January 2011

Computational Methods for Accelerated Discovery and Characterization of Genes in Emerging Model Organisms

Alan Kwan
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS
School of Engineering and Applied Science
Department of Computer Science and Engineering

Dissertation Examination Committee:
Gary D. Stormo, Chair
Susan K. Dutcher, Co-advisor
Jeremy Buhler
Patrick Crowley
William Smart
Ting Wang

Computational Methods for Accelerated Discovery and Characterization of Genes in Emerging Model Organisms

by

Alan Lechuen Kwan

A dissertation presented to the Graduate School of Arts & Sciences
of Washington University in partial fulfillment of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2011
Saint Louis, Missouri

ABSTRACT OF THE DISSERTATION

Computational Methods for Accelerated Discovery and Characterization of Genes in Emerging Model Organisms

by

Alan Lechuen Kwan

Doctor of Philosophy in Computer Science

Washington University in St. Louis, 2011

Research Advisors: Dr. Gary D. Stormo and Dr. Susan K. Dutcher

Cilia are evolutionarily conserved, complex, microtubule-based structures that protrude from many eukaryotic cells. In humans, cilia can be found on almost all cell types. The effect of abnormal or absent cilia has been established as the common underlying cause of a recently emerging class of genetic diseases collectively referred to as ciliopathies. The function and structure of cilia are conserved across all organisms with cilia. One of the most influential model systems used to study ciliopathies has been the ciliated green alga *Chlamydomonas* reinhardtii, an organism for which there is a sequenced genome with relatively few experimentally validated whole-gene annotations but in which the ciliogenesis process can be reliably induced. Experimental methods have been successful in identifying a handful of highly specific cilia disease genes in the alga, but high-throughput, automated computational analyses harbor the greatest potential to reveal a more comprehensive ciliopathy disease gene list. However,

in order for a genome to be informative for downstream computational analyses, it must first be accurately annotated.

This dissertation focuses on accelerating the accurate annotation of the *Chlamydomonas* genome using whole-genome and whole-transcriptome methodologies to identify human ciliopathy genes. Towards this end, we first develop a genefinder training method for *Chlamydomonas* that does not require whole gene annotations and demonstrate that this traning method results in a more accurate genefinder than any other genefinder for this alga. Next, we develop a new automated protein characterization method that facilitates the transfer of information across different protein families by extending simple homology categorization to identify new cilia gene candidates. Finally we perform and analyze high-throughput whole-transcriptome sequencing of *Chlamydomonas* at various timepoints during ciliogenesis to identify ~300 novel human ciliopathy gene candidates. Together these three methodologies complement each other and the existing literature to better elucidate a more complete and informative cilia gene catalog.

Acknowledgments

First and foremost, my most humble and numerous thanks are owed to Z Liu for all her unwavering support and belief in me, for tolerating the sacrifices that are necessary for a work of this nature; I am indebted to you for your love, patience and understanding. Many thanks are owed to my family, especially my parents and my sister upon whom I have rested my tired spirit countless times. To Ricky and Emma, thank you for your unquestioning love and loyal support.

To my Advisors: Thank you for the uniquely complementary education that you have provided me during my time here, for showing me how to approach a problem, the nuances of testability and validation, how to ask the right questions the right way, and above all else how to keep a line of investigation firmly grounded on fundamentals. To Susan, thank you for the opportunity you extended to me and the patience you showed while my undisciplined mind matured under your guidance. Thank you for holding me to a standard of scientific responsibility that I have been told repeatedly is very rare in the world today. To Gary, you have given me the quantitative and analytical tools that largely define who I have become. Thank you, humbly, both.

Many thanks are owed to Aaron Spivak for his kind, loyal and faultless friendship, for his insight and encouragement, for the beers and scotch and introducing me to Wilkes; you are a true **bro**ther to me. Thanks are owed to Drs. Huawen Lin and Alison Albee for their patience in showing and discussing with me further about proper experimental design and execution. Also thank you to the rest of the Stormo and Dutcher Labs for our, at times, thought-provoking conversations. My time here has shaped me and accelerated my *own* maturation and intellectual development and I thank you all for being integral parts of such a defining part of my life.

Alan Lechuen Kwan

Washington University in St. Louis
May 2011

Dedicated to my parents
Joseph L. and Jannie S. Kwan

Contents

Abstract ... ii

Acknowledgments ... iv

List of Tables ... viii

List of Figures ... ix

1 Introduction ... 1
- 1.1 Motivation for this research ... 1
- 1.2 Specific problems addressed ... 3
  - 1.2.1 Inadequate numbers of verified genes for genefinder training ... 3
  - 1.2.2 Limits of sharing protein information between similar proteins ... 5
  - 1.2.3 Repurposing high-throughput sequencing for gene annotation ... 6
- 1.3 Dissertation layout ... 7

2 Background and Significance ... 8
- 2.1 Molecular biology primer ... 8
- 2.2 Basic gene selection theory ... 11
- 2.3 Sequence similarity and protein function ... 12
- 2.4 The study of human ciliopathies in *Chlamydomonas* ... 13
- 2.5 Accelerating automated annotation of emerging model organisms ... 16
  - 2.5.1 An effective genefinder training for emerging model organisms ... 18
  - 2.5.2 Detecting coevolution for protein annotation ... 21
  - 2.5.3 High-throughput sequencing identifies new ciliopathy genes ... 24

3 An effective genefinder training method for emerging model organisms ... 26
- 3.1 Introduction ... 26
- 3.2 Results ... 30
  - 3.2.1 Constructing and evaluating a training-set from ESTs ... 30
  - 3.2.2 GreenGenie2 is more accurate than GeneMark.hmm-ES 3.0 ... 33
  - 3.2.3 GreenGenie2 models in *v3* complement the FGC ... 34
  - 3.2.4 GreenGenie2 is a robust, effective genefinder ... 42
- 3.3 Discussion ... 43
- 3.4 Summary ... 47
- 3.5 Methods ... 48
  - 3.5.1 Sequence datasets ... 48
  - 3.5.2 *Chlamydomonas* gene catalogs ... 49
  - 3.5.3 Programs ... 50
  - 3.5.4 Short-sequence prediction performance evaluation ... 51
  - 3.5.5 Interval overlap analysis ... 53
  - 3.5.6 PCR and RT-PCR ... 54
- 3.6 Supplemental Tables ... 55
  - 3.6.1 NCBI protein ID accessions for *gb140* dataset ... 55
  - 3.6.2 Primers to verify PASA assemblies ... 57
  - 3.6.3 Primers to verify partial overlap exons ... 58
  - 3.6.4 Primers to verify novel *gg2v3* exons ... 59
  - 3.6.5 Primers to verify novel *gg2v3* genes ... 60
  - 3.6.6 Primers to verify *FGC07* exclusive genes ... 60

4 Detecting co-evolution for protein annotation ... 61
- 4.1 Introduction ... 61
- 4.2 The Approach ... 65

4.3 Results and Analysis 75
4.4 Summary 88

5 Transcriptome sequencing reveals early ciliogenesis gene program 90
5.1 Introduction 90
5.2 Results 93
5.2.1 RNAseq generates reliable ciliogenesis sequence data 93
5.2.2 Timeseries analysis reveals early ciliogenesis regulation programs 96
5.2.3 RNAseq supports previous comparative genomics predictions 99
5.2.4 RNAseq identifies 211 novel ciliopathy gene candidates 107
5.3 Discussion 126
5.4 Summary 130
5.5 Methods 131
5.5.1 RNAseq of *Chlamydomonas* transcriptome during ciliogenesis 131
5.5.2 Expression profile clustering 132

6 Closing remarks and future directions 134

Appendix A Colorfy: heatmap visualization for multisequence alignments 137

Appendix B Table of 1400 genes up-regulated in ciliogenesis 139

References 175

Vita 181

List of Tables

Table 3.1: Categorization of 2384 PASA EST assembly gene models 31

Table 3.2: Analysis of PASA gene models: RT-PCR 32

Table 3.3: Comparing GreenGenie2 and GeneMark.hmm-ES 3.0 34

Table 3.4: Comparison of *gg2v3* and *FGC* catalog by interval overlap 37

Table 3.5: Validation of competing *gg2v3* and *FGC* gene models: RT-PCR 38

Table 3.6: Validation of mutually exclusive gene models in *gg2v3* and *FGC* 41

Table 4.1: Annotation of 82 *Chlamydomonas* cilia motility genes 79

Table 4.2: 55 human multicellularity genes predicted by APACE 82

Table 4.3: Lifecycle phenotypes of 28 disrupted *P. falciparum* genes 84

Table 4.4: Comparing APACE and Barker on co-crystallization data 87

Table 5.1: Peak timepoint of 1400 up-regulated genes 95

Table 5.2: Distribution of 1400 up-regulated genes in 16 patterns 96

Table 5.3: Annotation of 78 Human homologs of cilia gene candidates in *Chlamydomonas* previously identified by comparative genomics sorted by expression pattern 102

Table 5.4: Annotation of 144 of 188 Human homologs of novel *Chlamydomonas* cilia gene candidates that can be assigned an annotation group sorted by expression pattern 112

Table 5.5: Annotation of 44 of 188 Human homologs of novel *Chlamydomonas* cilia gene candidates involved in other processes or diseases sorted by expression pattern 123

List of Figures

Figure 2.1: Protein biosynthesis 10

Figure 3.1: Four classes of gene level interval overlaps 36

Figure 3.2: Histogram of partial exon overlap *gg2v3* models to *FGC07* 39

Figure 4.1: Alignment scores need to be normalized 64

Figure 4.2: Different evolutions of orthologous and paralogous genes 66

Figure 4.3: Protein evolution in sequence and function space 67

Figure 4.4: Three classes of similarity score distributions 71

Figure 4.5: A flow diagram of APACE 73

Figure 4.6: The phylogenetic tree determined by APACE 74

Figure 5.1: Distribution of maximum fold-change values 94

Figure 5.2: Sixteen principal expression profiles shown in pattern groups 97

Figure 5.3: Annotation group distribution of 188 Human cilia genes 110

Chapter 1

Introduction

1.1 Motivation for this research

Computational methods have been essential to extracting information from DNA sequences since the invention of sequencing methods in 1977. Emerging and future sequencing technologies are making the digitization of new genomes ever more accessible to researchers and clinicians (PETTERSSON *et al.* 2009). A new genome must first be accurately annotated before it is informative in downstream computational analyses. Current and emerging sequencing technologies output sequence data at an economy and volume that far outpaces the rate at which reliable annotation of sequence data can take place. If the growing disparity between the rates of sequencing and annotation is left unaddressed, then the output of ever more efficient and precise sequencing technologies will result only in a vast, accurate collection of unusable genome sequence data. The annotation of a genome spans the central dogma of molecular biology beginning at the identification of genes embedded in the genomic DNA, to individual protein characterization, through to how proteins function together

in protein interaction networks. In addition, recent innovations that capitalize on high-throughput sequencing technology have made it possible to capture the sequences of all the genes being used by an organism under a condition of interest. High-throughput, whole-transcriptome sequencing methods are potent tools for genome annotation because they can provide direct evidence of genes that are used by the organism that can, in turn, be used to better inform automated gene identification methods.

The main objective of this dissertation is to accelerate the accurate annotation of novel and existing sequence data. In particular, this work focuses on accelerating the accurate annotation of the *Chlamydomonas* genome to identify human genes underlying a recently emerging class of genetic diseases referred to collectively as ciliopathies. The etiology of ciliopathies has been attributed to the dysfunction, malformation or absence of cilia, which are complex organelles protruding from virtually all cell types in the human body (FLIEGAUF *et al.* 2007; TOBIN and BEALES 2009). Cilia are evolutionarily conserved, complex, microtubule-based structures that protrude from many eukaryotic cells. The function and structure of cilia are conserved across all organisms with cilia. One of the most influential model systems used to study ciliopathies has been the ciliated green alga *Chlamydomonas reinhardtii*, an organism for which there is a sequenced genome with relatively few experimentally validated whole-gene annotations, but in which the ciliogenesis process can be reliably induced. Experimental methods have been successful in identifying a handful of highly specific cilia disease genes in the alga, but high-

throughput, automated computational analyses harbor the greatest potential to reveal a more comprehensive ciliopathy disease gene list.

In this dissertation, we first develop a genefinder training method for Chlamydomonas that does not require whole gene annotations and demonstrate that this training method results in a more accurate genefinder than any other genefinder for this alga. Next, we develop a new automated protein characterization method that facilitates the transfer of information across different protein families by extending simple homology categorization to identify new cilia gene candidates. Finally we perform and analyze high-throughput whole-transcriptome sequencing of *Chlamydomonas* at various timepoints during ciliogenesis to identify ~300 novel human ciliopathy gene candidates. Together these three methodologies complement each other and the existing literature to better elucidate a more complete and informative cilia gene catalog.

1.2 Specific problems addressed

1.2.1 Inadequate numbers of verified genes for effective genefinder training

Annotation of a genomic sequence begins with determining the most complete set of accurate gene models. The completeness and accuracy of computational methods developed for genefinding rely on a comprehensive model of gene structure in the genomic sequence and the effective determination

of parameter values for that model based on a representative training set of known gene annotations. The variation of internal gene structure models has decreased dramatically as biological understanding of coding sequence architecture has matured. However, the existence of an adequate, representative set of experimentally based gene annotations on which to train parameter values is limited to a very small set of widely studied organisms. The ability and relevance of existing and future genefinding methods to accurately predict coding genes in less well annotated genomes will be severely limited without a reliable training set data source. One abundant source of experimental coding gene sequence data that exists in some abundance for almost all sequenced organisms is in the form of Expressed Sequence Tags (ESTs). While multiple ESTs represent a complete coding gene, individual ESTs cannot be directly used as training data for *de novo* genefinders. Thus, this part of the research focuses on:

i. The development of a novel genefinder training protocol using gene fragments exclusively.
ii. The application of this novel approach to train an existing genefinding method onto the *Chlamydomonas reinhardtii* genome.
iii. The performance evaluation of this training method by comparing the predictive accuracy of the newly trained genefinder compared to the traditional training method using available *Chlamydomonas* gene annotations.
iv. Experimental validation of novel predictions that result from the novel training method.

1.2.2 The limits of sharing protein information within a protein family

The automated characterization of novel genes and proteins has relied almost exclusively on sequence similarity, or "homology". The basis of such methods is that similar sequences will fold into similar functional conformations. Consequently, proteins can be grouped into families of similar sequences and knowledge about one member is extended throughout the family. The extent of characterization in this manner is thus dependent on the existence of knowledge about at least one protein in every protein family. Consequently, a large proportion of protein families remain uncharacterized beyond sequence similarity. An alternative approach organizes proteins by a phylogenetic profile comparison (PPC), or pattern of conservation. Due to their relative abundance, accessibility and size, bacterial genomes have been the main focus for most existing phylogenetic profile comparison methods, while the development of PPC methods for eukaryotic species remains largely unexplored. Hence, this part of the research addresses:

i. Development of a PPC method for eukaryotes scalable to the number of predicted eukaryotic proteomes that are and will become available.
ii. Incorporation of a weighting scheme that compensates for phylogenetic bias that is internally consistent with the sequence data.
iii. Validation of predicted characterizations in existing literature and resources.

1.2.3 Repurposing high-throughput sequencing for gene annotation of emerging organisms

Previous work has leveraged the fact that transcript abundance of many genes encoding known cilia components are greatly amplified in *Chlamydomonas* during ciliogenesis (LEFEBVRE and ROSENBAUM 1986). *Chlamydomonas* is a unicellular, green alga with genetics similar to yeast but for two cilia that are practically identical to cilia found in humans. *Chlamydomonas* is an ideal model organism for transcript abundance based cilia gene detection because ciliogenesis can be induced by pH-shock. When environmental pH is precipitously dropped, *Chlamydomonas* cells shed their cilia and ciliogenesis begins immediately once environmental pH is restored. The specific transcriptional induction of genes encoding many known cilia components during ciliogenesis have been widely reported and further underscore the efficacy and potential advantages of using *Chlamydomonas* as a model organism to study cilia and ciliogenesis. Predicted cilia genes are often validated by testing for up-regulation by quantitative expression assays 30 minutes into ciliogenesis *Chlamydomonas* (LI *et al.* 2004; PAZOUR *et al.* 2005; STOLC *et al.* 2005). While there is evidence that many genes involved in cilia do show some up-regulation, it is unclear how many false negatives result from expression testing at this timepoint alone. Moreover, the regulation program of cilia genes during ciliogenesis is not well understood. High-throughput transcriptome sequencing of *Chlamydomonas* at various timepoints during ciliogenesis will help us better understand these and other

factors necessary for proper functioning cilia. Thus, in this part of the dissertation we:

i. Perform high-throughput sequencing of the *Chlamydomonas reinhardtii* transcriptome at various timepoints during ciliogenesis.
ii. Evaluate the sensitivity and specificity of our high-throughput sequencing expression data with qRT-PCR
iii. Determine whether the *de facto* validation timepoint for up-regulation is correct by peak expression analysis.
iv. Investigate whether there is an early ciliogenesis gene regulation program.
v. Identify novel cilia gene candidates and potential, novel human ciliopathy genes.

1.3 Dissertation Layout

This dissertation is laid out as follows: Chapter 1 provides the motivation for, and a brief description of, each of the projects that make up this dissertation. Chapter 2 provides a brief introduction to molecular biology with a focus on aspects that are relevant to each of the projects that make up this dissertation. Chapters 3-5 are the independent projects. Chapters 3 and 4 are published and Chapter 5 is being submitted. Chapter 6 contains concluding remarks and future directions.

Chapter 2

Background and Significance

2.1 Molecular Biology Primer

A eukaryotic genome is made up of deoxyribonucleic acid (DNA) molecules, organized into multiple, linear chromosomes that reside in the nucleus. DNA is a polymer of nucleotides, a nucleic acid base (or base) bound to a phosphate-deoxyribose sugar group. In a living cell, DNA typically exists as two tightly coupled molecules in a double helix held together by hydrogen bonds between the internally oriented bases. There are four canonical bases that make up the alphabet underlying the language and grammar encoding all information necessary for the creation and maintenance of life: they are adenine (A), cytosine (C), guanine (G) and thymine (T) and when two DNA molecules interact, such as in a double helix, A pairs with T and C pairs with G.

Perhaps the most obvious elements encoded in DNA are the protein-coding genes, which are regions of the genome that are ultimately translated into functional proteins that perform the life giving functions necessary for the survival of an organism. Genes have many features that are recognized by DNA interacting cellular machinery, which include promoters, exons, introns and flanking untranslated regions that are essential for different regulatory mechanisms. Protein coding genes were among the first elements to be targeted for computational analysis because their characteristic features must necessarily be distinguishable from the surrounding DNA sequence in which they are embedded. Genefinders are machine learning methods that recognize gene features by training parameter values using known genes to predict the locations and structures of new genes as gene models. When genefinders are applied to entire genomes, the output is a predicted gene catalog. Genefinders have successfully annotated the genomes of species that have a comprehensive and representative training set of genes.

