of eukaryote diversity provide poor representations of plant genes. The PlantTribes database (Chapter 2) is a global classification of plant gene family space from all of the sequenced plant genomes and EST sets from more than 200 species. With the sequencing of more than ten plant genomes scheduled over the next two years, and dozens of additional plant genome projects being initiated (Table 1-1), there is a need for flexible plant-focused databases that provide a rich informatic toolset.

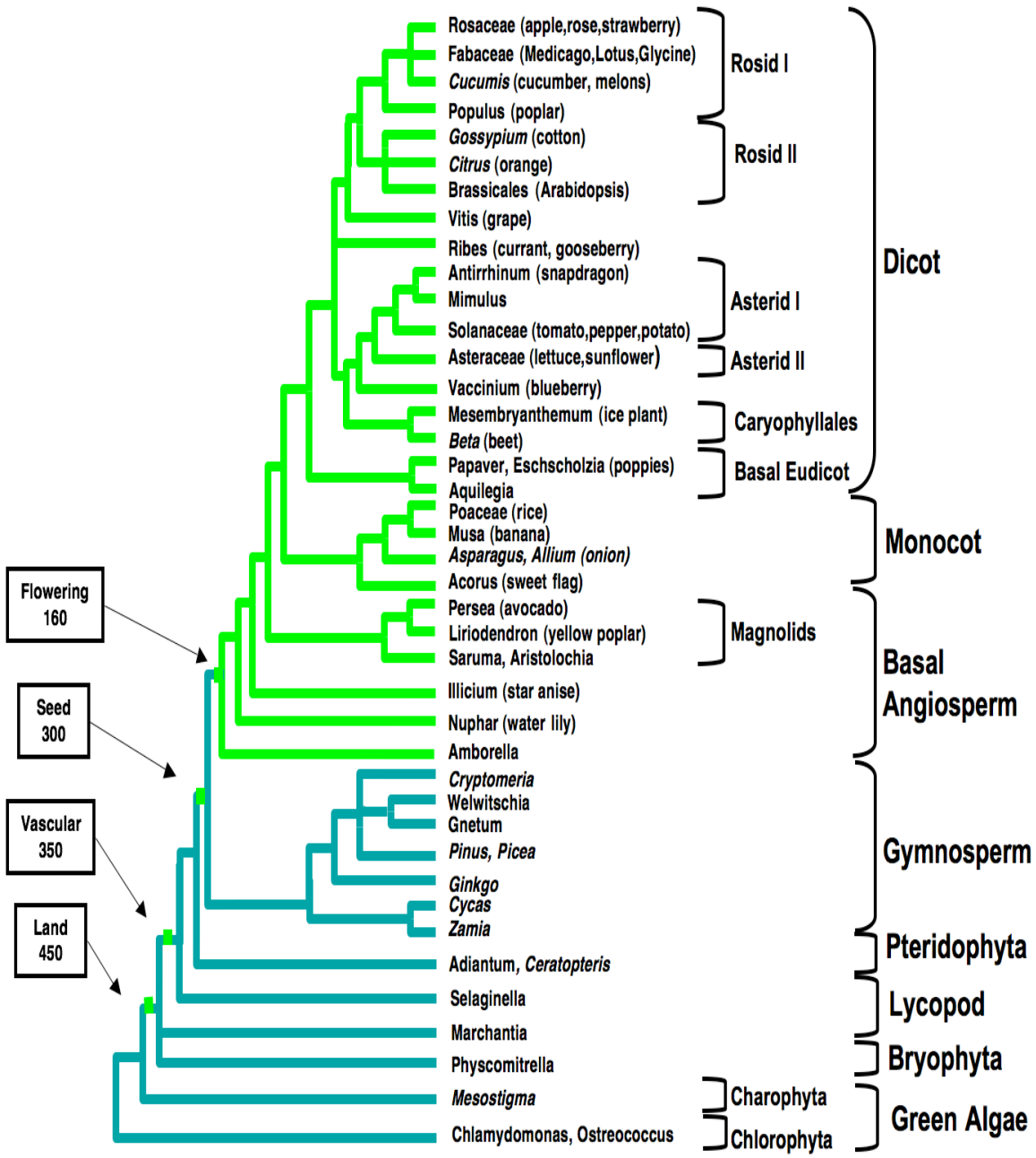

Figure **1-1**: Phylogenetic Tree of Viridiplantae. The light green branches are angiosperms, commonly referred to as flowering plants. The Angiosperms are broken up into the basal angiosperms, monocots, and dicots. The darker green branches are the sister species to the angiosperms, which include the gymnosperms, lycopods, bryophytes, green algae, etc. The approximate ages of the evolution of the major lineages are in the black boxes with arrows pointing to the appropriate nodes within the tree. This tree has been modified from Jansen et al (2007) and Leebens-Mack et al. (2006).

Table **1-1**: Genome Sequencing Progress of Viridiplantae. The status of each species is provided and is defined as either 'Finished' (F) or 'In Progress' (IP). For species with a completed genome sequence, the date of completion is also provided.

Species	**Clade**	**Status (F/IP)**
Arabidopsis thaliana	Rosid II	F - 2000
Orysa sativa (rice)	Monocot	F - 2003
Chlamydomonas reinhardtii	Green Algae	F - 2003
Populus trichocarpa	Rosid I	F - 2006
Physcomitrella patens	Bryophyta	F - 2007
Vitis vinifera (grape)	Rosid	F - 2007
Volvox cateri	Chlorophyta	F - 2007
Ostreococcus lucimarinus	Green Algae	F - 2007
Ostreococcus tauri	Green Algae	F – 2007
Selaginella moellendorffii	Lycopod	F - 2008
Carica papaya	Rosid II	F - 2008
Glycine max (soybean)	Rosid I	IP
Medicago trunculata	Rosid I	IP
Lotus japonicus	Rosid I	IP
Manihot esculenta (cassava)	Rosid I	IP
Prunus persica (peach)	Rosid I	IP
Arabidopsis lyrata	Rosid II	IP
Capsella rubella (pink shepherd's-purse)	Rosid II	IP
Eucalyptus grandis (Eucalyptus Tree)	Rosid II	IP
Gossypium (cotton)	Rosid II	IP
Theobroma cacao (cocoa)	Rosid II	IP
Mimulus guttatus (monkey flower)	Asterid I	IP
Solanum lycopersicum (tomato)	Asterid I	IP
Solanum tuberosum (potato)	Asterid I	IP
Aquilegia formosa	Basal Eudicot	IP
Acorus americanus	Monocot	IP
Brachypodium distachyon	Monocot	IP
Foxtail millet (Setaria italica)	Monocot	IP
Sorghum bicolor	Monocot	IP
Zea mays (corn)	Monocot	IP
Persea americana (avacado)	Basal Angiosperm	IP

PlantTribes offers a unique view of gene families and plant genomes that facilitate comparative analyses. For example, the database allows one to identify all gene families of a given size in a species and quickly assess the range of copy numbers for closely related genes in other plant genomes. Families that have remained stable in size, or have proliferated greatly in one genome compared to another can easily be identified. This type of analysis has aided interpretation of gene family stability and diversification in the face of gene and genome duplications. PlantTribes aids comparative analyses by serving as a scaffold of gene families into which users can sort their genes of interest. Search and query tools that allow the user to access this information, making it possible to investigate the evolution of plant genomes through analysis of the scaffold itself.

The PlantTribes database is based on the results of a series of controlled clustering experiments performed with MCL (Van Dongen 2000) clustering at three stringencies to produce sets of objective gene families from all of the sequenced plant genomes. Additional rounds of MCL clustering were used to identify super gene families (super tribes), and unified annotations were assigned to each tribe using common word patterns within the gene annotations. Additional information is connected to each sequence, including domain presence from NCBI's Conserved Domain Database (CDD) (Marchler-Bauer et al. 2002) and putative orthologous sequences. Following the assembly of this classification structure, we sorted into the tribe scaffold roughly 4 million unigene sequences, assembled from nearly 11 million sequences from dbEST, from the TIGR Plant Transcriptome Database (Childs et al. 2007). The annotations, sequences, tribe and super-tribe definitions, conserved domains, and sorted unigenes are loaded into a MySQL database with user searchable CGI scripts.

The PlantTribes database offers a unique and powerful view of plant genomes and evolution. Collaborators working on annotation and interpretation of gene models for the *Populus*, Papaya, and *Selaginella* Genome Projects have found the tribe results to be an

invaluable tool for gene family identification and annotation. Some of the major findings and results have been highlighted in the recent *Populus* (Tuskan et al. 2006) and *papaya* (Ming et al. 2008) genome sequence papers. Nearly a dozen published articles to date have relied on data extracted from PlantTribes including expression divergence following gene duplication (Duarte et al. 2006), identification of gene families for intensive phylogenetic analysis (Albert et al. 2005; Carlson et al. 2006; Kim et al. 2006; Leebens-Mack et al. 2006; Tuskan et al. 2006; Duarte 2007; Ming et al. 2008; Soltis et al. 2008), identification of microRNAs and their associated targets (Barakat et al. 2007; Barakat et al. 2007), and genome duplication history of basal angiosperms (Cui et al. 2006).

The first three plant genomes to be fully sequenced were *Arabidopsis thaliana* (AGI 2000), *Oryza sativa* (Goff et al. 2002; Yu et al. 2002), and *Populus trichocarpa* (Tuskan et al. 2006). *Populus* and *Arabidopsis*, both rosid eudicots, are separated by approximately 100 million years, whereas both shared a common ancestor with rice around 145 million years ago (Bell et al. 2005; Leebens-Mack et al. 2005). For these three predicted proteomes, I present the results of the gene family clustering and characterize the gene family space (Chapter 3). I introduce novel jackknife procedures used to explore the reliability of gene cluster delineation and membership assignment. The comparative analyses of these three plant species have helped to verify gene content in the predicted annotations and helped to elucidate the process of gene and genome duplication that have occurred throughout their respective evolutionary histories. The results provide insights into each proteome and the putative gene families provide a foundation for investigations of gene family evolution.

Next Generation sequencing platforms (454, Solexa, in particular) are of great interest to the genomics community because of the extremely low 'per base' cost of sequencing. Although the initial application for Next Generation sequencing was geared mainly toward genome re-sequencing, novel applications have emerged over the last few years. One of the most

compelling of these is the use of Next Generation platforms for transcriptome sequencing, and its applications for the study of expressed genes in organisms with both sequenced and unsequenced genomes. Next Generation transcriptomics is being widely considered as a replacement or complement to traditional (clone-based) EST sequencing, and even as a potential replacement for microarray-based expression studies. Indeed, several initial papers have shown that modest amounts of 454 sequencing can tag very large numbers of expressed genes, and indicate some promise for providing quantitative expression data. These studies have been largely descriptive, but they raise important questions about how much could be learned from more extensive studies, how to best design NG transcriptome studies, and the relative effectiveness of single or combined sequencing approaches.

In Chapter 5, I describe the first quantitative analysis and simulation study to address central questions regarding the use of Next Generation sequencing for transcriptome analysis. I adapted a proven simulation engine, ESTstat (Wang et al. 2004; Wang et al. 2005), which is used to simulate traditional capillary EST sequencing. The simulator considers relevant biological factors such as the relative distributions of transcript abundances and cDNA lengths. These biological factors can differ across species and different tissues within the same organism. The simulator also considers technical features such as read length, number of reads, and the distribution of sequence start sites. These technical factors are closely related to the sequencing platform and the approach used to build the cDNA library. The approach is easily adapted to transcriptomes of any organism, tissue, or combination of tissues. This claim is supported with surprising evidence that the most important biological variable in the model, the transcript abundance distribution, is very similar for different organisms and tissues.

The analyzed Next Generation datasets are drawn from original data of three species of plants and different library build approaches. The *Arabidopsis* transcriptome data is used to parameterize the model and develop quantitative predictions about the outcomes of small to very

large transcriptome studies using 454, Solexa, and traditional Sanger sequencing. I also examine the outcomes when combinations of these technologies are used. Many practical messages can be drawn from the analyses for design of new experiments, or for performing similar studies with other organisms.

References

AGI (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408(6814): 796-815.

Albert, V. A., D. E. Soltis, J. E. Carlson, W. G. Farmerie, P. K. Wall, D. C. Ilut, T. M. Solow, L. A. Mueller, L. L. Landherr, Y. Hu, M. Buzgo, S. Kim, M. J. Yoo, M. W. Frohlich, R. Perl-Treves, S. E. Schlarbaum, B. J. Bliss, X. Zhang, S. D. Tanksley, D. G. Oppenheimer, P. S. Soltis, H. Ma, C. W. DePamphilis and J. H. Leebens-Mack (2005). Floral gene resources from basal angiosperms for comparative genomics research. BMC Plant Biology 5: 5.

Barakat, A., K. Wall, J. Leebens-Mack, Y. J. Wang, J. E. Carlson and C. W. Depamphilis (2007). Large-scale identification of microRNAs from a basal eudicot (*Eschscholzia californica*) and conservation in flowering plants. Plant Jour 51(6): 991-1003.

Barakat, A., P. K. Wall, S. Diloreto, C. W. Depamphilis and J. E. Carlson (2007). Conservation and divergence of microRNAs in *Populus*. BMC Genomics 8: 481.

Bell, C. D., D. E. Soltis and P. S. Soltis (2005). The age of the angiosperms: a molecular timescale without a clock. Evolution Int J Org Evolution 59(6): 1245-58.

Blanc, G. and K. H. Wolfe (2004). Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. Plant Cell 16(7): 1679-91.

Blanc, G. and K. H. Wolfe (2004). Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell 16(7): 1667-78.

Bowman, J. L., S. K. Floyd and K. Sakakibara (2007). Green genes-comparative genomics of the green branch of life. Cell 129(2): 229-34.

Bremer, K., E. M. Friis and B. Bremer (2004). Molecular phylogenetic dating of asterid flowering plants shows early Cretaceous diversification. Syst Biol 53(3): 496-505.

Carlson, J. E., J. H. Leebens-Mack, P. K. Wall, L. M. Zahn, L. A. Mueller, L. L. Landherr, Y. Hu, D. C. Ilut, J. M. Arrington, S. Choirean, A. Becker, D. Field, S. D. Tanksley, H. Ma and C. W. Depamphilis (2006). EST database for early flower development in California poppy (*Eschscholzia californica* Cham., Papaveraceae) tags over 6000 genes from a basal eudicot. Plant Mol Biol.

Chaw, S. M., A. Zharkikh, H. M. Sung, T. C. Lau and W. H. Li (1997). Molecular phylogeny of extant gymnosperms and seed plant evolution: analysis of nuclear 18S rRNA sequences. Mol Biol Evol 14(1): 56-68.

Childs, K. L., J. P. Hamilton, W. Zhu, E. Ly, F. Cheung, H. Wu, P. D. Rabinowicz, C. D. Town, C. R. Buell and A. P. Chan (2007). The TIGR Plant Transcript Assemblies database. Nucleic Acids Res 35(Database issue): D846-51.

Cui, L., P. K. Wall, J. H. Leebens-Mack, B. G. Lindsay, D. E. Soltis, J. J. Doyle, P. S. Soltis, J. E. Carlson, K. Arumuganathan, A. Barakat, V. A. Albert, H. Ma and C. W. Depamphilis (2006). Widespread genome duplications throughout the history of flowering plants. Genome Res.

Duarte, J. M., L. Y. Cui, P. K. Wall, Q. Zhang, X. H. Zhang, J. Leebens-Mack, H. Ma, N. Altman and C. W. dePamphilis (2006). Expression pattern shifts following duplication indicative of subfunctionalization and neofunctionalization in regulatory genes of Arabidopsis. Molecular Biology And Evolution 23(2): 469-478.

Duarte, J. M., P.K. Wall, L.M. Zahn, J.H. Leebens-Mack, and C.W. dePamphilis. (2007). Utility of *Amborella trichopoda* and *Nuphar advena* ESTs for phylogeny and comparative sequence analysis. . Taxon.

Friis, E. M., K. R. Pedersen and P. R. Crane (2004). Araceae from the Early Cretaceous of Portugal: evidence on the emergence of monocotyledons. Proc Natl Acad Sci U S A 101(47): 16565-70.

Goff, S. A., D. Ricke, T. H. Lan, G. Presting, R. Wang, M. Dunn, J. Glazebrook, A. Sessions, P. Oeller, H. Varma, D. Hadley, D. Hutchison, C. Martin, F. Katagiri, B. M. Lange, T. Moughamer, Y. Xia, P. Budworth, J. Zhong, T. Miguel, U. Paszkowski, S. Zhang, M. Colbert, W. L. Sun, L. Chen, B. Cooper, S. Park, T. C. Wood, L. Mao, P. Quail, R. Wing, R. Dean, Y. Yu, A. Zharkikh, R. Shen, S. Sahasrabudhe, A. Thomas, R. Cannings, A. Gutin, D. Pruss, J. Reid, S. Tavtigian, J. Mitchell, G. Eldredge, T. Scholl, R. M. Miller, S. Bhatnagar, N. Adey, T. Rubano, N. Tusneem, R. Robinson, J. Feldhaus, T. Macalma, A. Oliphant and S. Briggs (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. japonica). Science 296(5565): 92-100.

Jansen, R. K., Z. Cai, L. A. Raubeson, H. Daniell, C. W. Depamphilis, J. Leebens-Mack, K. F. Muller, M. Guisinger-Bellian, R. C. Haberle, A. K. Hansen, T. W. Chumley, S. B. Lee, R. Peery, J. R. McNeal, J. V. Kuehl and J. L. Boore (2007). Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc Natl Acad Sci U S A 104(49): 19369-74.

Kim, S., P. S. Soltis, K. Wall and D. E. Soltis (2006). Phylogeny and domain evolution in the APETALA2-like gene family. Molecular Biology And Evolution 23(1): 107-120.

Leebens-Mack, J., L. A. Raubeson, L. Cui, J. V. Kuehl, M. H. Fourcade, T. W. Chumley, J. L. Boore, R. K. Jansen and C. W. depamphilis (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.

Leebens-Mack, J. H., K. Wall, Z. Zheng, D. Oppenheimer and C. W. dePamphilis (2006). A genomics approach to the study of floral developmental genetics: strengths and limitations. Developmental Genetics of the Flower. London, Elsevier Limited.

Marchler-Bauer, A., A. R. Panchenko, B. A. Shoemaker, P. A. Thiessen, L. Y. Geer and S. H. Bryant (2002). CDD: a database of conserved domain alignments with links to domain three-dimensional structure. Nucleic Acids Res 30(1): 281-3.

Merchant, S. S., S. E. Prochnik, O. Vallon, E. H. Harris, S. J. Karpowicz, G. B. Witman, A. Terry, A. Salamov, L. K. Fritz-Laylin, L. Marechal-Drouard, W. F. Marshall, L. H. Qu, D. R. Nelson, A. A. Sanderfoot, M. H. Spalding, V. V. Kapitonov, Q. Ren, P. Ferris, E. Lindquist, H. Shapiro, S. M. Lucas, J. Grimwood, J. Schmutz, P. Cardol, H. Cerutti, G. Chanfreau, C. L. Chen, V. Cognat, M. T. Croft, R. Dent, S. Dutcher, E. Fernandez, H. Fukuzawa, D. Gonzalez-Ballester, D. Gonzalez-Halphen, A. Hallmann, M. Hanikenne, M. Hippler, W. Inwood, K. Jabbari, M. Kalanon, R. Kuras, P. A. Lefebvre, S. D. Lemaire, A. V. Lobanov, M. Lohr, A. Manuell, I. Meier, L. Mets, M. Mittag, T. Mittelmeier, J. V. Moroney, J. Moseley, C. Napoli, A. M. Nedelcu, K. Niyogi, S. V.

Novoselov, I. T. Paulsen, G. Pazour, S. Purton, J. P. Ral, D. M. Riano-Pachon, W. Riekhof, L. Rymarquis, M. Schroda, D. Stern, J. Umen, R. Willows, N. Wilson, S. L. Zimmer, J. Allmer, J. Balk, K. Bisova, C. J. Chen, M. Elias, K. Gendler, C. Hauser, M. R. Lamb, H. Ledford, J. C. Long, J. Minagawa, M. D. Page, J. Pan, W. Pootakham, S. Roje, A. Rose, E. Stahlberg, A. M. Terauchi, P. Yang, S. Ball, C. Bowler, C. L. Dieckmann, V. N. Gladyshev, P. Green, R. Jorgensen, S. Mayfield, B. Mueller-Roeber, S. Rajamani, R. T. Sayre, P. Brokstein, I. Dubchak, D. Goodstein, L. Hornick, Y. W. Huang, J. Jhaveri, Y. Luo, D. Martinez, W. C. Ngau, B. Otillar, A. Poliakov, A. Porter, L. Szajkowski, G. Werner, K. Zhou, I. V. Grigoriev, D. S. Rokhsar and A. R. Grossman (2007). The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. Science 318(5848): 245-50.

Ming, R., S. Hou, Y. Feng, Q. Yu, A. Dionne-Laporte, J. H. Saw, P. Senin, W. Wang, B. V. Ly, K. L. Lewis, S. L. Salzberg, L. Feng, M. R. Jones, R. L. Skelton, J. E. Murray, C. Chen, W. Qian, J. Shen, P. Du, M. Eustice, E. Tong, H. Tang, E. Lyons, R. E. Paull, T. P. Michael, K. Wall, D. W. Rice, H. Albert, M. L. Wang, Y. J. Zhu, M. Schatz, N. Nagarajan, R. A. Acob, P. Guan, A. Blas, C. M. Wai, C. M. Ackerman, Y. Ren, C. Liu, J. Wang, J. Wang, J. K. Na, E. V. Shakirov, B. Haas, J. Thimmapuram, D. Nelson, X. Wang, J. E. Bowers, A. R. Gschwend, A. L. Delcher, R. Singh, J. Y. Suzuki, S. Tripathi, K. Neupane, H. Wei, B. Irikura, M. Paidi, N. Jiang, W. Zhang, G. Presting, A. Windsor, R. Navajas-Perez, M. J. Torres, F. A. Feltus, B. Porter, Y. Li, A. M. Burroughs, M. C. Luo, L. Liu, D. A. Christopher, S. M. Mount, P. H. Moore, T. Sugimura, J. Jiang, M. A. Schuler, V. Friedman, T. Mitchell-Olds, D. E. Shippen, C. W. dePamphilis, J. D. Palmer, M. Freeling, A. H. Paterson, D. Gonsalves, L. Wang and M. Alam (2008). The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). Nature 452(7190): 991-6.

Rensing, S. A., D. Lang, A. D. Zimmer, A. Terry, A. Salamov, H. Shapiro, T. Nishiyama, P. F. Perroud, E. A. Lindquist, Y. Kamisugi, T. Tanahashi, K. Sakakibara, T. Fujita, K. Oishi, I. T. Shin, Y. Kuroki, A. Toyoda, Y. Suzuki, S. Hashimoto, K. Yamaguchi, S. Sugano, Y. Kohara, A. Fujiyama, A. Anterola, S. Aoki, N. Ashton, W. B. Barbazuk, E. Barker, J. L. Bennetzen, R. Blankenship, S. H. Cho, S. K. Dutcher, M. Estelle, J. A. Fawcett, H. Gundlach, K. Hanada, A. Heyl, K. A. Hicks, J. Hughes, M. Lohr, K. Mayer, A. Melkozernov, T. Murata, D. R. Nelson, B. Pils, M. Prigge, B. Reiss, T. Renner, S. Rombauts, P. J. Rushton, A. Sanderfoot, G. Schween, S. H. Shiu, K. Stueber, F. L. Theodoulou, H. Tu, Y. Van de Peer, P. J. Verrier, E. Waters, A. Wood, L. Yang, D. Cove, A. C. Cuming, M. Hasebe, S. Lucas, B. D. Mishler, R. Reski, I. V. Grigoriev, R. S. Quatrano and J. L. Boore (2008). The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. Science 319(5859): 64-9.

Rydin, C., K. R. Pedersen, P. R. Crane and E. M. Friis (2006). Former diversity of *Ephedra* (Gnetales): evidence from Early Cretaceous seeds from Portugal and North America. Ann Bot (Lond) 98(1): 123-40.

Soltis, D. E., V. A. Albert, J. Leebens-Mack, J. D. Palmer, R. A. Wing, C. W. dePamphilis, H. Ma, J. E. Carlson, N. Altman, S. Kim, P. K. Wall, A. Zuccolo and P. S. Soltis (2008). The *Amborella* genome: an evolutionary reference for plant biology. Genome Biol 9(3): 402.

Troitsky, A. V., F. Melekhovets Yu, G. M. Rakhimova, V. K. Bobrova, K. M. Valiejo-Roman and A. S. Antonov (1991). Angiosperm origin and early stages of seed plant evolution deduced from rRNA sequence comparisons. J Mol Evol 32(3): 253-61.

Tuskan, G. A., S. Difazio, S. Jansson, J. Bohlmann, I. Grigoriev, U. Hellsten, N. Putnam, S. Ralph, S. Rombauts, A. Salamov, J. Schein, L. Sterck, A. Aerts, R. R. Bhalerao, R. P.

Bhalerao, D. Blaudez, W. Boerjan, A. Brun, A. Brunner, V. Busov, M. Campbell, J. Carlson, M. Chalot, J. Chapman, G. L. Chen, D. Cooper, P. M. Coutinho, J. Couturier, S. Covert, Q. Cronk, R. Cunningham, J. Davis, S. Degroeve, A. Dejardin, C. Depamphilis, J. Detter, B. Dirks, I. Dubchak, S. Duplessis, J. Ehlting, B. Ellis, K. Gendler, D. Goodstein, M. Gribskov, J. Grimwood, A. Groover, L. Gunter, B. Hamberger, B. Heinze, Y. Helariutta, B. Henrissat, D. Holligan, R. Holt, W. Huang, N. Islam-Faridi, S. Jones, M. Jones-Rhoades, R. Jorgensen, C. Joshi, J. Kangasjarvi, J. Karlsson, C. Kelleher, R. Kirkpatrick, M. Kirst, A. Kohler, U. Kalluri, F. Larimer, J. Leebens-Mack, J. C. Leple, P. Locascio, Y. Lou, S. Lucas, F. Martin, B. Montanini, C. Napoli, D. R. Nelson, C. Nelson, K. Nieminen, O. Nilsson, V. Pereda, G. Peter, R. Philippe, G. Pilate, A. Poliakov, J. Razumovskaya, P. Richardson, C. Rinaldi, K. Ritland, P. Rouze, D. Ryaboy, J. Schmutz, J. Schrader, B. Segerman, H. Shin, A. Siddiqui, F. Sterky, A. Terry, C. J. Tsai, E. Uberbacher, P. Unneberg, J. Vahala, K. Wall, S. Wessler, G. Yang, T. Yin, C. Douglas, M. Marra, G. Sandberg, Y. Van de Peer and D. Rokhsar (2006). The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 313(5793): 1596-604.

Van Dongen, S. (2000). A cluster algorithm for graphs. Technical Report INS-R0010.

Wang, J. P., B. G. Lindsay, L. Cui, P. K. Wall, J. Marion, J. Zhang and C. W. dePamphilis (2005). Gene capture prediction and overlap estimation in EST sequencing from one or multiple libraries. BMC Bioinformatics 6: 300.

Wang, J. P., B. G. Lindsay, J. Leebens-Mack, L. Cui, K. Wall, W. C. Miller and C. W. dePamphilis (2004). EST clustering error evaluation and correction. Bioinformatics 20(17): 2973-84.

Yoon, H. S., J. D. Hackett, C. Ciniglia, G. Pinto and D. Bhattacharya (2004). A molecular timeline for the origin of photosynthetic eukaryotes. Mol Biol Evol 21(5): 809-18.

Yu, J., S. Hu, J. Wang, G. K. Wong, S. Li, B. Liu, Y. Deng, L. Dai, Y. Zhou, X. Zhang, M. Cao, J. Liu, J. Sun, J. Tang, Y. Chen, X. Huang, W. Lin, C. Ye, W. Tong, L. Cong, J. Geng, Y. Han, L. Li, W. Li, G. Hu, X. Huang, W. Li, J. Li, Z. Liu, L. Li, J. Liu, Q. Qi, J. Liu, L. Li, T. Li, X. Wang, H. Lu, T. Wu, M. Zhu, P. Ni, H. Han, W. Dong, X. Ren, X. Feng, P. Cui, X. Li, H. Wang, X. Xu, W. Zhai, Z. Xu, J. Zhang, S. He, J. Zhang, J. Xu, K. Zhang, X. Zheng, J. Dong, W. Zeng, L. Tao, J. Ye, J. Tan, X. Ren, X. Chen, J. He, D. Liu, W. Tian, C. Tian, H. Xia, Q. Bao, G. Li, H. Gao, T. Cao, J. Wang, W. Zhao, P. Li, W. Chen, X. Wang, Y. Zhang, J. Hu, J. Wang, S. Liu, J. Yang, G. Zhang, Y. Xiong, Z. Li, L. Mao, C. Zhou, Z. Zhu, R. Chen, B. Hao, W. Zheng, S. Chen, W. Guo, G. Li, S. Liu, M. Tao, J. Wang, L. Zhu, L. Yuan and H. Yang (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). Science 296(5565): 79-92.

Chapter 2

PlantTribes: A gene family resource for comparative genomics in plants

Preface

This manuscript has been published in *Nucleic Acids Research*. The authors are P. Kerr Wall, Jim Leebens-Mack, Kai Müller, Dawn Field, Naomi S. Altman, and Claude W. dePamphilis. PKW designed, tested, and deployed the original database. PKW, JLM, and CWD wrote the manuscript and all authors suggested changes or edits, and approved the manuscript.

P. Kerr Wall[1], Jim Leebens-Mack[2,1], Kai Müller[3,1], Dawn Field[4], Naomi S. Altman[5], Claude W. dePamphilis[1*]

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

[2]Department of Plant Biology, University of Georgia, Athens, GA 30602, USA

[3]Nees Institute for the Biodiversity of Plants, University of Bonn, Meckenheimer Allee 170, 53115 Bonn, Germany

[4] Molecular Evolution and Bioinformatics Group, NERC Centre for Ecology and Hydrology, Mansfield Road, Oxford, OX1 3SR, UK

[5]Department of Statistics and The Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, PA 16802, USA

* corresponding author Claude dePamphilis: email cwd3@psu.edu; phone 814-863-641

Abstract

The PlantTribes database (http://fgp.huck.psu.edu/tribe.html) is a plant gene family database based on the inferred proteomes of five sequenced plant species: *Arabidopsis thaliana*, *Carica papaya*, *Medicago truncatula*, *Oryza sativa, and Populus trichocarpa*. We used the graph-based clustering algorithm MCL (Van Dongen 2000; Enright et al. 2002) to classify all of these species' protein-coding genes into putative gene families, called tribes, using three clustering stringencies (low, medium, and high). For all tribes, we have generated protein and DNA alignments and maximum likelihood phylogenetic trees. A parallel database of microarray experimental results is linked to the genes, which lets researchers identify groups of related genes and their expression patterns. Unified nomenclatures were developed, and tribes can be related to traditional gene families and conserved domain identifiers. SuperTribes, constructed through a second iteration of MCL clustering, connect distant, but potentially related gene clusters. The global classification of nearly 200,000 plant proteins was used as a scaffold for sorting approximately 4 million additional cDNA sequences from over 200 plant species. All data and analyses are accessible through a flexible interface allowing users to explore the classification, to place query sequences within the classification, and to download results for further study.

Introduction

A common goal of current plant genomics research is to establish an expandable platform for global classification and analysis of plant gene family-space. A large fraction of genes in plant genomes are the product of duplication and novel gene creation processes that have occurred within plants over their 500 million year history. Gene classifications that attempt to capture all of eukaryote diversity typically provide a poor representation of plant genes sets. With more than a dozen plant genomes scheduled for completion over the next two years, and many additional genome and transcriptome projects being initiated, there is a need for flexible, gene family focused databases that provide rich toolsets for comparative analyses of plant genomes. Comparative analyses of the modeled proteomes for sequenced genomes can help verify gene content and elucidate the process of gene duplication and functional diversification. Cross validation of gene models for available plant genomes and nucleotide sequence translations of EST sets for other plant species can be achieved through clustering and similarity analyses involving whole genome sequences and large EST sets (Dong et al. 2004; Rudd 2005; Hartmann et al. 2006; Childs et al. 2007).

The PlantTribes database is a global classification of genes from all of the five sequenced plant genomes: *Arabidopsis thaliana*, *Carica papaya* (papaya), *Medicago truncatula* (barrel medic, 60% sequenced), *Populus trichocarpa* (poplar), and *Oryza sativa* (rice). The database also contains unigene sets from the TIGR Plant Transcript Assemblies (Childs et al. 2007), which includes approximately 4 million sequences from more than 200 species, that facilitates a wide range of comparative study of plant genes and gene families. PlantTribes offers a unique view of objectively defined gene families that facilitates comparative analyses of plant genomes. For example, our database allows one to identify all gene families of a given size in a species and

quickly assess the range of copy numbers for closely related genes in other plant genomes. Families that have remained stable in size, or have proliferated greatly in one genome compared to another can easily be identified. In our own research, this type of analysis has aided interpretation of gene family stability and diversification in the face of gene and whole genome duplications (Albert et al. 2005; Cui et al. 2006; Leebens-Mack et al. 2006; Soltis et al. 2007). Integration of expression data, linked seamlessly to the tribe gene classification, will facilitate studies of expression divergence following gene duplication (Duarte et al. 2006). PlantTribes can aid comparative analyses by serving as a scaffold of gene families into which users can sort their genes of interest. We have devised search and query tools that allow users to access this information, making it possible to investigate the evolution of plant genomes through analysis of the scaffold itself and sequences sorted into the scaffold.

Database production

Sequences were downloaded from each of the five sequenced angiosperm species including 31,921 gene models from *Arabidopsis thaliana* (TAIR, version 7.0), 25,536 from *Carica papaya* (version 1.0, complete), 40,567 from *Medicago trunculata* (IMGA, version 1.0, 60% complete), 45,555 from *Populus trichocarpa* (JGI, version 1.0), and 66,710 from *Oryza sativa* (TIGR, version 5.0). The *Carica* and *Medicago* genome sequencing projects are underway; the data for these species were included with the protein scaffold and results for these species will go 'live' for public access following the publication of these genomes. As summarized in Figure 2-1, we compared the predicted proteins for all five species in an all-against-all blastp (e=1e-10, b=10000) using the NCBI blast package (Altschul et al. 1990). MCL clustering was then performed at low, medium, and high stringencies (Inflation, I = 1.2, 3.0, 5.0, respectively) to produce the sets of objectively defined gene families (tribes).

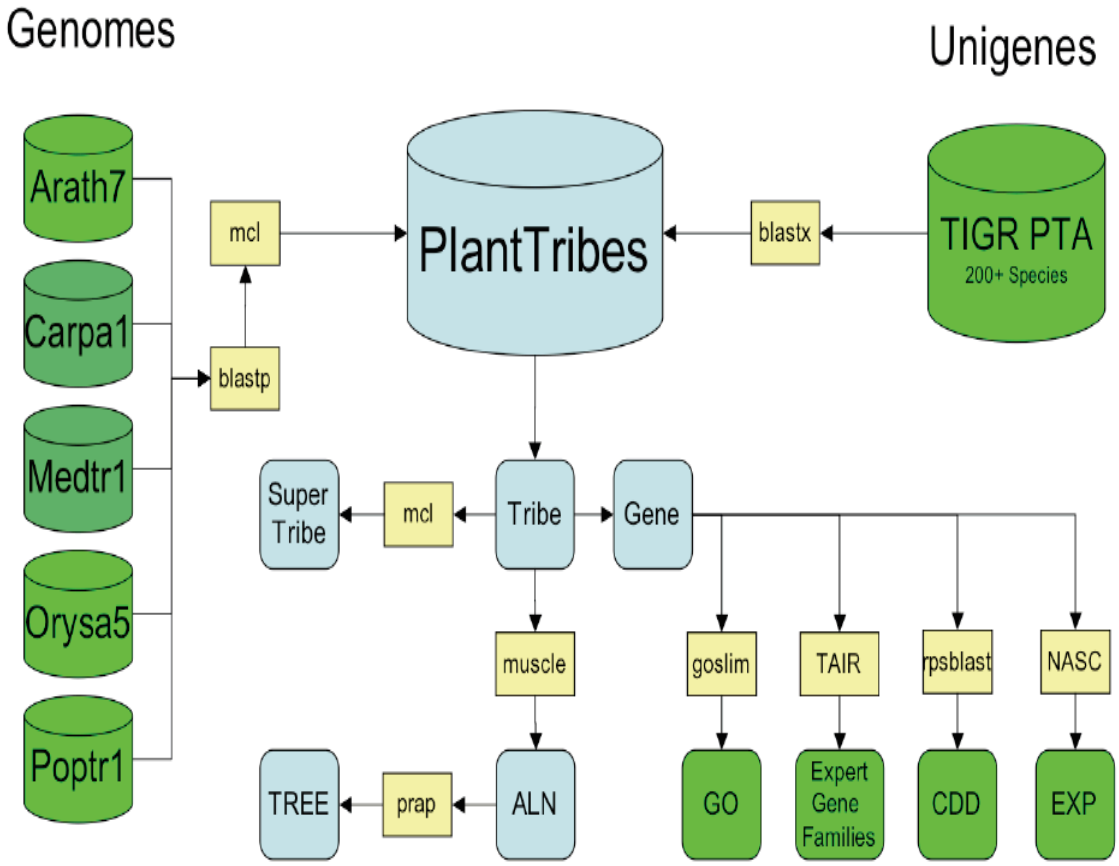

Figure **2-1**: PlantTribes Database Production.

Schematic diagram detailing the process of creating the PlantTribes database. External datasets are indicated in green, 'results' in blue, and software in yellow. First, an all-against-all blastp of five sequenced plant genomes is conducted with the results sent to MCL. Taxon abbreviations: Arath7 (*Arabidopsis thaliana*), Carpa (*Carica papaya*), Medtr1 (*Medicago truncatula*, currently 60% complete), Orysa5 (*Oryza sativa*), and Poptr1 (*Populus trichocarpa*). Darker green for *Carica* and *Medicago* indicate that although these genomes were included in the genome scaffold, tribe results for these species will not be accessible through the web interface of PlantTribes until the public release of these genomes. Tribes are produced at low, medium, and high stringencies and are annotated using Gene Ontology (GO), NCBI Conserved Domain Database (CDD), and expression data from NASCArrays (EXP). A second round of MCL clustering is performed on all tribes to group related tribes, called SuperTribes. For all tribes, protein and DNA alignments and maximum likelihood phylogenetic trees using prap are generated. Unigene sets from the TIGR Plant Transcriptome Assemblies are searched against the fully-sequenced genomes and are automatically sorted into respective tribes.

A second iteration of MCL was conducted in order to connect distant, but potentially related gene clusters which we define as SuperTribes. In order to construct SuperTribes, we computed both the average and minimum e-value between all pairs of tribes and used these as the input matrix for MCL. In addition, we ran MCL with low, medium, and high inflation values to generate SuperTribe clusters at the three different stringencies. In total there are 18 SuperTribe classifications for users to access and compare (i.e., 3 original tribe stringencies x 3 super tribe stringencies x 2 metrics ~ average/minimum evalue).

In order to annotate each tribe, we used additional information connected to all member genes according to the following criteria: gene ontology (GO), presence of domains, manually curated gene families, and common word patterns associated with the gene descriptions within a tribe. We downloaded the gene_ontology.v1.2.obo, goslim_plant.obo, and gene_association.TAIR flat files (Ashburner et al. 2000) and used the map2slim.pl script to create a GO slim database for the Arabidopsis genes in each tribe. To annotate our tribes by domain information, we downloaded NCBI's Conserved Domain Database (CDD) (Marchler-Bauer et al. 2002) and used the formatrpsdb (default parameters, with f=9.82, S=100.0) utility to index the domains. We then searched all protein sequences from the five genomes using rpsblast (default parameter, with evalue=1e-5). To annotate tribes according to manually curated gene families, we downloaded gene_family_tab_121906.txt from TAIR, which includes 996 gene families that include 8,331 genes. Finally, a perl script was used to extract all gene descriptions within a tribe, and determined the most common words within the tribe, keeping track of the relative position of each word, using only the top 5 words. Therefore, each tribe has a composite annotation defined by each of the four criteria.

The resulting constellation of gene family tribes was used as a scaffold for plant gene-space onto which roughly 4 million unigene sequences were sorted. These unigenes, derived from over 11 million ESTs, were downloaded from TIGR PTA (http://plantta.tigr.org). In

addition, we sorted the predicted proteomes of *Chlamydomonas reinhardtii* (green alga; JGI, version 3) and *Physcomitrella patens* (moss; JGI, version 1). We searched the 5 sequenced proteomes using a blastx search (evalue=1e-5) for the unigene sequences and a blastp (evalue=1e-1) search for the distantly related *Physcomitrella* and *Chlamydomonas* proteomes.

Phylogenetic analysis pipeline

A sequence alignment and phylogenetic analysis pipeline included the following steps. We generated fasta files of both amino acid and DNA sequences (CDS) for each tribe. Each amino acid file was aligned using the MUSCLE alignment program (Edgar 2004). We then forced the DNA sequences onto the amino acid alignments using custom perl/Bioperl scripts. Phylogenetic trees were built using a fast maximum likelihood ratchet approach (Morrison 2007) as newly implemented in PRAP (Müller 2004) v.2.0 for this study. PRAP generated command files that were handed over to PAUP (Swofford 2002). The heuristics involves (i) rapidly getting a starting tree not too far from the optimal score; (ii) move rapidly to a [near-] optimal tree island, (iii) getting the best tree within the island. Step (i) was achieved by calculating a BioNJ tree using LogDet distances, followed by one round of NNI and then one round of SPR branch swapping, optimizing the substitution model parameters between these steps. Similar to parsimony ratchet (Nixon 1999), step (ii) included alternating between branch swapping on the original matrix and branch swapping on a matrix with 25% of characters upweighted. Unlike Nixon's strategy for parsimony, SPR branch swapping was used, only ten iterations were performed, and during the weighted analyses, only one tree was saved. In particular for datasets with low levels of phylogenetic signal, this strategy was found to be more successful (Morrison 2007) than the strategies implemented in GARLI (Zwickl 2006) or RAxML (Stamatakis et al. 2005). To assess confidence in clades, bootstrapping was performed by executing PRAP-

generated command files in PAUP. Using optimized parameters from the likelihood ratchet search, SPR branch swapping was performed on the maximum likelihood topology for each bootstrapped data matrix, and the proportion of iterations in which a given clade was recovered was mapped onto the maximum likelihood tree using a strongly modified version of TreeGraph (Müller and Müller 2004). The latter program was also used to generate SVG trees that can be viewed via the web interface.

Understanding how gene expression patterns vary among gene family members will inform our understanding of evolutionary processes shaping plant gene function and genome structure. Characterization of changing gene expression following gene duplication and speciation events (Duarte et al. 2006),will improve as additional plant genomes are sequenced and genome-wide gene expression studies are preformed on a wide range of plant species (Shen et al. 2005). We aim to advance this research by placing gene expression data within a gene family context. To incorporate expression data into PlantTribes, we downloaded all AFFY expression data and associated descriptions of experiments, tissue, etc. from NASCArrays (Craigon et al. 2004). This has allowed us to link tribes with Arabidopsis genes to a curated expression dataset including 327 experiments conducted on more than 200 tissue types. Gene expression data for additional species will be added to future versions of PlantTribes as an ontology is developed to relate organs and developmental stages across plant species (Buzgo et al. 2004; Pujar et al. 2006; Ilic et al. 2007).

Data access and retrieval

All output discussed in the previous section was loaded into a MySQL database with user searchable cgi scripts. There are four main ways to search within the PlantTribes Database (Figure 2-2): 1) using a gene ID or annotation term for any of the Arabidopsis, rice, Populus, Medicago or papaya gene models; 2) using a CDD domain accession ID, name, or description; 3) running a BLAST search on a single sequence or file of user supplied sequences; or 4) querying the database of tribe characteristics. For example, all tribes with a minimum and/or maximum tribe size for each species or a threshold cumulative gene number can be retrieved with a simple query. HTML formatted search results include hyperlinks to sequence information, domain content and the tribes represented by each hit. All search results have links to the main page for each tribe within the database. Each tribe page includes the following information: unified annotation, stringency, SuperTribe identifier, the number of sequences from each species, all CDD entries for each of the genes in the tribe, a list of the genes in the tribe as well as each gene's tribe identifier at low, medium, and high stringency. Tribe stability can be readily examined through comparison of tribe membership at the different stringencies. Tools are also provided to view sequences from other species that have been sorted into each tribe and to view and/or download the sequences, alignments (constructed at both protein and DNA level), and phylogenetic trees in all major formats for each tribe.

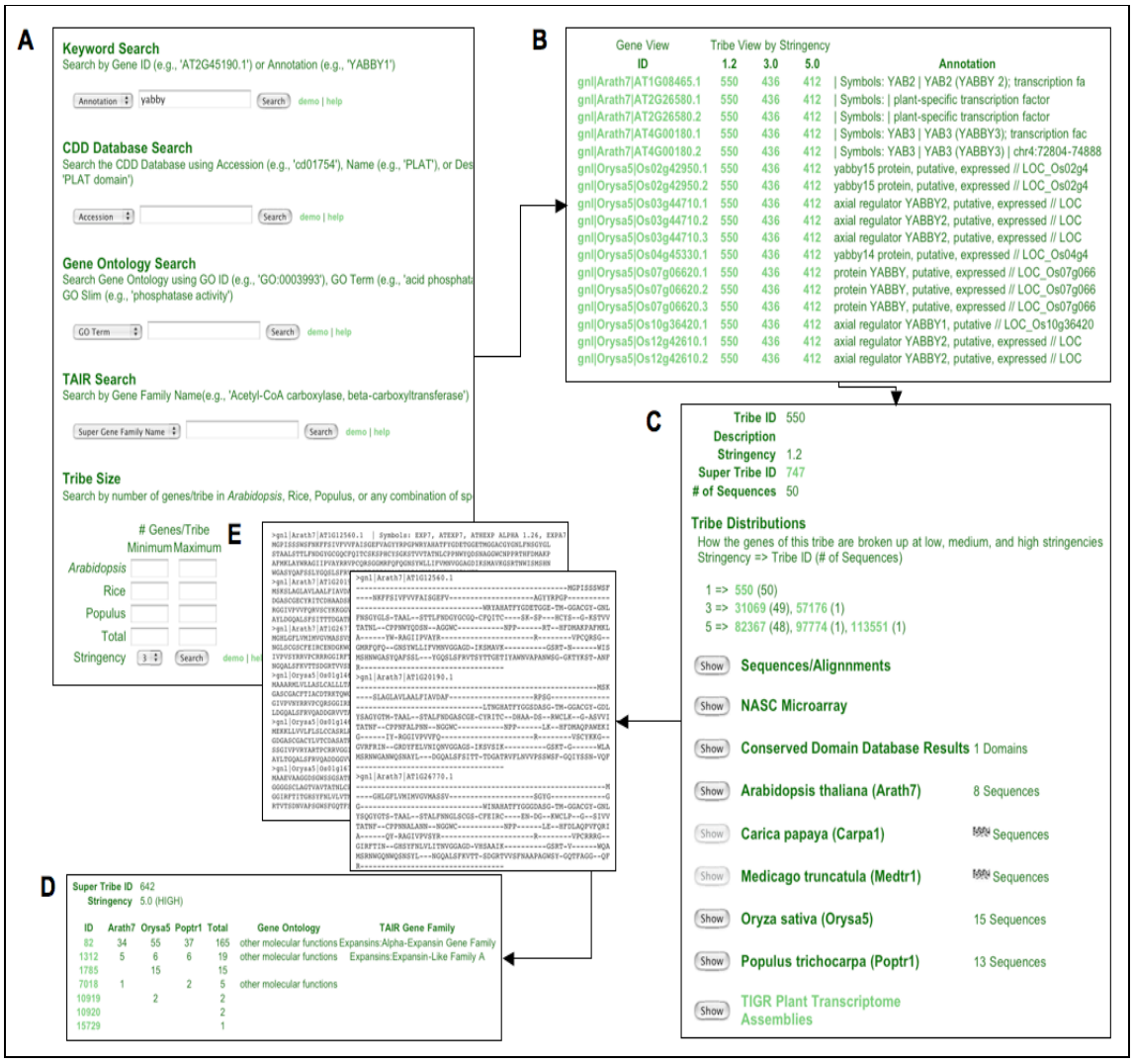

Figure **2-2**: Schematic Diagram describing navigation through the PlantTribes database.

A) A user can search by gene, domain, gene ontology, TAIR gene family annotations, and tribe size. B) All search results are linked into C) a tribe page with information about the tribe including tribe size distributions at low, medium, and high stringency MCL clustering, links to D) super tribe pages, domain information for all member genes of the tribe, a listing of all genes within the tribe, and E) a download/view area of additional data for each tribe including sequences, alignments, phylogenetic trees, and microarray expression data.

The utility of PlantTribes for gene family analyses

An important utility of PlantTribes is the ability to quickly find organism specific tribes. In poplar [27], the first sequenced tree, we were able to use this feature to identify tribes containing genes that are unique to that species (among those with sequenced genomes; criteria were no hits with e-values better than the 1.0E-10 threshold). These genome-specific gene models were quite distinct with no hits to any other sequences outside of their tribes. Genes with similar sequences, however, were found in the TIGR Plant Transcript Assembly database (e-values < 1.0E-10 in tblastx searches), suggesting that these may be expressed genes rather than annotation artifacts. A similar experimental strategy can be adopted to identify all organism-specific gene families as well as gene families that have been lost in one or more lineages.

Beyond the insights gained from the comparisons of the gene families, the PlantTribes database provides a useful scaffold for sorting new sequences. For example, the Floral Genome Project (Soltis et al. 2002; Albert et al. 2005; Soltis et al. 2007) has been sorting ESTs into the putative gene families defined by PlantTribes. Each unigene is searched against the fully sequenced plant proteomes using blastx, best hits are recorded and then used to tentatively place each unigene into a tribe. Further evaluation of tribe membership is facilitated by reports for each unigene showing the best hit(s) and the proportion of genes within each tribe with significant blast scores. This process has allowed us to immediately produce classifications of the genes we are finding in our EST data (Albert et al. 2005; Carlson et al. 2006; Kim et al. 2006; Duarte 2007; Soltis et al. 2007).

We have used the PlantTribes database to identify single copy tribes (genes with just 1 member from each species) whose memberships were stable across all three stringencies (Leebens-Mack et al. 2006). These shared single copy genes are more abundant than expected by chance, given the frequency of single copy genes in *Arabidopsis*, rice and Poplar, that have

resulted from gene death following gene and genome duplication in these lineages (Leebens-Mack et al. 2006). Identifying orthologs in EST sets from several basal angiosperms has also allowed us to infer lineage-specific substitution rate variation (Cui et al. 2006). The database has also aided the identification of paralogous pairs to explore gene duplication through angiosperm history (Duarte et al. 2006) and assess the frequency of expression shifts following gene duplication. In contrast, conserved gene expression in some single-copy genes suggests conserved function throughout angiosperm history (Leebens-Mack et al. 2006).

PlantTribes circumscribes objectively defined gene families, but we need to assess the degree to which the MCL clustering algorithm recovers evolutionary complete gene family clades. Using the expansin and MADS-box gene families as exemplars, we mapped tribe assignments at low, medium, and high stringencies onto previously reported phylogenies for these two well-studied gene families. We wanted to test the extent to which tribes represent "putative" gene families and investigate whether large tribes recovered at low stringency typically break up at higher stringencies into smaller tribes corresponding to subfamilies. We tested whether there is a strict nested relationship among tribes identified at low, medium and high stringencies, and if so, whether the nested pattern of relationships corresponds to the historical relationships and past gene duplication events as estimated in phylogenetic analyses.

Figure 2-3A contains the mapping of tribe membership from the three-way clustering to a phylogeny of the expansin superfamily (Sampedro and Cosgrove 2005). All of the expansin genes reported in the phylogeny from three expansin subfamilies, alpha, beta, and expansin-like, are found in only one tribe at low stringency. At medium stringency, the genes are broken into 2 tribes with one tribe containing the alpha and beta-expansins and the second tribe containing the expansin-like genes. At the highest stringency, the expansin-like alpha and expansin-like beta subfamilies are split from the main tribe containing the alpha and beta expansins.

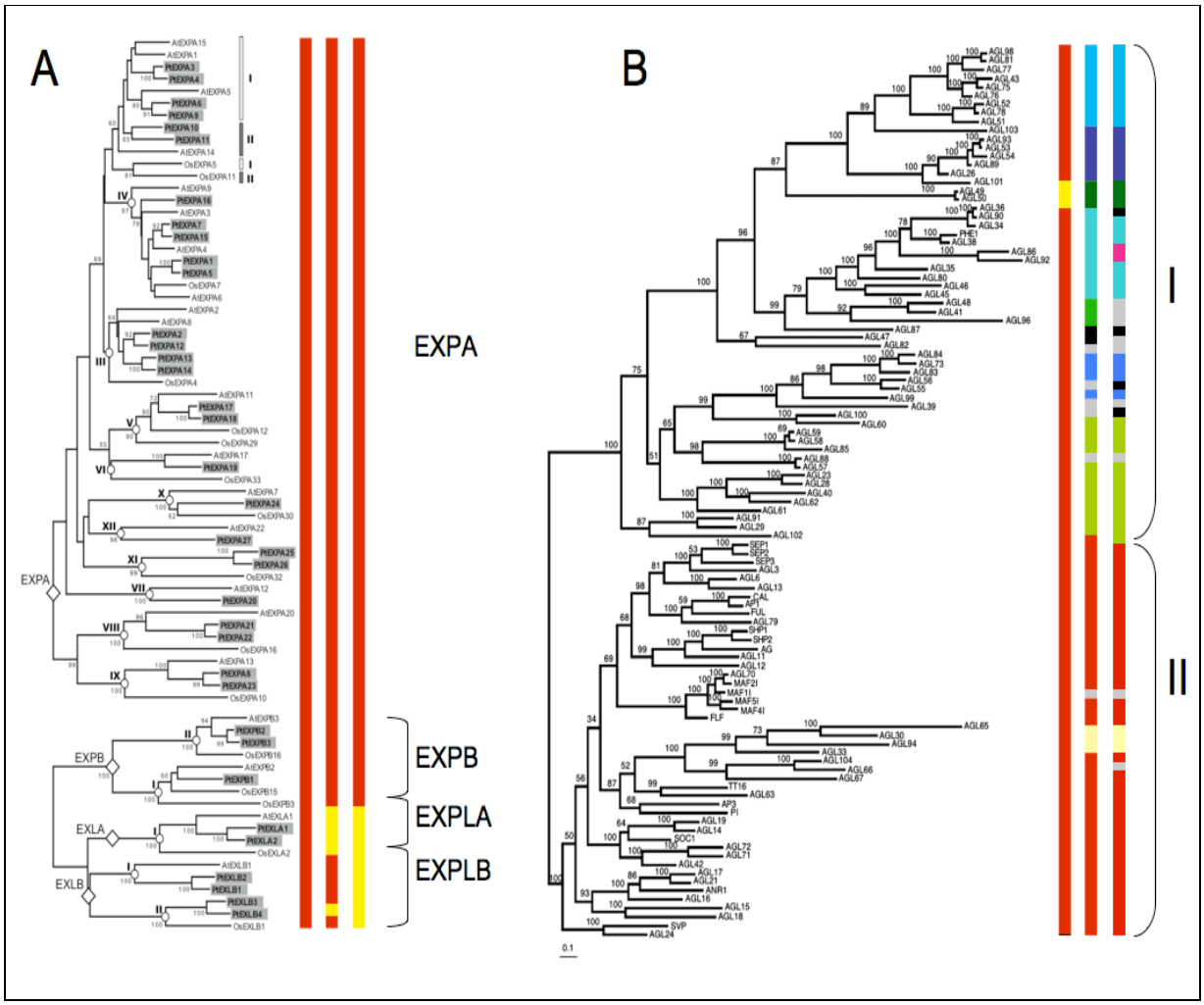

Figure **2-3**: Tribes as Gene Family Approximations.

Tribes with Expansin (A) and MADS box genes (B) formed at low, medium, and high stringencies in the three-species clustering are mapped onto a recently published gene phylogeny (Sampedro and Cosgrove, 2005; Martinez-Castilla and Alvarez-Buylla, 2003).

A) In the Expansin phylogeny, all genes are found in a single tribe at low stringency. At medium stringency, the genes are broken up into 2 tribes separating expansin-like A subfamily genes from all others expansin sub-families (tribe containing additional expansin-like genes not included in the original phylogeny). At high stringency, expansins are resolved as 2 tribes corresponding to the sub-families alpha + beta and expansin-like.

B) The MADS box genes (including type I & II) included in the phylogeny are in 2 tribes with all genes in 1 tribe except AGL49 and AGL50. At medium and high stringencies, well defined clades appear. The Type I genes break up into many more tribes than Type II genes, which is expected since Type I genes are more divergent among themselves. Within the Type II genes, AGL65, AGL30, AGL94 are broken out from the main tribe, which is to be expected since this group of genes are highly divergent Type II genes.

It would be desirable for tribes to generally correspond to monophyletic clades as was the case in the expansin superfamily. This would allow investigators to download and align sequences from a tribe with confidence that all genes in the alignment share a common ancestor and all extant descendents of that ancestral gene are included in the alignment. However, this is not always the case. Figure 3B contains the mapping of tribe membership from the three-way clustering to a well-accepted Arabidopsis MADS-box gene family phylogeny (Martinez-Castilla and Alvarez-Buylla 2003). At higher stringencies, small groups of related genes are peeled off of the low stringency clusters. As a consequence, the largest tribe identified at high stringency is paraphyletic with some divergent internal clades segregating into distinct tribes. Whereas nearly all type-II MADS-box genes were placed in a single tribe at low stringency, the complete MADS-box gene family (including type-I) was distributed among 14 tribes in the 5-species analysis. Even though the tribes at three stringencies may not always coincide with phylogenetic clades, the ability of tribe and supertribe analyses to capture a large number of related genes nevertheless provides an efficient starting point for investigations of gene and gene family diversification across complete genomes.

Conclusions and future perspectives

The PlantTribes database offers a unique and powerful view of plant genomes and evolution. Collaborators working on annotation and interpretation of gene models for the Poplar and papaya genomes found the tribe results to be an invaluable tool for gene family identification and annotation, and our results were highlighted in the recent Poplar genome sequence paper (Tuskan et al. 2006). More than 15 other published articles to date have relied on data extracted from PlantTribes including expression divergence following gene duplication, identification of novel functional motifs, identification of gene families for intensive phylogenetic analysis, and

genome duplication history of basal angiosperms. With many plant genome sequence projects in progress, formal comparative approaches such as PlantTribes will allow researchers to rapidly identify the best gene models, quickly determine errors in the initial annotations, identify new gene families, and increase the confidence in the limits and structure of existing gene families.

PlantTribes has been designed for ease of expansion and feature addition. As new genomes are sequenced, or large EST sets generated, PlantTribes will be continuously expanded to include these data. New features being developed presently include 1) a tool for the rapid incorporation of new query sequences into tribe alignments and phylogenies, 2) connecting the rapidly expanding microRNA database into PlantTribes so that genes that are putative targets of known or predicted miRNAs may be easily found, 3) expansion of the microarray database to include large-scale array experiments from basal angiosperms and other plants (Soltis et al. 2007) that will facilitate cross-species expression analyses, and 4) synteny-based tools to map genome duplications onto gene family phylogenies. As the number of sequenced genomes increases rapidly, the continued expansion of the PlantTribes database will facilitate a multitude of genome and gene family studies, particularly homology-based annotation, genome-scale analysis of multiple gene families, characterization of large gene families, and subsets of genes with common domain architectures.

Funding

National Science Foundation (DEB 0115684 to C.W.D. and J.H.L-M., DEB 0638595 to C.W.D. and J.H.L-M). K.F.M was supported by a scholarship from the Deutsche Telekom Stiftung.