The process by which a protein coding gene is expressed involves many steps that can be grouped into two processes: transcription and translation. Transcription of a gene into a primary transcript of ribonucleic acid (RNA) begins with the binding of transcription factors that recruit the transcriptional machinery to a region located before the start site of the protein coding region of the gene. RNA is a nucleic acid like DNA except the sugar group is a ribose and RNA contains the base uracil (U) in lieu of thymine (T). The product of transcription is a primary transcript of the desired gene. The primary transcript

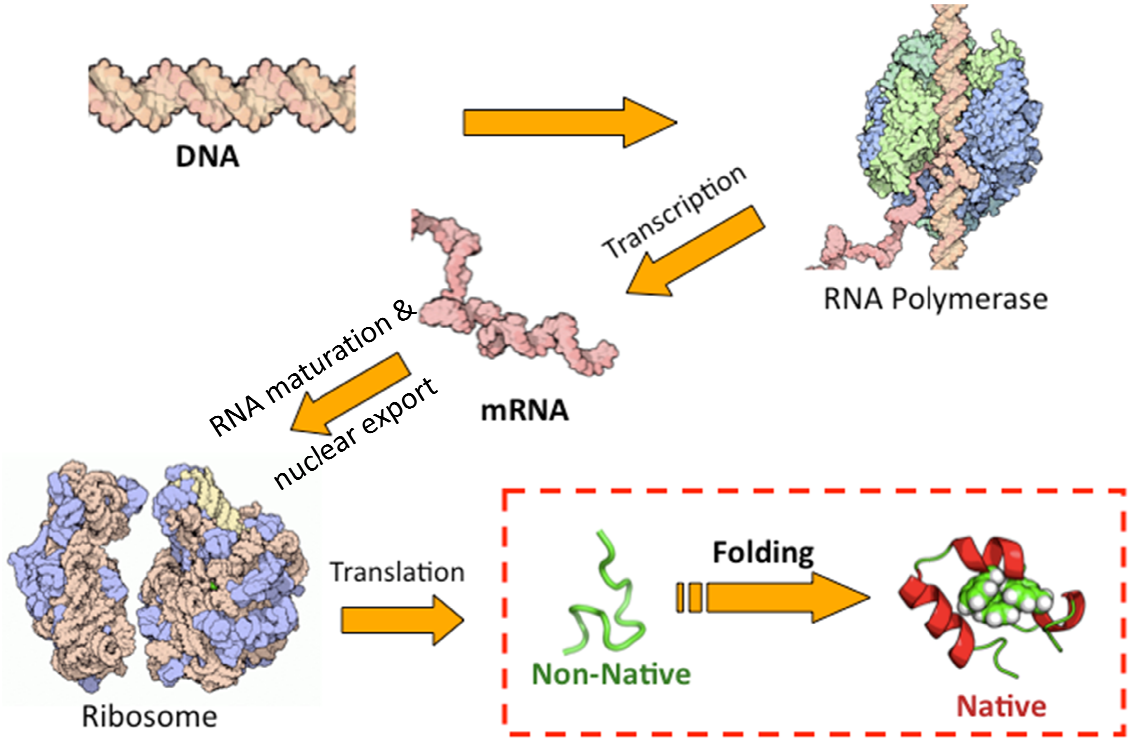

Figure 2.1 Protein biosynthesis. *DNA is first transcribed by the enzyme RNA polymerase, which undergoes matureation prior to export from the nucleus where translation of the mature mRNA into amino acid polymers is performed by ribosomes. The resulting protein may be subject to post-translational modification and must undergo folding, perhaps with assistance from other proteins, before taking its functional, native conformation.*

is then subject to post-transcriptional modifications, including the excision of regions corresponding to the introns of the desired gene, rejoining of the remaining exons and polyadenylation, which is the addition of adenines to the 3' end of the transcript and marks the mature messenger RNA transcript (mRNA) for export from the nucleus. Once in the cytoplasm, the mRNA is recognized and bound by components of the translation machinery. These components together read the mRNA sequence in non-overlapping, three letter words called codons from the 5' end to the 3' end, recruit the unique amino acid corresponding to each

codon and link those amino acids in the order they are indicated to synthesize the protein product of the gene.

The amino acids of a protein interact and fold the polymer on itself in a process called protein folding. Achieving the correct final conformation, or shape, is critical for the proper function of any protein. The sequence of a protein determines the order and proximity of the 20 different amino acids, which dictates the folding dynamics of the polymer. Thus the sequence determines the final conformation and efficacy of a protein. Multiple proteins interact and function in concert as complexes or in pathways on substrates to bring about the desired effects and the traits biologists observe as phenotypes.

2.2 Basic gene selection theory

Basic gene selection theory formulates the evolution of genes and genomes as functions over time driven by spontaneous changes on the gene and genome levels, which are checked by the various mechanisms of natural selection for the most advantageous complement of genes. The evolution of protein-coding genes may result from a single base change at a single position, or the insertion or deletion of an entire gene feature. If the event impacts the intended protein function, it will be acted upon by selective pressure and selected for or against depending on whether the change in function is deleterious or confers an advantage to the mutation-carrying organism (Nei *et al.* 2010).

Evolutionary events on the genome level may span entire genome duplications, select chromosome duplications or deletions and duplications or deletions of large contiguous DNA fragments along with all the genes they encode. Duplication events at this scale result in extra copies of genes and transcription promoting elements in a species that is perfectly viable with a single copy. Organisms with the excess protein product that results from extra copies of some genes will be selected for or against. Duplicated genes that result in excess proteins that are neither beneficial nor deleterious will be propagated to subsequent generations, free from the effects of selective pressure. Over time, the extra gene copies may go on to accumulate gene level mutations that result in the loss of the original function or gain of new functions that are themselves acted upon by selective pressure. If a new function confers an adequate advantage to the carrying organism, the mutated copy will eventually become fixed in the species population (Nei *et al.* 2010).

2.3 Sequence similarity and protein function

The transcription and translation processes entail that genes with similar sequences will be translated into proteins with similar sequences. Moreover, the relationship between protein sequence, structure and function implies that similar proteins will fold into similar conformations and are likely to perform similar functions and have similar characteristics. Indeed, one of the most successful and widely used computational methods in biology is designed

specifically to quantify the degree of similarity between input sequences to inform further analysis. Proteins and genes that have significantly similar sequences are homologs and their sequences are homologous.

Basic gene selection theory entails that there be two classes of homologous sequences. Orthologs are genes from two different species that share a common ancestral gene that existed as a single copy prior to the speciation event giving rise to the two species. So, orthologs are presumed to perform similar functions because they have presumably been subject to similar selective pressures. Paralogs are genes that evolve from separate, duplicate ancestral genes in a common ancestral species. As such, paralogs likely do not perform similar functions because while one preserves the original function, the other copy is free to evolve and thus be subject to different selective pressures.

2.4 The study of human ciliopathies using the model organism *Chlamydomonas reinhardtii*

Cilia are evolutionarily conserved, complex structures that protrude from most eukaryotic cells. These organelles are important components of a variety of signaling cascades including the canonical Wnt/β-catenin pathway (WALLINGFORD and MITCHELL 2011), the non-canonical Wnt/planar cell polarity (PCP) pathway (WALLINGFORD and MITCHELL 2011) and the sonic Hedgehog signaling (Shh) pathway (MURDOCH and COPP 2010). Cilia can be further divided

into primary cilia or motile cilia (FLIEGAUF *et al.* 2007). Primary cilia do not impart motion into the extracellular environment and act as thermo-, mechano- and chemosensory organelles (FLIEGAUF *et al.* 2007). Motile cilia actively move in the extracellular space and are responsible for transporting extracellular fluids or bodies over the surface of the cell like mucous in the respiratory tract or ova through the fallopian tubes. Motile cilia are also responsible for cell motility, as in the case of human spermatozoa. Defects in cilia can result in a wide array of developmental and physiological abnormalities. Cystic kidneys and liver disease are the most common clinical features of ciliopathies (TOBIN and BEALES 2009). Another common feature is a reversal in organ laterality (e.g. heart and stomach on right side, liver on the left side; *situs inversus*). Extra digits on the hands and/or feet (i.e. polydactyly), agenesis of the corpus callosum (i.e. failure to develop the component of the brain that connects the left and right hemispheres) and mental retardation often manifest together (FLIEGAUF *et al.* 2007; TOBIN and BEALES 2009). Other symptoms include retinal degeneration that ultimately results in blindness, abnormal brain development resulting in a brain that protrudes through the skull and death, infertility, chronic ear and airway infections, obesity and hypogenitalism, among others (FLIEGAUF *et al.* 2007; TOBIN and BEALES 2009). The growing list of recognized ciliopathies currently includes Bardet-Biedl syndrome (BBS), Meckel syndrome (MKS), Joubert syndrome (JBTS), Nephrophthisis (NPHP), Senior-Løken syndrome (SLSN), Jeune syndrome (JATD), Oro-facial-digital syndrome type 1 (OFD1), Ellis van Creveld syndrome (EVC), Alström syndrome (ALMS), primary ciliary dyskinesia (PCD; Kartageners Syndrome), polycystic kidney disease (PKD) and Cancer

(FLIEGAUF *et al.* 2007; TOBIN and BEALES 2009). Furthermore, aberrations in cilia disease gene orthologs tend to result in multisystemic abnormalities in multiple organisms and indicate a conserved, pervasive reliance of many physiological and developmental processes on the proper synthesis and function of cilia (TOBIN and BEALES 2009). Consequently, the identification, characterization and implication of human ciliopathy disease genes have greatly benefited from their study in model organisms (FLIEGAUF *et al.* 2007; TOBIN and BEALES 2009). One of the most influential model organisms for the study of ciliopathy disease genes is *Chlamydomonas reinhardtii* (FLIEGAUF *et al.* 2007), an organism in which ciliogenesis is readily induced.

Transcript abundance of most genes encoding known cilia components are greatly amplified in *Chlamydomonas* during ciliogenesis (STOLC *et al.* 2005). *Chlamydomonas* is a unicellular, green alga with genetics similar to yeast but for two cilia that are virtually identical to cilia found in Human. *Chlamydomonas* is an ideal model organism for transcript abundance based cilia gene detection because ciliogenesis can be induced by pH-shock. When environmental pH is precipitously dropped, *Chlamydomonas* cells shed their cilia and ciliogenesis begins immediately once environmental pH is restored. The specific transcriptional induction of genes encoding many known cilia components during ciliogenesis have been widely reported and further underscore the efficacy and potential advantages of using *Chlamydomonas* as a model organism to study cilia and ciliogenesis (LI *et al.* 2004; PAZOUR and WITMAN 2009; STOLC *et al.* 2005). A variety of methodologies have been successfully used to determine the

Chlamydomonas ciliome, including direct proteomic analysis (PAZOUR *et al.* 2005), comparative genomics (LI *et al.* 2004; MERCHANT *et al.* 2007) and microarrays (STOLC *et al.* 2005). Currently, between *Chlamydomonas* and other cilia model organisms, there are more than 650 genes models that have either experimental evidence of cilia or ciliogenesis involvement or that have predictions suggesting some cilia association (FLIEGAUF *et al.* 2007).

2.5 Accelerating accurate automated annotation

The yeast-like characteristics of *Chlamydomonas* genetics implies many advantages to using the green alga as a model system to study cilia and ciliopathies. However, relative to the yeast genome, the *Chlamydomonas* genome sequence is poorly annotated. Furthermore, as sequencing technologies continue to advance and drive down the costs of genome digitization, the annotation state of the *Chlamydomonas* genome is less an exception and increasingly the norm (VARSHNEY *et al.* 2010). Although experimental methods have been successful in identifying a handful of highly specific cilia disease genes in *Chlamydomonas*, it is high-throughput computational studies that harbor the greatest potential to rapidly elucidate a more comprehensive ciliopathy disease gene catalog, methods which depend on the availability of an accurately annotated genome.

Existing computational genome and proteome annotation methods rely heavily on direct experimental evidence to determine and characterize genes and

proteins in organisms of interest. Genefinders require whole-gene annotations to train statistical parameters. When there are inadequate numbers of experimentally determined whole-gene annotations, research communities either delay automated genefinding until more experimentally determined whole gene annotations become available, or use parameters trained from some other species. Protein characterization is the next step in the annotation process. Existing automated protein characterization methods confine novel information about protein function within protein families of adequately similar sequences. The extent of automated protein characterization made possible by such methods is largely dependent on existing knowledge about at least one member of every protein family. In the case where a protein of interest belongs to a poorly or uncharacterized protein family, the researcher cannot infer any more information from the entire protein annotation database.

Aside from sequencing whole genomes, high-throughput sequencing methodologies can also be used to quantify transcriptome changes in an organism under conditions of interest. *Chlamydomonas* undergoes ciliogenesis following pH-shock by greatly up-regulating transcript abundance of cilia genes (LEFEBVRE and ROSENBAUM 1986). High-throughput sequencing of the *Chlamydomonas* transcriptome would provide transcriptome-wide evidence of genes in the predicted catalog. High-throughput whole-transcriptome sequencing of *Chlamydomonas* during ciliogenesis would potentially reveal novel cilia association for genes that have no prior evidence of cilia involvement. Furthermore, sequencing of the transcriptome at successive timepoints would

facilitate time-series analysis of cilia genes, possibly revealing a ciliogenesis regulation program that would begin to forward our understanding not only of genes involved in cilia assembly, but their interactions and expression control.

The objective of this work is threefold. First, it develops and validates a computational strategy to train genefinders for annotation poor genomes using fragments of expressed genes. Second, this work develops and validates a novel computational method that facilitates the transfer of protein characterization information between dissimilar protein sequences. Finally, this work will analyze time-series transcriptome data of *Chlamydomonas* during induced ciliogenesis to identify new human ciliopathy gene candidates and determine whether there is a ciliogenesis gene regulation program that co-ordinates the proper assembly of cilia.

2.5.1 An effective genefinder training method for emerging model organisms

By analyzing the sequence composition of known genes, computational tools can be trained to recognize characteristic differences between coding and non-coding genomic sequence. Genefinders are computational methods that take a genomic sequence as input and outputs positions of the input sequence that are predicted to be boundaries of modeled gene features like exons, introns and untranslated regions (UTRs). The most widely available genefinders model gene structure using a generative statistical model called generalized Hidden Markov

models (gHMM). Typically, given a model of a gene structure that is parameterized by different gene features, a gHMM genefinder will fit values for model parameters based on observed characteristics of features in a set of high-confidence training gene models. Then, for a given genomic sequence, the genefinder will define gene feature boundaries based on the internal gene model and parameter values determined during the training phase.

In order to predict an accurate and complete catalog of gene models, genefinders must be trained on a large, representative set of experimentally supported gene annotations from the target species. Large repositories of whole-gene annotations are available for some species, but the vast majority remains annotation poor. While genefinders trained on gene models from one species can be used to predict gene models in other species, it is found that the accuracy suffers compared to models trained on genes from the same species (LI *et al.* 2003).

The number of available whole-gene annotations is far outnumbered by expressed sequence tag (EST) data. Libraries of ESTs are made up of sequence fragments templated by the mRNA of expressed genes from organisms under different environmental condition (e.g. stress, mating) or tissues (PARKINSON and BLAXTER 2009; VARSHNEY *et al.* 2010). Many organisms with few whole-gene annotations have sizable collections of ESTs because gene fragment libraries can be constructed from any organism that can be sequenced (PARKINSON and BLAXTER 2009). *Chlamydomonas*, for example, had 156 whole-gene annotations, but more than 165,000 ESTs at the time this work was conducted. The

incorporation of EST data into the genefinding process has been noted to have a positive effect on accuracy and has been used to successfully identify alternative protein products from the same gene, a result of alternative splicing of transcribed RNA (PARKINSON and BLAXTER 2009; WEI and BRENT 2006). However, there has been no report or evaluation of a genefinder trained solely on gene fragments, the most abundant coding sequence source available.

The greatest implication of an effective training protocol that is based exclusively on gene fragments is that it would expand the application space of automated genefinders to any species with a genome that can be sequenced. Furthermore, the main limitations of direct proteomic and whole transcriptome sequencing alternatives, in the cases where these methods are economically feasible, are that they depend on minimum protein levels or transcript abundance, which vary between different environments and conditions. Consequently, in order to identify a complete gene set, the exclusive use of these methods would require measurements from a wide variety of environmental and developmental conditions, to ensure that all transcribed regions have been included. One of great advantage of including genefinders into the annotation process is that their performance and reliability are not sensitive to protein abundance, transcript abundance or untested environmental conditions. Therefore, accurate genefinders are complementary to high-throughput, whole-transcriptome sequencing applications of next-generation sequencing technologies and to direct proteomic methods in defining a more complete and accurate gene catalog for a given species of interest. Hence, the implementation

and validation of a novel strategy to leverage gene fragment data for training the genefinder GreenGenie2 make up the first objective in this research and is presented in Chapter 3.

2.5.2 Detecting co-evolution of proteins for automated protein annotation

Predicted gene models can be conceptually translated using the genetic code that equates codons to unique amino acids. Presently, automation of protein characterization is largely limited to transferring existing knowledge between protein homologs. Implications of protein sequence similarity and functional similarity lead to homology based organization of proteins into protein families. Knowledge about one family member is presumed to be transferable to all other members. The extent of characterization made possible by this method depends on the number of families for which there is characterization data on at least one member of that family. The homology method of automatic annotation does not facilitate the transfer of information between dissimilar sequence families. As a result, a large proportion of protein families remain uncharacterized beyond sequence similarity.

Proteins function in pathways or bind together and form complexes to bring about traits and so rarely act alone. In order for a trait to be conserved through evolution, co-operating proteins that are responsible for the trait need to maintain functional compatibility. The phylogenetic profile comparison (PPC)

class of automated protein characterization methods organizes proteins by their patterns of occurrence across divergent proteomes. PPC methods operate on the premise that patterns of protein occurrence across diverse species sets evidence instances of protein co-evolution and that a common pattern is indicative of an interaction. The greater the diversity and number of species included in the analysis, the more specific the occurrence patterns become. The occurrence profile of a protein depends on the completeness and accuracy of the protein catalog for each species included in the analysis. These two observations further underscore the importance of expanding the application space of accurate automated genefinders (Chapter 3).

Phylogenetic profile comparison methods are based on the premise that there is strong selective pressure for proteins that functionally interact to be inherited together through speciation events. Early phylogenetic profiling annotation methods rely exclusively on sequence similarity when determining a bit vector of occurrence across different reference species as determined by a static cutoff E-value (ANANTHARAMAN and ARAVIND 2003; KARIMPOUR-FARD *et al.* 2007; PELLEGRINI *et al.* 1999; SUN *et al.* 2005). More advanced algorithms extended early methods by using real-valued vectors to capture more of the continuous nature of sequence similarity scores across multiple species and to correct for evolutionary bias in sequence similarity (JOTHI *et al.* 2007). Real-values are computed by normalizing raw similarity scores by imposing the branch lengths of some phylogenetic tree external onto the input data, biasing normalized scores towards the external tree. Many existing trees are derived from

single proteins or computational concatenation of multiple proteins into superproteins that are then aligned for a relative measure of evolutionary distance, further removing them from biological reality (ROGER and HUG 2006). Typically, after similarity scores have been normalized with respect to a phylogeny, clustering of proteins based on their normalized profiles is performed by profile comparison methods such as Hamming distance (PELLEGRINI *et al.* 1999), a measure of correlation (KARIMPOUR-FARD *et al.* 2007) or some measure of mutual information . The Hamming distance measure assumes sustained protein loss and gain as equally likely events in evolution, which is inconsistent with existing knowledge of eukaryotic evolution (ARAVIND *et al.* 2000). Mutual information comparison methods introduce parameters that are computationally and statistically convenient but have little if any biological basis (BARKER *et al.* 2007). Existing methods further assume that all proteomes are complete and correct. In reality, most eukaryotic protein catalogs are incomplete and contain some number of proteins that are not real or are mispredicted. Incomplete data confounds any attempt at reliably determining the presence or absence of a given protein in a given proteome and introduces noise into subsequent phases of existing methods. Ultimately, existing methods continue to rely on inference through direct homology, which exclude them from novel protein function discovery (JIANG). Thus, there is a need to develop a novel method that can help characterize knowledge poor protein families in eukaryotes by inferring information from other families based on patterns of co-occurrence, referred to as APACE, which is the second objective of this research and will be discussed in Chapter 4.

2.5.3 High-throughput transcriptome sequencing of *Chlamydomonas* identifies new ciliopathy disease gene candidates

Transcript abundance of most genes encoding known cilia components are greatly amplified in *Chlamydomonas* during ciliogenesis (LEFEBVRE and ROSENBAUM 1986; STOLC *et al.* 2005). *Chlamydomonas* is a unicellular, green alga with genetics similar to yeast but for two cilia that are practically identical to cilia found in humans. *Chlamydomonas* is an ideal model organism for transcript abundance based cilia gene detection because ciliogenesis can be induced by pH-shock. When environmental pH is precipitously dropped, *Chlamydomonas* cells shed their cilia and ciliogenesis begins immediately once environmental pH is restored. The specific transcriptional induction of genes encoding many known cilia components during ciliogenesis in *Chlamydomonas* have been widely reported and further underscore the efficacy and potential advantages of using this alga as a model organism to study cilia and ciliogenesis. High-throughput sequencing of the *Chlamydomonas* transcriptome at various timepoints during ciliogenesis would complement existing direct proteomic results (PAZOUR *et al.* 2005) because such a study would probe the entire transcriptome during ciliogenesis as a whole, facilitating not only the detection of genes that encode products inherent in the mature cilium, but also the genes that, while not intrinsic to the mature organelle, are essential for the initiation and regulation of ciliogenesis and cilia function. This methodology would also complement existing comparative genomics methods that have been applied to defining the complete cilia gene catalog. Comparative genomics methods must discard genes that have

an adequate degree of conservation in a non-ciliated species, a necessary practice to reduce the number of false positive genes that are conserved across ciliated species to conserve related traits or processes that are not specific to cilia (e.g. transcription or mitosis). Whole-transcriptome next-generation sequencing does not depend on gene conservation patterns and will compliment comparative genomics methods because of its capacity to include genes that are conserved in non-ciliated organisms, but remain essential for proper cilia biogenesis, structure and function (e.g. tubulins and kinesins). In Chapter 5, we utilize the recently updated *Chlamydomonas* genome assembly (v4) and gene models predicted on that assembly by the GreenGenie2 *Chlamydomonas* genefinder to present results of the first whole-transcriptome next-generation sequencing of *Chlamydomonas reinhardtii* during ciliogenesis.

Chapter 3

An effective genefinder training method for emerging model organisms

Note: Results in this chapter are published in Kwan AL, Li L, Kulp DC, Dutcher SK, Stormo GD: Improving Genefinding in *Chlamydomonas reinhardtii*: GreenGenie2. BMC Genomics 2009, 10:210.

3.1 Introduction

A complete genome sequence facilitates the identification of all the genes in an organism and helps determine the set of functions encoded by those genes as well as the regulation of their expression. The identification of protein-coding genes can be approached both experimentally and computationally and the combination of approaches leads to the most complete catalog of genes (HAAS *et al.* 2003). Expressed sequence tags (ESTs) provide experimental evidence for the

transcription of specific regions of the genome and significant similarity with known proteins in other organisms also provides evidence for the existence of a gene. However, both approaches have limitations that often preclude them from identifying the complete gene set. The exclusive use of the former would require a very large library of ESTs, obtained from a wide variety of environmental and developmental conditions, to ensure that all transcribed regions have been included. Identification based on homology will fail to identify genes that are novel to a particular species, or that are sufficiently diverged to make detection unreliable. *ab initio* genefinders provide a complementary gene identification method by predicting gene models based on the statistical characteristics of a representative set of protein-coding genes from the genome of interest.

Research using the unicellular green alga, *Chlamydomonas reinhardtii,* has provided important insights into many cellular processes that include cilia assembly and motility, basal body assembly and positioning, phototaxis, gametogenesis and fertilization, circadian rhythms, photosynthesis, starch metabolism, and cell wall assembly (BALL and DESCHAMPS 2009; DUTCHER 2009; HEGEMANN and BERTHOD 2009; PAZOUR and WITMAN 2009; ROCHAIX 2009; SNELL and GOODENOUGH 2009) . *Chlamydomonas* is amenable to genetic analysis using classical techniques of tetrad analysis and complementation as well as molecular techniques of transformation and RNA interference (HARRIS 2009).

The current catalog of genes for *Chlamydomonas reinhardtii* is based on a combination of experimental and computational approaches (MERCHANT *et al.* 2007) where 44% of the 15,143 models in the catalog are derived from *ab initio*

methods and the remainder use various evidence including similarity in other organisms and manual annotation. The inclusion of multiple *ab initio* genefinders gives rise to complementary predictions by providing alternative models that can be used for experimental validation and may lead to the determination of true gene structures. Taken together, multiple methods may yield multiple correct predictions for genes with multiple alternate splice variants and a complementing genefinder can also provide complete models for genes that are incomplete within an existing catalog and predict novel genes.

Ab initio genefinders employ models that capture the essential features of gene structure that include sequence characteristics that distinguish exons and introns that include codon bias and feature length distributions as well as signal sequences that correspond to the splice sites that separate them (BRENT 2007; STORMO 2000). Generalized hidden Markov models (gHMMs) are commonly used because gene structure can be represented in a probabilistic framework. Given a particular model of gene structure, the quality of predictions depends on the specific values assigned to the model parameters. Because these model parameters, such as codon bias and splice site patterns, vary between species, training a genefinder on a representative set of example genes from the target species is closely related to the accuracy of the resulting predictions. The original GreenGenie (LI *et al.* 2003) is a version of the Genie genefinder (KULP 2003) that was optimized for the prediction of genes in *Chlamydomonas*. The parameters for GreenGenie were obtained by training on only 71 genes with experimentally determined structure. GreenGenie provided more accurate predictions than other programs available at the time; it predicted 86 genes within 81Kb and

443Kb regions of *Chlamydomonas* genomic sequence and we extrapolated that number to predict between 12,215 and 16,414 genes in the *Chlamydomonas* genome. This prediction was recently corroborated (MERCHANT *et al.* 2007). GreenGenie facilitated gene identification in *Chlamydomonas* by many groups (DYMEK *et al.* 2004; MURAKAMI *et al.* 2005; WIRSCHELL *et al.* 2004).

To improve the quality of gene prediction in *Chlamydomonas*, we used the EST assembly tool, Program to Assemble Spliced Alignments (PASA) (HAAS *et al.* 2003), to assemble 167,613 *Chlamydomonas* EST sequences into protein coding gene models and trained the most recent version of the Genie *ab initio* genefinder (KULP 2003) on this larger set of *Chlamydomonas* gene models. The PASA pipeline begins by filtering and aligning input EST sequences onto a genome assembly. These ESTs alignments are then filtered further and clustered based on alignment compatibility. Finally, through a dynamic programming process, the EST alignment clusters are stitched into a set of consistent, non-overlapping EST assemblies (HAAS *et al.* 2003). PASA has been used for gene prediction in *Arabidopsis thaliana* (HAAS *et al.* 2003), *Drosophila melanogaster* and *Homo sapiens* (KENT *et al.* 2002). This larger training set improves the predictions made by the program, now called GreenGenie2, as determined on a set of 140 well-characterized *Chlamydomonas* genes that were not included in the training set and outperforms the most current published genefinder trained for *Chlamydomonas*. Importantly, GreenGenie2 complements the existing *Chlamydomonas* gene catalog (MERCHANT *et al.* 2007) by completing incomplete models and predicting new genes that were not previously identified.

3.2 Results

3.2.1 Constructing and evaluating a training-set of gene predictions from ESTs

PASA aligned 167,641 high-quality *Chlamydomonas* EST sequences onto the published genome assembly of *Chlamydomonas*, which is called *v3*, and assembled those alignments into 19,707 unique models. The set of PASA assembled models to be used for training were selected based on three criteria. First, the model must be complete; it must begin with an ATG codon and terminate with a stop codon (TAA, TAG or TGA). Second, the assembly must have a minimum open reading frame length of 270bp. Third, the PASA model must lack similarity to the *gb140* reference set of GenBank *Chlamydomonas* gene records (3.5.1; 3.6.1 for GenBank accessions) and known transposable elements (ftp://ftp.jgi-psf.org/pub/JGI_data/Chlamy/v3.0/CHLREP.fn.gz). These criteria reduce the 19,707 models to 2,384 models.

A similarity search of the 2,384 EST assembled models against the NCBI non-redundant database (NRdb) using NCBI BLAST ($E < 1.0 \times 10^{-3}$) was performed to assess the novelty of the assembled ESTs. 957 (40.1%) of the selected PASA assembled models align to an entry in NRdb (Table 3.1) and 482 (20.2%) of the remaining predictions have some overlap (see section 3.5.1) to models in the Frozen Gene Catalog (MERCHANT *et al.* 2007), which we will refer to as *FGC07* (see section 3.5.2). The remaining 945 (39.6%) complete PASA gene

Table 3.1 Categorization of the 2384 PASA EST assembly gene models

Class		N
Alignment to NCBI NRdb		957/2384
Absent from the NCBI NRdb		1427/2384
Exact overlap in *FGC07*		222/1427
Partial overlap in *FGC07*		260/1427
No overlap in *FGC07*		945/1427
	Single exon	835
	Tested via RT-PCR	13
	Verified via RT-PCR	10

models in *v3* are novel predictions identified by PASA EST assembly alone. We find that 835 of these novel models contain only a single exon. The quality of this large set of single-exon genes was evaluated by testing 13 randomly selected single exon models via RT-PCR. All 13 models yield product of the correct size with genomic DNA as the template and 10 of the 13 produce a fragment of the predicted size with cDNA as the template by RT-PCR (Table 3.2). Given that the final set of 2,384 PASA assembled models are derived directly from 167,641 *Chlamydomonas* EST records and screened to have a complete compliment of

Table 3.2 Analysis of PASA gene models: RT-PCR testing of 13 novel, single exon PASA gene assemblies

Assembly ID	Outcome
3146_3724	Present in cDNA
5172_6168	Present in cDNA
8132_9749	Present in cDNA
9104_10933	Present in cDNA
9866_11843	Present in cDNA
11161_13363	Present in cDNA
11240_13451	Present in cDNA
11709_14017	Present in cDNA
14828_17825	Present in cDNA
16095_19351	Present in cDNA
14105_16951	Not present in cDNA
15620_18773	Not present in cDNA
14205_17074	Not present in cDNA

Present: A product of the correct size was found in samples by RT-PCR
Not present: No product was obtained by RT-PCR
Assembly ID numbers can be downloaded from http://bifrost.wustl.edu/GreenGenie2
For primers used see section 3.6.2.

gene features, this set of models is likely to provide an improved training set to optimize the parameters of the GreenGenie2 genefinding program.

3.2.2 GreenGenie2 is more accurate than GeneMark.hmm-ES 3.0

One primary purpose of genefinders is to assist the user by accurately identifying genes in an isolated DNA segment that may be up to several kilobases in length. To evaluate the performance of GreenGenie2 on such short-sequence prediction queries we compared the performance statistics of GreenGenie2 and GeneMark.hmm-ES 3.0, the most recent, publicly available genefinder trained specifically for *Chlamydomonas* (LOMSADZE *et al.* 2005).

Short-sequence prediction sensitivity and specificity of GreenGenie2 and GeneMark.hmm-ES 3.0 were computed for the total predictions made by each genefinder using 140 genomic sequences obtained from the literature, referred to as *gb140* (see section 3.5.1). At the whole-gene level, GreenGenie2 performs considerably better than GeneMark.hmm-ES 3.0. GreenGenie2 achieves sensitivity and specificity values of 0.51 (±0.10) and 0.47 (±0.11) while GeneMark.hmm-ES 3.0 sensitivity and specificity values are 0.31 (±0.10) and 0.24 (±0.09) (Table 3.3). A two-proportion *z*-test indicates that both differences are statistically significant ($p < 0.001$; see section 3.5.4). At the exon level, GreenGenie2 outperforms GeneMark.hmm-ES 3.0 with sensitivity and specificity values of 0.83 and 0.83 as compared to the corresponding values of 0.79 and 0.74 when using GeneMark.hmm-ES 3.0 (Table 3.3). The improvements in predictive accuracy made by GreenGenie2 are most obvious with initial and terminal exons (Table 3.3). At the nucleotide level, the least stringent assessment

Table 3.3 Comparing GreenGenie2 and GeneMark.hmm-ES 3.0 in *gb140* catalog

		GreenGenie2		GeneMark.hmm-ES 3.0	
	N	**Sensitivity**	**Specificity**	**Sensitivity**	**Specificity**
Genes	140	0.51	0.47	0.31	0.24
Exons	1145	0.83	0.83	0.79	0.74
Init. Exons	133	0.65	0.60	0.50	0.40
Int. Exons	870	0.87	0.88	0.84	0.84
Term. Exons	133	0.82	0.75	0.78	0.63
Single Exon	7	0.71	0.62	0.00	0.00
Nucleotides	713682	0.93	0.92	0.91	0.89

of prediction performance, GreenGenie2 shows an improvement of 2-3% over the GeneMark.hmm-ES 3.0 predictions (Table 3.3). These results indicate that GreenGenie2 is an improved *ab initio* genefinder for *Chlamydomonas* and encouraged us to make whole-genome predictions on assembly *v3* and compare them to the *FGC07* catalog (MERCHANT *et al.* 2007) with the goal of identifying new genes and improving the accuracy of the current gene models.

3.2.3 GreenGenie2 models in *v3* complement the Frozen Gene Catalog

GreenGenie2 predictions on *Chlamydomonas* genome assembly *v3* were screened for a minimum coding length of 270bp and against significant

alignment to known transposable elements (see section 3.5.2). The final GreenGenie2 *v3* catalog, *gg2v3*, consists of 12,387 predictions. The identical criteria applied to the *FGC07* catalog leaves 12,320 predictions. All models were further classified as complete or incomplete based on the presence of start and stop codons (see section 3.5.2). All *gg2v3* models are complete by construction. Of the 12,320 models in *FGC07*, only 67.7% are complete; the remaining 3,981 models lack a start codon, a stop codon or both.

Given the possible bias towards single-exon models in the GreenGenie2 training set, a comparison of single-exon models between *gg2v3* and *FGC07* was performed. In *FGC07*, 7.0% of complete models are single-exon genes and a similar proportion is observed in *gg2v3* where 6.4% of the models are single-exon predictions. A two-proportion *z*-test (see section 3.5.4) indicates that there is no significant difference between the two proportions of single exon genes and that there is no bias towards the prediction of single-exon genes made by GreenGenie2.

The *gg2v3* gene catalog was compared to both the complete and incomplete partitions of *FGC07* (Table 3.4) using interval overlap analysis. This analysis compares two lists of coding sequence coordinates that index a common underlying genome sequence and categorizes each prediction as consistent or conflicting (Figure 3.1; see 3.5.5). Our analysis finds that 11% of the *FGC07* models are predicted identically in *gg2v3* and another 67% partially overlap with

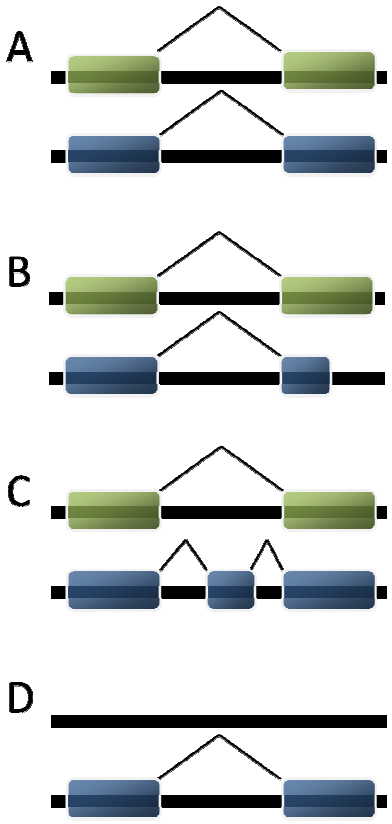

Figure 3.1 Diagram of four classes of gene level interval overlaps. *Interval overlap analysis identifies four classes of overlaps: identical (A), partial (B), novel exon (C), novel gene (D) (See section 3.5.5).*

gg2v3 models (Table 3.4). The remaining 22% of *FGC07* models have no overlap with *gg2v3* models. Additionally, there are 2,859 (23%) *gg2v3* models without interval overlaps to any model in *FGC07*. Predictions in *gg2v3* that have partial interval overlaps to *FGC07* models can be categorized into models with partially overlapping exons and models containing novel exons. Because Genie does not allow non-canonical splice sites, we determined the proportion of *FGC07* exons that partially overlap *gg2v3* exons with either canonical or non-canonical splice sites. Not all splice sites in *Chlamydomonas* follow the canonical rules (STARK *et al.* 2001). However, allowing non-canonical splice sites might improve the

Table 3.4 Comparison of *gg2v3* and *FGC07* catalog by overlap interval analysis

Complete *FGC07* models		**Incomplete *FGC07* models**	
Type of overlap	Count	Type of overlap	Count
Exact Overlap	1,324	Exact Overlap	0
Partial Overlap	5,425	Partial Overlap	2,826
No Overlap	1,574	No Overlap	1,149
Other	16	Other	16
Total	**8,339**	**Total**	**3,981**

Complete model: Any model that includes a starting ATG gene feature and terminates with a stop codon (TAA, TAG or TGA).
Incomplete model: Any model that lacks a start or stop codon or both.
Other: Models that interlaced overlaps and concatenated exact overlaps.

sensitivity slightly, the marginal gain would come with the cost of many additional false positives.

In total, 15% of the partially overlapping *FGC07* exons contain a non-canonical splice 5' site (GT) and 7% contain a non-canonical 3' splice site (AG). Therefore, about 20% of the non-identical, but overlapping exons between the *gg2v3* and *FGC07* catalogs are attributable to the fact that the GreenGenie2 model does not allow non-canonical splice sites. The set of partially overlapping models are of particular interest because they may include examples of alternative splicing as well as highlight incorrect models in each catalog. Each partially overlapping *gg2v3* gene model with three or more exons (N=6,885) was

Table 3.5 Validation of competing *gg2v3* and *FGC* gene models via RT-PCR

Models with alternate exon termini predicted in *gg2v3* and *FGC07*			**Novel exons predicted in *gg2v3* not present in *FGC07***	
***gg2v3* Gene ID**	**Support for *gg2v3***	**Support for *FGC07***	***gg2v3* Gene ID**	***gg2v3* support**
4t254	+	—	1t16	+
11t344	+	—	1t34	+
25t123	+	—	1t147	+
24t200	+	—	11t344	+
5t126	—	—	15t291	+
			30t106	+
			30t170	+
			3t257	—

+: A product of the correct size was found in samples by RT-PCR
—: No product was obtained by RT-PCR
**For primers see sections 3.6.3 and 3.6.4.*

compared to the corresponding *FGC07* model at the exon level. These exons were classified as initial, internal or terminal. The number of novel *gg2v3* exons and partially overlapping exons was determined (Figure 3.2). The four largest groups have 1) partial overlaps for all three exon types (N=761) and no new exons in the *gg2v3* model, 2) an alternative initial exon (N=480), 3) partially overlapping internal exons and both a novel initial and novel terminal exon