Acknowledgements

The authors thank our faculty, postdoctoral, and student colleagues in the Floral Genome Project, the Ancestral Angiosperm Genome Project, and the poplar and papaya genome projects for their enthusiastic support and use of PlantTribes through its initial stages of development. We would like to thank Hong Ma, John Carlson, and Victor Albert for invaluable discussions of the biological implications of PlantTribes. Finally, we thank Josh Marion, Tony Orenga, Severn Everett, Kevin Beckmann, Anthony Carroll, and Erik Wolcott for their assistance in the development of portions of the PlantTribes database and web interface.

References

Albert, V. A., D. E. Soltis, J. E. Carlson, W. G. Farmerie, P. K. Wall, D. C. Ilut, T. M. Solow, L. A. Mueller, L. L. Landherr, Y. Hu, M. Buzgo, S. Kim, M. J. Yoo, M. W. Frohlich, R. Perl-Treves, S. E. Schlarbaum, B. J. Bliss, X. Zhang, S. D. Tanksley, D. G. Oppenheimer, P. S. Soltis, H. Ma, C. W. DePamphilis and J. H. Leebens-Mack (2005). Floral gene resources from basal angiosperms for comparative genomics research. BMC Plant Biol 5: 5.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman (1990). Basic local alignment search tool. J Mol Biol 215(3): 403-10.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin and G. Sherlock (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25(1): 25-9.

Buzgo, M., D. E. Soltis, P. S. Soltis and H. Ma (2004). Towards a comprehensive integration of morphological and genetic studies of floral development. Trends Plant Sci 9(4): 164-73.

Carlson, J. E., J. H. Leebens-Mack, P. K. Wall, L. M. Zahn, L. A. Mueller, L. L. Landherr, Y. Hu, D. C. Ilut, J. M. Arrington, S. Choirean, A. Becker, D. Field, S. D. Tanksley, H. Ma and C. W. Depamphilis (2006). EST database for early flower development in California poppy (*Eschscholzia californica* Cham., Papaveraceae) tags over 6000 genes from a basal eudicot. Plant Mol Biol.

Childs, K. L., J. P. Hamilton, W. Zhu, E. Ly, F. Cheung, H. Wu, P. D. Rabinowicz, C. D. Town, C. R. Buell and A. P. Chan (2007). The TIGR Plant Transcript Assemblies database. Nucleic Acids Res 35(Database issue): D846-51.

Craigon, D. J., N. James, J. Okyere, J. Higgins, J. Jotham and S. May (2004). NASCArrays: a repository for microarray data generated by NASC's transcriptomics service. Nucleic Acids Res 32(Database issue): D575-7.

Cui, L., P. K. Wall, J. H. Leebens-Mack, B. G. Lindsay, D. E. Soltis, J. J. Doyle, P. S. Soltis, J. E. Carlson, K. Arumuganathan, A. Barakat, V. A. Albert, H. Ma and C. W. Depamphilis (2006). Widespread genome duplications throughout the history of flowering plants. Genome Res.

Dong, Q., S. D. Schlueter and V. Brendel (2004). PlantGDB, plant genome database and analysis tools. Nucleic Acids Res 32(Database issue): D354-9.

Duarte, J. M., L. Y. Cui, P. K. Wall, Q. Zhang, X. H. Zhang, J. Leebens-Mack, H. Ma, N. Altman and C. W. dePamphilis (2006). Expression pattern shifts following duplication indicative of subfunctionalization and neofunctionalization in regulatory genes of *Arabidopsis*. Molecular Biology And Evolution 23(2): 469-478.

Duarte, J. M., P.K. Wall, L.M. Zahn, J.H. Leebens-Mack, and C.W. dePamphilis. (2007). Utility of *Amborella trichopoda* and *Nuphar advena* ESTs for phylogeny and comparative sequence analysis. . Taxon.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32(5): 1792-7.

Enright, A. J., S. Van Dongen and C. A. Ouzounis (2002). An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30(7): 1575-84.

Hartmann, S., D. Lu, J. Phillips and T. J. Vision (2006). Phytome: a platform for plant comparative genomics. Nucleic Acids Res 34(Database issue): D724-30.

Ilic, K., E. A. Kellogg, P. Jaiswal, F. Zapata, P. F. Stevens, L. P. Vincent, S. Avraham, L. Reiser, A. Pujar, M. M. Sachs, N. T. Whitman, S. R. McCouch, M. L. Schaeffer, D. H. Ware, L. D. Stein and S. Y. Rhee (2007). The plant structure ontology, a unified vocabulary of anatomy and morphology of a flowering plant. Plant Physiol 143(2): 587-99.

Kim, S., P. S. Soltis, K. Wall and D. E. Soltis (2006). Phylogeny and domain evolution in the APETALA2-like gene family. Molecular Biology And Evolution 23(1): 107-120.

Leebens-Mack, J. H., K. Wall, Z. Zheng, D. Oppenheimer and C. W. dePamphilis (2006). A genomics approach to the study of floral developmental genetics: strengths and limitations. Developmental Genetics of the Flower. London, Elsevier Limited.

Marchler-Bauer, A., A. R. Panchenko, B. A. Shoemaker, P. A. Thiessen, L. Y. Geer and S. H. Bryant (2002). CDD: a database of conserved domain alignments with links to domain three-dimensional structure. Nucleic Acids Res 30(1): 281-3.

Martinez-Castilla, L. P. and E. R. Alvarez-Buylla (2003). Adaptive evolution in the *Arabidopsis* MADS-box gene family inferred from its complete resolved phylogeny. Proc Natl Acad Sci U S A 100(23): 13407-12.

Morrison, D. A. (2007). Increasing the efficiency of searches for the maximum likelihood tree in a phylogenetic analysis of up to 150 nucleotide sequences. Syst Biol 56(6): 988-1010.

Müller, J. and K. Müller (2004). TreeGraph: automated drawing of complex tree figures using an extensible tree description format. Molecular Ecology Notes 4: 786-788.

Müller, K. F. (2004). PRAP - computation of Bremer support for large data sets. Molecular Phylogenetics and Evolution 31: 780-782.

Nixon, K. C. (1999). The Parsimony Ratchet, a new method for rapid parsimony analysis. Cladistics 15: 407-414.

Pujar, A., P. Jaiswal, E. A. Kellogg, K. Ilic, L. Vincent, S. Avraham, P. Stevens, F. Zapata, L. Reiser, S. Y. Rhee, M. M. Sachs, M. Schaeffer, L. Stein, D. Ware and S. McCouch (2006). Whole-plant growth stage ontology for angiosperms and its application in plant biology. Plant Physiol 142(2): 414-28.

Rudd, S. (2005). openSputnik--a database to ESTablish comparative plant genomics using unsaturated sequence collections. Nucleic Acids Res 33(Database issue): D622-7.

Sampedro, J. and D. J. Cosgrove (2005). The expansin superfamily. Genome Biol 6(12): 242.

Shen, L., J. Gong, R. A. Caldo, D. Nettleton, D. Cook, R. P. Wise and J. A. Dickerson (2005). BarleyBase--an expression profiling database for plant genomics. Nucleic Acids Res 33(Database issue): D614-8.

Soltis, D. E., H. Ma, M. W. Frohlich, P. S. Soltis, V. A. Albert, D. G. Oppenheimer, N. S. Altman, C. Depamphilis and J. Leebens-Mack (2007). The floral genome: an evolutionary history of gene duplication and shifting patterns of gene expression. Trends Plant Sci 12(8): 358-67.

Soltis, D. E., P. S. Soltis, V. A. Albert, D. G. Oppenheimer, C. W. dePamphilis, H. Ma, M. W. Frohlich and G. Theissen (2002). Missing links: the genetic architecture of flowers [correction of flower] and floral diversification. Trends Plant Sci 7(1): 22-31; dicussion 31-4.

Stamatakis, A., T. Ludwig and H. Meier (2005). RAxML-III: a fast program for maximum likelihoodbased
inference of large phylogenetic trees. Bioinformatics 21: 456-463.

Swofford, D. L. (2002). PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4.

Tuskan, G. A., S. Difazio, S. Jansson, J. Bohlmann, I. Grigoriev, U. Hellsten, N. Putnam, S. Ralph, S. Rombauts, A. Salamov, J. Schein, L. Sterck, A. Aerts, R. R. Bhalerao, R. P. Bhalerao, D. Blaudez, W. Boerjan, A. Brun, A. Brunner, V. Busov, M. Campbell, J. Carlson, M. Chalot, J. Chapman, G. L. Chen, D. Cooper, P. M. Coutinho, J. Couturier, S. Covert, Q. Cronk, R. Cunningham, J. Davis, S. Degroeve, A. Dejardin, C. Depamphilis, J. Detter, B. Dirks, I. Dubchak, S. Duplessis, J. Ehlting, B. Ellis, K. Gendler, D. Goodstein, M. Gribskov, J. Grimwood, A. Groover, L. Gunter, B. Hamberger, B. Heinze, Y. Helariutta, B. Henrissat, D. Holligan, R. Holt, W. Huang, N. Islam-Faridi, S. Jones, M. Jones-Rhoades, R. Jorgensen, C. Joshi, J. Kangasjarvi, J. Karlsson, C. Kelleher, R. Kirkpatrick, M. Kirst, A. Kohler, U. Kalluri, F. Larimer, J. Leebens-Mack, J. C. Leple, P. Locascio, Y. Lou, S. Lucas, F. Martin, B. Montanini, C. Napoli, D. R. Nelson, C. Nelson, K. Nieminen, O. Nilsson, V. Pereda, G. Peter, R. Philippe, G. Pilate, A. Poliakov, J. Razumovskaya, P. Richardson, C. Rinaldi, K. Ritland, P. Rouze, D. Ryaboy, J. Schmutz, J. Schrader, B. Segerman, H. Shin, A. Siddiqui, F. Sterky, A. Terry, C. J. Tsai, E. Uberbacher, P. Unneberg, J. Vahala, K. Wall, S. Wessler, G. Yang, T. Yin, C. Douglas, M. Marra, G. Sandberg, Y. Van de Peer and D. Rokhsar (2006). The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 313(5793): 1596-604.

Van Dongen, S. (2000). A cluster algorithm for graphs. Technical Report INS-R0010.

Zwickl, D. (2006). GARLI, Genetic Algorithm for Rapid Likelihood Inference, version 0.942.

Chapter 3

Gene Family Characteristics of *Populus trichocarpa*, a model tree, and comparisons with the model systems of *Arabidopsis* and rice

Preface

This manuscript has been formatted for submission to *Tree Genomics*. The authors are P. Kerr Wall, Jim Leebens-Mack, Victor Albert, John E. Carlson, Naomi S. Altman, Hong Ma, and Claude W. dePamphilis. PKW performed all the analyses. PKW, JLM, and CWD wrote the manuscript and all authors suggested changes or edits, and approved the manuscript.

P. Kerr Wall[1], James H. Leebens-Mack[1,‡], Victor Albert[2], John E. Carlson[3], Naomi Altman[4], Hong Ma[1], and Claude W. dePamphilis[1§]

[1]Department of Biology and the Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, PA 16802, USA

[2]The Natural History Museums and Botanical Garden, University of Oslo NO-0318 Oslo, Norway

[3]School of Forest Resources and the Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, PA 16802, USA

[4]Department of Statistics, The Pennsylvania State University, University Park, PA 16802, USA

[§]Author for correspondence and materials: Claude dePamphilis (cwd3@ psu.edu)

Abstract

With rapidly growing numbers of whole genome and EST sequences in our public databases, sequence-based gene classification systems are providing foundations for gene annotation, functional genomics, and comparative investigations of gene and genome evolution. Here we present a classification system based on the inferred proteomes of three fully sequenced plant species, *Arabidopsis thaliana*, *Oryza sativa* (rice), and *Populus trichocarpa*. We use the sequence similarity-based clustering algorithm MCL (Van Dongen 2000; Enright et al. 2002) to classify all of the protein-coding genes in these species into putative gene families, called tribes. Classifications were constructed using three clustering stringencies and the strength of support for clusters was tested through jackknife analyses. Most tribes are stable and well supported across the three clustering stringencies. The addition of *Populus* to the classification increased the stability and support of the tribes generated, confirming the power of additional proteomes in resolving evolutionary relationships among protein coding genes. We show how global gene classification provides insights into the *Arabidopsis*, rice, and *Populus* genomes, gene family evolution, and the evolutionary dynamics of functional domains among gene families. In addition, this classification provides a scaffold for sorting protein sequences from additional plant species.

Introduction

Protein classification systems typically aim to group functionally or phylogenetically (historically) related protein-coding genes. Classification schemes vary in their underlying objectives and methodologies. Therefore, researchers must match the goals of their investigation with the objectives and strengths of the classification systems they use. Functional classification systems are based on experimental data or pattern signatures from domains, while evolutionary-based classifications typically use sequence similarity or formal phylogenetic reconstructions to infer historical relationships (Enright et al. 2002; Enright et al. 2003). Variation in resolution is evident in both functional and phylogenetic classification methods. For example, pattern finding algorithms typically identify more narrowly defined groups of functionally related proteins than those based on hidden Markov models (HMMs) (Eddy 1996; Eddy 1998), which are constructed with diverse training sets of protein sequences. Evolutionary classifications range from sets of putative orthologous proteins (Tatusov et al. 1997; Fulton et al. 2002; Lee et al. 2002; Krylov et al. 2003; Li et al. 2003; O'Brien et al. 2005) to large sets including distantly related paralogs (Enright et al. 2002; Enright et al. 2003; Roth et al. 2005; Hartmann et al. 2006). Whether the objective is functional or phylogenetic classification, the method for classification should match the desired level of resolution.

Comprehensive functional classification of the protein coding genes in *Arabidopsis* and other completed genomes is providing a solid framework for the classification of genes from all plant species. The strength of this framework is increasing with the quality of the *Arabidopsis* and rice gene annotations. Recent whole genome microarray analyses of the *Arabidopsis* transcriptome have added genes and regulatory sequences, while generally verifying the quality of the *Arabidopsis* genome annotation (Yamada et al. 2003; Ma et al. 2005; Rensink and Buell

2005). However, for newly sequenced genomes, functional classification of genes is usually incomplete and analyses of sequence similarity aids initial annotation.

The well-annotated *Arabidopsis* genome has also aided in the interpretation of the genome sequences published for two rice subspecies, *O. sativa japonica* (Goff et al. 2002) and *O. sativa indica* (Yu et al. 2002). While there are many differences between *Arabidopsis* and rice in gene sequence and genome structure characteristics, they do contain enough similarity for comparative genomics. Sequence similarity and gene clustering analyses have been used to leverage the *Arabidopsis* genome annotations for interpretation and annotation of the much larger rice genome (Rensink and Buell 2004; Yuan et al. 2005).

The recently sequenced *Populus trichocarpa* genome (Tuskan et al. 2006) now provides a new and formal framework for comparative analyses across these three fully sequenced plant genomes. *Arabidopsis* and rice will now help to leverage the existing knowledge of their genomes into the newer *Populus* genome. *Populus* and *Arabidopsis*, both rosid eudicots, are separated by approximately 100 million years, whereas both shared a common ancestor with rice around 145 million years ago (Bell et al. 2005; Leebens-Mack et al. 2005). Comparative analyses of the three predicted proteomes can help verify gene content in the three genomic models and elucidate the process of gene duplication and functional diversification. Cross validation of gene models for each genome and nucleotide sequence translations of EST sets for other plant species is achieved through clustering and similarity analyses involving whole genome sequences and large EST sets (Dong et al. 2004; Rudd 2005; Hartmann et al. 2006; Mueller et al. 2008; Wall et al. 2008). Phylogenetic analyses of gene families that have been sampled completely from these three species are providing insights into gene duplication and the extent to which parts of individual gene families have remained stable or independently expanded in different lineages (Nam et al. 2004; Samuga and Joshi 2004; Djerbi et al. 2005; Leseberg et al. 2006; Sampedro et al. 2006). Future comparative analyses will focus on how genes have evolved following the

divergence from the ancestral genomes of these three taxa and other plant genomes currently being sequenced.

Here we present the results of cluster analyses including the inferred *Arabidopsis*, rice, and *Populus* proteomes. We have used the graph-based MCL algorithm (Van Dongen 2000; Enright et al. 2002) to cluster these proteomes at low, medium, and high stringencies. We introduce novel jackknife procedures used to explore the reliability of gene cluster delineation and membership assignment. The results provide insights into each proteome and the putative gene families provide a foundation for investigations of gene family evolution. Phylogenetic relationships can be inferred for each putative gene family and functional classifications can be mapped on to the resulting phylogenies.

Results

Tribe size distributions

We compared the predicted proteins from each of the three sequenced angiosperm species: 28,952 from *Arabidopsis* (TIGR, version 5.0), 61,250 from rice (TIGR, version 3.0) and 45,555 from *Populus* (JGI, version 1.0). We performed MCL clustering at low, medium, and high stringencies (Inflation, I = 1.2, 3.0, 5.0, respectively) on each species by itself, all two-way comparisons, and finally, with all three species together (Table 3-1). The inflation value is the only required input argument to the MCL algorithm and influences the size of the output clusters (tribes). The lowest inflation value (I=1.2) forms larger and fewer tribes than tribes built using the highest inflation value (i.e., I=5.0), which creates smaller and more tribes. In the three-species clustering, there are 22,815, 35,875, and 40,385 tribes at low, medium and high stringencies; among these, 15,063 (11.6%), 24,035 (18.5%), 28,144 (21.7%) are singleton tribes

(1 gene/tribe), respectively. At the lowest stringency, the two-way *Arabidopsis-Populus* clustering contained the fewest singleton tribes (7.92%) while the rice-only clustering contained the most (21.53%). As observed previously by Enright et al (2003) for *Arabidopsis* and diverse other organisms, the frequency distribution of tribe sizes in this study follows a power law at all inflation values and over all species' clustering combinations (results not shown). This supports the conclusions of Enright et al. (2003) that the known space of protein sequence similarity represents a scale-free network, that MCL is not biased towards any specific tribe size, and that there are minimal effects from erroneous gene predictions, which are typically in the form of singleton tribes.

In order to assess the degree to which estimated gene clusters represent sets of related genes, we determined the least significant E-value (i.e., lowest similarity) between members within each tribe over all tribe space. The frequency distribution of least significant E-values within tribes (Figure 3-1) shows that the majority of tribes contain a least significant E-value above or near 1e-10. Furthermore, when gene pairs in a tribe had E-values that were less significant than 1e-10, they were placed in the same tribe through transitive similarity to one or more other genes in the tribe. The least significant E-value within each tribe decreased with increasing stringency indicating that higher inflation values tend to remove the genes with the lowest similarity to the other genes within the tribe. In order to investigate the relationship between tribe E-value and tribe size, we plotted the median least significant E-value against tribe sizes (Figure 3-2). These values increase with tribe size as well as with stringency.

Table **3-1**: The total number of tribes, number of tribes with two or more genes, and number of singletons for each clustering at each stringency.

The percentage of genes that are singletons (tribes with only 1 gene) increases with increasing stringency. At the lowest stringency, the *Arabidopsis* and *Populus* clustering has the fewest singletons while the rice-only clustering has the highest number of singletons.

			Non-Singleton Tribes		Singletons (1 gene/tribe)				
Clustering	Inflation	# Tribes	# Tribes	# Genes	# ATH	# OSA	# PTR	# TOT	%
ATH	L	7104	3030	22046	4074			4074	15.6
ATH	M	11634	3606	18092	8028			8028	30.7
ATH	H	13137	3525	16508	9612			9612	36.8
OSA	L	16300	3946	45019		12354		12354	21.5
OSA	M	23610	5083	38846		18527		18527	32.3
OSA	H	25815	4864	36422		20951		20951	36.5
PTR	L	11230	5200	39365			6030	6030	13.3
PTR	M	16450	7180	36125			9270	9270	20.4
PTR	H	18028	7614	34981			10414	10414	22.9
ATH-OSA	L	18086	6620	71892	1786	9680		11466	13.8
ATH-OSA	M	28635	9039	63762	4603	14993		19596	23.5
ATH-OSA	H	32691	8761	59428	6263	17667		23930	28.7
ATH-PTR	L	12333	6670	65846	1371		4292	5663	7.9
ATH-PTR	M	19859	9557	61207	3289		7013	10302	14.4
ATH-PTR	H	22626	10009	58892	4489		8128	12617	17.6
OSA-PTR	L	21431	7396	88761		9650	4385	14035	13.7
OSA-PTR	M	32806	10980	80970		14470	7356	21826	21.2
OSA-PTR	H	36510	11175	77461		16796	8539	25335	24.7
ATH-OSA-PTR	L	22815	7752	113853	1337	9612	4114	15063	11.7
ATH-OSA-PTR	M	35875	11840	104881	3124	14179	6732	24035	18.6
ATH-OSA-PTR	H	40385	12241	100772	4205	16176	7763	28144	21.8

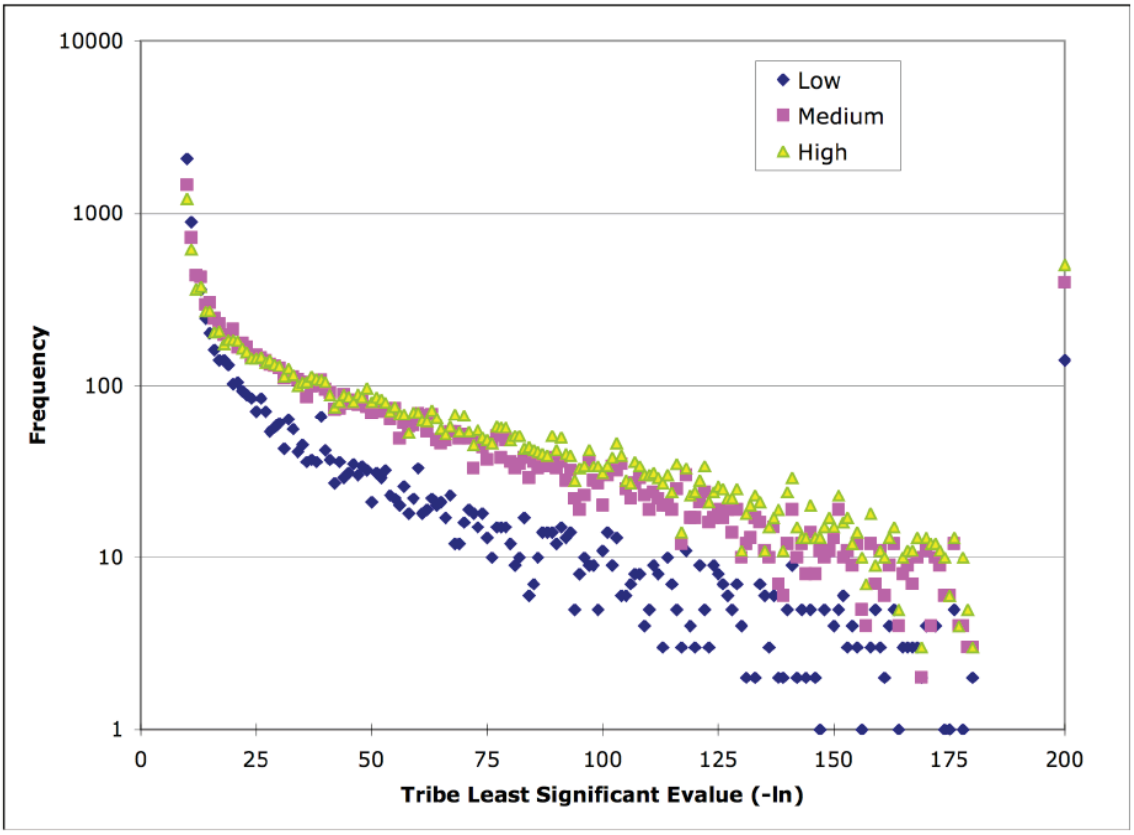

Figure **3-1**: Tribe e-value frequency distribution.

Within the three-way clustering, the least significant e-value within a tribe increases with increasing stringency. There are many tribes that have members without significant hits (e-10). Therefore, two tribe members can share significant hits with another member of the tribe but do not have significant hits with each other.

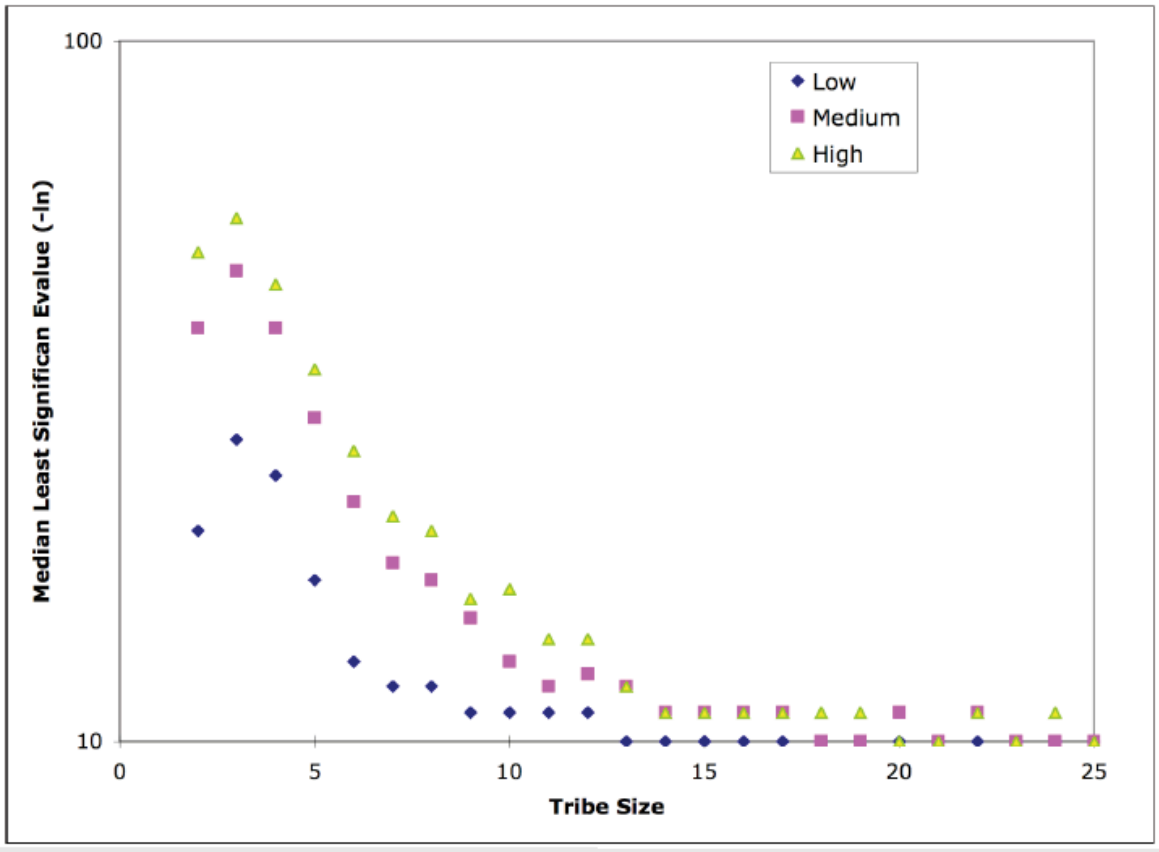

Figure **3-2**: Median tribe e-value as a function of tribe size.

As tribe sizes increase, tribes tend to have members with less significant e-values connecting them. As the stringency of the clustering is increased, the median least significant e-values within tribes also increase.

Many tribes are stable and strongly supported

The result of any cluster analysis provides summary descriptions of structured variation in a data set. However, investigators often want to interpret clusters as functionally or evolutionarily distinct entities. In this paper we are interpreting tribes as computational approximates of gene families or subfamilies. When equating tribes with gene families, we want to understand the degree of relatedness among genes within a tribe. We also would like to know whether genes in the same tribe share more recent common ancestry than genes in different tribes (i.e., the monophyly of the tribe). Alternatively, genes of related tribes might sometimes have a closer phylogenetic relationship than some genes in the same tribe(s). We propose that tribes with stable membership at all stringencies are most likely to be monophyletic. Therefore, we identified stable tribes that contained the same genes at low stringency clustering as they did at medium and high stringencies. In the 3-species clustering, we determined that 39.40% (3054/7752) of the tribes were stable across all three stringencies, which is up from 35.38% (1072/3030) and 36.83% (2438/6620) in the *Arabidopsis*-only and *Arabidopsis*-rice tribes respectively. These highly stable gene clusters appear to correspond to distinct gene families or subfamilies. For example, all 5 members of the YABBY gene family (e.g., YABBY1, At2g45190) in the annotated *Arabidopsis* genome were clustered with 7 rice and 13 *Populus* sequences at all three stringencies. Large tribes identified at low stringency were less likely to be stable across stringencies and were often subdivided into multiple high stringency tribes (Figure 3-3).