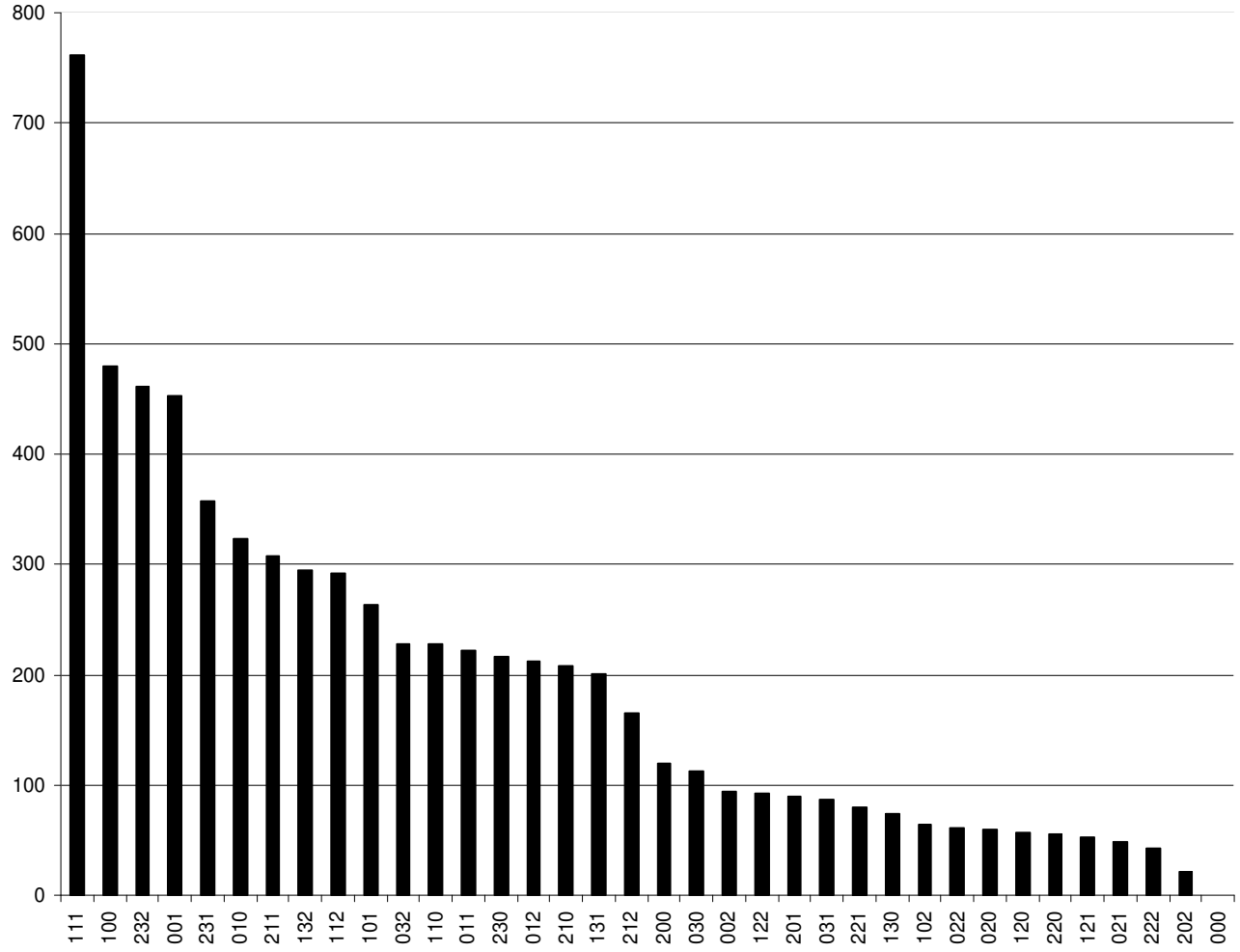

Figure 3.2 Histogram of partial exon overlap *gg2v3* models to *FGC07*. *Exon level interval overlap analysis identifies three types of exons in gg2v3 models with partial overlaps in FGC07: initial, internal and terminal. Each of the three exon types are represented in a three digit code. The rightmost digit corresponds to the terminal exon, the middle position corresponds to all internal exons and the leftmost digit corresponds to the initial exon. Each digit is assigned a value of 0, 1, 2 or 3. A value of 0 at a given position indicates that all exons of that type are exact for every gene in that category. A value of 1 indicates that there is one or more occurrence of partial exon overlap of exons in the position's exon type and no novel exons predicted in gg2v3. A value of 2 indicates that there is one or more occurrences of a whole new exon predicted in gg2v3 that is absent in the overlapping FGC07 model in the exon type corresponding to that position for all genes with that code and no partially overlapping exons between the two catalogs. A value of 3 indicates that there is one or more occurrences of both partially overlapping exons and extra exons in gg2v3 when compared to the model in FGC07 (e.g. gg2v3 models in the class 111 have one or more partially overlapping exons in FGC07 of all three exon types and no occurrences of extra exons predicted; gg2v3 models in the class 100 have exact exon matches across all exons in the model except for the initial exon).*

(N=461) and 4) an alternative terminal exon (N=453). Overall, 28% of these models have new exon splice sites and no new exons in the *gg2v3* model. Only 4% of the partially overlapping *gg2v3* models have only novel exons (Figure 3.2). A small number of each of the partially overlapping models was tested using RT-PCR (see Section 3.5.6). Figure 3.1B shows one type of model that has at least one exactly overlapping exon and at least one alternative exon terminus. No experimental support for any of the five *FGC07* models tested was found, but support for four of the five corresponding *gg2v3* models tested was found (Table 3.5). Figure 3.1C illustrates the second type that has at least one exactly overlapping exon and at least one additional exon in the *gg2v3* prediction that is absent from the *FGC07* model. We find support for seven of the eight predictions tested (Table 3.5).

Predictions in one catalog that have no overlapping counterpart in the other catalog (Figure 3.1D) make up a significant proportion of both *gg2v3* and *FGC07* and may represent substantive sets of true genes that reflect the complementarity of the two catalogs. Our analysis finds that 22% (N=2723) of complete *FGC07* models lack any overlap to models in *gg2v3* and that 23% (N=2,859) of *gg2v3* models do not have interval overlap with any complete or incomplete model in *FGC07*. A small sample of predictions that are exclusive to each catalog was tested by RT-PCR. Four of the five *gg2v3* predictions tested were supported by RT-PCR results (Table 3.6). Similarly, three of the five novel *FGC07* predictions

Table 3.6 Validation of mutually exclusive gene models in *gg2v3* and *FGC*: RT-PCR

Predictions exclusive to *gg2v3*		**Predictions exclusive to *FGC07***	
***gg2v3* Gene ID**	**Outcome**	***FGC07* Gene ID**	**Outcome**
3t69	+	141597	+
19t170	+	181956	+
30t189	+	184911	+
76t11	+	141023	—
69t65	—	180935	—

+: A product of the correct size was found in samples by RT-PCR
—: No product was obtained by RT-PCR
**For primers see sections 3.6.5 and 3.6.6.*

were supported by RT-PCR (Table 3.6). *in silico* analysis indicates that a majority of predictions exclusive to each catalog have EST or cross-species sequence similarity support or both. WU-BLASTP sequence similarity analysis indicates that 92.2% of gene models exclusive to *gg2v3* align to some protein in the Eukaryotic Clusters of Orthologous Genes database (KOG) (TATUSOV *et al.* 2003) or to some sequence in the *Chlamydomonas* EST database. Similarly, WU-BLASTP similarity analysis indicates that 94.5% of the *FGC07* exclusive models are supported by evidence in the KOG or *Chlamydomonas* EST databases.

3.2.4 GreenGenie2 is a robust, effective genefinder across different genome assemblies

Our results in the previous section indicate that GreenGenie2 whole-genome predictions complement *FGC07* (MERCHANT *et al.* 2007) and suggest the potential value of including GreenGenie2 models in the annotation of future *Chlamydomonas* assemblies, so we used GreenGenie2 to predict a whole-genome catalog from the latest assembly of the *Chlamydomonas* genome, denoted as *gg2v4*. Sequence analysis of the two *Chlamydomonas* genome assemblies reveals that *v4* contigs are seven times longer than *v3* contigs on average, which highlights improved continuity in the *v4* assembly compared to *v3* assembly. GreenGenie2 predicts 11,315 models in the *v4* assembly that satisfy the quality control constraints discussed previously. We mapped the *gg2v4* models onto *v3* scaffolds using BLAT (KENT 2002) to facilitate the interval overlap analysis of the *gg2v4* catalog with *gg2v3*. Only 20 of the *gg2v4* models do not have matches in the *v3* genome assembly. Conversely, 303 (2.4%) of the *gg2v3* models do not have matches on the *v4* assembly, which indicates a loss of some sequences in *v4* compared to *v3*. 82.5% of the *gg2v4* models (N=9,184) map completely to a unique locus in *v3* and likely represent loci that are shared between the *v3* and *v4* genome assemblies. 77% of these models are identical to models in *gg2v3* despite the large changes in the genome contigs that are used for prediction. 21% of them have partial overlaps and only 1% is novel in the *gg2v4* model set. Of the 17.1% of the *gg2v4* models that do not map entirely to a single *v3* locus, most of them (73%) have matches to two or more *v3* loci, and the remainder contains additional sequences that do not occur on any *v3* locus. The results indicate that

the *gg2v4* predictions from the updated *v4* assembly are typically the same as the predictions on the shorter genome contigs of *v3*, which suggests that the predictions are not overly sensitive to the length of the contigs used as input. Furthermore, models that either were previously split across multiple contigs or were missing from the *v3* assembly explain most of the differences. In both cases it appears that the updated *v4* assembly has led to improved accuracy of the predicted gene catalog.

3.3 Discussion

Determining genomic and EST sequence allows for the identification of the protein coding genes of a particular organism. We have used the information obtained from EST sequences to train the *ab initio* genefinder Genie (KULP 2003) on a filtered group of PASA assembled models that have both a start codon and a stop codon (complete) to create an accurate *ab initio* genefinder for the GC-rich genome of the green alga *Chlamydomonas reinhardtii*.

The Program to Assemble Spliced Alignments (PASA) (HAAS *et al.* 2003) was used to assemble *Chlamydomonas* EST sequences that were pre-aligned to the *v3 Chlamydomonas* genome assembly. This training set of 2,384 PASA assembled gene models has extensive biological evidence. Interval overlap analysis and homology search indicate that a majority of the PASA predictions align either to an existing *Chlamydomonas* gene model (21%) or have homologs in other organisms (40%). 39% of the PASA models are novel. Support for 10 of

13 novel predictions tested with RT-PCR suggests the potential for using the assembly of pre-aligned EST data as a primary basis of gene modeling, rather than as a supplementary source of predictive information.

One primary application of *ab initio* genefinders is to accurately predict genes within short genomic sequences. Such short-sequence queries are often regions where the user has knowledge of a gene, but depends on the *ab initio* genefinder to predict, confirm or correct the exon level structure of the gene. To test the short-sequence prediction accuracy of GreenGenie2, we compared the predictions of GreenGenie2 to the predictions of the most current, publicly available *ab initio* genefinder trained for *Chlamydomonas*, GeneMark.hmm-ES 3.0 [18] on a set of 140 *Chlamydomonas* genomic sequences. Each of these genomic sequences contains a single known GenBank reference *Chlamydomonas* mRNA and the corresponding upstream (average length: 564bp) and downstream (average length: 731bp) flanking regions. Sensitivity and specificity of the two genefinders was determined by comparing the prediction from each genefinder against the reference GenBank annotation. Comparing the predictions on the gene level, GreenGenie2 is significantly more sensitive and specific (Table 3.3; $p < 0.001$) than GeneMark.hmm-ES 3.0. Results also indicate that GreenGenie2 outperforms GeneMark.hmm-ES 3.0 across all four types of exons (initial, internal, terminal and single), in particular, the initial and terminal exons.

Another application of *ab initio* genefinders is the prediction of whole-genome gene catalogs. GreenGenie2 was used to predict a whole genome gene catalog on *Chlamydomonas* genome assembly *v3* and this catalog, *gg2v3*, was

compared to the existing *FGC07* gene models by interval overlap analysis. The two catalogs predict a similar number of genes and a significant number of the models are identical. However, the two catalogs differ in several ways. First, there are a substantial proportion of complete *FGC07* gene models that overlap but are not identical to *gg2v3* models (54%). Exon level analysis of partially overlapping *gg2v3* models shows that there are multiple causes (Figure 3.2). The four most frequent causes include partial exon overlap devoid of any new exons in *gg2v3*, models that are identical except in the initial exon, models where GreenGenie2 predicts entirely new initial and terminal exons and models that are identical except in the terminal exon. The third class reflects our observation that 32% of *FGC07* models are incomplete. This analysis illustrates the range of complementarity that exists between the two catalogs. RT-PCR analysis found support for four out of five *gg2v3* models (Figure 3.1B; Table 3.5), but failed to provide support for any of the five *FGC07* models tested. In addition, seven of eight randomly selected *gg2v3* models with additional exons that are absent from their *FGC07* counterparts were validated by RT-PCR (Figure 3.1C; Table 3.5). Although the number of genes tested is small, the results suggest that GreenGenie2 complements the existing catalog by successfully identifying and correcting gene models that may be incorrect in the current *Chlamydomonas* annotation. Second, there is a set of *gg2v3* predictions (N = 2,859) that is absent from *FGC07*, and a set of *FGC07* predictions (N = 2,723) that is absent from *gg2v3*. We tested five randomly selected models from each set of exclusive predictions using RT-PCR and found support for four *gg2v3* models and support for three of the *FGC07* models tested. Furthermore, BLASTP alignment and EST

alignment reveal that there is extensive support for almost all predictions that are absent from just *gg2v3* (93.8%) or absent from just *FGC07* (92.2%). These results indicate that each prediction method complements the other by identifying potentially true genes that are missing from the other catalog. Finally, GreenGenie2 completes 2,261 incomplete *FGC07* models, which demonstrates another benefit of including GreenGenie2 whole-genome predictions into current and future *Chlamydomonas* gene catalogs.

The average contig length from assembly *v3* to assembly *v4* increases seven-fold, which indicates a greater degree of assembly continuity. The robustness of our genefinder was tested across more continuous genome assemblies by using GreenGenie2 to predict a whole-genome gene catalog with the *v4* genome assembly. If GreenGenie2 predictions were sensitive to the exact genome assembly used, and in particular if they varied substantially when the length of the genomic contigs changed, it would indicate unreliability in the predictions. However, we find that 77% of the *gg2v4* models are identical to models in *gg2v3*, and most of the remainder overlaps significantly with the *gg2v3* models. A large fraction of the differences are models where the *gg2v4* predictions extend or merge models in *gg2v3* based on the longer contiguous sequences in *v4*. These results are consistent with improvements in the updated assembly of *v4* and with GreenGenie2 providing reliable predictions on a more contiguous genome assembly. Overall, GreenGenie2 performance on short-sequence and whole-genome predictions suggest that optimizing *ab initio* genefinding parameters on the assembly of a large collection of pre-aligned gene

fragments is a rapid, low-cost and effective method by which *ab initio* genefinders can be established.

3.4 Summary

In this chapter, the *ab initio* genefinder Genie was trained on a large set of complete PASA predicted gene models assembled from available *Chlamydomonas* EST sequence data. Short-sequence performance analysis indicates that GreenGenie2 is more accurate than the most recent *Chlamydomonas* genefinder in the literature (LOMSADZE *et al.* 2005). Interval overlap analysis between the GreenGenie2 *v3* whole-genome catalog and the *FGC07* catalog reveals that GreenGenie2 complements the current *Chlamydomonas* gene catalog (MERCHANT *et al.* 2007) by accurately predicting new *v3* gene models that are incomplete, incorrect or absent in *FGC07*. When GreenGenie2 was applied to the latest available *Chlamydomonas* genome assembly and the predicted *v4* models were mapped back onto *v3* scaffolds, GreenGenie2 appears to be robust against a seven-fold improvement in assembly continuity. These results illustrate a potential new application of EST sequence data to gene prediction and underscore the value of including the predictions of a fast, accurate *ab initio* genefinder like GreenGenie2 into present and future catalogs. We have made the GreenGenie2 genefinder described in this study available online. The submission form is available at http://bifrost.wustl.edu/cgi-bin/greengenie2/greenGenie2.

3.5 Methods

3.5.1 Sequence datasets

This study uses the *Chlamydomonas* genome assembly version 3 (ftp://ftp.jgi-psf.org/pub/JGI_data/Chlamy/v3.0/Chlre3.allmasked.gz). Genome assembly version 4 (*v4*) was obtained from Alan Kuo at the Joint Genome Institute. Sequences longer than 1Mb are pre-processed into shorter sequences prior to annotation by GreenGenie2. Pre-processing involves the removal of stretches of ambiguous nucleotides longer than 50bp and treating the prefix and suffix as independent sequences. This pre-processing is advantageous for computational efficiency but to preserve maximal continuity in the assembly, all splitting events were chosen to minimize the final number of sequences. We found that requiring a minimum length of greater than 50bp greatly increased the necessary number of splitting events. The *v3* assembly was split from 1,557 sequences totaling 120,186,811 bases (~77.2Kb/sequence) into 1,636 sequences totaling 120,076,271 bases (~73.4Kb/sequence) following the removal of 110,540 ambiguous positions. The *v4* assembly was split from 88 sequences totaling 112,305,447 bases (~1.3Mb/sequence) into 218 sequences totaling 111,935,880 bases (~513.5Kb/sequence) following the removal of 369,567 ambiguous positions.

A total of 140 experimentally verified *Chlamydomonas* annotations from GenBank (http://www.ncbi.nlm.nig.gov/Taxonomy/Browser/wwwtax.cgi?id=3055) constitute a reference set for short sequence analysis and are referred to as *gb140*

(see section 3.6.1 for listing). Initially, 222 GenBank records were retrieved by identifying records that indicated experimentally determined gene structure by direct sequencing of a complete cDNA and the genomic DNA and thus were not generated by automated assembly methods. The records were then filtered to remove genes with misannotated or missing start sites (N=17), non-canonical splice sites (N=46), misannotated or missing termination sites (N=6) or open reading frames that are not multiples of three (N=13). The included upstream and downstream flanking regions averaged 534bp and 731bp in length, respectively. The 167,613 EST records used to construct the PASA EST assemblies are from GenBank (http://www.ncbi.nlm.nig.gov/Taxonomy/Browser/wwwtax.cgi?id=3055). All PASA EST assemblies were screened for significant alignment (BLAST E-value < 1.0×10^{-20}) to *gb140* before training to remove any bias in the subsequent short-sequence performance evaluation.

3.5.2 *Chlamydomonas* gene catalogs

Three *Chlamydomonas* whole-genome catalogs were evaluated in this study: the GreenGenie2 whole-genome prediction on assembly *v3* (http://bifrost.wustl.edu/greengenie2/), the GreenGenie2 whole-genome prediction on assembly *v4* (http://bifrost.wustl.edu/greengenie2/) and the Frozen Gene Catalog (*FGC07*) from Merchant *et al.* (MERCHANT *et al.* 2007) (transcript file: ftp://ftp.jgi-psf.org/pub/JGI_data/Chlamydomonas_reinhardtii/v3.1/Chlre3_1.GeneCatalog_2007_09_13.transcripts.fasta.gz; model file: ftp://ftp.jgi-

psf.org/pub/JGI_data/Chlamydomonas_reinhardtii/v3.1/Chlre3_1.GeneCatalog_2007_09_13.gff.gz). Prior to further analysis all models from all catalogs were screened for a minimum coding length of 270bp and lack of significant alignment to known transposable elements (ftp://ftp.jgi-psf.org/pub/JGI_data/Chlamy/v3.0/CHLREP.fn.gz). The choice of 270bp as a minimum coding length is somewhat arbitrary, but there are very few verified genes shorter than this in *Chlamydomonas*. In *Sacharomyces cerevisiae*, recent studies show that there are about 200 genes (5%) that are less than 90 amino acids or 270bp (KASTENMAYER *et al.* 2006). However in a genome that is 2/3 G+C like *Chlamydomonas*, prediction of genes 270bp long or shorter will occur with a probability of 0.12. This probably in yeast about is about ten-fold lower (0.013). Thus, the inclusion of predicted genes that are less than 270bp is likely to increase the number of falsely predicted genes greatly. Many models in *FGC07* lack a start codon, a stop codon or both are thus considered incomplete models.

3.5.3 Programs

Seven publicly available programs are used in this study. They are PASA [2] (http://pasa.sourceforge.net), Genie (KULP 2003), GeneMark.hmm-ES 3.0 (LOMSADZE *et al.* 2005) (http://opal.biology.gatech.edu/GeneMark/eukhmm.cgi), BLAT (KENT 2002), WU-BLAST (ALTSCHUL *et al.*), NCBI-BLAST (http://blast.ncbi.nlm.nih.gov) and Primer3 (ROZEN and SKALETSKY 2000) (http://frodo.wi.mit.edu/). EST sequence assembly was performed using PASA (Program to Assemble Spliced Alignments).

The initial EST alignments were performed by PASA using the built-in GMAP algorithm option (WU and WATANABE 2005). The GreenGenie2 program is based on the latest version of the Genie genefinder (KULP 2003) (http://brl.cs.umass.edu/Research/GenePredictionWithConstraints). Genie implements a general hidden Markov model (gHMM) to predict protein-coding regions in genomic DNA. The most recently published gHMM genefinder trained specifically for *Chlamydomonas* is GeneMark.hmm-ES 3.0 (LOMSADZE *et al.* 2005), which is used in this study as the short-sequence performance benchmark for GreenGenie2. Unless otherwise stated, all sequence alignments were performed using WU-BLAST and significant alignments are those with BLAST E-value < 1.0×10^{-5}. PASA EST assembly alignment to the NCBI non-redundant database (NRdb) was conducted by NCBI using NCBI-BLAST (default BLAST E-value < 1.0×10^{-3}). Alignment of *v4* models onto *v3* was performed using BLAT with the -fine and -maxIntron=5000 program options invoked. All primers used in this study were designed using Primer3 (ROZEN and SKALETSKY 2000).

3.5.4 Short-sequence prediction performance evaluation

The evaluation of predictions requires independent and high quality annotated test sequences against which predictions are compared to determine sensitivity and specificity statistics and a quantitative evaluation of prediction accuracy. When comparing the predicted genes for a given test sequence to the reference annotation of that sequence, the predicted structure can be evaluated at three different levels: nucleotide accuracy, exonic accuracy and whole gene accuracy (ROGIC *et al.* 2001). Whole gene accuracy is the most stringent level

because a prediction is correct only when the prediction matches the reference at every position; a single mismatched exon boundary is an error and renders the entire prediction incorrect. Nucleotide accuracy is the least stringent level; each individual nucleotide is either correctly or incorrectly labeled as coding or non-coding. At each level, predictions are classified as either true positive, false positive, true negative or false negative. True positives and true negatives are those regions where the predicted structure agrees with the reference annotation in coding and non-coding regions respectively. Conversely, false positives and false negatives are those regions where the predicted structure does not agree with reference annotations in non-coding and coding regions respectively. Sensitivity is defined as the ratio of true positives to actual positives. Greater sensitivity on the gene level indicates that the prediction method being evaluated misses fewer genes. Specificity is defined as the proportion of all predictions that are true positives. Greater specificity at the gene level indicates that there are fewer wrong predictions being made by the prediction method under evaluation. By determining the different relative ratios of each of the four categories above, it is possible to gauge the inherent accuracy of a set of predictions and to compare the predictive performance across different sets of gene predictions. Short-sequence prediction performance of GreenGenie2 is performed by submitting the genomic sequences corresponding to each of the 140 reference annotations in *gb140* to both GreenGenie2 and GeneMark.hmm-ES 3.0. Each sequence yields a single set of predictions from each of the gene-predictors. Standard averaged sensitivity and specificity ratios are computed on the nucleotide, exon and gene levels by the Tally.pl and BaseCounts.pl, utilities that are included as a part of the

Genie software package (http://brl.cs.umass.edu/Research/GenePredictionWithConstraints). Statistical significance of differences between two ratios is computed by a two-proportion *z*-test that compares the corresponding ratios for a given confidence level from each of the two independent predictions. All such comparisons in this study are computed using a confidence level of 0.99.

3.5.5 Interval overlap analysis

Whole-genome predictions are compared using interval overlap analysis of predicted models and evaluated for accuracy and complementarity. The interval overlap analysis of gene features is performed by directly comparing two lists of coding sequence coordinates indexed on a common genome assembly. Coding nucleotides are classified as either overlapping or not overlapping. A coding nucleotide is overlapping if and only if that position is annotated as coding in both predicted models, otherwise the nucleotide is not overlapping. Exons are classified into three classes: exact, partial and novel. An exon for which every nucleotide is aligned is classified as an exact overlap. An exon that is not classified as an exact overlap but has at least thirty consecutive bases that overlap is classified as a partial overlap. An exon that is neither exact nor partial is classified as extra in the original catalog and absent in the other catalog. A gene is classified into three classes: exact, partial and novel. A gene for which every exon is classified as exact is classified as exact. A gene for which every exon is classified as novel is classified as novel. All other genes are classified as partial, which indicates that the two predictions overlap but are not identical. Differing

predictions between two catalogs can then be targeted for subsequent testing via RT-PCR and other *in silico* validation methods.

3.5.6 PCR and RT-PCR

A small subset of novel predictions with non-exact overlaps was tested by RT-PCR. Two classes of predictions were tested: predictions that overlap but are not exact and predictions that are exclusive to each catalog. To verify exons whose intron boundaries do not agree between two catalogs, one primer aligns to the overhanging region of each of the two partially aligned exons and the other primer aligns to a nearby exon that is exactly overlapping between the two catalogs. RT-PCR with these primers unambiguously indicates which prediction (if either) is correct, or whether both predicted genes are correct and arise from alternative splicing. The designed primers were also used in genomic DNA PCR to verify that they amplify the correct regions of interest. For genomic DNA PCR, crude *Chlamydomonas* DNA was prepared. A toothpick-tip-full of *Chlamydomonas* cells was lysed in 10 μL lysis buffer (10 mM Tris-HCl, pH 8.8, 50 mM KCl, 2 mM $MgCl_2$, 0.1% Triton-100, 1mg/mL proteinase K) at 58°C for 1 hr followed by 95°C 30 min to denature the proteinase K. Cell debris was collected by a 10 sec centrifugation and 0.5 μL of the supernatant were used in a 10 μL PCR reaction. Total RNA from wild-type vegetative *Chlamydomonas* cells was prepared as previously described (LIN and GOODENOUGH 2007). Total RNA (30μg) was treated with 2 units of RNase-free DNase I (New England Biolabs, Ipswitch, MA) to remove contaminating genomic DNA from the sample. One μg of total RNA was used for cDNA synthesis with or without the addition of

SuperScript II reverse transcriptase (Invitrogen, Carlsbad, CA) in a 20 µl reaction. The same reaction mix without reverse transcriptase serves as the control for the presence of genomic DNA contamination. 0.5 µL of cDNA synthesis products was used in a 10µl PCR reaction with RedTaq DNA polymerase (Sigma, St. Louis, MO) according to the manufacturer's protocol. PCR conditions used were the following: 95°C 2 min, followed by 30 cycles of 95°C 15 sec, 53°C 15 sec, and 72°C 1 min, and ending at 72°C for 2 min.

3.6 Supplemental Tables

3.6.1 Supplemental Table 1 NCBI Protein ID codes for *gb140*

AAB23258.2	AAM01186.2	CAA37638.1	AAD28474.1
AAK77552.1	AAR04931.1	AAC49416.1	AAK06774.1
AAF43040.1	CAD60538.1	AAG30934.1	AAM15777.1
AAK32150.1	AAQ12259.1	AAK68064.1	AAO45104.1
AAR20884.1	AAT37069.1	AAK70872.1	AAO48940.1
AAY86155.1	AAC03784.1	AAK70874.1	AAQ16277.2
CAA48233.1	AAG45420.1	AAK84866.1	ABC02019.1
AAA82610.1	AAO25117.1	AAL31495.1	AAG29840.1
AAB71841.1	AAK14648.1	AAN77901.2	AAT40991.1
CAE17329.1	AAM19664.1	AAS07042.1	AAN87017.1
AAB71840.1	AAQ95705.1	AAS89977.1	AAQ19847.1
AAK01720.1	AAA57316.2	CAA65356.1	AAA84971.1
AAK82666.1	AAB39840.1	CAC19676.1	AAP30010.1
AAR23425.1	AAC49887.1	ABC49916.1	AAR82947.1

AAF65221.1	AAC49888.1	AAP12520.1	CAB56598.1
AAL75576.1	AAD45352.1	AAP12521.1	AAK38270.1
AAC08533.1	CAD24295.1	AAM44041.1	AAK54060.1
AAC08534.1	AAO86687.1	AAG45421.1	AAF34540.1
AAD39433.1	AAD27871.1	AAK37411.1	AAG33634.1
AAG40000.1	AAF17595.1	AAM88388.1	AAM23012.1
AAP21826.1	AAF73174.1	AAY56335.1	
AAQ83687.1	AAM15771.1	CAA41039.1	
AAD10324.1	AAR82949.1	AAW67003.1	
AAM88387.2	AAD27849.1	ABG38184.1	
CAF25319.1	AAL28128.1	AAM18057.1	
AAK32117.1	ABK56835.1	AAQ16626.1	
AAL35726.1	CAA44066.1	AAD55941.1	
AAL79816.1	AAB60274.1	ABK34486.1	
AAF36402.1	AAO53242.1	AAG37909.1	
ABB88568.1	AAP57169.1_v1	AAP83163.1	
AAB95196.1	AAP57169.1_v2	CAD32174.1	
AAM23259.1	AAY56333.1	AAT38474.1	
AAM23262.1_v1	AAY56334.1	AAT38475.1	
AAM23262.1_v2	AAL37900.1	AAB00730.2	
AAM44130.1_v1	AAP85534.1	AAD52203.1	
AAM44130.1_v2	CAE46409.1	AAC37438.2	
AAQ55462.1	AAO61143.1	AAD38856.1	
AAS07044.1	AAL73208.1	AAM43910.1	
AAG33633.1	AAN01224.1	AAD50464.1	
AAK77219.1	AAC27525.1	AAK14341.1	

3.6.2 Supplemental Table 2 Primers to verify PASA assemblies

Assembly ID	Left Primer	Right Primer	Pred. Length

3146_3724	GCC GCA ACA CTG TTT GTG TA	AAA GCA TGT GTC CCC TCG T	138
5172_6168	TGC ACT AAG TCC GAA CAC GA	CCA TGT AGG CGG GAG AGT AA	143
8132_9749	AGA GCA AGC GAG TTC GAG AG	GTG AGC AAA GGC ACT TAG GC	136
9104_10933	GCC GAA ATT CCA AGT CAA GA	TGC CTG GTG TAA TCG TGG TA	168
9866_11843	CCA AGT GCC ACT CCA TAG C	ATC GTG GAC TGA GCG GTG T	130
11161_13363	CCC ACA AAC ACA TGA GAA TCC	TCC AGT GCA GTT CCA TCT GA	169
11240_13451	CGG AGT GAC CAA TAG GGT TC	CAC CTC GAG GCT TAG CTG TC	149
11709_14017	ACC ACA CCT TTT TGC GGT AA	GAT GCA GTG TGG CAG AGG TA	139
14828_17825	GTC TGG TAG CTT CCG AGC AG	ACC CCC TCA GGA ACG TGT AT	139
16095_19351	TAC TAC GAT GCG GAT GTG GA	GGA TTT GGT TCA GGG AGG AG	150
14105_16951*	AGA CAT GAA CGT CCC CTC AC	CAG CGC AAC TCT GAC AGA CA	158
15620_18773*	GGT TGT ATA CGC TGC TGC TG	GGC AAA GCC TAC ACA GCT TC	150
14205_17074*	TCT TCT CGT TTA GCG CGT TT	CGC ACG CTA TAC GTC TCT CC	147

*failed to yield predicted product

3.6.3 Supplemental Table 3 primers to verify partial overlap exons

Gene ID	Left Primer	Right Primer	Pred.Length
4t254	ACA ACG GCA CCA TCA TCA AT	GCC GGT TAC GGT GAT GTT	123
4t254_143087†	CTA CAA CGG CAC CAT CAT CA	AGC CAG CGT GCC GTA CTC	103
11t344	GTT CTG CTG CCT CTG GTC AT	GTC CCA CTC GAC CCT CCT	100
11t344_169877†	GTT CTG CTG CCT CTG GTC AT	TTG ATT GCG TCA ATG GAA AC	105
25t123	GTG TCC ATC TGC CTG CAC	TTC AGC GGG CAC ACA TTT AC	90
25t123_104389†	GTG TCC ATC TGC CTG CAC	TGT GCA CTT GCA ATG GAG TAT	106
24t200	AGA TGA TTG TGT TCC GAC AGG	GGC GTC GCT TAC GTC CAG	104
24t200_195571†	CCC CTC CTA CCA GAT GAT TG	GTT TGG GTG AAA GCG GAC T	100
5t126*	ATC TCT TCA CGG CAC CTT C	TGT GTG CAG GTA AGG GTG AG	148
5t126_186782†*	ATG TCT TCA CGG CAC CTT C	GGG GAT GGC TGT CAT GTA CT	143

*failed to yield predicted product
†*gg2v3* gene id and corresponding protein ID in *FGC07*

3.6.4 Supplemental Table 4 Primers to verify novel *gg2v3* exons

Gene ID	Left Primer	Right Primer	Pred. Length
1t16	GCG TAT CGC CCA AAT GAA	GCG GTG ATG ATG TGT TTG TC	100
1t34	ACG AGG ACG ACT ACG ACG AC	GTC CTT GAG AAG GCG GAA C	102
1t147	CTG GTG TCC GTG TAC ATT GC	TCG GGT GCC ATC CAG TAG	198
11t344	ACC GAC TGC GAA GAC TGT G	CCT TGC TCT GCA GCA ACC	107
15t291	CCT GAC GCC TAC GAC AAG TT	GGA ACA CGG ACT CCA GAG C	128
30t106	ACA ACC AGT CGC AGA AGG AG	CTG TCC ACA GCT CTG ACG TG	181
30t170	CAT TGG AGA CCA GGA CGA G	GTC TCG CGT GTG AGT GTT TG	106
3t257*	GTC ACC GCG GAC CTA CTG	GAC TCT CAG CAG CTT CTC TCG	140

*failed to yield predicted product

3.6.5 Supplemental Table 5 Primers to verify novel *gg2v3* genes

Gene ID	Left Primer	Right Primer	Pred. Length
3t69	CAG CTC CAC CAA CAA CGA G	ATC ACC ACC AGC TTG CTG TC	115
19t170	GCT GGT GCT GGT GTT AAA TG	GTG TCC GCT AGC CGC TTA AT	136
30t189	ATC AGC CTG GAG GAG CTG	TGA CAC CGT GGA TCT TAC ACA	119
76t11	CCT GGG CTG GGA CTT TTC	GTC CTG GTA GCG CTC ACA TC	110
69t65*	AAC TCC GGG AGC TTT ACA CA	TTT GGA CCA AGA CCT GAA GC	108

*failed to yield predicted product

3.6.6 Supplemental Table 6 Primers to verify *FGC07* exclusive genes

Gene ID	Left Primer	Right Primer	Pred. Length
141597	GTG CAA CTC GGC CTG GAT	GTG GGC GAG AAT GTG GTT AG	103
181956	CCT GAA CTG CAT CAT CCA CA	ATC ATG ACC TCA CGC GTC TC	152
184911	GCG CAG GCA TTA CAG GTC	GGA GCC TCC TGG TGA TGA G	112
141023*	GTG GAT CCC GAG GCT GTC	ATG CCG ACA TCG TGA ACT G	104
180935*	GTG CTG TCC AGG CAA AGG	TGC TAG CAG CTC TGA CAC CT	168

*failed to yield predicted product

Chapter 4

Detecting co-evolution for protein annotation

Note: Portions of the results in this chapter are published in Kwan AL, Dutcher SK, Stormo GD: Detecting Coevolution of Functionally Related Proteins for Automated Protein Annotation. Proc. 2010 IEEE Int. Conf. on Bioinformatics and Bioengineering, pp. 99-105.

4.1 Introduction

The relationship between the genes and the observable traits of a given organism is mediated by the function of the protein products of the genes in question. Interactions between individual amino acids are conserved across instances. Therefore, proteins that have similar sequences also fold in a similar

manner and presumably have similar functions. This relationship between structure and function is the basis of sequence similarity-based protein annotation methods. These homology methods infer knowledge about a new protein from knowledge about a known protein with a sufficiently similar amino acid sequence. The organization of proteins into so-called protein families facilitates the association of new proteins with known protein families by sequence similarity, which facilitates the transfer of knowledge from existing annotations to novel proteins. The extent of automated protein characterization made possible by such methods is largely dependent on existing knowledge about at least one protein in every protein family. As a result, a large proportion of protein families remain uncharacterized beyond sequence similarity (JAROSZEWSKI *et al.* 2009; KARIMPOUR-FARD *et al.* 2007).

The fact that proteins rarely act in isolation suggests an extended annotation approach where the function of a known protein can inform the user about the function of a novel protein based on its functional *context* (PELLEGRINI *et al.* 1999). The phylogenetic profile comparison (PPC) class of automated protein characterization methods operates on the premise that members of protein networks co-evolve to preserve functional compatibility and that similar patterns of protein occurrence across sets of diverse species evidence instances of protein co-evolution (PAZOS *et al.* 2005). Typically, a PPC method proceeds as follows: for each protein in a proteome of interest, the presence or absence of an orthologous sequence is determined in each of the reference proteomes that a user has selected, and an occurrence profile of each protein is constructed. This is followed by a pair-wise occurrence profile comparison step. Proteins occurrence

profile pairs that satisfy some criterion of occurrence profile similarity are predicted to have co-evolved to maintain functional compatibility. PPC methods tend to differ in how orthologs are detected and how occurrence profiles are compared. For ortholog detection, certain methods use a similarity score cutoff to determine the existence of an ortholog (KARIMPOUR-FARD *et al.* 2007; LI *et al.* 2004; LI *et al.* 2005; SUN *et al.* 2005), while other methods use pre-computed ortholog clusters (COKUS *et al.* 2007). Each method has its strengths and drawbacks. For profile comparison, reported schemes range from simple Hamming distance (PELLEGRINI *et al.* 1999) to phylogeny-based maximum-likelihood methods complete with an internal model of gene evolution (BARKER *et al.* 2007). Combinations of the more straightforward solutions to both problems have made existing methods particularly applicable to prokaryotic proteomes (BARKER *et al.* 2007; KARIMPOUR-FARD *et al.* 2007; SUN *et al.* 2005), while the development of PPC methods focusing on eukaryotic species remains largely unexplored (BARKER *et al.* 2007; LI *et al.* 2004).

PPC methods aim to characterize proteins by extracting information for a protein of interest from its compatibility context by leveraging the strength of the association relating co-evolution and profile similarity. The use of reference at varying evolutionary distances to the proteome of interest is integral for the successful application of any PPC method. Varying evolutionary distances between species inherently introduces evolutionary biases into sequence similarity scores that confound accurate profile construction. Thus, it is imperative to normalize similarity scores for any variation in the underlying evolutionary distances between a focus species and each reference species (Figure 4.1). While

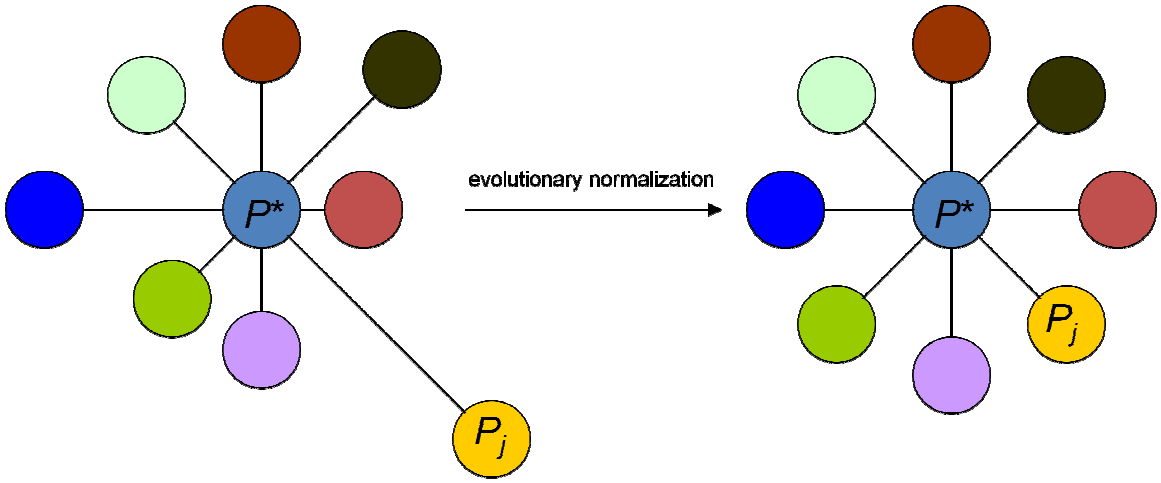

Figure 4.1 Alignment scores need to be normalized for differences in evolutionary distance. *Reference proteomes P_j evolve away from a focus proteome P* at different rates and for different lengths of time (left). Normalization of similarity scores equalize evolutionary distances (rate × time) between P* and all P_j facilitates proper comparison of sequence similarity scores across multiple P_j (right).*

gene-evolution events like horizontal gene transfer may justify the use of convenient profile comparison approaches, like Hamming distance, as found in existing methods, the same approaches to profile comparison are less applicable within the context of eukaryotic phylogenetics. Other profile comparison schemes rely on many assumptions about eukaryotic gene and species evolution that do not accurately reflect known biology.

One of the few methods to focus on eukaryotic systems is described in (Li *et al.* 2005) in which a PPC method called Procom is presented. Procom works by determining the set of proteins in a given focus proteome that has a detected ortholog in every species classified as positive for a trait of interest and no detectable orthologs species classified as negative for the same trait of interest (Li *et al.* 2004). An ortholog is detected if the BLASTP E-value of the best-hit to a given focus protein in a given reference proteome is less than the significance

cutoff value of 1E-10. Li and coworkers (Li *et al.* 2004) demonstrate the effectiveness of Procom by identifying and characterizing novel cilia proteins in the biflagellate, green alga *Chlamydomonas reinhardtii* (Li *et al.* 2004), in which the trait of interest is the presence or absence of cilia. *Homo sapiens* is the species positive for the trait and *Arabidopsis thaliana*, an unciliated angiosperm, is the negative species. Procom was used to define the well-established Flagellar and Basal Body proteome (Li *et al.* 2004). Among many other cilia and basal body related proteins, Procom is responsible for the characterization of *BBS5*, a new Bardet-Biedl Syndrome disease gene.