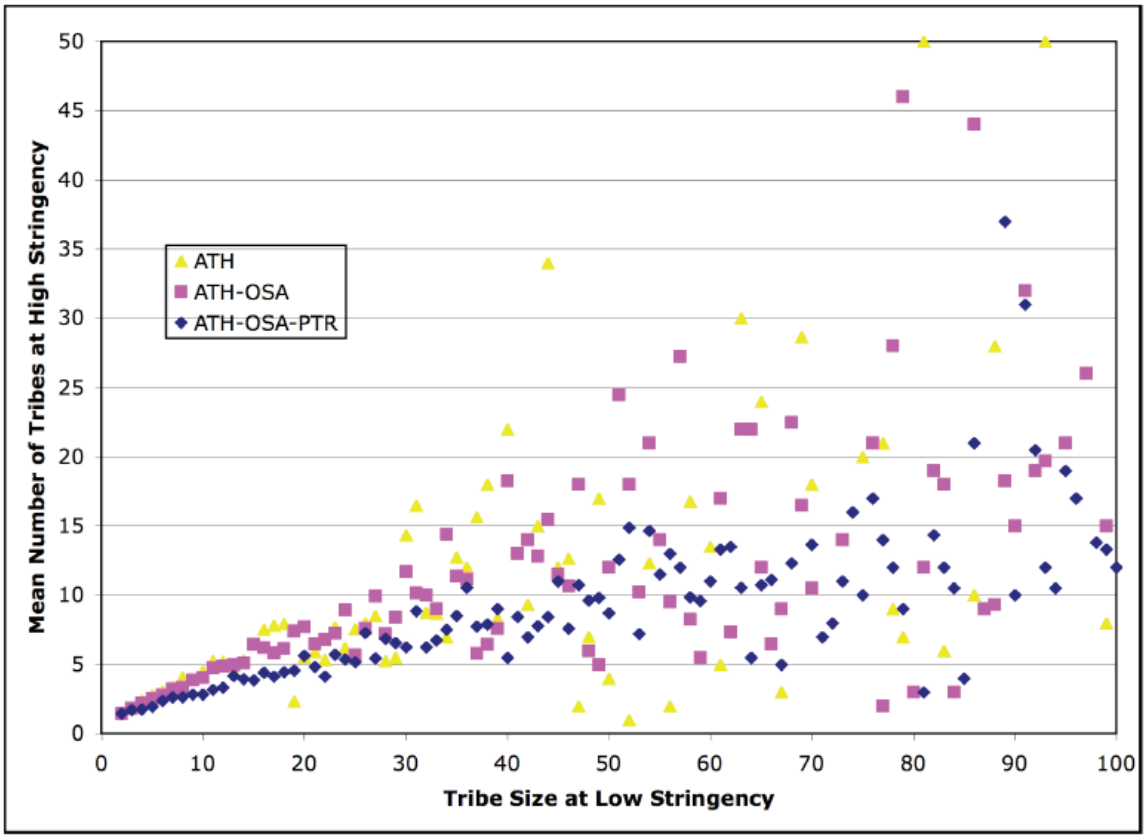

Figure **3-3**: Tribe size in relation to clustering stringency for one, two, and three species tribe analyses .

Each dot represents the tribe size at low stringency versus the mean number of tribes at the highest stringency. Large tribes formed at low stringency tend to be broken up into many tribes at high stringency. For smaller tribes (i.e., tribes with 0-20 genes), the addition of Populus creates more stable tribes.

We also assessed the stability of clusters at a single stringency using support values obtained through a jackknife resampling analysis (see Methods). Stable tribes tended to stay intact when a fraction of member genes were not sampled in a jackknife pseudoreplicate. Across all datasets at medium stringency, at least 75% of the tribes had jackknife support values of 90% or better and 90% of the tribes have at least 72% jackknife support (Figure 3-4). In addition, tribe support appears to increase as genomes are added to the classification. For example, 90% of the tribes are supported at 82% jackknife values when rice is clustered with *Arabidopsis*, and at 87% jackknife values with the addition of *Populus* in the three-genome classification. When support values were calculated across all three stringency levels, jackknife support values were generally higher at the low-stringency analysis (results not shown). Although this result may be counter intuitive, since different genes are removed in each jackknife sample, missing genes may have had a greater effect on clustering of jackknife replicates in higher stringency analyses. Mean jackknife support values tended to drop as tribe size increased (Figure 3-5). However, there are 16 tribes at the medium stringency with 100 or more genes that have jackknife values of at least 90. These clusters may represent recent, rapid and extensive gene duplication events within a few gene families.

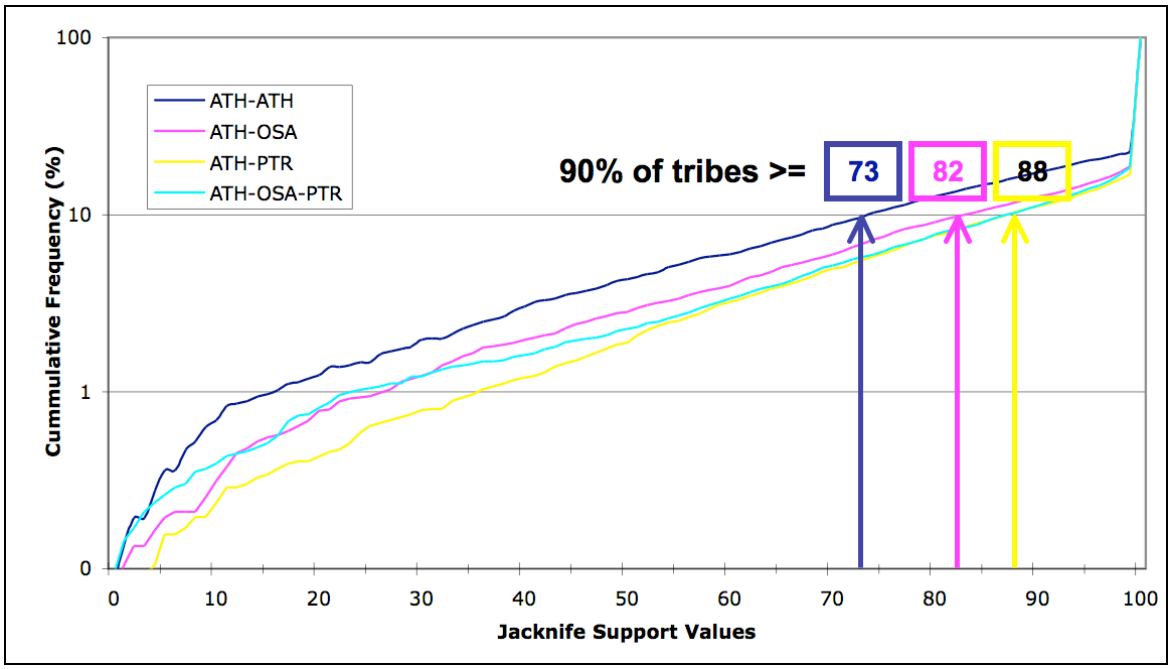

Figure **3-4**: Cumulative frequencies of jackknife support values for one, two, and three-species tribe analyses.

Approximately 83% of the *Arabidopsis*-Rice tribes have jackknife support values of 90 or greater. The addition of *Populus* increases tribe support values, with nearly 87% of the 3-species tribes having jackknife support values of 90 or greater.

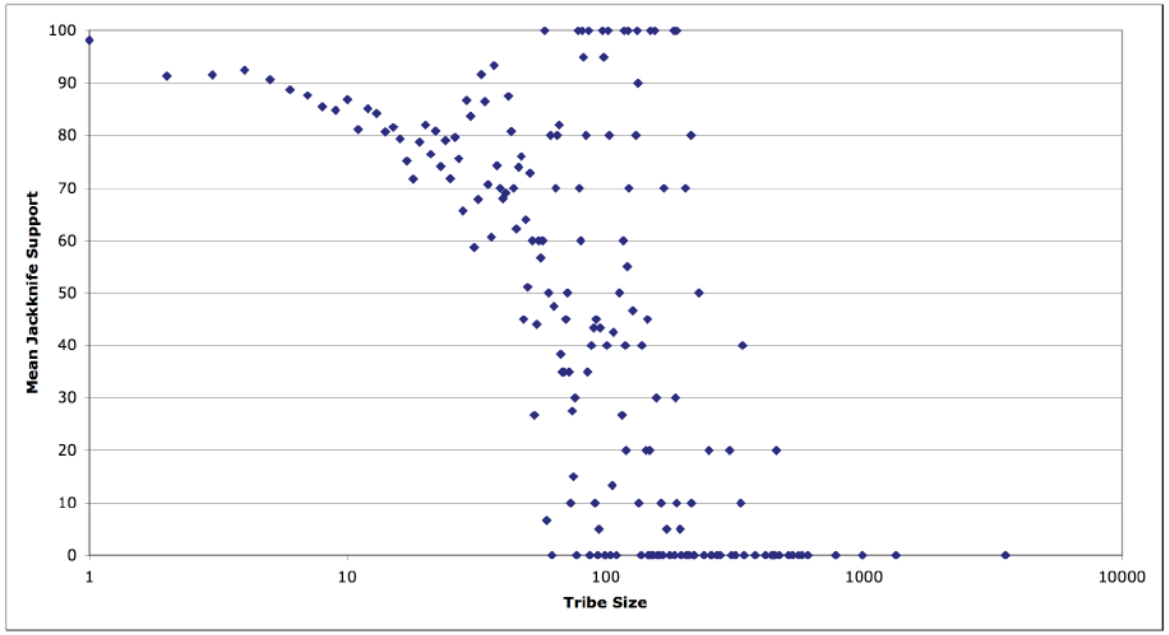

Figure **3-5**: Jackknife support values in relation to tribe size.

Jackknife support values tend to decrease with increasing tribe size. However, there are 16 tribes with at least 100 genes that have jackknife values of at least 90.

The distribution of *Arabidopsis*, rice, and *Populus* genes among tribes elucidates the pattern of genome expansion

In order to obtain a rough estimate of the timing of gene duplication events within tribes, the source species for best BLASTP matches were scored for each *Arabidopsis,* rice, and *Populus* gene. Ignoring any bias that may be introduced by differences in amino acid usage, we would expect that if the gene duplication rate were high, the best match to most genes would be another gene from the same species. Within each species, the highest percentage of genes had best hits to its own species (Table 3-2). The high percentage of *Populus* genes with best hits to other *Populus* genes is consistent with a hypothesized recent genome duplication; furthermore, *Populus* has retained a higher percentage of duplicates since this genome duplication than *Arabidopsis* or rice have since genome duplications in their respective lineages. In order to elucidate the impact of polyploidy in the lineages leading to *Arabidopsis* and *Populus*, we plotted (Figure 3-6) the number of *Arabidopsis* genes versus the number of *Populus* genes for every tribe. The slope of 1.86 indicates that each tribe contains nearly two *Populus* genes for every *Arabidopsis* gene at all tribe sizes.

Table **3-2**: Percentages of genes within a genome that have best BLASTp values against each genome.

The columns 'ATH', 'OSA', 'PTR' indicate the genomes of *Arabidopsis*, rice, and *Populus*, respectively. The column 'MULTI' includes genes that have best hits against genes from multiple genomes while the column 'SELF' include singletons that do not have significant blastp values against any other genes.

	ATH	OSA	PTR	MULTI	SELF
ATH	43.3	3.5	41.3	6.5	5.4
OSA	3.8	62.4	13.2	3.4	17.2
PTR	13.0	3.7	67.0	7.1	9.2

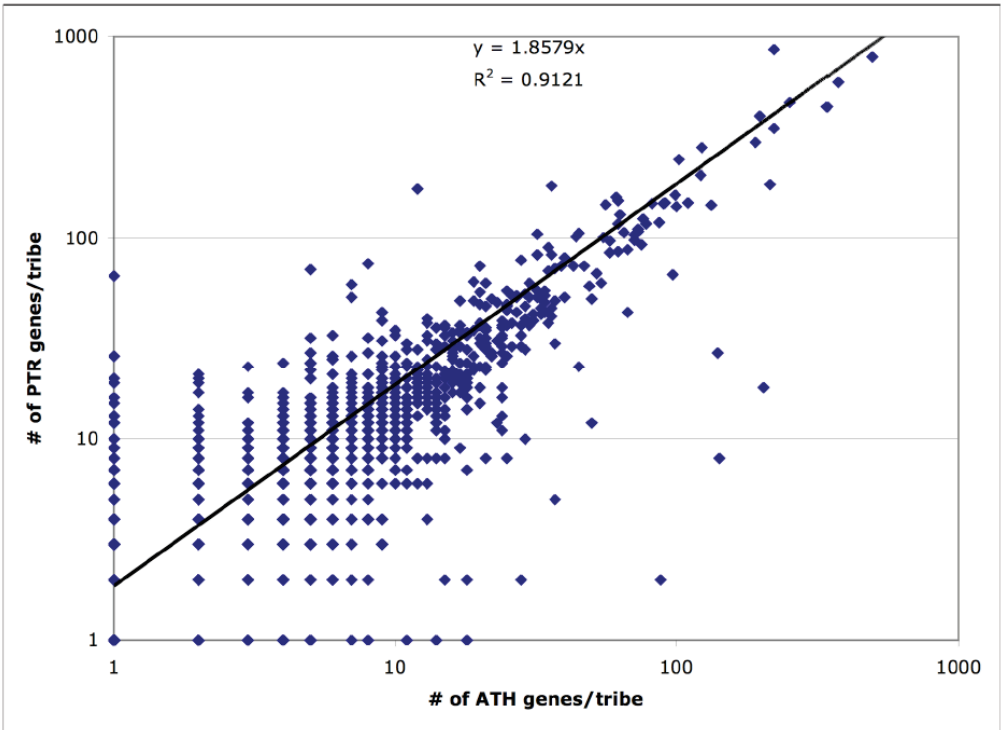

Figure **3-6**: Number of *Populus* and *Arabidopsis* genes per tribe in all tribes with at least one gene from each species.

In the *Arabidopsis* and *Populus* clustering, most tribes contain more *Populus* than *Arabidopsis* genes. However, there are many tribes that are dominated by *Arabidopsis* genes. Linear regression line with slope 1.86 indicates that on average, there are 1.86 *Populus* genes per *Arabidopsis* gene.

Even considering that all three of these species have undergone at least one recent genome duplication, we still find many tribes with the same number of genes from each species. Excluding 1058 single ortholog tribes (tribes with 1 member from each species), there are 187 tribes (1347 genes) that contain the same number of genes from each species. This group included 1 Rboh-related tribe with 10 genes from each species, 1 RRM-related tribe with 7 genes from each species, 2 tribes with 6 genes from each species, 3 tribes with 5 genes from each species, 8 tribes with 4 genes from each species, 29 tribes with 3 genes from each species, and 142 tribes with 2 genes from each species. We performed formal phylogenetic analyses (see Methods) on this selected group of tribes and found 9 tribes with strict 1:1:1 orthologous relationships throughout the tree including 1 tribe with 4 genes from each species and 8 tribes with 2 genes from each species. All genes within the tribe having strict orthologous relationships with 4 genes from each species are annotated in *Arabidopsis* as 'expressed protein' and warrant further investigation.

Organism-specific tribes were observed at all stringencies in the three-species clustering. Excluding singletons at the lowest stringency, there were 237 *Arabidopsis*-specific tribes (1,137 genes), 1,264 rice-specific tribes (7,778 genes), and 736 *Populus*-specific tribes (2,649 genes). Some of these "species-specific" tribes included genes sharing significant sequence similarity with genes from other species, placed in other tribes; others were clearly distinct. Excluding singletons, 648 out of 736 *Populus*-specific tribes included 2,023 genes with no detected homology to any other genes outside of their own tribes. There were 20 *Populus*-specific tribes with at least 10 genes each. The largest *Arabidopsis*-specific tribes included an Ulp1-related family that contains 175 genes and two zinc-finger related families that contain 29 and 15 genes each. The largest rice-specific tribes included many retroelement families with the largest containing 573 genes. There are also 32 rice-specific F-box related tribes including 7 with at least 20 genes each. We also determined the tribes with genes that were specific to only 2 of

the 3 species and found 37 tribes (153 genes) specific to *Arabidopsis* and rice, 466 tribes (2,253 genes) specific to *Arabidopsis* and *Populus*, and 205 tribes (2,193 genes) specific to *Populus* and rice.

Tribe classification is not skewed by highly abundant and promiscuous domains

We also used the PFAM database to define all known protein domains in each of the genes from these three species. Of the 8296 domains in the PFAM database (version 20.0), 2122 (25.6%) were observed in 68% of the annotated *Arabidopsis* proteome. Excluding singleton tribes and tribes with no known PFAM domain, 2302/2643 (87.1%) of the high stringency *Arabidopsis*-only tribes are comprised of members that all share the most common domain within the tribe. This percentage increased from 1389/1966 (71.1%) and 2215/2599 (85.2%) in the low stringency and medium stringency *Arabidopsis*-only tribes, respectively. Therefore, as the clustering stringency is increased, the number of tribes with members that all contain the most common domain in the tribe also increased.

An initial claim for the potential strength of MCL clustering is that it is not overly sensitive to multi-domain protein families and the presence or absence of promiscuous domains. To test this idea, we determined the number of genes and number of high stringency *Arabidopsis*-only tribes where each PFAM domain was present (Table 3-3). As was found by Enright et al. (2002), the most common domains were observed in many tribes, suggesting that the MCL clustering algorithm depends on the overall sequence similarity, but is not strongly influenced by strong matches over conserved or convergent domains that reside in distantly related genes (i.e., multi-domain families). In addition to the most highly abundant domains as noted by Enright et al. (2002), we were also interested in using a functionally-based estimate of promiscuous domains. We used the InterDom database (Ng et al. 2003) to identify promiscuous domains,

defined there in as domains that have over 50 interacting partners. We plotted the top 250 PFAM domains found in the most high stringency *Arabidopsis*-only tribes by the number of genes and by the number of tribes where each domain was present, grouping the domains into two categories: promiscuous and non-promiscuous (Figure 3-7). With 104 of the promiscuous domains that are found in *Arabidopsis* among these top 250 most abundant domains, we found that the linear slope of the promiscuous domains is higher than the non-promiscuous domain slope. This result indicates that the promiscuous domains are found in a greater number of tribes than their non-promiscuous, but highly abundant, counterparts. For example, the Helicase_C (PF00271) is a promiscuous PFAM domain that is found in 147 *Arabidopsis* genes across 30 high stringency tribes including 19 non-singleton tribes with 14 of these non-singleton tribes containing a more common PFAM domain among the members of the tribe. Therefore, although domains help to define tribe membership, highly abundant domains and/or functionally defined promiscuous domains do not unite unrelated sequences into larger clusters.

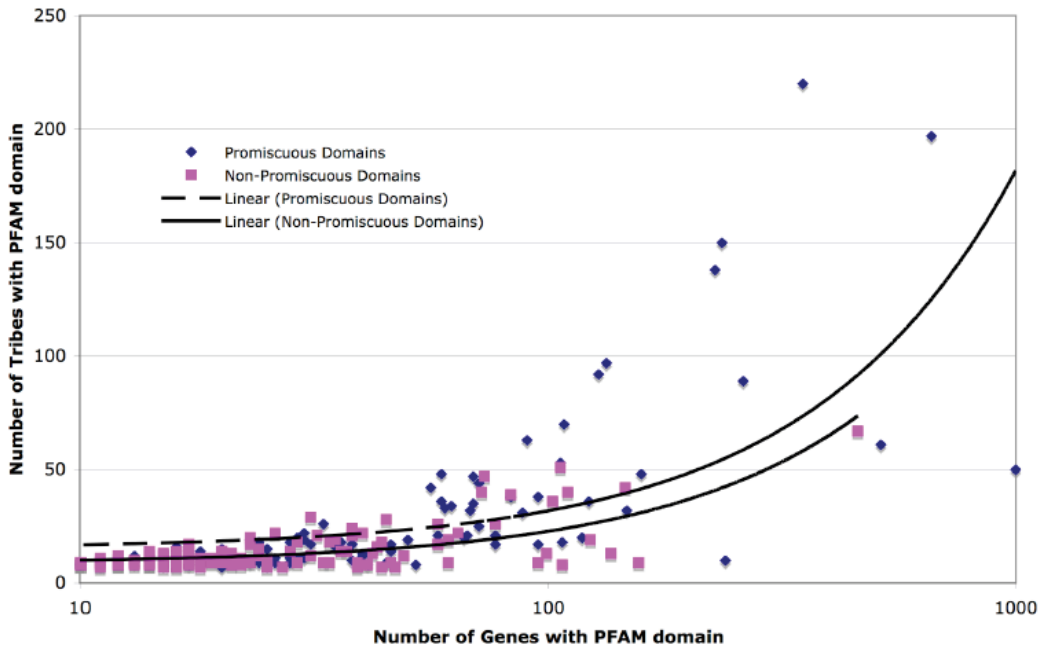

Figure **3-7**: Promiscuous versus non-promiscuous PFAM domains defined by InterDom.

Each dot represents a PFAM domain plotted with the number of tribes versus the number of genes that it is found in. The top 250 PFAM domains found in the most number of tribes are plotted in two groups, 104 promiscuous domains, defined by InterDom, and 146 non-promiscuous but highly abundant domains. The linear regression line for the promiscuous domains is higher than the non-promiscuous domain line indicating that promiscuous domains are generally found in a higher number of tribes and do not bias MCL clustering.

Table **3-3**: The PFAM domains present in the *Arabidopsis* only tribes.

Promiscuous domains such as Pkinase, F-box, and Leucine Rich Repeats do generally not define tribes. Promiscuous domains (column "P"), indicated by "*", are found in many genes across many tribes. Promiscuous domains are defined as domains with more than 50 interacting partners from the InterDom Database (Ng *et al.*, 2003).

PFAM ID	# GENES	# TRIBES	P	NAME	DESC
PF00097	350	220	*	zf-C3HC4	Zinc finger, C3HC4 type (RING finger)
PF00646	659	197	*	F-box	F-box domain
PF00076	235	150	*	RRM_1	RNA recognition motif. (a.k.a. RRM, RBD, or RNP dom
PF00400	227	138	*	WD40	WD domain, G-beta repeat
PF00010	133	97	*	HLH	Helix-loop-helix DNA-binding domain
PF00096	128	92	*	zf-C2H2	Zinc finger, C2H2 type
PF00249	261	89	*	Myb_DNA-binding	Myb-like DNA-binding domain
PF00226	108	70	*	DnaJ	DnaJ domain
PF01535	459	67		PPR	PPR repeat
PF00515	90	63	*	TPR_1	Tetratricopeptide repeat
PF00560	514	61	*	LRR_1	Leucine Rich Repeat
PF00023	106	53	*	Ank	Ankyrin repeat
PF00234	106	51		Tryp_alpha_amyl	Protease inhibitor/seed storage/LTP family
PF00069	999	50	*	Pkinase	Protein kinase domain
PF00036	158	48	*	efhand	EF hand
PF00098	59	48	*	zf-CCHC	Zinc knuckle
PF00403	73	47		HMA	Heavy-metal-associated domain
PF00642	69	47	*	zf-CCCH	Zinc finger C-x8-C-x5-C-x3-H type (and similar)
PF00628	71	44	*	PHD	PHD-finger
PF00170	56	42	*	bZIP_1	bZIP transcription factor
PF00847	146	42		AP2	AP2 domain
PF03106	72	40		WRKY	WRKY DNA -binding domain
PF04043	110	40		PMEI	Plant invertase/pectin methylesterase inhibitor
PF02362	83	39		B3	B3 DNA binding domain
PF00168	95	38	*	C2	C2 domain
PF00612	83	38	*	IQ	IQ calmodulin-binding motif
PF00153	59	36	*	Mito_carr	Mitochondrial carrier protein
PF00319	102	36		SRF-TF	SRF-type transcription factor (DNA-binding and dimer
PF01344	122	36	*	Kelch_1	Kelch motif
PF00561	69	35	*	Abhydrolase_1	alpha/beta hydrolase fold
PF00240	62	34	*	ubiquitin	Ubiquitin family
PF00085	60	33	*	Thioredoxin	Thioredoxin
PF00271	147	32	*	Helicase_C	Helicase conserved C-terminal domain
PF00514	68	32	*	Arm	Armadillo/beta-catenin-like repeat
PF00106	88	31	*	adh_short	short chain dehydrogenase
PF05678	31	29		VQ	VQ motif
PF00856	45	28		SET	SET domain
PF00583	33	26	*	Acetyltransf_1	Acetyltransferase (GNAT) family
PF02519	77	26		Auxin_inducible	Auxin responsive protein
PF04564	58	26		U-box	U-box domain
PF00702	71	25	*	Hydrolase	haloacid dehalogenase-like hydrolase
PF04535	38	24		DUF588	Domain of unknown function (DUF588)
PF00072	40	22		Response_reg	Response regulator receiver domain
PF00627	30	22	*	UBA	UBA/TS-N domain
PF02902	64	22		Peptidase_C48	Ulp1 protease family, C-terminal catalytic domain
PF04396	26	22		DUF537	Protein of unknown function, DUF537
PF00046	58	21	*	Homeobox	Homeobox domain
PF00149	67	21	*	Metallophos	Calcineurin-like phosphoesterase
PF00481	77	21	*	PP2C	Protein phosphatase 2C
PF00643	32	21		zf-B_box	B-box zinc finger

Discussion

Organizing protein coding genes obtained from whole genome sequences is an important first step in phylogenomics analyses aimed at understanding the evolution of genome structure (Paterson et al. 2003; Hillier et al. 2004; Bourque et al. 2005) and gene diversification and function (Eisenstein et al. 2000; Sjolander 2004). In this paper, we describe an application of MCL clustering to the complete genome sequences of three flowering plants and demonstrate the power of this approach to probe gene family structures at the genome scale. We used the MCL algorithm to cluster the three fully sequenced plant genomes of *Arabidopsis*, rice, and *Populus* under low, medium, and high stringencies, and generated jackknife support values for all of the resulting tribes. This approach has made it possible to explore stability, support, hierarchical classification information, and confirmation of gene family space in plant proteomes.

One of the challenges of some approaches to protein clustering is the handling of highly abundant and promiscuous domains. These domains usually have over 50 interacting partners and are found in many genes within a proteome. An initial claim for the potential strength of MCL clustering is that it should not be overly sensitive to the presence and absence of such domains (Enright et al. 2003). For example, the PKinase (PF00069), F-box (PF00646), and Leucine Rich Repeat (PF00560) PFAM domains are found in 999, 659, and 514 proteins in *Arabidopsis* respectively. These domains are found in 50, 197, and 61 *Arabidopsis*-only tribes, respectively, providing evidence that MCL does not have difficulties with highly abundant domains. Furthermore, whether promiscuous domains are defined based on their abundance (Enright et al. 2003) or based on the number of interacting partners (Figure 3-7), these domains are generally found in a greater number of tribes. The domain architecture includes both the composition and location of the domains within proteins. This architecture, not the presence or absence of highly abundant and/or promiscuous domains, help define tribe space.

In our analysis, tribes are units of classification that are based on objective statistical analysis and may approximate gene families. Gene family designations are typically subjective as they are defined usually by the presence of one or more functional domains and can vary greatly between families in the degree of sequence similarity. Different specialists may segregate members of the same group of genes in very different ways, depending on the criteria used to define the family. For example, gene family pages at The *Arabidopsis* Information Resource (TAIR), http://www.*Arabidopsis*.org/info/genefamily/genefamily.html, illustrate a variety of classification approaches, some including super families. Our analysis identifies 40% of all tribes whose classification is unambiguous and stable, with identical sets of genes identified under all stringencies. However, depending on the clustering stringency and the evolutionary distance of the genomes used in the clustering, many tribes actually represent smaller clades within gene families or larger super families. This was evident in the MADS-box gene family (results not shown), where the members of the family were found in 5 main tribes at low stringency and more than 15 tribes at high stringency. Therefore our Tribes approach allows the user to impose the same rules of classification to all proteins and to systematically explore the impact of changing the classification rules.

Virtually all flowering plants have polyploidy histories (Blanc and Wolfe 2004; Cui et al. 2006), including the three species studied here. Genome duplications can provide an opportunity for rapid expansion of gene families that prove to be biologically important for the organism. For example, tribes involved in cellulose biosynthesis (e.g., cesA) and lignin biosynthesis (e.g., cytochrome P450s) have duplicate copies of genes in *Populus* relative to *Arabidopsis* (Tuskan et al. 2006). However, it is also surprising to find that many gene families show relative stability of size in the face of these independent genome duplications within each lineage. We found nearly 1,250 tribes consisting of approximately 4,500 genes that had equal numbers of genes from each of the three species including 9 tribes with strict orthologous relationships throughout the

phylogeny. These results suggest that natural selection could be acting to regulate gene family size more strongly in some gene families than others.

The addition of the *Populus* predicted proteome to the *Arabidopsis* and rice tribes reduced uncertainty in the clustering of the predicted *Arabidopsis* proteome in two ways. First, inclusion of *Populus* increased the stability of tribe classification as measured by changes in tribe membership at different stringencies. This addition decreased the mean number of tribes formed at low stringency that are broken up into multiple tribes at high stringency (Figure 3-1). Second, the addition of the *Populus* genome also improved tribe jackknife support. For example, in the *Arabidopsis*-*Populus* cluster analysis, tribes had the lowest frequency of jackknife support values below 50, and the *Arabidopsis*-*Populus* and the 3-way clustering had the lowest frequency of jackknife support values below 90 (Figure 3-2). The density of genes within clusters is expected to increase with the addition of taxa. Therefore, we expect that jackknife support values for gene clusters will increase as more angiosperm genomes are sequenced and added to cluster analyses.

An improvement in the assignment of genes to gene clusters will lead to improvements in our ability to assess gene models produced by automated annotation algorithms. Whereas the placement of gene predictions in multi-gene clusters provides a form of validation, gene models identified as singletons in cluster analyses may be mis-annotations. The addition of *Populus* to our cluster analyses decreased the number of singletons relative to the *Arabidopsis*-rice and single species analyses (Table 3-1). For example, at the lowest clustering stringency, the *Arabidopsis*-*Populus* tribes had the fewest percentage of singleton tribes (7.9%) which was nearly half that found in the *Arabidopsis*-rice tribes (13.8%).

Variation among species in the number of singletons may reflect differences in the quality of current annotations, differences in the gene birth and death process or both. The low stringency three-species cluster analysis identified 1,337, 9,612, and 4,114 singletons in *Arabidopsis*, rice and *Populus*, respectively. We searched these singletons, defined by MCL,

back against the three full proteomes, and removed proteins that had any significant matches to any other protein except itself, thus removing 77, 654, and 158 genes from *Arabidopsis*, rice, and *Populus*, respectively. In order to determine the validity of the singletons from each species, we ran TBLASTX for each gene against the TIGR plant gene indices database (Quackenbush et al. 2000; Quackenbush et al. 2001; Lee et al. 2005). All 1260 singletons from *Arabidopsis* had highly significant hits (1.0E-10) in the unigene database. For rice, 7232/8958 (80.7%) of the singletons had significant hits while only 2107/3956 (53.3%) of the *Populus* singletons had significant hits. Singletons without transcript verification should be suspect; however, we should expect these percentages to improve (increase) with each new version of these genomes, and as more global gene expression data becomes available for rice and *Populus*, as has been the case with *Arabidopsis*.