This chapter presents a new PPC method called APACE (Automated Protein Annotation by Coordinate Evolution) based on a novel similarity score normalization process and ortholog detection approach that automatically clusters proteins without requiring any additional profile comparison scheme. Our novel normalization function adjusts sequence similarity scores to equalize the evolutionary distance between a focus species and each reference species (Figure 4.1). Furthermore, the APACE is able to organize proteins into co-evolving groups without any additional profile comparison scheme.

4.2 The Approach

In this section, the input to any PPC method is taken to be a set of N+1 proteomes consisting of a focus proteome P^* and a set of N reference proteomes labeled P_1,

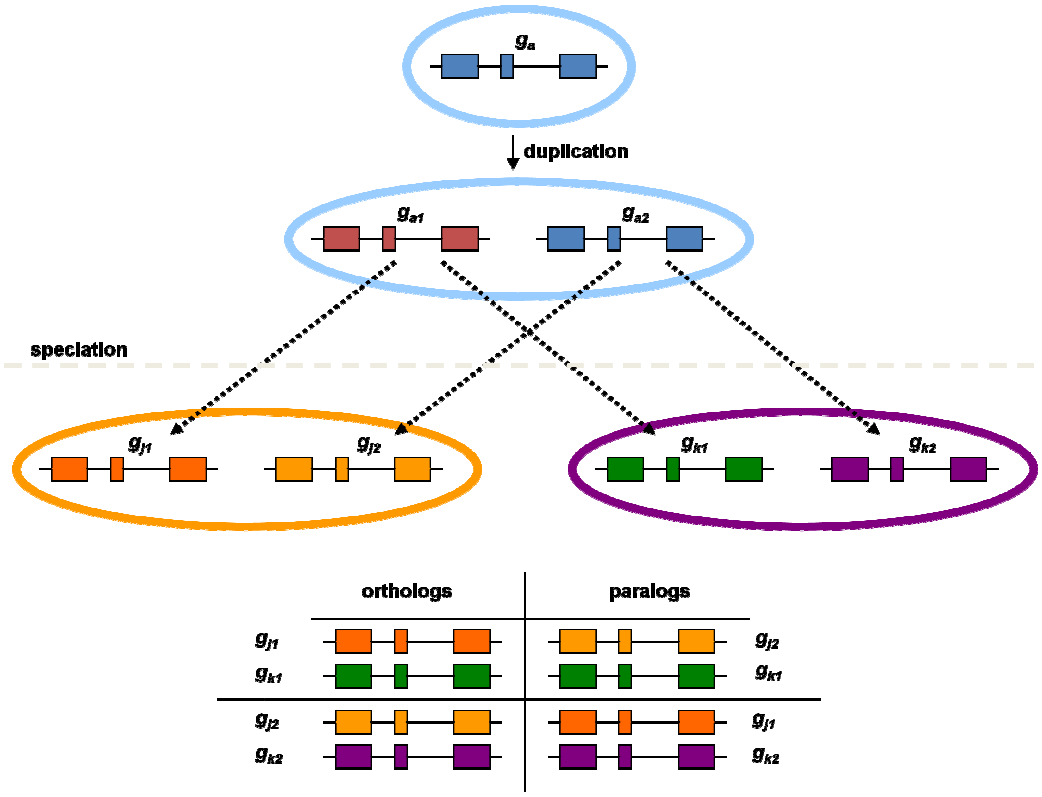

Figure 4.2 Different evolutions of orthologous and paralogous sequences

P_2, ..., P_N. Proteome P^* is made up of the appropriate number of individual proteins p_i encoded by gene g_i in genome G^*. The protein in reference proteome P_j that is most similar in sequence to a given protein of interest p_i in P^* is the "best-hit to p_i from P_j" and is denoted by p_{ij}. The degree of sequence similarity between p_i and p_{ij} is quantified by a similarity score s_{ij}.

Orthologs are genes g_{j1} and g_{k1} from two different species J and K that evolve from a common ancestral gene g_{a1} through speciation from their last common ancestral species (Figure 4.2). Proteins encoded by orthologous genes are assumed to retain the same function in J as in K; that is, orthologous sequences are constrained to mutate within a functionally equivalent sequence space. Paralogs g_{j1}, g_{k2} and g_{j2},

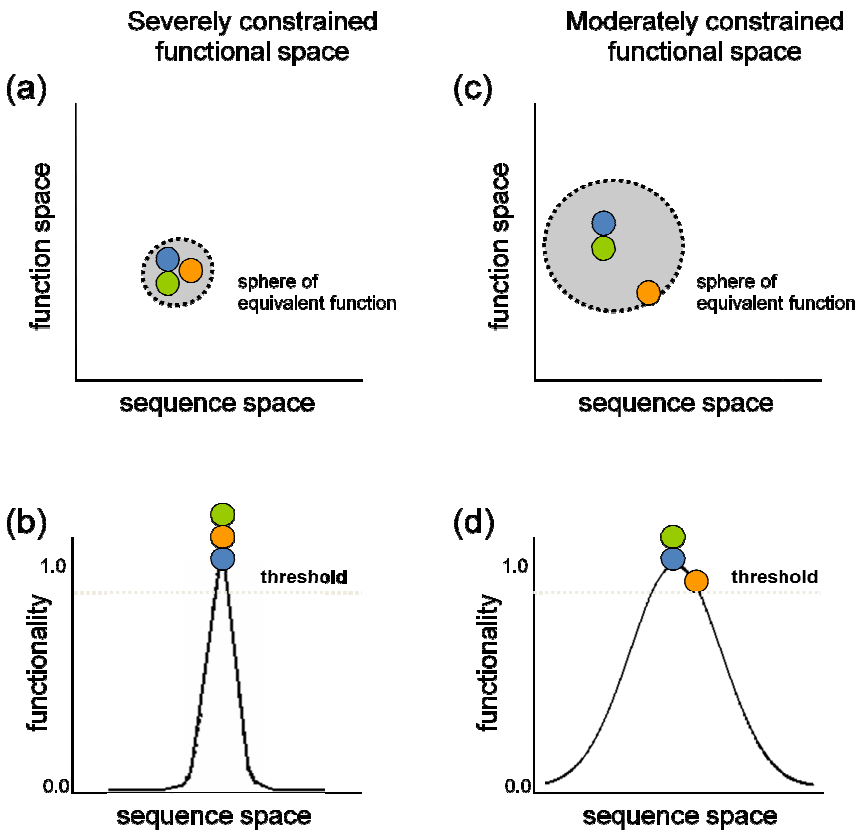

Figure 4.3 Protein evolution in terms of sequence and function space. *Different proteins have function-specific spheres of equivalent function in function-sequence space, which determines the sequence space within which a protein may mutate and still retain the same function as the ancestral sequence.*

g_{k1} are genes that evolve from duplicate ancestral genes g_{a1} and g_{a2} and often do not retain the same function across species; that is, paralogs are free to evolve outside of their functionally equivalent sequence space (Figure 4.3). This relationship between orthologs and paralogs implies that protein p_i will be more similar in sequence to an ortholog than to any other paralogous sequence in an arbitrary reference proteome P_j, which suggests that a superset of all orthologs to a protein of interest p_i in a set of reference species can be generated by identifying the proteins p_{ij} over each of the N reference proteomes P_j. This set is a superset of

the orthologs of p_i because not every P_j necessarily contains an ortholog of p_i. In the case where a P_j does not contain an ortholog to p_i, the best-hit p_{ij} would still be returned. Note that such cases are not necessarily handled correctly by reciprocal-best-hit ortholog detection schemes. Orthologs can be extracted from a superset of best-hits provided that there is a way to separate orthologous best-hits from paralogous best-hits.

The sequence, structure and function of a given protein are intimately related characteristics. Evolving proteins can be viewed as moving points within a sequence-function space in which every biological action performed by a protein defines an associated sphere of equivalent function. In this space, perturbations in the sequence that do not result in a loss-of-function place an extant protein point within the functional sphere of the ancestral protein (Figure 4.3a). Perturbations in a sequence that greatly affect function place an extant protein point outside the functional sphere of the ancestral protein. To visualize this relationship we plot functionality against sequence space (Figure 4.3b). Proteins that occupy steep functionality curves cannot diverge significantly from a functional ancestral sequence without falling below some equivalency threshold of functionality (Figure 4.3b). Other genes encode proteins that can withstand greater degrees of perturbations will result in orthologs that mutate within a more relaxed sequence space (Figure 4.3c). In terms of functionality, functions with moderate sequence constraints allow for a larger variety of protein sequences to carry out equivalent function (Figure 4.3d). In such a case, there arise some sequences that result in conformations more functionally favorable than other

sequences, but as in the first case, at a certain point, the degree of functionality drops below some critical threshold and the original function is lost.

Orthologs are proteins from different species that carry out the same function by remaining within the functional sphere of the common ancestral protein. Proteins under greater functional constraints likely have a lower tolerance to sequence perturbations than proteins under fewer constraints. This further reduces the degree of sequence space within which functionally equivalent orthologs can evolve. Paralogs, by contrast, are free to evolve outside this doubly constrained sequence space suggesting the assumption that, over the magnitudes of evolutionary time considered by our method, paralogous proteins have had ample opportunity to evolve into the wider sequence space and no longer occupy the same functional space. The sphere of equivalent function is specific to each individual protein; that is, different proteins evolve at different rates and perform functions that are tolerant of different degrees of perturbation. Thus, while one may define a cutoff similarity score for each protein individually, it is impossible to correctly define a universal cutoff similarity score for every protein.

Evolutionary distance between the two source species and functional equality are the two principal factors that influence the degree of similarity between any two proteins. PPC methods rely on the accurate detection of functionally equivalent orthologs across many species. Therefore, a critical step in any PPC pipeline ought to be the normalization of similarity scores for the effects of different evolutionary distance. Proteins that are least tolerant of sequence perturbations are presumably the most functionally constrained. For these

proteins, we expect a high degree of inter-ortholog similarity and any observed dissimilarity is primarily a reflection of the evolutionary distance between the two source species. Following this rationale, sequence similarity scores of the most widely conserved proteins across multiple phyla have been used to infer branch lengths of the phylogenetic trees. Our method extends this rationale by equating the evolutionary distance between P^* and each P_j as the average s_{ij} of the most widely conserved proteins in P^* and each P_j. Our normalization determines a normalization factor r_j for each reference proteome P_j that is inversely related to the distance between P^* and P_j. The idea is to calculate an adjusted similarity score a_{ij} as the product of s_{ij} and r_j, which equalizes the evolutionary distance between P^* and every P_j (Figure 4.1).

Orthologous sequences can be extracted from a set of best-hit sequences containing both orthologous and paralogous sequences by leveraging the observation that orthologs always evolve with a relatively more constrained sequence space; the orthologous best-hits can be distinguished from paralogous best-hits by their greater degree of similarity to the reference protein p_i than paralogous best-hits. Discriminating orthologs from paralogs in a set of best-hits can be intuitively interpreted as a simple clustering problem. A best-hit set can be represented as a list of similarity scores for each p_{ij} for each P_j sorted in decreasing order of adjusted similarity according to a_{ij}, suggesting that the problem be solved by some flavor of k-means. The critical observation here is that the sorted similarity scores for every p_i in P^* will be a mixture of three classes of score distributions. Ideally, similarity scores from orthologous best-hits will form a

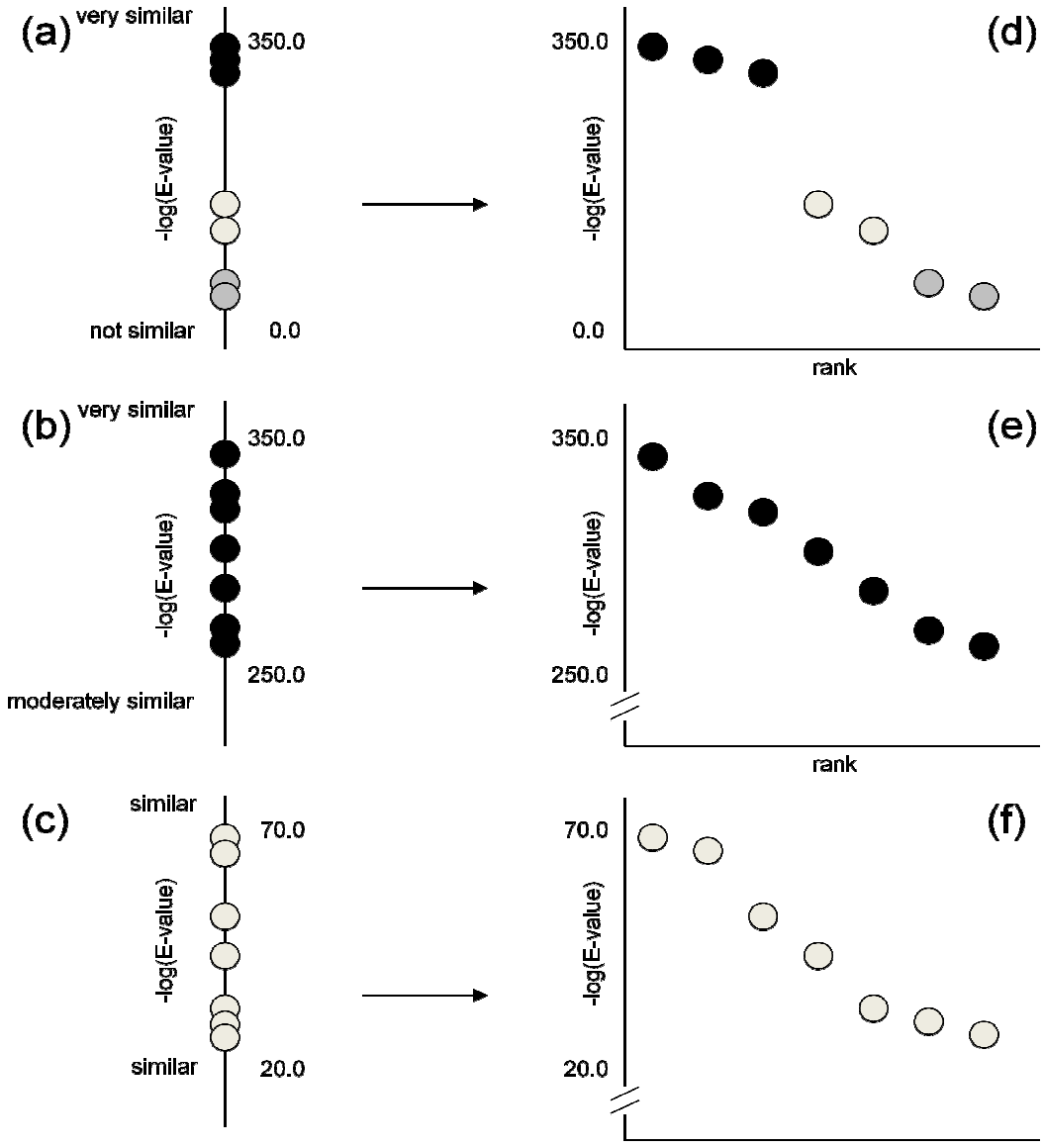

Figure 4.4 Three classes of similarity score distributions in 1D and their corresponding 2D spreads. *(a) Orthologs form a distinct cluster of very similar sequences and the 2D spread has a distinct ortholog-paralog gap; (b) Orthologs exist across all species forming a single high-scoring cluster and the 2D spread is roughly linear across the 2D spread; (c) Orthologs and paralogs are not easily distinguished in 1D while in the 2D spread an inflection point can still be detected.*

cluster of high scoring best-hits within the set (Figure 4.4a). Proteins that are widely conserved across all reference species are another class of score distributions that form a single cluster with similarity scores scattered over a

small range in a roughly uniform manner (Figure 4.4b). A third class of score distributions arise when proteins have orthologous and paralogous best-hits scores that are not as clearly defined as the ideal first class and yet not as uniformly distributed as the second class. This mixture of score distributions precludes an *a priori* determination of the requisite constant *k* for a *k*-means solution for all proteins in P^* (Figure 4.4c).

We propose a novel solution to this problem by defining a 2D "spread" for each list of 1D sorted scores. The 2D spread of any sorted list is constructed by introducing an axis of decreasing rank that is orthogonal to the native axis of real valued similarity scores (Figures 4.4d-f). The units on the new axis are the rank of the score within the scores for a given p_i. Because any list of scores has an implicit ranking, the property used to cluster the scores is the inter-cluster versus intra-cluster differences in 1D. Constructing the 2D spread of a list of scores translates a large difference between two scores into a line segment with a steep negative slope (Figure 4.4d) and a small difference between two scores into a line segment with a shallow negative slope and (Figure 4.4e). Thus, clusters of orthologs will form approximately linear sub-profiles that begin at the first rank position (leftmost position on the rank axis) in a 2D spread. The problem of determining whether a group of scores form a cluster in the 1D list as a whole can be reduced to determining the rank after which there is an inflection point in the 2D spread. Our method determines the appropriate inflection point by computing the second forward derivative of the 2D spread at every rank. To mitigate the effects of spurious noise in the spreads, the method takes the averaged second derivative over rank *t*, *t*+1 and *t*+2 as the smoothed second-forward derivative at *t*. The

method selects the leftmost rank t^* with a smoothed second forward derivative that is greater than or equal to zero to be the lowest ranked species with an ortholog to p_i.