The formal comparison of the new *Populus* genome to the mature genomes of *Arabidopsis* and rice has helped to rapidly annotate *Populus* and further improve the existing annotations of the two previous sequenced genomes. This comparison has provided important ways to assess information about the content of each genome and provided evidence in the form of mutual support of genes across these three taxonomic lineages. The addition of the *Populus* genome has decreased the percentage of singletons, increased the number of stable tribes, and increased the jackknife support in the three-way classification compared to the previous two-way classification. With many genome sequence projects in progress, formal comparative approaches are needed to rapidly identify the best gene models, quickly determine errors in the initial annotations, identify new gene families, and increase the confidence in the limits and structure of existing gene families.

Methods

Proteome clustering

The predicted protein sequences from the fully sequenced genomes of *Arabidopsis thaliana* (version 5.0, ftp://ftp.tigr.org/pub/data/a_thaliana/) and *Oryza sativa* (rice) (version 3.0, ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/o_sativa/) were downloaded from The Institute for Genomic Research (TIGR). These proteomes contained a total of 28,952 and 61,250 proteins, respectively. We eliminated multiple versions of loci, reducing the number of sequences to 26,207 and 57,915 respectively. The predicted proteome of *Populus trichocarpa* was downloaded from the Joint Genome Institute (JGI) and included 45,555 proteins. The combined non-redundant protein set consisting of 129,677 proteins was included in an all-against-all BLASTp search (Altschul et al. 1990) with an E-value cutoff of e-10. The resulting blast reports were then used to generate a similarity matrix from the transformed E-values. Finally, the similarity matrix was used to perform MCL clustering at low, medium, and high stringencies (inflation of 1.2, 3.0, and 5.0, respectively) for each species and all subsequent combinations of species.

Tribe stability and support

The resulting MCL output from the proteome clustering step was then used to determine the extent of the stability and support for each tribe. Perl scripts were written to parse the MCL output and determine the number of stable tribes, defined as tribes that contain the same membership of genes at the low, medium, and high stringencies, and non-stable tribes, defined as those containing a different complement of genes (usually less) at the medium and high stringencies compared to the low stringency clustering. To evaluate the support for each tribe

with at least two members, a total of 100 jackknife replicates were generated by randomly deleting roughly 33% of the genes in the original matrix. For each pseudoreplicate, the matrix of transformed e-values was re-clustered using MCL and tribe membership was compared to the original set of tribes. Each tribe was scored as being recovered in a pseudoreplicate when all sampled genes were placed in a single cluster and no other genes were included in that cluster. The jackknife analysis was performed on all combinations of proteomes and all stringencies.

Phylogenetic analysis

We performed phylogenetic analysis on selected tribes from the high stringency 3-species clustering that contained exactly the same number of genes from each of the three species to infer whether the grouping agreed with a strict definition of orthology. We used the program MUSCLE (Edgar 2004) with default parameters to align the proteins of these selected tribes. Maximum parsimony (MP) phylogenetic analysis was performed using PAUP 4.0b* (Swofford 2002). Tree searching was performed using 200 random sequence additions and tree bisection reconnection branch swapping. MP bootstrapping was performed with 100 replicates using heuristic searches using the same sequence additions as above.

PFAM domain analysis

Domain definitions for *Arabidopsis* were downloaded from The *Arabidopsis* Information Resource (TAIR, ftp://ftp.arabidpsis.org/home/tair/Proteins/Domains/all.domains.txt). A Perl script was written to parse out only the PFAM definitions and calculate the number of genes from *Arabidopsis* that contained each of the PFAM domains. The script also calculated the number of *Arabidopsis*-only tribes at low stringency that contained each PFAM domain. To evaluate

promiscuous domains, defined as domains with more than 50 interacting partners, we downloaded the entire InterDom (Ng et al. 2003) database at http://interdom.lit.org.sg/download/interdom_v1.2.zip, which contained all of the domain-domain interactions in a tab delimited table. A perl script was written to parse the file and determine all PFAM domains with more than 50 interacting partners. The promiscuous domain information was then added to the gene and tribe counts for each PFAM domain.

Super tribes

In order to identify tribes that may be distantly related to each other, we used an additional round of MCL clustering on all of the tribe output. We determined the most significant E-value between members of each tribe and constructed a matrix using the –log(E-value) of each tribe as the metric. We then performed MCL clustering at low, medium, and high stringencies on each original tribe output to identify related or super tribes.

Acknowledgements

The authors thank undergraduate interns Josh Marion and Michael Frederick for help with database and website programming, Drs. Ali Barakat and Liying Cui, and Jill Duarte and Barbara Bliss for comments on the manuscript and the many users of the PlantTribes for suggestions about the database and its features. This work was supported by the Floral Genome Project (NSF Plant Genome award DBI-0115684), Functional Genomics of Woody Perrennial Flowering (NSF Plant Genome award DBI-0501890), and the Ancestral Angiosperm Genome Project.

References

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman (1990). Basic local alignment search tool. J Mol Biol 215(3): 403-10.

Bell, C. D., D. E. Soltis and P. S. Soltis (2005). The age of the angiosperms: a molecular timescale without a clock. Evolution Int J Org Evolution 59(6): 1245-58.

Blanc, G. and K. H. Wolfe (2004). Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell 16(7): 1667-78.

Bourque, G., E. M. Zdobnov, P. Bork, P. A. Pevzner and G. Tesler (2005). Comparative architectures of mammalian and chicken genomes reveal highly variable rates of genomic rearrangements across different lineages. Genome Res 15(1): 98-110.

Cui, L., P. K. Wall, J. H. Leebens-Mack, B. G. Lindsay, D. E. Soltis, J. J. Doyle, P. S. Soltis, J. E. Carlson, K. Arumuganathan, A. Barakat, V. A. Albert, H. Ma and C. W. Depamphilis (2006). Widespread genome duplications throughout the history of flowering plants. Genome Res.

Djerbi, S., M. Lindskog, L. Arvestad, F. Sterky and T. T. Teeri (2005). The genome sequence of black cottonwood (*Populus trichocarpa*) reveals 18 conserved cellulose synthase (CesA) genes. Planta 221(5): 739-46.

Dong, Q., S. D. Schlueter and V. Brendel (2004). PlantGDB, plant genome database and analysis tools. Nucleic Acids Res 32(Database issue): D354-9.

Eddy, S. R. (1996). Hidden Markov models. Curr Opin Struct Biol 6(3): 361-5.

Eddy, S. R. (1998). Profile hidden Markov models. Bioinformatics 14(9): 755-63.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32(5): 1792-7.

Eisenstein, E., G. L. Gilliland, O. Herzberg, J. Moult, J. Orban, R. J. Poljak, L. Banerjei, D. Richardson and A. J. Howard (2000). Biological function made crystal clear - annotation of hypothetical proteins via structural genomics. Curr Opin Biotechnol 11(1): 25-30.

Enright, A. J., V. Kunin and C. A. Ouzounis (2003). Protein families and TRIBES in genome sequence space. Nucleic Acids Res 31(15): 4632-8.

Enright, A. J., S. Van Dongen and C. A. Ouzounis (2002). An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30(7): 1575-84.

Fulton, T. M., R. Van der Hoeven, N. T. Eannetta and S. D. Tanksley (2002). Identification, analysis, and utilization of conserved ortholog set markers for comparative genomics in higher plants. Plant Cell 14(7): 1457-67.

Goff, S. A., D. Ricke, T. H. Lan, G. Presting, R. Wang, M. Dunn, J. Glazebrook, A. Sessions, P. Oeller, H. Varma, D. Hadley, D. Hutchison, C. Martin, F. Katagiri, B. M. Lange, T. Moughamer, Y. Xia, P. Budworth, J. Zhong, T. Miguel, U. Paszkowski, S. Zhang, M. Colbert, W. L. Sun, L. Chen, B. Cooper, S. Park, T. C. Wood, L. Mao, P. Quail, R. Wing, R. Dean, Y. Yu, A. Zharkikh, R. Shen, S. Sahasrabudhe, A. Thomas, R. Cannings, A. Gutin, D. Pruss, J. Reid, S. Tavtigian, J. Mitchell, G. Eldredge, T. Scholl, R. M. Miller, S. Bhatnagar, N. Adey, T.

Rubano, N. Tusneem, R. Robinson, J. Feldhaus, T. Macalma, A. Oliphant and S. Briggs (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. japonica). Science 296(5565): 92-100.

Hartmann, S., D. Lu, J. Phillips and T. J. Vision (2006). Phytome: a platform for plant comparative genomics. Nucleic Acids Res 34(Database issue): D724-30.

Hillier, L. W., W. Miller, E. Birney, W. Warren, R. C. Hardison, C. P. Ponting, P. Bork, D. W. Burt, M. A. Groenen, M. E. Delany, J. B. Dodgson, A. T. Chinwalla, P. F. Cliften, S. W. Clifton, K. D. Delehaunty, C. Fronick, R. S. Fulton, T. A. Graves, C. Kremitzki, D. Layman, V. Magrini, J. D. McPherson, T. L. Miner, P. Minx, W. E. Nash, M. N. Nhan, J. O. Nelson, L. G. Oddy, C. S. Pohl, J. Randall-Maher, S. M. Smith, J. W. Wallis, S. P. Yang, M. N. Romanov, C. M. Rondelli, B. Paton, J. Smith, D. Morrice, L. Daniels, H. G. Tempest, L. Robertson, J. S. Masabanda, D. K. Griffin, A. Vignal, V. Fillon, L. Jacobbson, S. Kerje, L. Andersson, R. P. Crooijmans, J. Aerts, J. J. van der Poel, H. Ellegren, R. B. Caldwell, S. J. Hubbard, D. V. Grafham, A. M. Kierzek, S. R. McLaren, I. M. Overton, H. Arakawa, K. J. Beattie, Y. Bezzubov, P. E. Boardman, J. K. Bonfield, M. D. Croning, R. M. Davies, M. D. Francis, S. J. Humphray, C. E. Scott, R. G. Taylor, C. Tickle, W. R. Brown, J. Rogers, J. M. Buerstedde, S. A. Wilson, L. Stubbs, I. Ovcharenko, L. Gordon, S. Lucas, M. M. Miller, H. Inoko, T. Shiina, J. Kaufman, J. Salomonsen, K. Skjoedt, G. K. Wong, J. Wang, B. Liu, J. Wang, J. Yu, H. Yang, M. Nefedov, M. Koriabine, P. J. Dejong, L. Goodstadt, C. Webber, N. J. Dickens, I. Letunic, M. Suyama, D. Torrents, C. von Mering, E. M. Zdobnov, K. Makova, A. Nekrutenko, L. Elnitski, P. Eswara, D. C. King, S. Yang, S. Tyekucheva, A. Radakrishnan, R. S. Harris, F. Chiaromonte, J. Taylor, J. He, M. Rijnkels, S. Griffiths-Jones, A. Ureta-Vidal, M. M. Hoffman, J. Severin, S. M. Searle, A. S. Law, D. Speed, D. Waddington, Z. Cheng, E. Tuzun, E. Eichler, Z. Bao, P. Flicek, D. D. Shteynberg, M. R. Brent, J. M. Bye, E. J. Huckle, S. Chatterji, C. Dewey, L. Pachter, A. Kouranov, Z. Mourelatos, A. G. Hatzigeorgiou, A. H. Paterson, R. Ivarie, M. Brandstrom, E. Axelsson, N. Backstrom, S. Berlin, M. T. Webster, O. Pourquie, A. Reymond, C. Ucla, S. E. Antonarakis, M. Long, J. J. Emerson, E. Betran, I. Dupanloup, H. Kaessmann, A. S. Hinrichs, G. Bejerano, T. S. Furey, R. A. Harte, B. Raney, A. Siepel, W. J. Kent, D. Haussler, E. Eyras, R. Castelo, J. F. Abril, S. Castellano, F. Camara, G. Parra, R. Guigo, G. Bourque, G. Tesler, P. A. Pevzner, A. Smit, L. A. Fulton, E. R. Mardis and R. K. Wilson (2004). Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432(7018): 695-716.

Krylov, D. M., Y. I. Wolf, I. B. Rogozin and E. V. Koonin (2003). Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. Genome Res 13(10): 2229-35.

Lee, Y., R. Sultana, G. Pertea, J. Cho, S. Karamycheva, J. Tsai, B. Parvizi, F. Cheung, V. Antonescu, J. White, I. Holt, F. Liang and J. Quackenbush (2002). Cross-referencing eukaryotic genomes: TIGR Orthologous Gene Alignments (TOGA). Genome Res 12(3): 493-502.

Lee, Y., J. Tsai, S. Sunkara, S. Karamycheva, G. Pertea, R. Sultana, V. Antonescu, A. Chan, F. Cheung and J. Quackenbush (2005). The TIGR Gene Indices: clustering and assembling EST and known genes and integration with eukaryotic genomes. Nucleic Acids Res 33(Database issue): D71-4.

Leebens-Mack, J., L. A. Raubeson, L. Cui, J. V. Kuehl, M. H. Fourcade, T. W. Chumley, J. L. Boore, R. K. Jansen and C. W. depamphilis (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.

Leseberg, C. H., A. Li, H. Kang, M. Duvall and L. Mao (2006). Genome-wide analysis of the MADS-box gene family in *Populus trichocarpa*. Gene 378: 84-94.

Li, L., C. J. Stoeckert, Jr. and D. S. Roos (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13(9): 2178-89.

Ma, L., C. Chen, X. Liu, Y. Jiao, N. Su, L. Li, X. Wang, M. Cao, N. Sun, X. Zhang, J. Bao, J. Li, S. Pedersen, L. Bolund, H. Zhao, L. Yuan, G. K. Wong, J. Wang, X. W. Deng and J. Wang (2005). A microarray analysis of the rice transcriptome and its comparison to *Arabidopsis*. Genome Res 15(9): 1274-83.

Mueller, L. A., A. A. Mills, B. Skwarecki, R. M. Buels, N. Menda and S. D. Tanksley (2008). The SGN comparative map viewer. Bioinformatics 24(3): 422-3.

Nam, J., J. Kim, S. Lee, G. An, H. Ma and M. Nei (2004). Type I MADS-box genes have experienced faster birth-and-death evolution than type II MADS-box genes in angiosperms. Proc Natl Acad Sci U S A 101(7): 1910-5.

Ng, S. K., Z. Zhang, S. H. Tan and K. Lin (2003). InterDom: a database of putative interacting protein domains for validating predicted protein interactions and complexes. Nucleic Acids Res 31(1): 251-4.

O'Brien, K. P., M. Remm and E. L. Sonnhammer (2005). Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Res 33(Database issue): D476-80.

Paterson, A. H., J. E. Bowers, D. G. Peterson, J. C. Estill and B. A. Chapman (2003). Structure and evolution of cereal genomes. Curr Opin Genet Dev 13(6): 644-50.

Quackenbush, J., J. Cho, D. Lee, F. Liang, I. Holt, S. Karamycheva, B. Parvizi, G. Pertea, R. Sultana and J. White (2001). The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. Nucleic Acids Res 29(1): 159-64.

Quackenbush, J., F. Liang, I. Holt, G. Pertea and J. Upton (2000). The TIGR gene indices: reconstruction and representation of expressed gene sequences. Nucleic Acids Res 28(1): 141-5.

Rensink, W. A. and C. R. Buell (2004). *Arabidopsis* to rice. Applying knowledge from a weed to enhance our understanding of a crop species. Plant Physiol 135(2): 622-9.

Rensink, W. A. and C. R. Buell (2005). Microarray expression profiling resources for plant genomics. Trends Plant Sci 10(12): 603-9.

Roth, C., M. J. Betts, P. Steffansson, G. Saelensminde and D. A. Liberles (2005). The Adaptive Evolution Database (TAED): a phylogeny based tool for comparative genomics. Nucleic Acids Res 33(Database issue): D495-7.

Rudd, S. (2005). openSputnik--a database to ESTablish comparative plant genomics using unsaturated sequence collections. Nucleic Acids Res 33(Database issue): D622-7.

Sampedro, J., R. E. Carey and D. J. Cosgrove (2006). Genome histories clarify evolution of the expansin superfamily: new insights from the poplar genome and pine ESTs. J Plant Res 119(1): 11-21.

Samuga, A. and C. P. Joshi (2004). Differential expression patterns of two new primary cell wall-related cellulose synthase cDNAs, PtrCesA6 and PtrCesA7 from aspen trees. Gene 334: 73-82.

Sjolander, K. (2004). Phylogenomic inference of protein molecular function: advances and challenges. Bioinformatics 20(2): 170-9.

Swofford, D. L. (2002). PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4.

Tatusov, R. L., E. V. Koonin and D. J. Lipman (1997). A genomic perspective on protein families. Science 278(5338): 631-7.

Tuskan, G. A., S. DiFazio, S. Jansson, J. Bohlmann, I. Grigoriev, U. Hellsten, N. Putnam, S. Ralph, S. Rombauts, A. Salamov, J. Schein, L. Sterck, A. Aerts, R. R. Bhalerao, R. P. Bhalerao, D. Blaudez, W. Boerjan, A. Brun, A. Brunner, V. Busov, M. Campbell, J. Carlson, M. Chalot, J. Chapman, G. L. Chen, D. Cooper, P. M. Coutinho, J. Couturier, S. Covert, Q. Cronk, R. Cunningham, J. Davis, S. Degroeve, A. Déjardin, C. dePamphilis, J. Detter, B. Dirks, I. Dubchak, S. Duplessis, J. Ehlting, B. Ellis, K. Gendler, D. Goodstein, M. Gribskov, J. Grimwood, A. Groover, L. Gunter, B. Hamberger, B. Heinze, Y. Helariutta, B. Henrissat, D. Holligan, R. Holt, W. Huang, N. Islam-Faridi, S. Jones, M. Jones-Rhoades, R. Jorgensen, C. Joshi, J. Kangasjärvi, J. Karlsson, C. Kelleher, R. Kirkpatrick, M. Kirst, A. Kohler, U. Kalluri, F. Larimer, J. Leebens-Mack, J. C. Leplé, P. Locascio, Y. Lou, S. Lucas, F. Martin, B. Montanini, C. Napoli, D. R. Nelson, C. Nelson, K. Nieminen, O. Nilsson, V. Pereda, G. Peter, R. Philippe, G. Pilate, A. Poliakov, J. Razumovskaya, P. Richardson, C. Rinaldi, K. Ritland, P. Rouzé, D. Ryaboy, J. Schmutz, J. Schrader, B. Segerman, H. Shin, A. Siddiqui, F. Sterky, A. Terry, C. J. Tsai, E. Uberbacher, P. Unneberg, J. Vahala, K. Wall, S. Wessler, G. Yang, T. Yin, C. Douglas, M. Marra, G. Sandberg, Y. Van de Peer and D. Rokhsar (2006). The Genome of Black Cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 313(5793): 1596-1604.

Van Dongen, S. (2000). A cluster algorithm for graphs. Technical Report INS-R0010.

Wall, P. K., J. Leebens-Mack, K. F. Muller, D. Field, N. S. Altman and C. W. dePamphilis (2008). PlantTribes: a gene and gene family resource for comparative genomics in plants. Nucleic Acids Res 36(Database issue): D970-6.

Yamada, K., J. Lim, J. M. Dale, H. Chen, P. Shinn, C. J. Palm, A. M. Southwick, H. C. Wu, C. Kim, M. Nguyen, P. Pham, R. Cheuk, G. Karlin-Newmann, S. X. Liu, B. Lam, H. Sakano, T. Wu, G. Yu, M. Miranda, H. L. Quach, M. Tripp, C. H. Chang, J. M. Lee, M. Toriumi, M. M. Chan, C. C. Tang, C. S. Onodera, J. M. Deng, K. Akiyama, Y. Ansari, T. Arakawa, J. Banh, F. Banno, L. Bowser, S. Brooks, P. Carninci, Q. Chao, N. Choy, A. Enju, A. D. Goldsmith, M. Gurjal, N. F. Hansen, Y. Hayashizaki, C. Johnson-Hopson, V. W. Hsuan, K. Iida, M.

Karnes, S. Khan, E. Koesema, J. Ishida, P. X. Jiang, T. Jones, J. Kawai, A. Kamiya, C. Meyers, M. Nakajima, M. Narusaka, M. Seki, T. Sakurai, M. Satou, R. Tamse, M. Vaysberg, E. K. Wallender, C. Wong, Y. Yamamura, S. Yuan, K. Shinozaki, R. W. Davis, A. Theologis and J. R. Ecker (2003). Empirical analysis of transcriptional activity in the *Arabidopsis* genome. Science 302(5646): 842-6.

Yu, J., S. Hu, J. Wang, G. K. Wong, S. Li, B. Liu, Y. Deng, L. Dai, Y. Zhou, X. Zhang, M. Cao, J. Liu, J. Sun, J. Tang, Y. Chen, X. Huang, W. Lin, C. Ye, W. Tong, L. Cong, J. Geng, Y. Han, L. Li, W. Li, G. Hu, X. Huang, W. Li, J. Li, Z. Liu, L. Li, J. Liu, Q. Qi, J. Liu, L. Li, T. Li, X. Wang, H. Lu, T. Wu, M. Zhu, P. Ni, H. Han, W. Dong, X. Ren, X. Feng, P. Cui, X. Li, H. Wang, X. Xu, W. Zhai, Z. Xu, J. Zhang, S. He, J. Zhang, J. Xu, K. Zhang, X. Zheng, J. Dong, W. Zeng, L. Tao, J. Ye, J. Tan, X. Ren, X. Chen, J. He, D. Liu, W. Tian, C. Tian, H. Xia, Q. Bao, G. Li, H. Gao, T. Cao, J. Wang, W. Zhao, P. Li, W. Chen, X. Wang, Y. Zhang, J. Hu, J. Wang, S. Liu, J. Yang, G. Zhang, Y. Xiong, Z. Li, L. Mao, C. Zhou, Z. Zhu, R. Chen, B. Hao, W. Zheng, S. Chen, W. Guo, G. Li, S. Liu, M. Tao, J. Wang, L. Zhu, L. Yuan and H. Yang (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). Science 296(5565): 79-92.

Yuan, Q., S. Ouyang, A. Wang, W. Zhu, R. Maiti, H. Lin, J. Hamilton, B. Haas, R. Sultana, F. Cheung, J. Wortman and C. R. Buell (2005). The institute for genomic research Osa1 rice genome annotation database. Plant Physiol 138(1): 18-26.

Chapter 4

Comparison of next generation sequencing technologies for transcriptome characterization

Preface

This manuscript will be submitted to BMC Genomics. The authors are P. Kerr Wall[1], Jim Leebens-Mack[2], André Chanderbali[3], Abdelali Barakat[4], Haiying Liang[4], Lena Landherr[1], Lynn P. Tomsho[5], Yi Hu[1], John E. Carlson[4], Hong Ma[1], Stephan Schuster[5], Douglas E. Soltis[3], Pamela S. Soltis[6], Naomi Altman[7], and Claude W. dePamphilis. PKW, JLM, NA, and CWD designed the simulation study. PKW, JLM, and CWD wrote the manuscript and all authors suggested changes or edits, and approved the manuscript. PKW carried out the data analysis and simulation study. AC, AB, HL, LL, LT, YH generated the data for all GS-20 experiments, including the tissue extraction, library construction and optimization. PKW, JLM, HM, SS, DES, PSS, NA, CWD conceived of the study, and participated in its design. PKW, JLM, AC, AB, DES, PSS, NA, CWD coordinated and helped to draft the manuscript. All authors read and approved the final manuscript.

P. Kerr Wall[1], Jim Leebens-Mack[2], André Chanderbali[3], Abdelali Barakat[4], Haiying Liang[4], Lena Landherr[1], Lynn P. Tomsho[5], Yi Hu[1], John E. Carlson[4], Hong Ma[1], Stephan Schuster[5], Douglas E. Soltis[3], Pamela S. Soltis[6], Naomi Altman[7], and Claude W. dePamphilis[1§]

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

[2]Department of Plant Biology, University of Georgia, Athens, GA 30602, USA

[3] Department of Botany, University of Florida, P.O. Box 118526, Gainesville, FL, 32611, USA

[4] The School of Forest Resources, Department of Horticulture, and Huck Institutes of the Life Sciences, Pennsylvania State University, 323 Forest Resources Building, University Park, PA 16802, USA.

[5] Center for Comparative Genomics, Center for Infectious Disease Dynamics, The Pennsylvania State University

ity, University Park, PA 16802, USA

[6] Florida Museum of Natural History, University of Florida, P.O. Box 117800, Gainesville, FL, 32611, USA

[7] Department of Statistics and The Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, PA 16802, USA

[§]Corresponding author

Email addresses:

PKW: pkerrwall@psu.edu
JLM: jleebensmack@plantbio.uga.edu
AC: achander@botany.ufl.edu
AB: aub14@psu.edu
HL: hliang@clemson.edu
LL: lll109@psu.edu
LT: lap153@psu.edu
YH: yxh13@psu.edu
JEC: jec16@psu.edu
HM: hxm16@psu.edu
SS: scs@bx.psu.edu
DES: dsoltis@botany.ufl.edu
PSS: psoltis@flmnh.ufl.edu
NA: naomi@stat.psu.edu
CWD: cwd3@psu.edu

Abstract

Background

We have developed a simulation approach to help determine the optimal mixture of sequencing methods for most complete and cost effective transcriptome sequencing. We compared simulation results for traditional capillary sequencing with "Next Generation" (NG) ultra high-throughput technologies, specifically, the 454-GS20, 454-FLX, and Solexa (Illumina) platforms. The simulation model was parameterized using mappings of 130,000 cDNA sequence reads to the *Arabidopsis* genome. We also generated 454-GS20 sequences and *de novo* assemblies for the basal eudicot California poppy (*Eschscholzia californica*) and the magnoliid avocado (*Persea americana*) using a variety of methods for cDNA synthesis.

Results

The *Arabidopsis* reads tagged more than 15,000 genes, including new splice variants and extended UTR regions. Of the total 134,791 reads (13.8 MB), 119,518 (88.7%) mapped exactly to known exons, while 1,117 (0.8%) mapped to introns, 11,524 (8.6%) spanned annotated intron/exon boundaries, and 3,066 (2.3%) extended beyond the end of annotated UTRs. Sequence-based inference of relative gene expression levels correlated significantly with microarray data. As expected, NG sequencing of normalized libraries tagged more genes than non-normalized libraries, although non-normalized libraries yielded more full-length cDNA sequences. The *Arabidopsis* data were used to simulate additional rounds of GS20, FLX, Solexa, and traditional EST sequencing, and various combinations of each. Our simulations suggest a

combination of FLX and Solexa sequencing (or other similar combinations) for optimal transcriptome coverage at modest cost.

Conclusions

In terms of sequence coverage alone, the NG sequencing platforms are a dramatic advance over capillary-based sequencing, but NG sequencing also presents significant challenges in assembly and sequence accuracy due to short read lengths, method-specific sequencing errors, and the absence of physical clones that can be used in future experiments. These problems may be overcome by hybrid sequencing strategies using a mixture of sequencing methodologies, by new assemblers, and by sequencing more deeply. Sequencing and microarray outcomes from multiple experiments suggest that our simulator will be useful for guiding NG transcriptome sequencing projects in a wide range of organisms.

Background

Sequencing technology has made great advances over the last 30 years since the development of chain-terminating inhibitor-based technologies (Sanger et al. 1977). Traditional sequencing approaches require cloning of DNA fragments into bacterial vectors for amplification and sequencing of individual templates using vector-based primers. This approach was adapted for cDNA libraries (Adams et al. 1991) and, with the advent of capillary sequencing, became suitable for high-throughput sequencing of large samples of transcripts, termed Expressed Sequence Tags (ESTs). ESTs have become an invaluable resource for gene discovery, genome annotation, alternative splicing, SNP discovery, molecular markers for population analysis, and expression analysis in animal, plant, and microbial species (Bouck and Vision 2007). Other approaches for analyzing transcriptomes include serial analysis of gene expression (SAGE) (Velculescu et al. 1995), massively parallel signature sequencing (MPSS) (Brenner et al. 2000), and microarrays (Kulesh et al. 1987; Schena et al. 1995). These approaches, which involve the sequencing or hybridizing of small concatamers of cDNA derived from mRNA by reverse transcription, have been used successfully in analyzing the expression of genomes (transcriptomes) at a very large scale, usually from species with a sequenced genome or an existing and extensive EST data set. Although several alternatives have been described since the emergence of EST sequencing projects, none has yet totally supplanted the use of bacterial vectors and Sanger sequencing.

In 2005, two new sequencing technologies were introduced both based on sequencing by synthesis, which promised to replace or enhance traditional sequencing methods. The 454 system (www.454.com), using pyrosequencing technology (Margulies et al. 2005), and the Solexa system (www.illumina.com), which detects fluorescence signals (Porreca et al. 2007). Both

execute millions of sequencing reactions in parallel, producing data at ultrahigh rates (Bentley 2006). Although read lengths are much shorter with these new methods than with capillary sequencing, averaging 100–230 bp and 20-35 bp for 454 and Solexa, respectively, both platforms generate sufficient data to completely re-sequence bacterial genomes in a single run (Margulies et al. 2005; Shendure et al. 2005; Goldberg et al. 2006; Poinar et al. 2006). In the past year, Applied Biosystems has introduced their SOLiD sequencer (www3.appliedbiosystems.com), another short-read 20-35 bp platform. Recent studies have reported success with 454 sequencing of chloroplast genomes (Cai et al. 2006; Moore et al. 2006), small RNAs (Lu et al. 2006; Barakat et al. 2007; Barakat et al. 2007; Lu et al. 2007), and transcriptomes of organisms with (Bainbridge et al. 2006; Cheung et al. 2006; Weber et al. 2007) or without (Vera et al. 2007) extensive genomic sequence information. These Next Generation (NG) sequencing methods promise a cost-effective means of either deeply sampling or fully sequencing an organism's transcriptome, with even small experiments tagging a very large number of expressed genes. However, prior transcriptome sequencing studies have been largely exploratory, only hinting at the potential for NG transcriptome sequencing at different scales. There is a great need for quantitative studies and analysis tools that help investigators optimally design NG sequencing experiments to address specific goals.

A complete solution to this problem would involve realistic models for each technology, accounting for the cost of library generation and data collection, the characteristics of cDNA libraries, transcript abundance distributions, read length distributions, and the error rates in sequence generation and assembly. The present study focuses on the first four of these issues to provide estimates of theoretical coverage of complex transcriptomes with varying scales and types of DNA sequencing experiments. In earlier publications (Wang et al. 2004; Wang et al. 2005), we developed a robust simulation approach to model traditional capillary transcriptome sequencing, which incorporates distributions of the relative start site of cDNA sequences as a

function of cDNA length, the read length distribution, and the transcript abundance distribution. We have now adapted this simulation approach to model the specific characteristics of NG sequencing. The results from this study should help researchers working with these new and exciting technologies.

The present study has several goals. First, we report empirical comparisons of 454 pyrosequencing and capillary-based transcriptome sequencing from the model plant, *Arabidopsis thaliana*, and two non-model plant species, the basal eudicot *Eschscholzia californica* (California poppy) and the magnoliid *Persea americana* (avocado). We use these results to examine the effects of library preparation procedures, specifically, normalized versus non-normalized and random versus oligo-dT primed libraries. We then introduce a simulation approach, based on the GS20 sequencing results, to predict the outcome of additional GS20 transcriptome sequencing experiments while accounting for critical features in cDNA library construction. We then use the GS20 simulation results to extrapolate results for 454FLX and Solexa platforms, in order to estimate technology-specific sequencing characteristics. Finally, we report on simulated experiments aimed at characterizing the optimal mixture of methods for most complete and cost-effective transcriptome sequencing with one or more sequencing technologies.

Results

Next Generation Transcriptome sequencing of *Arabidopsis* floral tissue

A half plate of GS20 sequencing from an *Arabidopsis* random-primed cDNA library generated 134,791 reads totaling 13.8 MB with an average length of 102.2 bp. The reads were assembled into 82,281 unigenes, which included 8,188 contigs with an average length of 147 bp and 74,093 singleton reads (Table 4-1). We mapped 122,344 (90.8%) reads to the TAIR 7

Arabidopsis genome annotation (Table 4-2 and see Methods). Of the total mapped reads, 88.7% were located within 15,539 genic regions and 2.1% were located in intergenic regions. Within the genic regions, 119,518 (88.7%) reads mapped exactly to known exons, while 1,117 (0.8%) and 11,524 (8.6%) reads mapped to introns and intron/exon boundaries, respectively. Also, 3,066 (2.3%) of the reads included in the genic regions extended current boundaries of known genes while 302 reads combined two annotated genes or marked areas of the genome with overlapping genes. There were 12,447 (6.7%) reads that did not have a significant BLASTn match to any location within the genome. There were 1,085 genes that had more than 20 reads per locus, and the 10 most highly expressed genes (Table 4-3A), included two subunits of the photosynthetic protein RuBisCo, as well as *TASTY*, *TGG1*, and *PDF1*. These "top ten" transcripts had read counts ranging from 190 to 586 reads with the RuBisCO small subunit 1A being most highly represented. At this shallow sequencing depth, 2 non-overlapping contigs, with lengths of 357 and 240 bp, mapped to the RuBisCO small subunit 1A gene.

Despite low overall transcriptome coverage, one-half plate of *Arabidopsis* GS20 sequence data returned 27 fully sequenced cDNAs, as well as 292, 628, and 1008 genes at 90%, 80%, and 70% coverage, respectively. These results demonstrate that nominal amounts of 454 sequencing can generate complete or nearly complete sequences for an appreciable number of genes, especially those that are small and highly expressed. Another very promising result is the improved annotation of genes for both model and non-model species. For example, although the *Arabidopsis* genome has been largely sequenced since 2000 (AGI 2000), the half plate of GS20 extended the untranslated regions (UTRs) of roughly 3,066 genes and mapped new transcript boundaries of 8,662 genic regions. These regions are possibly new splice variants of previously annotated genes. Finally, 2,826 transcripts were mapped to 2,096 unique intergenic regions. These transcripts might represent un-annotated protein-coding genes or non-coding RNA sequences that have not previously been sampled in traditional cDNA libaries.

Table **4-1**: Sequencing Statistics of analyzed libraries.

Read, Contig, Singleton, and Unigene Counts (n), mean sequence lengths ($\bar{x}$), and total amount of sequence data (MB) for 454 GS20 libraries analyzed. Species codes are Ath (*Arabidopsis thaliana*), Pam (*Persea americana*, avocado), Eca (*Eschscholzia californica*, California poppy). cDNA library production method indicated in parentheses. Read lengths based on number of Q20 equivalent bases produced, after trimming and cleaning with the program seqclean (http://compbio.dfci.harvard.edu/tgi/software/); normalized library original read length was 100.1 prior to trimming normalization adapter**.**

	Ath (random)		Pam (normal)		Eca (oligo-dT)		Eca (random)		Eca (combined)	
Type	**n**	$\bar{x}$	**n**	$\bar{x}$	**n**	$\bar{x}$	**n**	$\bar{x}$	**n**	$\bar{x}$
Reads	134,791	102.2	269,057	85.9	251,716	98.9	307,836	98.2	559,552	98.6
Contigs	8,188	147.0	22,303	107.3	18,339	148.5	14,242	146.9	30,603	159.1
Singletons	74,093	101.6	211,882	85.0	64,931	99.9	61,031	99.1	89,982	99.5
Unigenes	82,271	106.1	234,185	90.6	83,270	107.7	75,273	105.1	120,585	106.9
MB	13.8		23.1		24.9		30.2		55.1	

Table **4-2**: *Arabidopsis* 454 Reads Mapped to the annotated genome.

All 454 reads were mapped (BLAST-n, default parameters) to the genome. TAIR XML files were parsed to obtain exon structure and location within the genome. Percentages were calculated for each class of sequence type. The number of genes does not equal the summation of gene components because there are some genes that are hit by multiple reads in different sections of the gene. The percent for each gene component is the percent of total reads.

Sequence Type	Reads	Genes	Total (%)
Genes	119,518	15,539	88.7
Exon	103,509	14,754	76.8
Intron	1,117	877	0.8
Intron/Exon	11,524	5,973	8.6
Extended UTR	3,066	1,635	2.3
Overlapped Genes	302	177	0.2
Intergenic	2,826	2,096	2.1
No Hit	12,447		9.2
Total	134,791		100.0

Table **4-3**: Top 10 Most Frequently Detected unigenes in 454 cDNA libraries of *Arabidopsis*, *Eschscholzia*, and *Persea*.

A. *Arabidopsis* 454 GS20 Reads were mapped to the annotated TAIR cDNA and protein datasets using BLASTn and BLASTx, respectively. Column headers shown are contig name (Contig), contig length (Len), number of reads per contig (Reads), percent coverage (Cov), *Arabidopsis* gene identifier (AGI), E-value, and *Arabidopsis* gene annotation (Annotation). Column headers are same for Tables 4-3B-D.

B. Top 10 Most Frequently Detected unigenes in California poppy flower bud oligo-dT library. Each unigene was searched (BLASTn or BLASTx, with cutoff e-5) against the *Arabidopsis* proteome and *Arabidopsis* cDNA datasets. Ribosomal RNA and contaminants such as putative endophytes removed from this list.

C. Top 10 Most Frequently Detected unigenes in California poppy flower bud random-primed library.

D. Top 10 Most Frequently Detected unigenes in Persea americana flower bud normalized library.

A. *Arabidopsis*

Contig	Len	Reads	Cov	AGI	Len	Evalue	Annotation
08061	357	586	34.8	AT1G67090	1025	0.0	RuBisCO small subunit 1A (RBCS-1A) (ATS1A)
00035	1326	541	96.8	AT1G54040	1370	0.0	TASTY, ESP (EPITHIOSPECIFIER PROTEIN)
08724	1653	391	90.0	AT5G26000	1836	0.0	TGG1 (THIOGLUCOSIDE GLUCOHYDROLASE1)
08295	1175	278	94.6	AT2G42840	1242	0.0	PDF1 (PROTODERMAL FACTOR 1)
08670	310	258	31.5	AT5G38410	984	4e-175	RuBisCO small subunit 3B (RBCS-3B) (ATS3B)
00011	240	229	23.4	AT1G67090	1025	9e-43	RuBisCO small subunit 1A (RBCS-1A) (ATS1A)
00660	640	219	76.9	AT2G21660	832	2e-157	ATGRP7 (Cold, Circadian Rhythm, RNA Binding 2)
07960	927	215	52.6	AT5G60390	1764	0.0	elongation factor 1-alpha / EF-1-alpha
04760	1157	206	82.3	AT3G12145	1406	0.0	FLR1 (FLOR1); enzyme inhibitor
08550	373	190	100	ATCG00220	105	3e-53	PSBM, PSII low MW protein

B. *Eschscholzia* randomly-primed

Contig	Len	Reads	Cov	AGI	Len	Evalue	Annotation
19682	387	850	83.2	AT5G39170	465	2e-7	Unknown protein
19707	2089	784	100	AT1G70370	1878	0	BURP domain-containing protein / polygalacturonase
18128	151	707	10.0	AT3G47550	1505	0.02	C3HC4-type RING finger family protein
19695	308	678	100	AT5G52160	288	1e-15	protease inhibitor/seed storage/lipid transfer protein
19793	940	608	100	AT2G36830	753	6e-102	GAMMA-TIP (Tonoplast intrinsic protein gamma)
18734	849	485	100	AT3G16640	504	7e-52	TCTP (Translationally Controlled Tumor Protein)
00048	2823	450	80.0	AT5G35750	3528	0	AHK2 (Arabidopsis Histidine Kinsase 2)
18697	144	450	24.7	AT4G06746	584	0.31	RAP2.9 (related to AP2 9); transcription factor
19623	2638	421	81.4	AT2G01830	3240	0	WOL (CYTOKININ RESPONSE 1)
19622	120	415	6.4	AT1G23800	1866	1	ALDH2B7 (Aldehyde dehydrogenase 2B7)

C. *Persea* flower bud normalized library

Contig	Len	Reads	Cov	AGI	Len	Evalue	Annotation
15341	109	4296	6.9	AT4G03930	1575	0.23	Pectinesterase
15345	162	4274	12.0	AT3G59430	1353	0.33	Unknown protein
15162	315	852	19.7	AT5G26670	1596	0.18	Pectinacetylesterase, putative
15258	606	726	53.3	AT3G12340	1137	0.1	FK506 binding / peptidyl-prolyl cis-trans isomerase
14312	182	682	56.2	ATMG00030	324	2e-77	ORF107A
15290	2020	674	100	AT1G70370	1878	0	BURP domain-containing protein / polygalacturonase
15208	1052	514	100	AT2G36830	753	7e-102	GAMMA-TIP (Tonoplast intrinsic protein gamma)
14424	2660	480	75.4	AT5G34750	3528	2e-162	AHK2 (ARABIDOPSIS HISTIDINE KINASE 2)
15320	1146	437	48.3	AT5G02500	2373	0	HSC70-1 (heat shock cognate 70 kDa protein 1)
15269	304	417	5.8	AT2G47410	5221	0.2	Nucleotide binding

D. *Eschscholzia* (California poppy) Oligo dT library

Contig	Len	Reads	Cov	AGI	Len	Evalue	Annotation
15603	133	37	9.8	AT1G59830	1357	0.005	PP2A-1 (protein phosphatase 2A-2)
18074	139	32	10.4	AT1G14270	1343	0.3	CAAX amino terminal protease family protein
8473	176	27	10.4	AT4G17890	1688	0.1	AGD8, UBP20 (Ubiquitin-specific Protease 20)
14132	213	26	7.3	AT2G40820	2907	1.9	Proline-rich family protein
15140	237	26	48.5	AT2G41430	489	2e-13	ERD15 (Early Responsive To Dehydration 15)
4395	144	25	6.4	AT1G45545	2259	0.08	Similar to unknown protein
15762	102	24	3.5	AT1G01950	2901	0.2	Armadillo/beta-catenin repeat family protein
10833	112	20	6.4	AT3G03640	1747	0.001	GLUC (Beta-glucosidase homolog)
18760	253	19	59.0	AT4G14270	429	2e-04	Protein containing PAM2 motif
18306	208	18	48.5	AT4G14270	429	8e-05	Protein containing PAM2 motif

Transcriptome sequencing of *Eschscholzia californica* using oligo-dT and random-primed libraries

Two full plates (over 559,000 total reads) of GS20 sequencing was performed on the emerging model basal eudicot, *Eschscholzia californica* (Carlson et al. 2006; Wege et al. 2007), including one plate from a 454 library of oligo-dT primed cDNA and one plate from a 454 library of random hexamer-primed cDNA. The library of oligo-dT primed cDNA generated 251,716 reads totaling 24.9 MB with an average length of 98.9 bp. The reads assembled into 83,270 unigenes, including 18,339 contigs with an average length of 148.5 bp and 64,931 singletons (Table 4-1). The library of random-primed cDNA generated 307,836 reads totaling 30.2 MB with an average length of 98.2 bp. The reads assembled into 75,273 unigenes, including 14,242 contigs with an average length of 146.9 bp and 61,031 singleton reads (Table 4-1). Finally, we assembled both plates, which resulted in 120,585 unigenes, including 30,603 contigs with an average length of 157.0 bp and 89,892 singleton reads (Table 4-1).

As expected, the most obvious difference between the oligo-dT and random-primed cDNA sequences was the representation of rRNA genes. Additional rounds of mRNA purification, however, could have reduced the level of rRNA "contamination". We also examined the relative start positions of the reads from each library by mapping the reads to the proteome of *Arabidopsis* (Figure 4-1B). The relative start positions are defined as the start position of the best *Arabidopsis* HSP divided by the length of the best protein match. As expected, the oligo-dT library had a greater 3' bias than the random primed library. The unigenes from both libraries mapped to 6,498 unique *Arabidopsis* genes, with 4,066 of the transcripts found in both. The level of redundancy observed between these two plates (just 62.6%) suggests that many more genes would be discovered with additional sequencing.

Transcriptome sequencing in a normalized library of *Persea americana*

One plate of GS20 sequencing was performed on a normalized library for *Persea americana*, an emerging model for the magnoliids (Chanderbali et al. 2007). The plate generated 298,055 reads totaling 29.8 MB with an average length of 100.1 bp. We then trimmed the adaptors used in the normalization step, which reduced the total number of reads to 269,057 with an average sequence length of 85.9 bp. Trimming the adaptors reduced the total amount of sequence by more than 6 MB, bringing the total to 23.1 MB. The reads assembled into 234,185 unigenes, including 22,303 contigs with an average length of 107.3 bp and 211,882 singleton reads (Table 4-1).

To determine the success of the normalization step, we plotted the relative frequency of the number of reads per gene, using *Arabidopsis* as a reference (Figure 4-1B). Compared to the other library methods used in this study, the normalized *Persea* library (solid blue line) contained the largest number of genes with fewer than five reads per gene and the fewest number of genes with more than 5 reads per gene. The gene with the highest number of mapped reads was a protein phosphatase with 37 reads. In contrast, the most highly represented genes in the poppy non-normalized libraries had over 1000 reads mapping to specific *Arabidopsis* genes. Hence, the normalization step was successful. Note that the *Persea* library, constructed using the Trimmer-Direct Kit (Evrogen) with amplification of full-length cDNAs (Clontech's SMART technology), also has the least amount of 3' bias in read start positions (Figure 41-A).

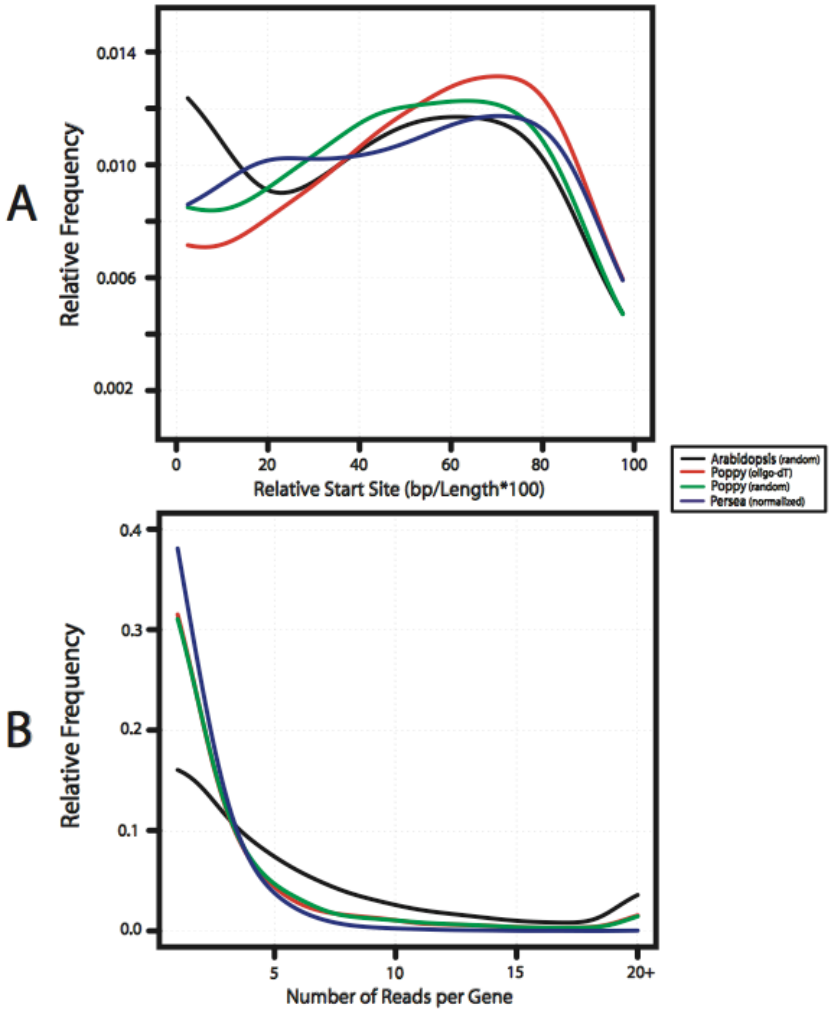

Figure **4-1**: Distributions of relative start sites and number of reads per gene.

A. Start site distributions of 454 sequences for for each species in this study including random, oligo-dT, and normalized oligo-dT libraries. Sequencing start sites are calculated as the start position, defined by BLASTn (*Arabidopsis*) or BLASTx (*Eschscholzia*, *Persea*) hit divided by the cDNA or protein length and expressed as percentage of the gene length.

B. Distribution of the number of reads from each library mapped to an *Arabidopsis* gene, defined by best BLASTn or BLASTx hit of each read to the TAIR genes. Species abbreviations are ATH (*Arabidopsis thaliana*), ECA (*Eschscholzia californica*), and PAM (*Persea americana*).

Correlation of observed *Arabidopsis* transcript frequencies with microarray data

Of the 21,707 genes included on the *Arabidopsis* Affymetrix (AFFY) microarray, 13,790 had at least one read mapped to its cDNA sequence. For these genes, we used AFFY microarray expression values generated from inflorescence tissue in the same *A. thaliana* ecotype (Zhang et al. 2005) to compare with the number of 454 reads for each gene. The comparison revealed that 1,907 genes that were detected above normalized expression level 50 with the AFFY chip were not detected in the 454 sequences, while 1,375 genes were detected in 454 reads, but were below expression level 50 with AFFY data (a common cutoff for reliable detection with the AFFY system). An additional 1,717 genes detected by 454 reads were not included as probes on the AFFY gene chip. A moderate correlation was observed between microarray expression values and number of 454 reads (Figure 4-2 with $r = 0.67$, $r^2 = 0.444$, $p < 0.0001$).

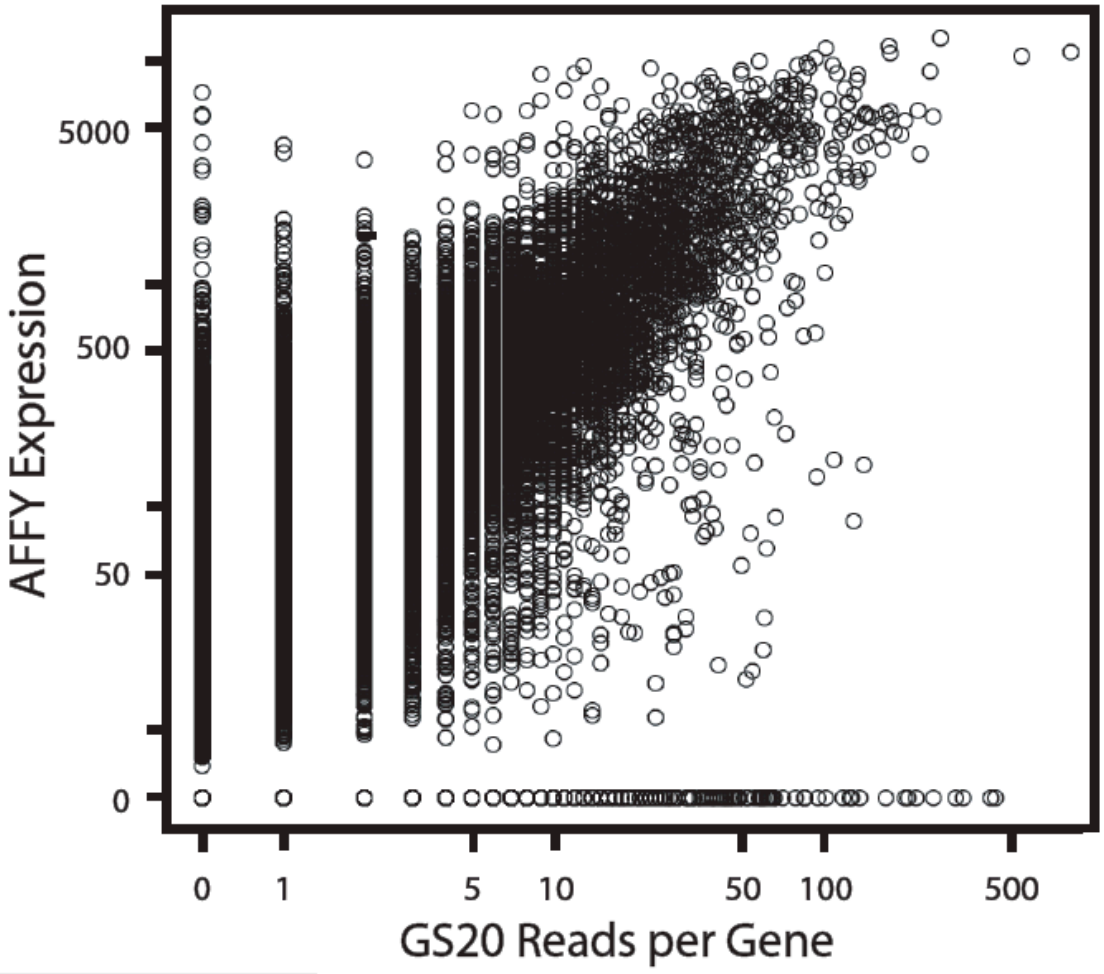

Figure **4-2**: Correlation of gene expression with number of transcripts.

Linear Regression comparing number of 454 Reads with Affymetrix (AFFY) gene chip expression values for Arabidopsis young inflorescence. Each symbol represents a single gene, with many genes having overlapping counts. Correlation between the two measures of gene expression is highly significant ($r = 0.67$, $r^2 = 0.444$, $p < 0.0001$).

Next Generation transcriptome simulation study

A primary goal of large-scale transcriptome sequencing is to identify and obtain full-length sequences of all of the expressed genes in an organism or tissue. A researcher will typically begin with RNAs isolated from a tissue of interest or a collection of tissues from the entire organism. The researcher may use tissue from a particular developmental stage or assay gene expression under a range of experimental conditions (e.g., light/temperature/water/nutrient stress, gene knock out). Each of the new NG technologies (e.g., 454-GS20/FLX, Solexa) produces data with characteristics that can be evaluated and compared to each other and traditional capillary sequencing.

In order to predict the expected outcomes of varied amounts of sequencing effort using a blend of technologies, we developed a predictive model based on the simulation engine of ESTstat (Wang et al. 2004; Wang et al. 2005). Inputs to the model include four distribution profiles that reflect information about the cDNA library or sequencing technology: 1) the transcript abundance profile, a transcriptome-specific frequency distribution of the number of tags of different genes in the entire transcriptome, 2) the distribution of cDNA lengths 3) the distribution of sequencing start sites, and 4) the distribution of read lengths after removal of vector and low quality data. The first three of these reflect library specific features, while the fourth is mostly dependent upon the sequencing technology. The ESTstat simulation model has been tested under a variety of situations and found to robustly predict the outcomes of future sequencing experiments. Although ESTstat can estimate and correct assembly errors *in silico* without reference to a known genome sequence, we were able to map each read to its known location on the *Arabidopsis* genome to assess and correct assembly error.

We used the results from our GS20 sequencing to simulate different levels of sequencing coverage for each of the NG and capillary technologies. For each technology, we considered

both non-normalized and perfectly normalized libraries, in which the expression level of every gene is made identical. Actual normalization experiments should therefore fall somewhere between non-normalized and perfectly normalized, depending on the normalization method, RNA quality, and success of the normalization procedure (see Materials and Methods for more detail). We used the following parameters to help evaluate the different sequencing platforms: transcriptome coverage, percentage of all expressed genes that were tagged, percentage of singletons, number of unigenes, mean unigene length, and the percentage of all expressed genes that were sequenced completely (i.e. 100% covered; Figures 4-3A-F).

Figure **4-3**: Simulation results for different Next Generation sequencing technologies.

(Figure on next page)

Simulation results illustrating predicted outcomes for different transcriptome sequencing technologies with a complex library expressing ca. 18,000 genes. Left column illustrates predicted outcomes as a function of MB of sequence, right column gives predicted outcomes as a function of estimated sequencing cost (see text for cost assumptions, which do not include varied costs for RNA isolation and library preparation). Each simulated data set was used to calculate: A) percent of transcriptome sequenced with at least one read and not necessarily in one contiguous sequence, B) number of genes tagged, C) number of unigenes obtained, D) mean unigene length (bp), E) percent of reads that are singleton sequences, and F) the number of genes with 100% coverage. Each technology is represented by a different line color, with solid lines indicating non-normalized libraries and dashed lines indicating theoretically perfectly normalized libraries. EST5 = 5' capillary sequence (black); GS20 = 454 GS20 (green); GSFLX = 454 GSFLX (blue); SOL = Solexa (red). The following prices (per MB) were used in the calculations: EST5 ($1330), GS20 ($240), GSFLX ($90), and SOL ($4). For several of the measures, the Solexa result is hidden under the topmost line. Additional details provided in text. Figure on next page.

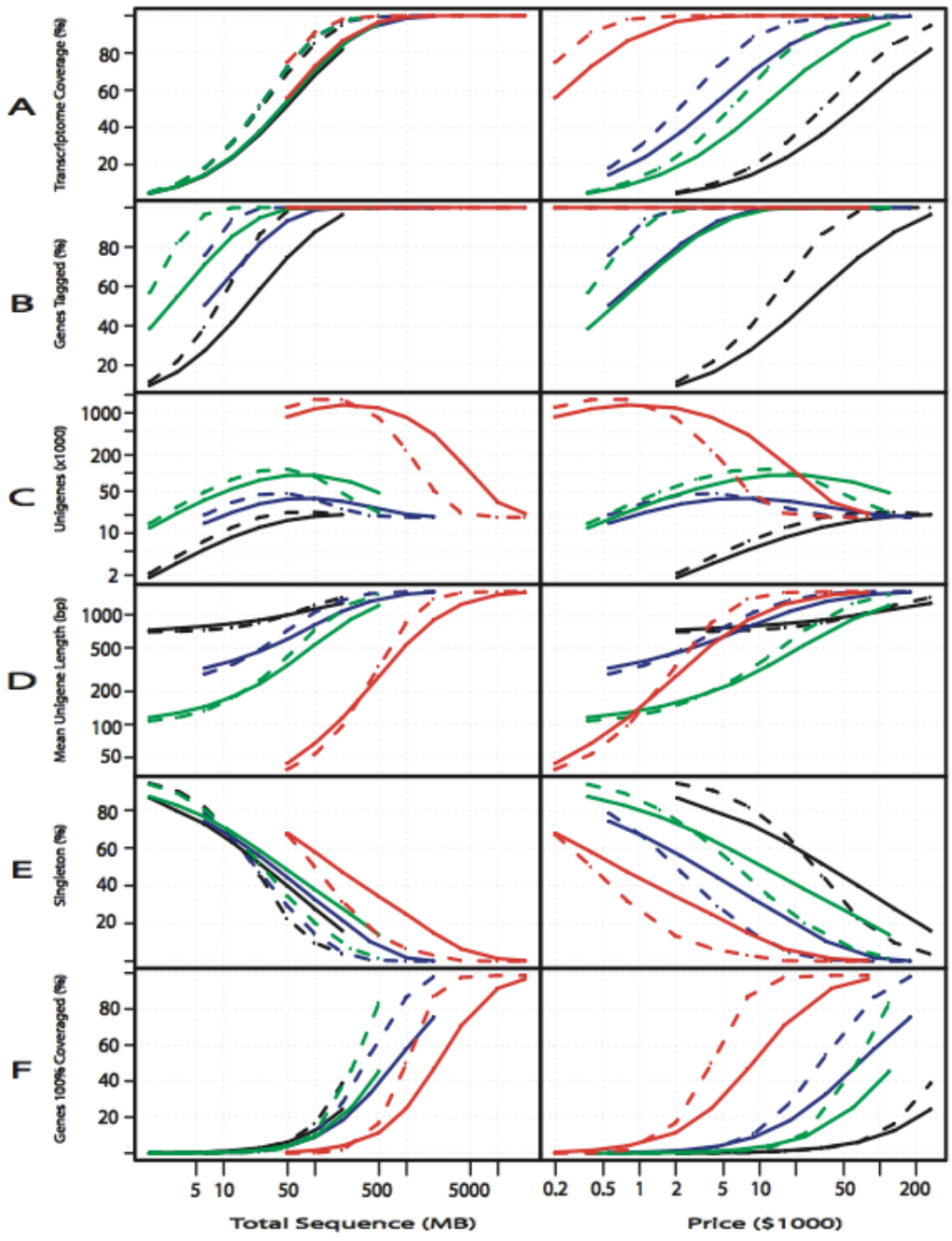

Figure **4-3**: Simulation results for different Next Generation sequencing technologies. (Legend on previous page)

Transcriptome coverage (Figure 4-3A) is a direct indicator of the sequencing depth of an organism's transcriptome. We define the transcriptome coverage as the total non-redundant number of bases from sampled genes that are included in at least one EST, divided by the sum of cDNA lengths for all expressed genes (including both detected and undetected genes in the transcriptome). In this study, the 15,276 detected genes and randomly sampled 3,007 undetected genes (estimated using ESTstat, see Materials and Methods) sum to 18,283 genes, with an expected total cDNA length of 29.8 MB. The transcriptome coverage, as a function of the total number of sequenced bases (MB), differs only slightly for all technologies. However, when the amount of sequence is low (1-500 MB), the transcriptome coverage is greater in the normalized libraries (dashed lines) compared to the non-normalized libraries (solid lines) for each technology. Theoretically, perfect normalization will equalize the level of expression for all genes, without any other impact on library quality, and thus will increase the coverage of genes that are randomly sampled. Using the distributions of cDNA length, read length, and sequencing start sites obtained in these experiments, we estimate that traditional 5' capillary sequencing of a non-normalized library will cover approximately 14%, 52%, and 82% of the transcriptome with 6.25, 50, and 200 MB of sequencing, respectively. For a normalized library, the percentage increases to 18%, 69%, and 95% with the same amounts of sequence. The same pattern was observed for the NG technologies but with higher levels of transcriptome coverage. For example, the GS20 technology is estimated to cover 15%, 54%, and 88% of the transcriptome for a non-normalized library and 18.2%, 72%, and 98% of the transcriptome for a normalized library at 6.25, 50, and 200 MB of sequencing. The lower coverage of capillary-based EST sequencing given the same number of sequenced bases is attributed to biases implicit in the cDNA cloning process. The FLX is estimated to cover 15%, 54%, and 88% for the non-normalized library and 18%, 72%, 98% for a normalized library at the same intervals. Finally, the Solexa platform is

estimated to cover 55% and 87% for the non-normalized library and 75% and 98% for the normalized library for 50 and 200 MB, respectively. Given that one plate of sequence data from the Solexa platform is estimated at 1,000 MB, we chose 50 MB (1/20 of a plate) as the first interval to be simulated, and we excluded all intervals less than 50 MB.