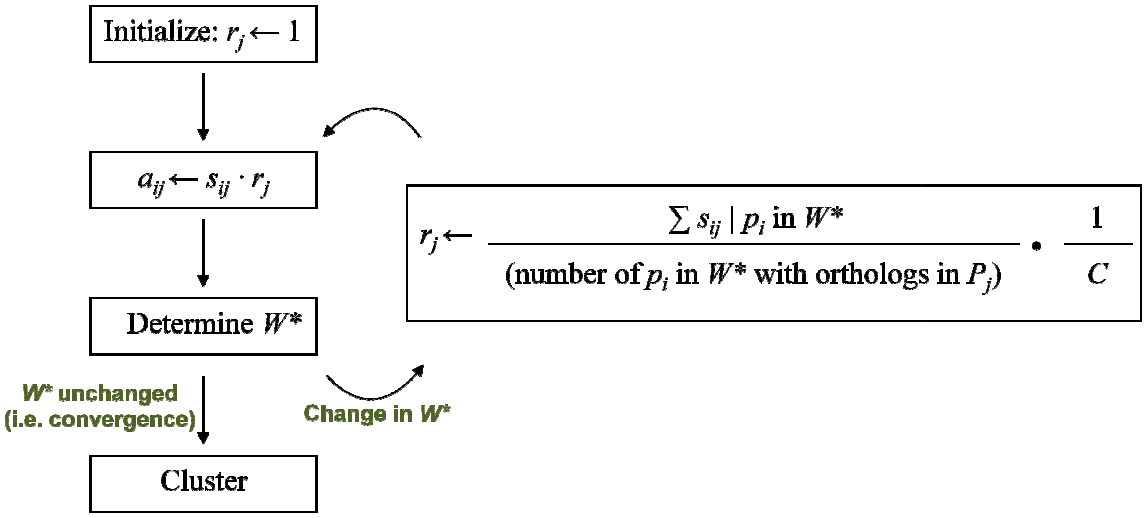

Figure 4.5 A flow diagram of APACE

Our method, called APACE, proceeds as follows: First, the method determines a normalization factor r_j for each reference proteome P_j that is inversely related to the distance between P^* and P_j. These factors are computed by determining the multiplicand that equalizes the average best-hit similarity score of most widely conserved proteins in P^* between all P_j. Let the set W^* be the subset of proteins in P^* with orthologs in at least N - ε reference proteomes for some small ε. W^* is the set of widely conserved proteins in P^*. The algorithm incrementally converges on the appropriate value for each r_j by initiating each r_j to 1 and use our profile inflection-point ortholog detection method (see above) to

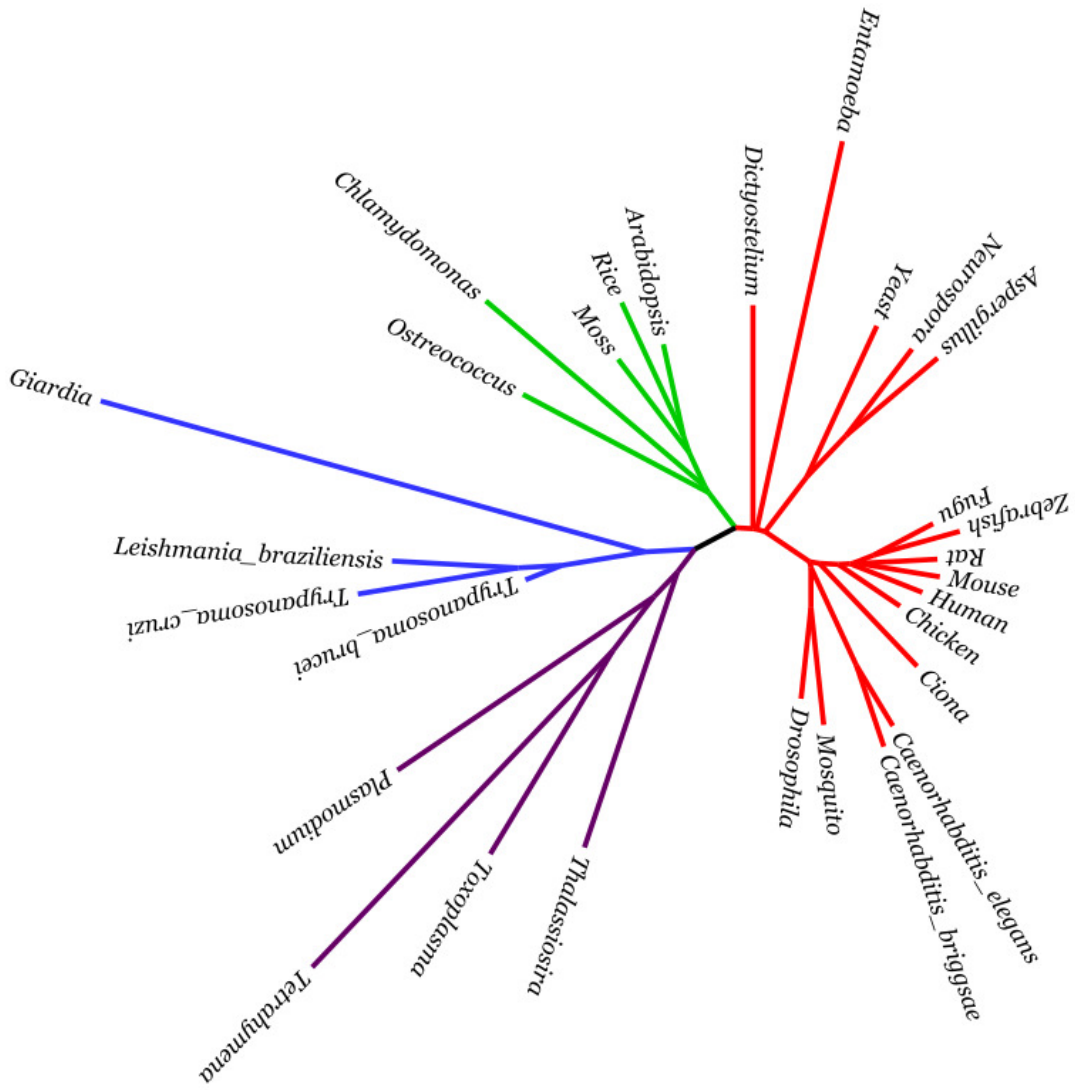

Figure 4.6 The phylogenetic tree determined by APACE. *A rendering of the phylogenetic tree of eukaryotes using FigTree (RAMBAUT 2007) constructed by FastME using interspecies distances computed based on the normalization factors r_j as determined by the normalization phase of APACE. The normalization factors facilitate the quantification of previously unresolved branches (KEELING et al. 2005) and the resulting tree topologically recapitulates the composite deep eukaryote tree presented in (KEELING et al. 2005). Represented supergroups are 'Unikonts' (red), Plantae (green), Excavates (blue) and Chromalveolates (purple).*

determine an initial W^* with unadjusted scores. Every r_j is then reassigned the ratio of the average s_{ij} of every p_i in W^* for each P_j to some globally constant value (i.e. the unit branch length in the balanced star topology). Every a_{ij} of every

protein p_i is then recalculated with the updated r_j and the next iteration begins (Figure 4.5). Convergence is reached when the composition of W^* remains unchanged or a maximum number of iterations has been reached. Empirically, convergence is reached within four iterations for the 29 eukaryotic test species analyzed in this study. After convergence is reached, the similarity scores are already appropriately adjusted to normalize evolutionary distance and the phylogenetic profile clustering of p_i by the occurrence of functionally equivalent orthologs is performed using the inflection point analysis method described above. The profile of each p_i is determined as the species from rank "1" down to and including the leftmost rank in the 2D spread with a smoothed second derivative greater than zero. Proteins that exhibit the same phylogenetic profile are predicted to have co-evolved, presumably to maintain functional compatibility.

4.3 Results and Analysis

APACE introduces two novel methods for its analyses of proteomic data from multiple eukaryotic species. The first is the normalization of similarity scores for differences in evolutionary distance between a focus proteome P^* and each of the N reference proteomes P_j. Unlike conventional methods that normalize similarity scores based on branch lengths of phylogenetic trees constructed using a single gene or a small set of genes, APACE makes use of as many widely conserved proteins as possible in determining a species-specific normalization

factor r_j that is inversely related to the evolutionary distance between P^* and each P_j. In the current example of 29 species, the interspecies distances are computed based on approximately 1,600 highly conserved proteins, varying in a species specific manner. The second novel method introduced by APACE is in how orthologous sequences are detected from a superset of best-hit sequences from each of the reference species from the 2D spread of a list of best-hit similarity scores. To validate the normalization approach introduced by APACE, we ask whether the phylogenetic tree generated from the inverse values of APACE normalization factors is able to topographically recapitulate the unresolved, deep eukaryotic phylogeny tree recently presented in (KEELING *et al.* 2005). To demonstrate the flexibility and robustness of the novel ortholog detection and phylogenetic profile construction approach introduced by APACE, we present results of five analyses designed to identify proteins with specific functional classifications. First we present a small-scale example query comparing APACE to the method documented in (LI *et al.* 2005) for identifying proteins involved in cilia motility with distantly related species. We demonstrate the scalability of APACE in comparison with existing methods, we ask both APACE and Procom (LI *et al.* 2005) to identify a list of human proteins that have co-evolved in a large number of multicellular metazoan and plant species. We demonstrate the robustness of APACE against false positives when using small numbers of more closely related query species by focusing solely on the malaria causing *Plasmodium falciparum* and the toxoplamosis causing *Toxoplasma gondii*, both members of the Apicomplexa phylum (Figure 4.6) to identify proteins that are important to the life cycles of these pathological parasites. We demonstrate the

accuracy of APACE by comparing interaction predictions and experimentally determined co-crystallization pairs from the BioGRID database (STARK *et al.* 2011). Finally, we directly demonstrate the ability of APACE to relate multiple proteins with dissimilar sequences by investigating whether APACE is able to predict cargo proteins of human RAB vesicular transport machinery.

To validate our normalization procedure, we analyzed 29 eukaryotic organisms, which span multiple phyla and supergroups (KEELING *et al.* 2005). We construct a 29 by 29 matrix of distances from the average similarities r_j as described in (FENG and DOOLITTLE 1997) and use the minimal evolution with ordinary least squares tree-building method FastME described in (DESPER and GASCUEL 2002) to build a tree from the distance matrix and compare our generated tree to a published composite tree (KEELING *et al.* 2005) (Figure 4.6). We confine our comparison to the gross topology because the tree in (KEELING *et al.* 2005) contains unresolved branches. A comparison shows that both trees share identical topology from the most general level (i.e. supergroups) down to the most specific level presented in (KEELING *et al.* 2005). The identical topologies of the generated tree and the reference tree indicate that our method successfully computes normalization factors for 29 widely divergent eukaryotic species. To test whether the topological identity is due to the strong bias in the 29 test species for species belonging to the 'Unikonts' supergroup (KEELING *et al.* 2005), we removed closely related species from metazoa leaving the same number of 'Unikonts' species as there are species from Plantae. The topological identity of the resultant "unbiased" tree remains unchanged (data not shown) and demonstrates that our

normalization method is robust against different species biases and sizes of different reference species sets.

We evaluate our predictions by comparing APACE to Procom (LI *et al.* 2005). We target *C. reinhardtii* proteins responsible for cilia motility in our comparison with Procom because one class of ciliopathies results from immotile cilia (e.g. primary cilia dyskinesia). For this demonstration, the focus species is *C. reinhardtii*; species with motile cilia that we include in our analysis are *Homo sapiens*, *Danio rerio* (zebrafish) and *Physcomitrella patens* (a moss) and species without motile cilia included in our analysis are *Arabidopsis thaliana*, *Caenorhabditis elegans* (nematode), *Oryza sativa* (rice) and *Saccharomyces cerevisiae* (yeast).

Procom identifies 50 proteins in *C. reinhardtii* that have orthologs in all three positive species and none of the negative species and APACE identifies 65 proteins that have coevolved in species considered. We find that APACE and Procom agree for 33 proteins; 21 have cilia related annotations. Ten of the 21 are known cilia proteins by mass spectroscopy of isolated cilia (PAZOUR *et al.* 2005), four were identified previously by Li *et al.* (LI *et al.* 2004) and the remainder are likely to be involved in cilia motility as they have mutant motility phenotypes (Table 4.1). The remaining eight proteins have no previous association with cilia or cilia motility. APACE identifies 32 proteins that are not in the Procom output (Table 4.1). Three-quarters (N=24) have existing ciliary or cilia motility associations; they include *ODA7*, recently implicated in primary ciliary dyskinesia (DUQUESNOY *et al.* 2009), and *PF13,* a chaperone of dynein arms (OMRAN *et al.*

2008). This group also includes a novel *C. reinhardtii* ortholog of human BLU/ZMYND10 which is a member of the chromosome 3p21.3 candidate tumor suppressor gene cluster (YAU *et al.* 2006), which may lend further support to the recent hypotheses of a role for cilia in cancer. Two proteins are completely novel with no existing annotations; none of the remaining six proteins identified exclusively by APACE are overtly unrelated to cilia motility (Table 4.1). Procom identifies 17 putative cilia motility proteins absent from APACE output. Given existing annotations, about half (N=9) have existing ciliary or cilia motility associations, while the remaining eight proteins seem unlikely to be involved with cilia motility given existing annotations (Table 4.1). These results demonstrate that APACE is able to contribute novel characterizations of proteins that are complimentary to existing methods and that it identifies fewer known negatives than Procom (LI *et al.* 2004; LI *et al.* 2005).

Table 4.1 Annotation of 82 *Chlamydomonas* cilia motility gene candidates

Gene ID	APACE	Procom	FAP	FBB/MOT	MUT	Other
c2_t817	x				BUG21	
c15_t340	x				DHC11	
merc07tr_126616	x				DHC4	
c2_t684	x				DHC5	
c2_t1137	x		FAP106			
c12_t138	x		FAP52			
c1_t1171	x			FBB7	FBB7	
c14_t356	x				IDA2	
c9_t632	x				MBO2	
merc07tr_175396	x			MOT45		
c1_t606	x				ODA7	
c11_t206	x				PF2	
c3_t444	x				POC4	
c14_t408	x				DHC7	
c3_t926	x					
c6_t636	x					
s18_t58	x					
c11_t128	x					
c12_t1305	x				PF13	

c16_t318	x					
c9_t63	x					
merc07tr_187155	x					
c1_t1271	x					
c8_t16	x					ZMYND10
c5_t31	x					
merc07tr_134599	x				DHC6	
c3_t1260	x		FAP57			
merc07tr_189109	x		FAP59			
c14_t480	x		FAP94			
c11_t321	x			MOT17		
s83_t2	x			MOT40		
c2_t1145	x				PF16	
c12_t747	x	x			BOP5	
c3_t752	x	x	FAP146			
c3_t728	x	x	FAP147			
c10_t16	x	x	FAP178			
c3_t243	x	x	FAP184			
c9_t345	x	x	FAP198			
merc07tr_106450	x	x	FAP250			
c12_t374	x	x	FAP253			
merc07tr_154904	x	x	FAP263			
c16_t849	x	x	FAP73			
c7_t275	x	x	FAP82			
c7_t117	x	x		FBB10		
c8_t36	x	x		FBB11		
c16_t824	x	x		FBB18		
merc07tr_132143	x	x		FBB18		
c1_t1383	x	x			IDA7	
c1_t1343	x	x		MOT16		
merc07tr_176821	x	x		MOT4		
merc07tr_116240	x	x				
c14_t191	x	x			PSL3	
c2_t371	x	x				RIB172
c6_t876	x	x			RSP3	
c7_t369	x	x			RSP9	
c12_t670	x	x				SAS6
c7_t489	x	x			TWI1	
c10_t248	x	x				
c10_t748	x	x				
c16_t509	x	x				
c3_t541	x	x				
c6_t864	x	x				
merc07tr_117479	x	x				
merc07tr_172110	x	x				
c6_t645	x	x			VFL3	
c2_t389		x				CPLD42
c2_t619		x	FAP100			
s22_t48		x	FAP116			
c12_t770		x	FAP194			
c2_t1144		x	FAP2			
c11_t319		x		FBB9		

c6_t783		x				HBP1
c9_t127		x		MOT39		
c17_t890		x		MOT47		
c10_t307		x				SMP10
c12_t292		x				THY28
c13_t572		x				
c13_t574		x		FBB6		
c2_t1149		x				SAS10
c6_t771		x				
c6_t782		x				
merc07tr_120200		x				

FAP: Identified by two or more peptides in direct proteomic study of Chlamydomonas *cilia (*PAZOUR *et al. 2005)*
*FBB/MOT: Previously associated to cilia by comparative genomics (*LI *et al. 2004;* MERCHANT *et al. 2007)*
MUT: Genes with mutant lines in Chlamydomonas *or* Mouse *that show cilia defects*
Other: Annotations with no known cilia association

Next, we compared APACE to Procom using 29 eukaryotic proteomes to identify human proteins that have orthologs in only multicellular species (*Anopheles gambiae*, *Arabidopsis thaliana*, *Caenorhabditis briggsae*, *Caenorhabditis elegans*, *Ciona intestinalis*, *Danio rerio*, *Drosophila melanogaster*, *Gallus gallus*, *Mus musculus*, *Physcomitrella patens*, *Oryza sativa*, *Rattus norvegicus* and *Takifugu rubripes*), but that are absent from unicellular species (*Aspergillus nidulans*, *Chlamydomonas reinhardtii*, *Dictyostelium discoideum*, *Entamoeba histolytica*, *Leishmania braziliensis*, *Neurospora crassa*, *Ostreococcus tauri*, *Plasmodium falciparum*, *Saccharomyces cerevisiae*, *Tetrahymena thermophila*, *Toxoplasma gondii*, *Trypanosoma brucei* and *Trypanosoma cruzi*) with the aim of identifying proteins that are essential for multicellularity. Our method identifies

55 proteins while Procom is unable to detect any proteins conserved in multicellular species exclusively (Table 4.2). Interestingly, the set identified by our method mostly consists of extracellular matrix degradation and regulatory proteins. The family of metalloproteinase is the most strongly represented group in the set (N=10) and have been implicated in multiple tissue remodeling of many physiological and pathological processes such as morphogenesis (WISEMAN *et al.* 2003), angiogenesis (RUNDHAUG 2005), wound healing/tissue repair (GABISON *et al.* 2005), cirrhosis (LICHTINGHAGEN *et al.* 2001), arthritis (KONTTINEN *et al.* 1999) and metastasis (KURAHARA *et al.* 1999). The majority of other proteins identified contain transcription factor, signaling or transmembrane domains, potentially highlighting more specific functional subclasses that are integral for the development, maintenance and pathology of multicellular organisms (Table 4.2).

Table 4.2 Fifty-five human multicellularity gene candidates predicted by APACE

GENE	NAME	APACE	PROCOM
ENSG00000148584	A1CF	x	
ENSG00000168397	ATG4B	x	
ENSG00000183778	B3GALT5	x	
ENSG00000176022	B3GALT6	x	
ENSG00000109956	B3GAT1	x	
ENSG00000112309	B3GAT2	x	
ENSG00000149541	B3GAT3	x	
ENSG00000176383	B3GNT4	x	
ENSG00000108588	CCDC47	x	
ENSG00000113722	CDX1	x	
ENSG00000116254	CHD5	x	
ENSG00000095485	CWF19L1	x	
ENSG00000008283	CYB561	x	
ENSG00000134698	EIF2C4	x	
ENSG00000139641	ESYT1	x	
ENSG00000117868	ESYT2	x	
ENSG00000205318	GCNT6	x	
ENSG00000120251	GRIA2	x	
ENSG00000164418	GRIK2	x	
ENSG00000125944	HNRNPR	x	

ENSG00000197576	HOXA4	x	
ENSG00000170166	HOXD4	x	
ENSG00000121774	KHDRBS1	x	
ENSG00000131773	KHDRBS3	x	
ENSG00000140950	KIAA1609	x	
ENSG00000179528	LBX2	x	
ENSG00000166670	**MMP10**	**x**	
ENSG00000110347	**MMP12**	**x**	
ENSG00000137745	**MMP13**	**x**	
ENSG00000102996	**MMP15**	**x**	
ENSG00000156103	**MMP16**	**x**	
ENSG00000198598	**MMP17**	**x**	
ENSG00000008516	**MMP25**	**x**	
ENSG00000137675	**MMP27**	**x**	
ENSG00000149968	**MMP3**	**x**	
ENSG00000118113	**MMP8**	**x**	
ENSG00000112664	NUDT3	x	
ENSG00000173598	NUDT4P1	x	
ENSG00000147162	OGT	x	
ENSG00000185129	PURA	x	
ENSG00000151962	RBM46	x	
ENSG00000133135	RNF128	x	
ENSG00000082996	RNF13	x	
ENSG00000113269	RNF130	x	
ENSG00000108523	RNF167	x	
ENSG00000133318	RTN3	x	
ENSG00000141485	SLC13A5	x	
ENSG00000100678	SLC8A3	x	
ENSG00000121067	SPOP	x	
ENSG00000144228	SPOPL	x	
ENSG00000167881	SRP68	x	
ENSG00000135316	SYNCRIP	x	
ENSG00000176769	TCERG1L	x	
ENSG00000170638	TRABD	x	
ENSG00000103489	XYLT1	x	

We identify synapomorphic proteins specific to the Apicomplexa phylum to evaluate the robustness of our algorithm to small numbers of closely related query species. APACE identifies 650 genes from the malaria-causing organism, *Plasmodium falciparum* that have co-evolved exclusively with *Toxoplasma gondii*. The RMgm database (RMgmDB) is a repository of *P. falciparum* genes that researchers have attempted to disrupt and records observed parasite lifecycle phenotypes that result (JANSE *et al.* 2011). RMgmDB documents gene

disruption attempts have been made on genes encoding only 28 of the 650 proteins predicted by APACE as essential for Apicomplexa species. Twenty of the 28 attempts were successful and 16 of these have severe lifecycle phenotypes, which translates to a sensitivity measure of 80% (16/20) (Table 4.3). In contrast, the entire RMgmDB database consists of 213 genes that have been selected by experimental researchers as potential targets for disruption, of which 152 have been successfully knocked out (JANSE *et al.* 2011). Lifecycle disruption is observed for 119 out of 152 successful gene perturbations or a sensitivity of 78% (119/152) and indicates that, of the tested genes, APACE identifies a set that is comparably enriched for essential Apicomplexa lifecycle genes to human selection and demonstrates a potentially powerful new application of PPC computational methods for effective, automated drug target discovery.

Table 4.3 Lifecycle phenotypes of 28 disrupted *P. falciparum* genes

Gene	Lifecycle stage	Phenotype
MAL13P1.301	Fertilization/Ookinete	Ookinetes unable to penetrate walls of mosquito midgut wall cells; motility of ookinetes reduced by 90%.
PF14_0672	Fertilization/Ookinete	>94% reduction in oocyst development from ookinetes.
PF11_0147	Fertilization/Ookinete	Male gametocytes cannot produce gametes; no fertilization.
PFI1145w	Fertilization/Ookinete	No oocysts formed; ookinetes cannot invade midgut epithelial cells in mosquitoes.
PFD0430c	Liver stage	Reduced infectivity
PFF1420w	Liver stage	90% reduction in infectivity of sporozoites in liver.
PF14_0723	Oocyst	Sporozoite formation in oocyst is blocked.