Transcriptome coverage differs substantially among the various technologies at the same cost. However, the cost used in this analysis refers only to the actual sequencing costs and not the preprocessing costs such as library preparation and normalization. The Solexa platform rapidly approaches 100% coverage primarily because the cost of sequencing is substantially smaller per MB (simulations for Solexa were based on $4000/plate at 1,000 MB/plate). Solexa is followed by GS20, FLX, and conventional EST sequences. It is estimated that traditional capillary sequencing would reach 100% transcriptome coverage at more than 200 MB and at a cost of over $200,000. While Solexa sequencing is the most economical technology for deep coverage of transcriptomes, *de novo* assembly of short Solexa sequences for non-model species remains an unresolved challenge.

A second indicator of the depth of transcriptome sequencing is the *percentage of genes tagged* (Figure 4-3B). A gene is considered tagged if it has been sampled with at least one read. The percentage of genes tagged increases with both amount of sequencing and price. For a non-normalized traditional library, we estimate that 27%, 75%, and 96% of the genes will be tagged in our sample transcriptome with 6.25, 50, and 200 MB of sequencing. For a normalized library, the percentage increases to 39%, 98%, and 100% with the same amounts of sequence. As expected, this percentage increases when the sequencing is done with any of the NG technologies. The cost of gene tagging also differs substantially among the various sequencing technologies. The Solexa platform tags essentially 100% of the expressed genes with less than one plate of sequence ($4000). Solexa is followed by GS20, FLX, and conventional EST sequences. Capillary sequencing would approach 100% genes tagged at more than 200 MB and over $200,000.

The *number of unigenes* (Figure 4-3C) - including singletons and contigs - has typically been used to estimate the number of transcribed genes in a tissue. With small amounts of sequencing, the number of unigenes is similar to the number of sequences, but with more sequencing multiple reads are observed for each gene (increasing redundancy), and the rate of discovery for new genes falls off. At a particular point in the sequencing process (peaks in Figure 4-3C), the number of unigenes will begin to decrease as disconnected reads coalesce into contigs covering entire genes, and eventually the unigene number approaches the number of genes expressed in the library. The rate at which multiple reads for a gene coalesce into a single contig is a function of read length. With the capillary technology, each read is large compared to the NG reads. With a non-normalized library similar to the model library, we will reach the peak unigene number at more than 200 MB of sequencing. With a normalized library, we reach the peak at approximately 100 MB and decrease gradually with an additional 100 MB of sequence. However, we still do not reach the estimated 18,000 genes expressed in the *Arabidopsis* floral library. For the FLX technology, the maximum number of unigenes occurs at roughly 100 MB and 50 MB for the non-normalized and normalized libraries, respectively. However, because the FLX sequences are two to three times shorter than the traditional sequences, the peak is reached with roughly double the number of unigenes (38,000 and 46,000, respectively). For the GS20 platform, the peaks occur at nearly the same levels (approximately 100 MB) as the FLX platform, but since these reads are half as long as FLX reads, the GS20 produces more than twice the number of unigenes (92,000 and 115,000) for both library types. The Solexa platform produces many more unigenes at all levels of sequencing and the peak occurs at approximately 200 MB for both library types (1.3 and 1.7 million reads).

The *mean unigene length* (Figure 4-3D) is an important statistic if the goal of the transcriptome sequences is to perform multi-gene phylogenetic or molecular evolutionary analyses. In this case, researchers would like full-length sequences for many expressed genes,

not just small fragments of expressed genes. In the *Arabidopsis* genome, the average transcript length is approximately 1,500 bp (1,436 for all transcripts and 1,628 bp for only the transcripts predicted to be expressed in this library). Therefore, a researcher would like to sequence enough of a library to produce contiguous sequences with average lengths of all genes in the library. We calculated the unigene length in two different ways. First, we used the mean length of all unigenes, although this estimate lowers the mean length for the shorter sequences in the NG technologies. Second, we calculated the mean length of only the longest unigenes for each gene (Figure 4-3D). All NG technology and library type combinations require greater depth of sequencing to reach the same level as its traditional counterpart. When we examine the mean unigene length in relation to price, the traditional sequencing produces the longest unigenes until approximately $5,000 worth of sequencing. This is approximately 4-5 MB of capillary sequencing and 6,000-8,000 reads. At this point, the NG technologies begin to generate enough sequences to assemble longer unigenes at a lower cost.

The *percentage of singleton reads* (Figure 4-3E) reflects sequencing depth and the likelihood that a given read will assemble to form a contig with other reads. A singleton is defined as a single read that does not contain enough overlap in length to be combined with other reads from the same transcribed gene. The percentage of singletons is also inversely proportional to the levels of redundancy in the library. Therefore, additional sequencing usually reduces the percentage of singletons. This is the case for capillary sequencing, where the percentages of singletons are 73%, 40%, and 16% for non-normalized and 81%, 23%, and 4% for normalized libraries at the 6.25, 50, and 200 MB levels, respectively. For the GS20, these values change to 76%, 48%, and 25% for non-normalized libraries and 80%, 34%, and 7% for normalized libraries at the same levels. For the FLX, the percentage of singletons changes to 74%, 44%, and 22% for non-normalized and to 78%, 29%, and 5% for normalized libraries at the same levels. Finally, for Solexa, the percentage of singletons is predicted to be around 68%, 47%, and 25% for non-

normalized and 67%, 32%, and 7% for normalized libraries at the 50, 200, and 1000 MB sequence intervals, respectively.

The final parameter used to evaluate and compare the technologies is the *percentage of genes with 100% coverage* (Figure 4-3F). As with mean unigene length, gene coverage can be calculated using all of the unigenes per gene, or by using only the longest unigene. The smaller reads from the NG technologies might cover all the regions within a gene. However, many of the reads for a gene will not have sufficient overlap to assemble into a contiguous sequence. Although we calculated both estimates, we use the percentage of gene coverage based on the longest unigene for comparisons to other platforms. In relation to amount of sequencing (MB), the capillary, GS20, and FLX technologies have similar percentages. The Solexa platform requires more data (MB of sequencing) to fully sequence a similar number of genes. For example, the FLX generates unigenes that completely cover roughly 18% and 58% of the total genes with 200 MB and 1000 MB of sequence data. The same amounts of Solexa sequencing would fully sequence 4% and 25% of the genes. However, the FLX experiment would cost approximately $18,000 and $90,000, whereas the Solexa data could be generated for roughly $800 and $4,000. Finally, with capillary sequencing, 200 MB would need to be sequenced at $250K to fully cover 25% of the genes.

Combinations of traditional and NG sequencing

Analyses of genome sequencing projects suggest that optimal genome assemblies can be obtained through a combination of traditional and NG technologies (Goldberg et al. 2006). In order to investigate the combination of these new technologies for transcriptome sequencing, we examined the addition of NG sequences to traditional capillary sequences (Figure 4-3A-C) and the combinations of NG sequences alone (Figure 4-3D-F). All of the indicators from the previous section dramatically improved with the addition of small amounts of NG sequences. Among the various combinations of technologies, there is little difference in most of the indicators used in the previous section. For example, the percentage of genes tagged approaches 100% with very small amounts of NG sequences. Therefore, to evaluate the various combinations of technologies, we compared three of the statistics described above: mean unigene length, transcriptome coverage, and percent of genes 100% covered.

The addition of NG sequences to traditional capillary sequences increased each of these three indicators at most sequence increments (Figure 4-4A-C). Only the addition of one plate of Solexa and all GS20 plate increments decreased the mean unigene length (Figure 4-4A). The addition of four plates of FLX increased the mean unigene length to 1327 and 1380 bp with 3.25 and 50 MB and of traditional sequences, respectively. At these same increments, transcriptome coverage would increase from 94% to 95% (Figure 4-4B), while the percent of genes 100% covered would increase from 33% to 38% (Figure 4-4C). The addition of this amount of FLX would increase the total cost of sequencing from $40K to $102,000. However, sequencing only four plates of FLX, assuming perfect assembly, could in theory generate 1323-bp unigenes at under $40,000, with approximately 94% transcriptome coverage and covering 37% of the genes 100% covered. Adding four plates of Solexa to four plates of FLX would generate 1466 bp

unigenes at just over $50,000 (Figure 4- 4D). This amount of sequencing would cover 100% of the transcriptome (Figure 4-4E) and fully sequence 84% of the genes (Figure 4-4F). Under these conditions, the primary advantage of including Sanger sequences would be the improvement of assembly through the inclusion of long individual reads, and simplification of downstream experiments with physical clones.

Figure **4-4**: Simulation results for combinations of Next Generation sequencing technologies.

(Figure on next page)

Simulation results illustrating predicted outcomes from combined sequencing technologies with a complex library expressing ca. 18,000 genes. A-C) Combinations include 3.125, 12.5, or 50 MB 5' Sanger sequencing plus 0 to 4 plates of GSFLX and/or Solexa sequence. Each technology or combination of technologies is represented by a different line color, with black indicating Sanger alone, and blue, red, and green lines indicating the addition of GSFLX, Solexa, and GSFLX+Solexa, respectively. The square (n=1), triangle (n=2), and diamond (n=4) shaped points on each line indicate the number of plates added for each technology. Results shown for non-normalized libraries. A) Mean length of longest Unigene per gene (bp), B) Transcriptome Coverage (%), and C) Number of Genes with 100% coverage as a function of total sequence (MB, left panel) and estimated sequencing cost ($1000, right panel) for the different technologies and combinations of technologies. Abbreviations and cost functions are as described in Fig. 3. D-F) Combinations include 100, 200, or 400 MB of GSFLX non-normalized or normalized sequencing plus 0 to 4 plates of Solexa sequence. Each technology or combination of technologies is represented by a different line color, with black and red lines indicating GSFLX alone with non-normalized and normalized libraries, respectively. Green and pink lines indicate the addition of Solexa non-normalized and normalized sequences, respectively. The square (n=1), triangle (n=2), and diamond (n=4) shaped points on each line indicate the number of plates added for each technology. D) Mean length of longest Unigene per gene (bp), E) Transcriptome Coverage (%), and F) Number of Genes with 100% coverage as a function of total sequence (MB, left panel) and estimated sequencing cost ($1000, right panel) for the different technologies.

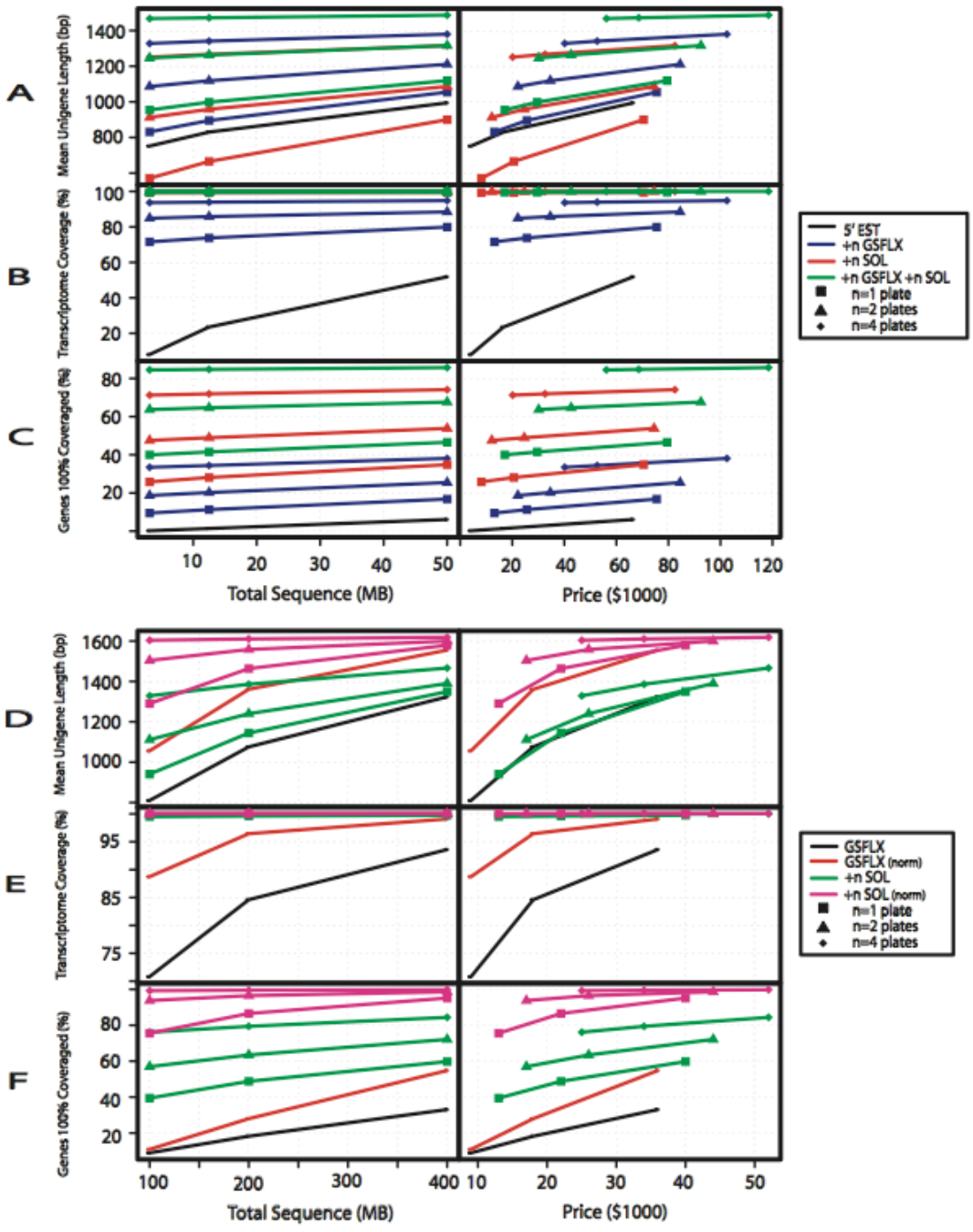

Figure **4-4**: Simulation results for combinations of Next Generation sequencing technologies. (Legend on previous page)

Discussion

NG transcriptome sequencing

Next Generation sequencing has great potential for accurate transcriptome characterization because of the large amount of data obtained at considerably lower costs compared to traditional methods. Although the cost of traditional sequencing (over $1000/MB) has continued to decrease over the last decade, the lower cost of NG sequencing ($250/MB for GS20, $90/MB for FLX, and $5/MB for Solexa) will dramatically improve transcriptome sequencing in future research. The overall yield and value of NG sequencing is evident in the amount of sequence data obtained in each run. We identified a large number of uniquely tagged gene sequences in each of our three cDNA libraries (*Arabidopsis*, poppy, and *Persea*). With only a small amount of sequencing (one-half plate on the GS20) in *Arabidopsis*, we tagged more than 15,000 genes and completely or nearly completely sequenced several hundred of the highly expressed genes. Even with a very modest amount of data by NG sequencing standards, many of these sequences extended the annotated untranslated regions (UTRs) and redefined intron/exon boundaries, including evidence of alternative splicing. We also identified more than 2,000 transcripts that were not previously annotated in the *Arabidopsis* genome. These may define new genes or transcribed non-coding regions such as miRNA or other small RNA. In any event, these results illustrate the utility of NG transcriptome sequencing for genome annotation.

Although our data were limited to *Arabidopsis* inflorescence, we expect similar results from other organisms and tissue. To assess the similarity of the *Arabidopsis* inflorescence transcriptome with other transcriptomes, we considered the distribution of intensities of the perfect match probes from several Affymetrix experiments involving various tissues and

organisms. We examined *Arabidopsis* inflorescence, leaf, and root on ATH1 arrays (Zhang et al. 2005), human skeletal muscle (Haslett et al. 2002) on the hgu95av2 array, *Caenorhabditis elegans* (whole worm) (O'Rourke et al. 2006) on the Celegans array, *Drosophila melanogaster* (whole fly) (Wang et al. 2004), and Saccharomyces cerevisiae (yeast) on ammonium sulfite nitrogen source (Usaite et al. 2006). Small differences are observed in the expression profiles, consistent with some samples having different proportions of genes expressed more or less highly, but overall, the distribution of expression intensities is very similar for all of the samples (Fig 4-5A). These differences among samples are on the same scale, and sometimes smaller than, the variations seen among replicate samples from *Arabidopsis* inflorescence. Because the tissue-specific expression profile is the one method-independent input to the simulation model, we can expect similar predictions for transcriptomes from different sources.

NG sequencing simulation studies and comparisons of platforms

Simulations studies help researchers predict outcomes of expensive or time consuming experiments that cannot be readily performed in the near term. For transcriptome sequencing, simulation studies have allowed researchers to conduct in silico experiments of systems that would be costly and time-consuming to do in the lab. We have developed a simulation approach to understand the advantages of each of the NG technologies in comparison with traditional capillary sequencing. Although all technologies eventually converge at similar points with regard to unigene length, transcriptome coverage, and percentage of genes fully sequenced (1,444 bp, 18,283 genes, and 100% for *Arabidopsis*), the NG technologies offer huge advances, most notably in the amount of sequence generated at considerably lower costs. Even though small NG experiments will tag a very large fraction of the transcriptome, it will commonly be in the form of thousands of disconnected fragments of genes, with relatively few full-length cDNAs. Thus, NG

technologies are very effective for tagging sequences from fully sequenced species. However, researchers sequencing transcripts from novel species with few genomic resources or from species that are evolutionarily distant from a model system, might face several challenges when evaluating the data. The problems might become amplified with 25-30 bp reads generated by the Solexa system or other short-read platforms. The benefits of normalization are most evident in traditional sequencing, although some benefits, which include longer unigenes, are apparent in the NG technologies. However, the cost of normalization, and the potential for loss of closely related genes from the dataset, might outweigh the potential benefits.

Although NG sequencing does outperform traditional sequencing in many areas, the problems in assembly cannot be underestimated. Solexa and SOLiD sequences, currently less than 40 bp, will pose problems in assembly of unigenes, especially for short segments of genes that may be present in several genes. For example, only 81.7% of 15-mers are unique in the *Populus* transcriptome (PTR, Figure 4-5B). This leaves nearly 7 million 15-mers not unique to the *Populus* transcriptome, including 43,069 15-mers that are present in at least 10 different genes (results not shown). Until methods are developed to deal with this large fraction of sequence fragments that might lead to mis-assembled unigenes, researchers will not be able to use the Solexa or SOLiD technologies alone for transcriptome sequencing in non-model species. Research into genome assembly strategies with these short sequences is currently under investigation (Dohm et al. 2007; Jeck et al. 2007; Warren et al. 2007) and will hopefully become part of transcriptome assembly. The addition of Solexa sequences to longer NG sequences or traditional capillary sequences should help assemble larger unigenes. These longer sequences will have a much higher confidence in their uniqueness. The combination of technologies for transcriptome sequencing is analogous to genome shotgun sequencing which uses varying sizes of clones.

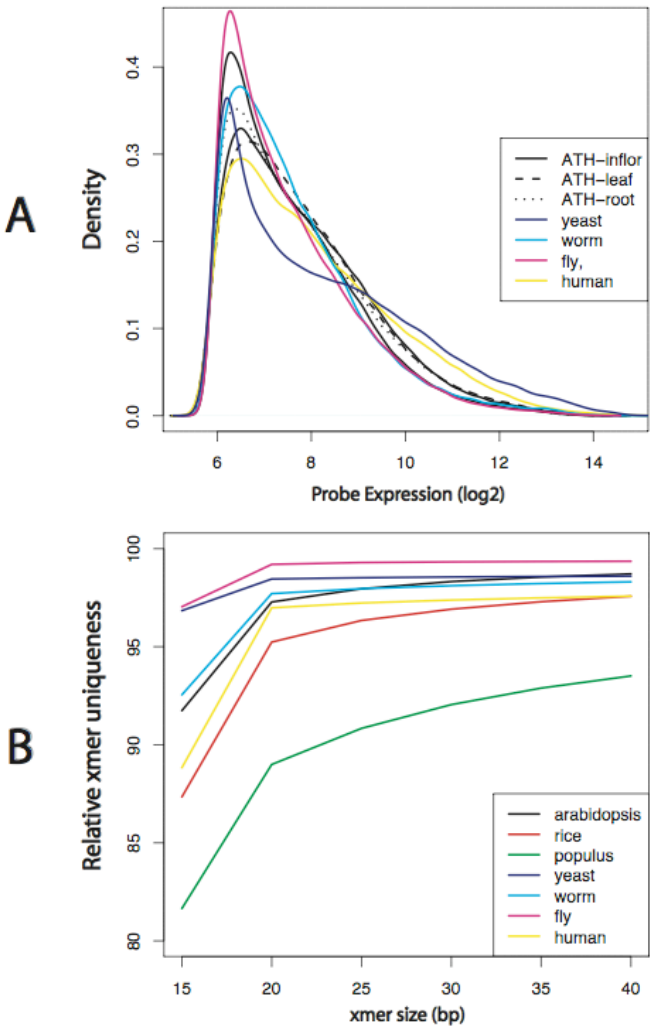

Figure **4-5**: Probe expression distributions and relative uniqueness of varying x-mer sizes.

A. Probe expression distribution for different tissues of *Arabidopsis* and various other species, as indicated. Data shown are the smoothed histograms of the log2 (perfect match expression) values for each sample taken from microarray datasets (see text). Datasets are adjusted to the same fifth percentile.

B. Relative uniqueness of varying x-mer sizes in full cDNA collections from several sequenced genomes. For each species, specified by a different colored line, we determined all DNA x-mers of various sizes (x = 15, 20, 25, 30, 35, or 40 bp). Relative x-mer uniqueness is number of x-mers unique to only one gene divided by the total number of x-mers in the transcriptome. The level of uniqueness increases with size of x-mer, and varies with the organism.

In order to evaluate the robustness of our simulator to both different organisms and tissues, we compared our simulated results against the poppy and avocado transcriptomes generated for this study and against two recently published 454 transcriptomes (Vera et al. 2007; Weber et al. 2007). Vera et al. (2007) performed de novo assembly of 454 transcriptome sequence in the butterfly *Melitaea cinxia*. The RNAs were isolated from a genetically diverse pool of larvae, pupae, and adults. The authors sequenced two plates of GS20, and after trimming and cleaning there were 518,079 reads (approximately 50 MB) with an average length of 110bp. The reads assembled into 108,297 unigenes that included 59,945 singleton reads (55.4% singletons). The mean unigene length of 149 bp is the summation of 197 bp for all contigs plus the 110 bp average for the singleton reads. From our simulation results, we would predict roughly 91,000 unigenes, 48% singletons, and an average unigene length of 177 bp. Weber et. al sequenced 2 GS20 plates of cDNA derived from above-ground tissues of 8-day old light-grown Arabidopsis seedlings. The reads tagged an estimated 17,499 cDNAs, which is nearly identical to what our simulations would predict for that amount of GS20 data. For poppy, each of the observed assembly characteristics (Table 4-1) are very close to the predicted values for this amount of GS20 sequence. For example, the average unigene numbers for each of the poppy libraries are 83,370 and 75,273, which are very close to the estimated 76,000 unigenes from the simulation. For avocado, the observed number of unigenes and percent singletons (234,000 and 90%) is considerably larger than predicted for a normalized library sequenced to this depth. We do not know if this unexpected large number of unigenes in avocado is due to a larger underlying transcriptome size, sequence error causing false misassembly (Wang et al. 2004), or some other unknown factor. For each of these comparisons, we applied the distributions of read lengths, sequence start sites, and transcript abundance frequencies previously observed from Arabidopsis. Therefore, although we have not actually fine-tuned these specific outcomes with the true (and unknown) transcript distribution profiles for each of species and tissue, the observed outcomes

are very close to the model predictions. This is particularly true when considering the above uncertainties associated with de novo assembly, differences in tissue sources, and technical variation that may be expected from run to run.

Analysis of gene expression by NG sequencing

NG sequencing is potentially a direct and cost-effective way to obtain genome scale expression information from organisms that lack a genome sequence and comprehensive microarray platform. Digital expression data obtained by direct sequencing is not dependent on gene models, comprehensive genome data, or understanding of alternative splice forms. The half plate of GS20 sequencing in *Arabidopsis* showed a moderate correlation ($r=0.67$, $r^2 = 0.444$, $p < 0.0001$) between the number of reads and microarray expression values generated from the same tissue and ecotype. Solexa and SOLiD have the potential to increase this correlation, since with millions of sequence reads per experiment, they will ultimately have a large dynamic range similar to traditional microarray experiments. Correct mapping to specific genes may be problematic for short Solexa or SOLiD reads (Figure 4-5B), when no genome sequence is available. Enriching for 3' UTRs, however, should improve assignment accuracy and increase efficiency of massively parallel sequencing for assessing gene expression levels (Eveland et al. 2007; Torres et al. 2007). In research on an organism without a sequenced genome, the value-added expression evidence could be very important in the early stages of developing its transcriptome, an advantage for GS20 or FLX NG technologies. Even in organisms that have comprehensive microarrays, the probe designs are usually dependent on and built with information from early genome assemblies. For example, the current Arabidopsis Affymetrix ATh1 array (Hennig et al. 2003; Redman et al. 2004) contains probes for approximately 22,000 genes, approximately 5000-6000 genes fewer than the current annotation.

NG sequencing can be scaled to suit different project goals

NG sequencing technologies are a highly flexible set of platforms that can be used alone or in combination to best suit the research at hand. Small-scale experiments (e.g., 1/16 – 1 plate) provide a wealth of information including the tagging of many or most expressed genes, microsatellite markers, full-length sequencing of highly expressed genes, and modest expression level information. Since there can be multiple lanes in NG plates (up to 16), and multiple bar-coded libraries can be sequenced on a single plate, researchers might fully sequence a small number of highly expressed genes with very little cost or time investment. Low-copy, highly expressed genes might be quite useful for phylogenetic analysis or markers for population level studies. Even small-scale experiments will tag a large fraction of genes in a transcriptome. These tags can be used for building microarray probes (Vera et al. 2007) and enhancing microarray design. Small studies might sequence the highly expressed genes from many different tissues in the same sequencing runs without the need for bar-coding. Small experiments might also be sufficient to provide rich information for genome annotations in pre-draft and early draft forms. For example, a single run plate will tag nearly every transcribed gene and help identify UTR and intron/exon boundaries. However, experiments on this scale sample only a small fraction of the actual transcriptome and the assembly is often in many small pieces. Moderate transcriptome studies (e.g., 2-5 plates) have the potential to sequence more than 50% of the transcriptome. They will provide small annotation datasets, identify new genes in an organism, further extend genic regions, and help with alternative splicing, especially in sequenced genomes. With deeper sequencing (e.g. 6-20 plates), researchers attain a level of transcriptome that has never been possible before due to the higher cost of earlier technologies. Not only will these studies sequence more than 90% of the transcriptome, the coverage per gene will approach traditional

sequencing. This should allow researchers to use these genes to identify pathways, determine tissue-specific expression for lowly expressed genes, and will be critical for genome annotation.

Conclusion

NG technologies are revolutionizing EST sequencing and applications that revolve around gene expression. Another important consideration in NG transcriptome sequencing is the efficiency and flexibility in library construction. NG library construction costs less, takes less time, and does not produce physical clones that must be stored. If the goal is sequencing full-length genes, non-normalized libraries will yield a larger number of full-length sequences in small sequencing experiments compared to normalized libraries. As sequencing quantities increase, this relationship reverses, and normalized libraries will capture more full-length cDNA sequences. There is also a trade-off in the cost of normalization versus the cost of sequencing. We have shown the feasibility of using a simulation approach to quantitatively evaluate the different platforms and the various combinations of platforms. Currently, the low per-base cost of Solexa sequencing suggests that it may be the most efficient method of transcriptome characterization for sequenced genomes (e.g. Figure 4-3), but in the absence of a reference genome, the problem of de novo assembly of the short Solexa reads has not yet been resolved. Under these circumstances, a blend of Solexa and GS-FLX sequencing may be optimal (Figure 4-4).

This is the first simulation study to address some of the technology-specific characteristics found in several NG sequencing technologies. Our approach focuses on the critical questions of data production and coverage, which differ dramatically between methods and experimental scales. By extrapolating the results of the GS20 simulations, we are able to predict outcomes with various NG methods and combinations. A next step will be to develop

realistic models of error in sequencing and assembly, and to provide tools to allow any sets of assumptions about read length and cost to be examined. Future studies should be able to build upon this first simulation study, while accounting for some or all of the additional and complex issues in transcriptome experiments.

Methods

RNA preparation

Arabidopsis thaliana (cv: Landsberg) plants used in this study were grown in a culture chamber at 23 temp and 40% humidity with 18 hours light / 6 hours dark. RNA isolation from *Arabidopsis* plants was performed with the RNA Aqueous-Midi kit (Ambion, inc; catalog: #1911) following the manufacturer's recommendations with modifications as previously reported (Carlson et al. 2006). California poppy total RNA was prepared from pre-meiotic flower buds using TRIzol reagent (Invitrogen) according to the manufacturer's recommendations. Stages of flower development were defined as described previously (Baker et al. 2005). RNA quality and quantity were checked using a Bioanalyzer (Agilent, inc). *Persea americana* pre-meiotic flower buds (Stages 6-7), in which all floral organs are present but stamens and carpels are immature (Buzgo et al. 2007), were collected from a tree cultivated on the Gainesville campus of the University of Florida (Kim 1135; voucher deposited at FLAS). Total RNA was isolated using a combination of the CTAB DNA extraction protocol (Doyle JJ and Doyle 1987) and the RNeasy Plant Mini kit (Qiagen) as previously described for basal angiosperms (Kim et al. 2004). RNA integrity was verified with a Bioanalyzer (Agilent Inc.).

mRNA purification and 454 library construction for *Arabidopsis thaliana* and *Eschscholzia californica*

Messenger RNA was extracted from total RNA using Poly(A)Purist™ mRNA Purification Kit (Ambion, Inc., catalog # 1916) according to the manufacturer's recommendation. mRNA quality was checked with a Bioanalyzer (Agilent, Inc). cDNAs were prepared using the ZAP-cDNA® Synthesis Kit (Stratagene) according to manufacturer's instructions, except that 2 micrograms of mRNA was used rather than the recommended 5 ug. For *Arabidopsis*, two cDNAs were prepared, the first by oligo-dT priming and the second using the random hexamer primers provided in the kit. For California poppy, only random hexamer priming was used to prepare ds cDNA. 454 libraries were constructed from the cDNAs and sequenced using the approach described by {Margulies, 2005 #11}. One half plate, one, and two plates were sequenced from Arabidopsis, Persea, and California poppy, respectively, using the 454-G20 sequencer according to manufacturers protocols (Roche, Inc).

Normalized cDNA library construction in *Persea americana*

Messenger RNA was isolated from 250 ug total RNA using the Poly(A)Purist™ mRNA Purification Kit (Ambion, Inc., catalog # 1916) according to the manufacturer's protocol. Approximately 1 ug high quality mRNA, verified through Bioanalyzer as above, was used to construct a normalized cDNA library using the Trimmer-Direct Kit (Evrogen), which combines a modification of SMART cDNA preparation (Zhu et al. 2001) with DSN-normalization technology (Zhulidov et al. 2004). Specifically, first strand cDNA was reverse transcribed using a 3' adapter (CDS-3M; Evrogen) that anneals to poly(A) RNA tails and a second adaptor, BD SMART™ Oligo IV (Clontech), that anneals to the 5’ dC tails created by MMLV reverse transcriptase, and serves as an extended template for the first strand synthesis. Double-stranded

cDNA was synthesized and simultaneously amplified with the BD SMART™ 5' PCR Primer (Clontech) that anneals to both adaptors through 13 PCR cycles of 95°C for 7 sec; 66°C for 30 sec; 72°C for 6 min on a BioRad thermocycler. Approximately 1.2 ug of double stranded cDNA was purified with the Wizard PCR Purification Kit (Promega) followed by ethanol precipitation, and normalized according to the Trimmer-Direct protocol. Normalized cDNA was subjected to two rounds of single primer PCR amplifications exploiting the complementarity of the cDNA ends (primer sites) to suppress short fragment amplification (Lukyanov et al. 1995) and enrich the cDNA pool with full length transcripts. The first PCR amplification was conducted for 18 cycles of 95°C for 7 sec; 65°C for 20 sec; 72°C for 6 min. First amplification products with efficient normalization were diluted 10-fold and subjected to 12 PCR cycles of 95°C for 7 sec; 64°C for 20 sec; 72°C for 4 min. Approximately 20 ug of normalized amplified cDNA were thus obtained.

Sequence analysis

454 reads for all species were assembled using the 454 Newbler Assembler {Margulies, 2005 #11}. Using the program seqclean (http://compbio.dfci.harvard.edu/tgi/software/), we vector trimmed and quality trimmed both the original read files and the contig files generated by the Newbler assembler. We parsed the 454ReadStatus.txt file to determine the singleton reads, which did not assemble with any other reads. For each library, we used then the contig and singleton files to generate a unigene file. Finally, we calculated the number of sequences, the mean length of all sequences, and the total MB for each of the 4 file types (read, contig, singleton, and unigene). We also performed an additional assembly step using cap3 with 95% identity and 30bp overlap (default for all other parameters).

Genome mappings for *Arabidopsis* were determined using the best BLASTn (Altschul et al. 1990) match of each individual read versus the TAIR7 genome annotation. We wrote Perl

scripts to parse the TAIR xml files containing chromosome locations for all genes in the current annotation. We assigned the following categories to each read: Exon, intron, intron/exon, extended UTR, overlapping genes, intergenic, or no hit. All reads mapping to known gene locations were also mapped to the TAIR cDNA dataset. We calculated the start and stop positions for all reads on each cDNA and determined the reads/gene distribution based on all *Arabidopsis* genes tagged.

In order to evaluate the sequences from the three libraries without a sequenced genome, we used BLASTx and BLASTn searches against the TAIR protein and cDNA datasets, respectively. We parsed the BLAST output to determine the location and e-value to the best Arabidopsis gene. For each best hit, we also determined the length of the protein and/or cDNA, as well as the annotation of the gene. We used the length of the read, the start and stop locations on the gene, and the length of the best hit to calculate each Arabidopsis gene's percent coverage. Gene coverage is defined as the number of bases or amino acids covered by at least 1 read divided by the length of the gene. We estimated the gene coverage using both relaxed and strict definitions. For the relaxed definition, we considered all reads or contigs mapped to a gene, thus allowing for a non-contiguous definition of the gene's coverage. In the strict case, we only considered the longest unigene or read mapped to an individual gene.

Simulation studies

Using a half plate of GS20 sequences, we developed an approach to simulate additional rounds of sequencing *in silico* as follows. We mapped 118,485 reads to 15,276 genes in the TAIR cDNA dataset using the similarity search program BLASTn. We selected the best high scoring pair (HSP) for each read and its corresponding cDNA. Since there are multiple versions of loci for 3,156 genes, we used the version for each gene with the largest number of reads

mapped to its sequence. Using this approach, we created the following distributions: Read length, read count, and cDNA start site distributions. The read length distribution is comprised of all read lengths sampled. The read count distribution contains the number of reads mapped to each gene. For example, the distribution contained 3149, 2238, and 1672 genes with 1, 2, and 3 reads/gene respectively. The most highly expressed gene (AT1G67090, see Table 4-3A) had 586 reads, or about 0.5% of total reads.

We then estimated 3,007 zero class genes by providing the reads/gene profile to the program ESTstat (Wang et al. 2004). Zero class genes are defined as genes that have not been sampled (sequenced) but are presumed to be present in the library, but typically expressed at low levels. Failure to account for these genes would bias many of the simulation estimates. We then randomly sampled the zero class genes from the remaining genes not originally sampled. The addition of the zero class genes to the 15,276 transcribed genes totaled 18,283 genes. For the simulation, we used the number of reads per gene distribution to randomly select reads from particular genes based on the number of times they were sequenced. For example, genes that were sequenced only once had a much lower chance of being selected for a plate of simulated reads than genes represented by hundreds or thousands of reads. Finally, we collected all the start sites for each read against its corresponding cDNA to generate a start site distribution. Based on previous work (Wang et al. 2004; Wang et al. 2005), we assumed that start sites are dependent on gene length. Based on the quartiles of the Arabidopsis cDNA lengths, we grouped the start sites into four groups: 1) 1-1000 bp, 2) 1001-1500 bp, 3) 1501-2000 bp, and 4) greater than 2000 bp. Using the gene length, we created relative start site distributions dependent on gene length.

The four libraries used in this study were all sequenced using a GS20 machine. In order to compare the GS20 technology to other NG technologies, we also simulated FLX, Solexa, and traditional capillary sequences. For FLX, we used all of the same distributions for the GS20 simulation, except the read length distribution. Since the average read length in our GS20 runs

was approximately 100 bp, and the FLX has been reported to generate 250 bp reads, we multiplied the read length of each randomly chosen read by 2.5 for the FLX simulations. In the Solexa simulations, we used a random start site distribution with a 25 bp read length average. We used the same gene expression distribution as the GS20 for all technologies.

To simulate traditional capillary sequences, we downloaded 48,130 ESTs from four different *Arabidopsis* libraries: flower buds, green siliques, roots, and above ground organs 2-6 week old. We partitioned the ESTs into two groups, 5' and 3' based on the annotations located in the fasta header. We then randomly selected 5,000 ESTs from both groups, mapped the transcripts to the *Arabidopsis* cDNA dataset using BLASTn, and generated start site distribution based on cDNA length. We used a 750 bp read length, except when the randomly chosen gene or the length of the gene minus the randomly chosen start site was less than 750 bp. In these cases, we used the gene length or the distance from the start site to the end of the gene as the read length.

In order to compare all technologies, we used current and conservative estimates of the amount (MB) and price ($) of sequencing with each technology. For GS20, an average plate costs $6000/plate and the plate generates 25 MB of data ($240/MB). Since NG machines can be partitioned into smaller segments, we simulated 1/16, 1/8, ¼, ½, 1, 2, 4, and 10 plates for all three technologies. For the GS20, this came to 1.56, 3.12, 6.25, 12.5, 25, 50, 100, 250, and 500 MB increments. For FLX, the cost was calculated using $9000/plate and 100 MB of data ($90/MB), with 6.25, 12.5, 25, 50, 100, 400, 1000, and 2000 MB increments. For Solexa, we used $4000/plate and 1 GB of data ($4/MB) with 50, 100, 200, 500, 1000, 2000, 4000, 10000, and 20000 MB increments. Finally, we converted the cost of traditional capillary sequencing, which is normally calculated per EST (read), by using the conventional $1/EST with 750 bp length ($1330/MB). This included 1.56, 3.12, 6.25, 12.5, 25, 50, 100, and 200 MB increments.

To examine the effects of normalization in next generation transcriptome sequencing, we simulated normalized sequencing for each of the above technologies. We assumed perfect normalization, and changed the gene expression distribution to be equal for all genes. Therefore, of the 18,261 genes estimated to be in the library, each gene has the exact same probability of being chosen as every other gene in the dataset. Although normalization can be technically difficult and requires more labor to generate, we excluded these costs and assumed that there were no additional library costs with normalization sequencing.

To compare the expected simulation results for each technology and combinations of technologies, we calculated the following parameters: percent transcriptome, the number of genes tagged, the total number of unigenes (contigs plus singletons), the mean length of the longest unigene per gene, and the number of genes covered by reads of at least 90% of the length of the gene and in only one unigene.

Acknowledgements

We thank the Huck Institutes of the Life Sciences and the Eberly College of Penn State for supporting the cost of 454 sequencing of *Arabidopsis* and poppy, Richard Meisel for valuable discussion of *Drosophila* transcriptomes, and Tony Omeis and the Biology Department Greenhouses (PSU) for cultivating the poppy plants. We would also like to thank Webb Miller, Jiping Wang, and Bruce Lindsay for ideas on transcript assembly and simulation strategies. This research was supported in part by NSF Plant Genome Awards DBI-0115684 (The Floral Genome Project) and DEB 0638595 (The Ancestral Angiosperm Genome Project).

References

Adams, M. D., J. M. Kelley, J. D. Gocayne, M. Dubnick, M. H. Polymeropoulos, H. Xiao, C. R. Merril, A. Wu, B. Olde, R. F. Moreno and et al. (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252(5013): 1651-6.

AGI (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408(6814): 796-815.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman (1990). Basic local alignment search tool. J Mol Biol 215(3): 403-10.

Bainbridge, M. N., R. L. Warren, M. Hirst, T. Romanuik, T. Zeng, A. Go, A. Delaney, M. Griffith, M. Hickenbotham, V. Magrini, E. R. Mardis, M. D. Sadar, A. S. Siddiqui, M. A. Marra and S. J. Jones (2006). Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. BMC Genomics 7: 246.

Baker, C. C., P. Sieber, F. Wellmer and E. M. Meyerowitz (2005). The early extra petals1 mutant uncovers a role for microRNA miR164c in regulating petal number in *Arabidopsis*. Curr Biol 15(4): 303-15.

Barakat, A., K. Wall, J. Leebens-Mack, Y. J. Wang, J. E. Carlson and C. W. Depamphilis (2007). Large-scale identification of microRNAs from a basal eudicot (*Eschscholzia californica*) and conservation in flowering plants. Plant J 51(6): 991-1003.

Barakat, A., P. K. Wall, S. Diloreto, C. W. Depamphilis and J. E. Carlson (2007). Conservation and divergence of microRNAs in Populus. BMC Genomics 8: 481.

Bentley, D. R. (2006). Whole-genome re-sequencing. Curr Opin Genet Dev 16(6): 545-52.

Bouck, A. and T. Vision (2007). The molecular ecologist's guide to expressed sequence tags. Mol Ecol 16(5): 907-24.

Brenner, S., M. Johnson, J. Bridgham, G. Golda, D. H. Lloyd, D. Johnson, S. Luo, S. McCurdy, M. Foy, M. Ewan, R. Roth, D. George, S. Eletr, G. Albrecht, E. Vermaas, S. R. Williams, K. Moon, T. Burcham, M. Pallas, R. B. DuBridge, J. Kirchner, K. Fearon, J. Mao and K. Corcoran (2000). Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat Biotechnol 18(6): 630-4.

Buzgo, M., A. S. Chanderbali, S. Kim, Z. Zheng, D. G. Oppenheimer, P. S. Soltis and D. E. Soltis (2007). Floral developmental morphology of *Persea americana* (Avocado, Lauraceae): The oddities of male organ identity. International Journal of Plant Sciences 168(3): 261-284.

Cai, Z., C. Penaflor, J. V. Kuehl, J. Leebens-Mack, J. E. Carlson, C. W. dePamphilis, J. L. Boore and R. K. Jansen (2006). Complete plastid genome sequences of Drimys, Liriodendron, and Piper: implications for the phylogenetic relationships of magnoliids. BMC Evol Biol 6: 77.

Carlson, J. E., J. H. Leebens-Mack, P. K. Wall, L. M. Zahn, L. A. Mueller, L. L. Landherr, Y. Hu, D. C. Ilut, J. M. Arrington, S. Choirean, A. Becker, D. Field, S. D. Tanksley, H. Ma and C. W. dePamphilis (2006). EST database for early flower development in California poppy (*Eschscholzia californica* Cham., Papaveraceae) tags over 6,000 genes from a basal eudicot. Plant Mol Biol 62(3): 351-69.

Chanderbali, A. S., V. Albert, V. Ashworth, M. T. Clegg, L. R. E., D. E. Soltis and P. S. Soltis (2007). *Persea americana* (avocado): bringing ancient flowers to fruit in the genomics era. Bioessays: in press.

Cheung, F., B. J. Haas, S. M. Goldberg, G. D. May, Y. Xiao and C. D. Town (2006). Sequencing *Medicago truncatula* expressed sequenced tags using 454 Life Sciences technology. BMC Genomics 7: 272.

Dohm, J. C., C. Lottaz, T. Borodina and H. Himmelbauer (2007). SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing. Genome Res 17(11): 1697-706.

Doyle JJ and J. Doyle (1987). A rapid DNA isolation from small amount of fresh leaf tissue. Phytochemical Bulletin 19: 11-15.

Eveland, A. L., D. R. McCarty and K. E. Koch (2007). Transcript Profiling by 3'UTR Sequencing Resolves Expression of Gene Families. Plant Physiol.

Goldberg, S. M., J. Johnson, D. Busam, T. Feldblyum, S. Ferriera, R. Friedman, A. Halpern, H. Khouri, S. A. Kravitz, F. M. Lauro, K. Li, Y. H. Rogers, R. Strausberg, G. Sutton, L. Tallon, T. Thomas, E. Venter, M. Frazier and J. C. Venter (2006). A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. Proc Natl Acad Sci U S A 103(30): 11240-5.

Haslett, J. N., D. Sanoudou, A. T. Kho, R. R. Bennett, S. A. Greenberg, I. S. Kohane, A. H. Beggs and L. M. Kunkel (2002). Gene expression comparison of biopsies from Duchenne muscular dystrophy (DMD) and normal skeletal muscle. Proc Natl Acad Sci U S A 99(23): 15000-5.

Hennig, L., M. Menges, J. A. Murray and W. Gruissem (2003). *Arabidopsis* transcript profiling on Affymetrix GeneChip arrays. Plant Mol Biol 53(4): 457-65.

Jeck, W. R., J. A. Reinhardt, D. A. Baltrus, M. T. Hickenbotham, V. Magrini, E. R. Mardis, J. L. Dangl and C. D. Jones (2007). Extending assembly of short DNA sequences to handle error. Bioinformatics 23(21): 2942-4.

Kim, S., M.-J. Yoo, V. A. Albert, J. S. Farris, P. S. Soltis and D. E. Soltis (2004). Phylogeny and diversification of B-function MADS-box genes in angiosperms: Evolutionary and functional implications of a 260-million-year-old duplication. American Journal of Botany 91(12): 2102-2118.

Kulesh, D. A., D. R. Clive, D. S. Zarlenga and J. J. Greene (1987). Identification of interferon-modulated proliferation-related cDNA sequences. Proc Natl Acad Sci U S A 84(23): 8453-7.

Lu, C., K. Kulkarni, F. F. Souret, R. MuthuValliappan, S. S. Tej, R. S. Poethig, I. R. Henderson, S. E. Jacobsen, W. Wang, P. J. Green and B. C. Meyers (2006). MicroRNAs and other small RNAs enriched in the *Arabidopsis* RNA-dependent RNA polymerase-2 mutant. Genome Res 16(10): 1276-88.

Lu, C., B. C. Meyers and P. J. Green (2007). Construction of small RNA cDNA libraries for deep sequencing. Methods 43(2): 110-7.

Lukyanov, K. A., G. A. Launer, V. S. Tarabykin, A. G. Zaraisky and S. A. Lukyanov (1995). Inverted terminal repeats permit the average length of amplified DNA fragments to be regulated during preparation of cDNA libraries by polymerase chain reaction. Anal Biochem 229(2): 198-202.

Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y. J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. Alenquer, T. P. Jarvie, K. B. Jirage, J. B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley and J. M.

Rothberg (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature 437(7057): 376-80.

Moore, M. J., A. Dhingra, P. S. Soltis, R. Shaw, W. G. Farmerie, K. M. Folta and D. E. Soltis (2006). Rapid and accurate pyrosequencing of angiosperm plastid genomes. BMC Plant Biol 6: 17.

O'Rourke, D., D. Baban, M. Demidova, R. Mott and J. Hodgkin (2006). Genomic clusters, putative pathogen recognition molecules, and antimicrobial genes are induced by infection of *C. elegans* with *M. nematophilum*. Genome Res 16(8): 1005-16.

Poinar, H. N., C. Schwarz, J. Qi, B. Shapiro, R. D. Macphee, B. Buigues, A. Tikhonov, D. H. Huson, L. P. Tomsho, A. Auch, M. Rampp, W. Miller and S. C. Schuster (2006). Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. Science 311(5759): 392-4.

Porreca, G. J., K. Zhang, J. B. Li, B. Xie, D. Austin, S. L. Vassallo, E. M. LeProust, B. J. Peck, C. J. Emig, F. Dahl, Y. Gao, G. M. Church and J. Shendure (2007). Multiplex amplification of large sets of human exons. Nat Methods 4(11): 931-6.

Redman, J. C., B. J. Haas, G. Tanimoto and C. D. Town (2004). Development and evaluation of an Arabidopsis whole genome Affymetrix probe array. Plant J 38(3): 545-61.

Sanger, F., S. Nicklen and A. R. Coulson (1977). DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74(12): 5463-7.

Schena, M., D. Shalon, R. W. Davis and P. O. Brown (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270(5235): 467-70.

Shendure, J., G. J. Porreca, N. B. Reppas, X. Lin, J. P. McCutcheon, A. M. Rosenbaum, M. D. Wang, K. Zhang, R. D. Mitra and G. M. Church (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309(5741): 1728-32.

Torres, T. T., M. Metta, B. Ottenwalder and C. Schlotterer (2007). Gene expression profiling by massively parallel sequencing. Genome Res.

Usaite, R., K. R. Patil, T. Grotkjaer, J. Nielsen and B. Regenberg (2006). Global transcriptional and physiological responses of *Saccharomyces cerevisiae* to ammonium, L-alanine, or L-glutamine limitation. Appl Environ Microbiol 72(9): 6194-203.

Velculescu, V. E., L. Zhang, B. Vogelstein and K. W. Kinzler (1995). Serial analysis of gene expression. Science 270(5235): 484-7.

Vera, J. C., C. W. Wheat, H. W. Fescemyer, M. J. Frilander, D. L. Crawford, I. Hanski and J. H. Marden (2007). Rapid transcriptome characterization for a non-model organism using *454* pyrosequencing. Molecular Ecology(in press).

Wang, J., L. Kean, J. Yang, A. K. Allan, S. A. Davies, P. Herzyk and J. A. Dow (2004). Function-informed transcriptome analysis of Drosophila renal tubule. Genome Biol 5(9): R69.

Wang, J. P., B. G. Lindsay, L. Cui, P. K. Wall, J. Marion, J. Zhang and C. W. dePamphilis (2005). Gene capture prediction and overlap estimation in EST sequencing from one or multiple libraries. BMC Bioinformatics 6: 300.

Wang, J. P., B. G. Lindsay, J. Leebens-Mack, L. Cui, K. Wall, W. C. Miller and C. W. dePamphilis (2004). EST clustering error evaluation and correction. Bioinformatics 20(17): 2973-84.

Warren, R. L., G. G. Sutton, S. J. Jones and R. A. Holt (2007). Assembling millions of short DNA sequences using SSAKE. Bioinformatics 23(4): 500-1.

Weber, A. P., K. L. Weber, K. Carr, C. Wilkerson and J. B. Ohlrogge (2007). Sampling the *Arabidopsis* transcriptome with massively parallel pyrosequencing. Plant Physiol 144(1): 32-42.

Wege, S., A. Scholz, S. Gleissberg and A. Becker (2007). Highly efficient virus-induced gene silencing (VIGS) in California poppy (*Eschscholzia californica*): an evaluation of VIGS as a strategy to obtain functional data from non-model plants. Ann Bot (Lond) 100(3): 641-9.

Zhang, X., B. Feng, Q. Zhang, D. Zhang, N. Altman and H. Ma (2005). Genome-wide expression profiling and identification of gene activities during early flower development in *Arabidopsis*. Plant Mol Biol 58(3): 401-19.

Zhu, Y. Y., E. M. Machleder, A. Chenchik, R. Li and P. D. Siebert (2001). Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. Biotechniques 30(4): 892-7.

Zhulidov, P. A., E. A. Bogdanova, A. S. Shcheglov, L. L. Vagner, G. L. Khaspekov, V. B. Kozhemyako, M. V. Matz, E. Meleshkevitch, L. L. Moroz, S. A. Lukyanov and D. A. Shagin (2004). Simple cDNA normalization using kamchatka crab duplex-specific nuclease. Nucleic Acids Res 32(3): e37.

Chapter 5

Conclusions and Future Directions

The research in this thesis details several aspects of comparative genomics in plants. The PlantTribes database, a comprehensive plant gene family database, has been designed for expansion of genomes and the addition of new features. As new genomes are sequenced or large EST sets produced, PlantTribes will be continuously expanded to include these data. However, many novel features should be developed that can be implemented in future releases of the database. For example, there is a need for a tool that can rapidly incorporate new query sequences into tribe alignments and phylogenies. I have also been involved with several small RNA projects, most notably the sequencing and analysis of microRNAs (miRNAs) from several plant lineages (Barakat et al. 2007; Barakat et al. 2007). miRNAs are small RNAs (sRNA) ~21 nucleotides in length that negatively control gene expression by cleaving or inhibiting the translation of target gene transcripts. Therefore, tools are needed to connect the rapidly expanding microRNA database into PlantTribes. Researchers would easily be able to identify the putative target regions from several plant lineages. Although the PlantTribes database contains an extensive set of microarray experiments from *Arabidopsis* (Craigon et al. 2004), there is a need to incorporate microarray data from large-scale array experiments that facilitate cross species expression analyses. Finally, because of the extensive polyploidy history of plants, there is a need for synteny-based tools to map genome duplications onto gene family phylogenies.

As more fully sequenced plant genomes become available over the next few years, there is a need for more comprehensive global gene family analyses. Although there are more than ten sequenced plant genomes to date, many interesting biological questions remain. For example, one such question concerns the factors that determine whether a plant lineage will evolve into a

tree? A tree is defined as a woody perennial, that is six meters or more in height at maturity, and that usually contains a single, unbranched trunk for several meters above ground, with a definite crown (Harlow et al. 1979). The plant scientific community has sequenced three woody plant genomes over the last few years, *Populus trichocarpa* (Poplar), *Vitis vinifera* (grape, a vine), and *Carica papaya* (papaya). These woody species can now be compared to the herbaceous species of *Arabidopsis thaliana*, *Orysa sativa* (rice), and *Medicago trunculata*, using *Physcomitrella patens* and *Selaginella moellendorffii* as outgroups. Some of the same strategies that I used in Chapter 3 to compare the genomes of *Arabidopsis*, rice, and *Populus*, could be implemented into a larger global gene family analysis. One of the main objectives would be to determine which gene families have expanded or contracted in the tree species compared to the non-tree species.

Finally, as sequencing technologies continue their rapid advancements, there will be a need for a more comprehensive simulation study on the use of Next Generation (NG) technologies for transcriptome sequencing. A complete solution to this problem would involve realistic models for each technology. These models would include the cost of library generation and data collection, the characteristics of cDNA libraries, transcript abundance distributions, read length distributions, and the error rates in sequence generation and assembly. The present study (Chapter 5) focuses on the first four of these issues to provide estimates of theoretical coverage of complex transcriptomes with varying scales and types of DNA sequencing experiments. I developed a robust simulation approach to model transcriptome sequencing, which incorporates distributions of the relative start site of cDNA sequences as a function of cDNA length, the read length distribution, and the transcript abundance distribution. However, a more complex model would include method-dependent sequencing and assembly errors that normally occur with all sequencing technologies. Until there is sufficient experimental data in the public domain for all of the NG technologies, as well as technologies not included in the current study (e.g., ABI

SOLiD), the results from this study should still help researchers working with these new and exciting technologies.

References

Barakat, A., K. Wall, J. Leebens-Mack, Y. J. Wang, J. E. Carlson and C. W. Depamphilis (2007). Large-scale identification of microRNAs from a basal eudicot (*Eschscholzia californica*) and conservation in flowering plants. Plant Journal 51(6): 991-1003.

Barakat, A., P. K. Wall, S. Diloreto, C. W. Depamphilis and J. E. Carlson (2007). Conservation and divergence of microRNAs in *Populus*. BMC Genomics 8: 481.

Craigon, D. J., N. James, J. Okyere, J. Higgins, J. Jotham and S. May (2004). NASCArrays: a repository for microarray data generated by NASC's transcriptomics service. Nucleic Acids Res 32(Database issue): D575-7.

Harlow, W., E. Harrar and F. White (1979). Textbook of Dendrology. New York, McGraw-Hill.

VITA

P. Kerr Wall

Education

1993–1999	B.S., Chemical Engineering and B.S., Zoology Louisiana State University, Baton Rouge, LA
2004–2008	PhD, Biology, Pennsylvania State University, University Park, PA

Professional Experience

1999–2001	Software Developer, VedaLabs, Inc., Baton Rouge, LA Designed, developed, and deployed significant components in a B2C digital media distribution system utilizing Java, SQL, and XML-messaging technologies
2001–present	Research Support Associate, Pennsylvania State University, University Park, PA Lead computational biologist on several NSF-funded Plant Genome Projects, including the Floral Genome and the Ancestral Angiosperm Genome Projects

Selected Publications

Wall PK, Leebens-Mack JH, Muller K, dePamphilis CW (2008). PlantTribes: A gene family database for comparative genomics in plants. Nucleic Acids Research (NAR) Database Issue.

Barakat A, **Wall PK**, Diloreto S, dePamphilis CW, Carlson JE (2007). Conservation and divergence of microRNAs in Populus. BMC Genomics. 8(1):481.

Barakat A, **Wall PK**, Leebens-Mack J, Carlson J, dePamphilis CW (2007). Large-scale identification of microRNAs from a basal eudicot (*Eschscholzia californica*) and conservation in flowering plants. The Plant Journal. 51(6):991.

Cui L, **Wall PK**, Leebens-Mack JH, Lindsay BG, Soltis DE, Doyle JJ, Soltis PS, Carlson JE, Arumuganathan K, Barakat A, Albert V, Ma H, dePamphilis CW (2006). Widespread genome duplications throughout the history of flowering plants. Genome Research. 16(6):738.

Duarte JM, Cui L, **Wall PK,** Zhang Q, Zhang X, Leebens-Mack J, Ma H, Altman N, and dePamphilis CW (2006). Expression Pattern Shifts Following Duplication Indicative of Subfunctionalization and Neofunctionalization in Regulatory Genes of Arabidopsis. Molecular Biology and Evolution. 23(2):469.

Kim S, Soltis PS, **Wall PK,** Soltis DE (2006). Phylogeny and domain evolution in the APETALA2-like gene family. Molecular Biology and Evolution. 23(1):107.

Molbak L, Tett A, Ussery DW, **Wall PK,** Turner S, Bailey M, Field D (2003). The plasmid genome database. Microbiology. 149(11):3043.