PFA0260c	Oocyst	Sporozoite formation in oocyst is blocked.
PFC0495w	Oocyst	Sporozoite formation in oocyst is blocked.
PFL1315w	Oocyst	>98% reduction of infectivity in mosquitoes.
PF14_0067	Oocyst	Sporozoite formation in oocyst is blocked.
PF14_0532	Oocyst	Sporozoite formation in oocyst is blocked.
MAL7P1.92	Sporozoite	Sporozoites do not invade mosquito salivary gland; cannot transmit to host.
PFI0550w	Sporozoite	Sporozoites do not invade mosquito salivary gland; cannot transmit to host.
PF14_0346	Sporozoite	Not infectious in host.
PF13_0201	Sporozoite	>97% reduction of infectivity in mosquitoes.
PFE1340w	N/A	No phenotype described
PFE0825w	N/A	No phenotype described
PFC0166w	N/A	No phenotype described
PF13_0289	N/A	No phenotype described
PF11_0381	N/A	Gene modification not successful
PF14_0495	N/A	Gene modification not successful
PFE0165w	N/A	Gene modification not successful
PF11_0395	N/A	Gene modification not successful

PFL1370w	N/A	Gene modification not successful
PF08_0108	N/A	Gene modification not successful
PFE0870w	N/A	Gene modification not successful
PFE0340c	N/A	Gene modification not successful

To further evaluate the ability of APACE to identify proteins in a functional context, we compare interacting protein partners in *Saccharomyces cerevisiae* as evidenced by co-crystallization data in the Biological General Repository for Interaction Databases (BioGRID) (STARK *et al.* 2011). Co-crystallization of two proteins is the gold-standard experimental evidence for the direct binding interaction of the proteins involved. To measure sensitivity and specificity of APACE, it is necessary to define a negative set of *S. cerevisiae* protein interaction pairs. A single, replicable instance of an interaction is sufficient to establish a positive protein pair. We define a known positive test set using the 245 proteins pairs with co-crystallization evidence in BioGRID. A negative interaction pair can only be defined using two proteins that do not interact directly under any circumstance. Therefore, in order to establish a negative protein pair with absolute certainty one would have to test every possible experimental condition and environment for the interaction in question, which is infeasible. We address this negative test set problem by adopting the definition of negative interaction provided in Barker *et al.* (BARKER *et al.* 2007) where each protein of a given pair have well established functions in unrelated biological processes and are

therefore presumably unlikely to interact with one another. In total, this results in 450,000 negative protein interaction pairs from the 6,700 proteins that make up the *S. cerevisiae* proteome (BARKER *et al.* 2007). Overall sensitivity of APACE is 47%, in contrast to the 0% of the method by Barker *et al.* (BARKER *et al.* 2007). The specificity of APACE is 92%, similar to the 97% of Barker *et al.* (BARKER *et al.* 2007) (Table 4.4).

Table 4.4 Comparing APACE and Barker on co-crystallization data

	Sensitivity	**Specificity**
APACE	47%	92%
Barker *et al.*	0%	97%

Known positives (N=245): S. cerevisiae proteins that have co-crystallization data in BioGRID. Known negatives(N=450,000): S. cerevisiae proteins predicted to not interact (BARKER et al. 2007).

The Rab family of proteins helps sort different protein cargo in the cell. One of the greatest advantages of a successful PPC method over conventional protein family methods is that PPC methods are capable of capturing associations between groups of interacting proteins that have dissimilar amino acid sequences (Figure 4.7). We test the ability of APACE to relate dissimilar proteins by searching for non-Rab human proteins that have co-evolved with Rab proteins to see if there is a functional enrichment in this set. We found that 57 Rab proteins exhibit 38 unique conservation profiles and that of the 38 conservation profiles, nine are shared by more than one Rab protein. APACE associates Rab6A, Rab11A, Rab11B, Rab35 and Rab41 with 137 non-Rab proteins, which are

enriched for cytokinesis annotations. Hence, APACE suggests that these Rabs are involved in cytokinesis of cells.

4.4 Summary

In this chapter, we describe a new, scalable method, APACE, for the characterization of proteins by their common phylogenetic profile through a new, effective PPC approach that does not require orthologous and paralogous proteins to be identified in a preprocessing step. Furthermore, our method is able to avoid the use of alignment significance cutoffs to distinguish orthologs from paralogs. Instead, the method recognizes that the set of best-hits to a protein of interest will always be a superset of the orthologs to that protein. The task of distinguishing ortholog from paralog in a set of best-hits for every protein in a proteome reduces to a k-means clustering problem where *k* cannot be determined *a priori*. Our method uses a novel clustering method that redistributes similarity scores in one dimensional space onto a second dimension to separate orthologs and paralogs within a best-hit protein set. We demonstrate that the method is able to determine interspecies evolutionary distances that corroborate the most recent deep eukaryotic phylogenies, that our approach can successfully detect a set of proteins involved in cilia motility that compliments existing approaches, that our method can be applied to larger problem spaces where existing methods often fail, that our approach is robust to closely related species and is comparable to human selection for drug target discovery in Apicomplexa, that APACE is more sensitive

than existing methods in identifying interacting protein pairs and that APACE is successful at associating multiple proteins with highly dissimilar amino acid sequences. The online submission form for APACE can be found at http://bifrost.wustl.edu/APACE.

Chapter 5
Whole-transcriptome sequencing reveals an early ciliogenesis gene program

5.1 Introduction

Cilia are complex organelles protruding out from virtually all cell types in the human body. Cilia dysfunction has both physiological implications, such as renal cysts, hepatobiliary disease, cognitive impairment, retinal degeneration, obesity and skeletal bone defects, as well as developmental effects, including laterality defects, polydactyly, agenesis of the corpus callosum and posterior fossa defects (FLIEGAUF *et al.* 2007; TOBIN and BEALES 2009). Abnormal formation or function of these structures has been implicated as an underlying cause of many syndromes and disorders that have traditionally been recognized as disjoint conditions. The list of recognized ciliopathies continues to grow and currently include Bardet-Biedl syndrome (BBS), Meckel syndrome (MKS), Joubert

syndrome (JBTS), Nephrophthisis (NPHP), Senior-Løken syndrome (SLSN), Jeune syndrome (JATD), Oro-facial-digital syndrome type 1 (OFD1), Ellis van Creveld syndrome (EVC), Alström syndrome (ALMS), primary ciliary dyskinesia (PCD), and polycystic kidney disease (PKD) (FLIEGAUF *et al.* 2007; TOBIN and BEALES 2009). Furthermore, mutations in cilia disease genes tend to result in multisystemic abnormalities in multiple organisms and indicate a conserved, pervasive reliance of many physiological and developmental processes on the proper synthesis and function of cilia (TOBIN and BEALES 2009). The identification, characterization and implication of human ciliopathy disease genes has greatly benefited from their study in model organisms such as the green alga *Chlamydomonas reinhardtii* (PAZOUR and WITMAN 2009), the nematode *Caenorhabditis elegans* (EFIMENKO *et al.* 2006), and mouse (OSTROWSKI *et al.* 2002).

In this study, we take advantage of the fact that transcript abundance of most genes that encode known cilia components are greatly increased in *Chlamydomonas* during ciliogenesis. *Chlamydomonas* is a unicellular, green alga with genetics similar to yeast except it has two cilia that are highly similar to cilia found in humans. *Chlamydomonas* is an ideal model organism for transcript abundance based cilia gene detection because ciliogenesis can be induced by pH-shock. When environmental pH is precipitously dropped, *Chlamydomonas* cells shed their cilia and ciliogenesis begins immediately once environmental pH is restored. The specific transcriptional induction of genes encoding many known cilia components during ciliogenesis have been widely reported and further

underscore the efficacy and potential advantages of using *Chlamydomonas* as a model organism to study cilia and ciliogenesis (PAZOUR and WITMAN 2009). In this work, we use the v4 *Chlamydomonas* genome assembly (MERCHANT *et al.* 2007) and gene models predicted on that assembly by the GreenGenie2 *Chlamydomonas* genefinder (Chapter 3) to present results of the first whole-transcriptome next-generation sequencing of *Chlamydomonas reinhardtii* during ciliogenesis.

This approach is complementary to direct proteomic results (PAZOUR *et al.* 2005) because it probes the entire transcriptome during ciliogenesis as a whole, thereby facilitating not only the detection of genes that encode products present in the mature cilium, but also the proteins that, while not intrinsic to the mature organelle, are essential for the initiation and regulation of ciliogenesis and cilia function. Our results also complement existing comparative genomics methods that have been applied to defining the complete cilia gene catalog. Comparative genomics methods must discard genes that have an adequate degree of conservation in a non-ciliated species. This policy is necessary to reduce the number of false positive genes that are conserved across ciliated species to conserve related traits or processes that are not specific to cilia (e.g. transcription or mitosis) (KWAN *et al.* 2010; LI *et al.* 2004; MERCHANT *et al.* 2007). Whole-transcriptome next-generation sequencing does not depend on gene conservation patterns and will compliment comparative genomics methods because of its capacity to include genes that are conserved in non-ciliated organisms but remain

essential for proper cilia biogenesis, structure and function (e.g. tubulins, kinesins).

5.2 Results

5.2.1 RNAseq generates reliable transcriptome-wide ciliogenesis dataset

Illumina sequencing of mRNA isolated from pre-shock, 3, 10, 30 and 60 minutes into ciliogenesis produced a total of 99.4 million 36-mer single-end reads, for an average of 19.9 million reads per timepoint sample. This equates to 3.58Gb or a 32-fold coverage of the 112Mb *Chlamydomonas* genome. TopHat (TRAPNELL *et al.* 2009) was used to align the reads onto the *Chlamydomonas* v4 genome assembly (MERCHANT *et al.* 2007) and Cufflinks (TRAPNELL *et al.* 2010) was used to compute expression levels of 11,315 GreenGenie2 assembly v4 gene models (KWAN *et al.* 2009) with the default settings except for a maximum false discovery rate set at 1E-5. Expression values calculated by Cufflinks are reported in terms of fragments per kilobase transcribed per million reads mapped (FPKM) (TRAPNELL *et al.* 2010). In five independent sets of RNAseq sequencing (pre-shock, 3, 10, 30 and 60 minutes), ~83% of RNAseq reads align to the v4 *Chlamydomonas* genome assembly and 98% of GreenGenie2 predicted models show detectible expression in at least one timepoint sample. Any gene with a timepoint to pre-shock expression value ratio of 2.5 or greater is considered an

up-regulated gene. In total, there are 1400 predicted genes that are up-regulated at one or more timepoints (Appendix B). We find that most (83.1%; N=1163) genes are up-regulated by four to sixteen fold, the maximum fold-change ranges from our lower limit of 2.5 times basal levels to as much as 147 times basal level (Figure 5.1).

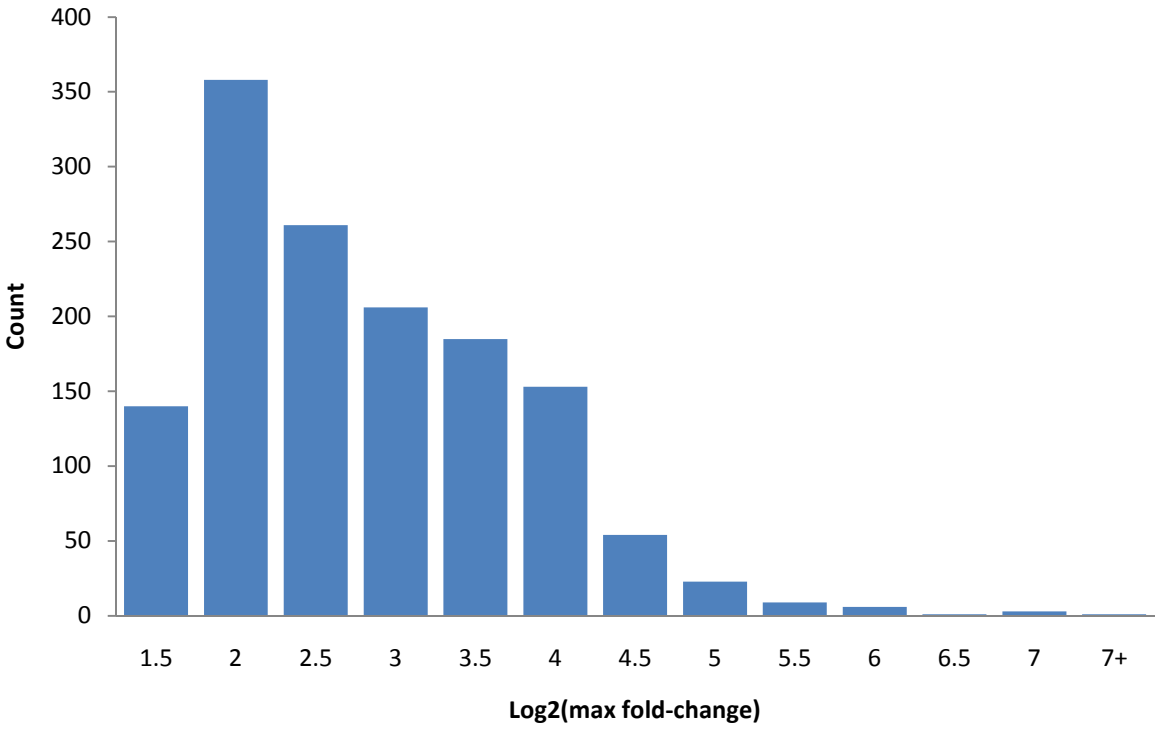

Figure 5.1 Distribution of maximum fold-change values

To evaluate the reliability of our data, we combine the qRT-PCR data from two previous studies to define a reference expression dataset of 201 genes that are up-regulated (N=91), down-regulated (N=29), or showed no change (N=81) when measured at the reference timepoint of 30 minutes by quantitative real-time RT-PCR (qRT-PCR) (Li *et al.* 2004; Pazour *et al.* 2005). Sensitivity is the

Table 5.1 Peak timepoint of 1400 up-regulated genes

Timepoint	Genes that show peak expression (N)	Genes that show peak expression (%)
3-minute	77	5.5%
10-minute	576	41.1%
30-minute	362	25.9%
60-minute	385	27.5%
TOTAL	**1400**	**100.0%**

proportion of reference up-regulated genes that are detected as such by RNAseq. We find that our RNAseq data are in agreement with 81 of 91 reference up-regulated genes, which equates to a sensitivity of 89.0%. Specificity is the proportion of reference genes that are not up-regulated, which are measured as such in the RNAseq data. We find that our RNAseq data agree with reference genes that do not show up-regulation in 97 out of 101 instances, which indicates a specificity of 96.0%. Given that all samples are from the same starting population and that they each underwent the same conditions and treatments, these performance results can be extended to all other recorded timepoints. We looked at the breakdown of peak expression in our timeseries to find that 5.5% (N=77) of up-regulated genes show peak abundance at the 3-minute timepoint, 41.1% (N=576) peak at the 10-minute timepoint, 25.9% (N=362) peak at the 30-minute timepoint, and the remaining genes (27.5%; N=385) peak at the 60-minute timepoint (Table 5.1).

Table 5.2 Distribution of 1400 up-regulated genes in 16 expression patterns

	Per pattern (N)	Per pattern (%)	Per group (N)	Per group (%)
Arch1	446	31.9%	543	38.8%
Arch2	97	6.9%		
Stag-3	68	4.9%	497	35.5%
Stag-10	71	5.1%		
Stag-30	170	12.1%		
Stag-60	188	13.4%	127	9.1%
Pulse-3	55	3.9%		
Pulse-10	43	3.1%		
Pulse-30	29	2.1%		
UT1	34	2.4%	110	7.9%
UT2	28	2.0%		
UT3	28	2.0%		
UT4	11	0.8%		
UT5	9	0.6%		
Hump1	50	3.6%	76	5.4%
Hump2	26	1.9%		
Ambiguous	36	2.6%	36	2.6%
Outliers	11	0.8%	11	0.8%
TOTAL	1400	100.0%	1400	100.0%

5.2.2 Timeseries analysis reveals early ciliogenesis regulation programs

We performed principal expression profile discovery by adapting the method from Brady *et al.* (BRADY *et al.* 2007) to determine a set of principal regulation profiles for genes that are up-regulated during the first 60 minutes of cilia regeneration in *Chlamydomonas*. All 1400 profiles (Appendix B) were included in the profile discovery process and 16 principal regulation profiles are identified. Note that this is significantly smaller than the 81 (that is, three possible outcomes for each measured timepoint, or 3^4) profiles that are

mathematically possible over four timepoints. The most common principal expression profile is *Arch1* and represents 31.9% (N=446) of up-regulated genes (Table 5.2). This profile is shaped like an arch (Figure 5.2A). The second most common principal expression profile is the pattern *Stag-60* (N=183; 13.4%) (Table 5.2), which shows no measurable up-regulation until an increase in mRNA abundance at the 60-minute mark (Figure 5.2B). A similar expression pattern is observed as the third most common principle expression profile, *Stag-30* (N=170; 12.1%) (Table 5.2). *Stag-30* exhibits no significant up-regulation until 30 minutes. This elevated transcript abundance is sustained through the 60-minute timepoint (Figure 5.2B). The fourth most common profile is *Arch2* (N=97; 6.9%)(Table 5.2), which is another arch-like pattern (Figure 5.2A). The fifth most common principal expression profile is *Stag-10* and is another delayed profile that first shows significant up-regulation at 10 minutes (N=71; 5.1%)(Table 5.2),

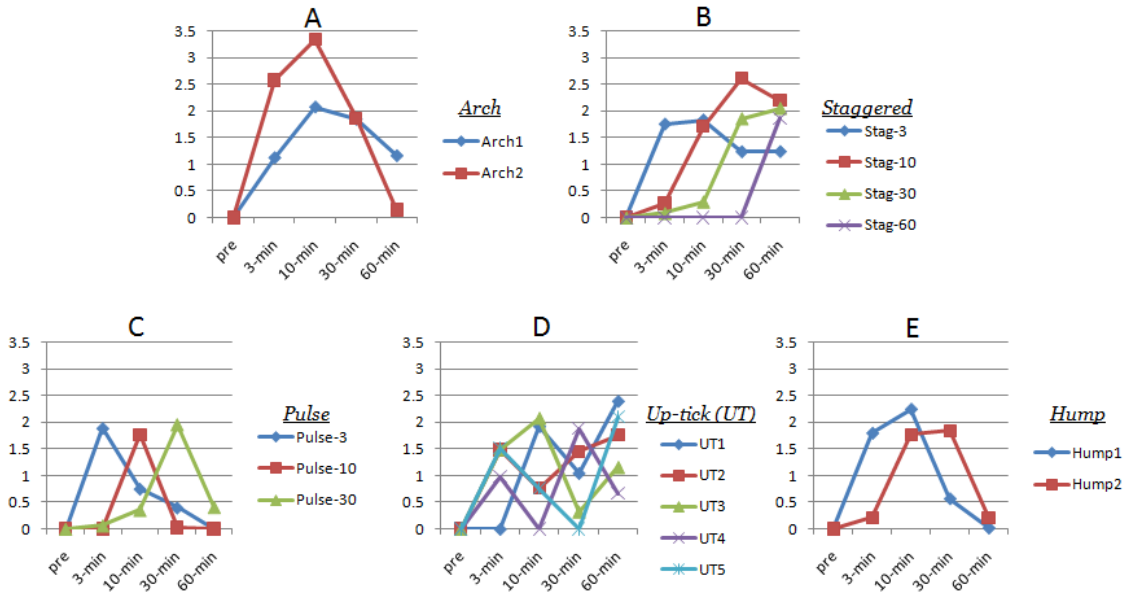

Figure 5.2 Sixteen principal expression profiles shown in pattern groups

which is sustained. The sixth most common profile is *Stag-3* (4.9%; N=68)(Table 5.2), the 3-minute analog of the three other principal expression profile that completes the *Staggered* pattern group (Figure 5.2B). The remaining 35.7% (N= 360) up-regulated genes are categorized into three pattern groups: *pulse*, *up-tick* (*UT*) and *hump* (Figures 5.2C-E). We observe pulse patterns for 9.1% (N=127) of the up-regulated genes (Table 5.2). This pattern group is characterized by a significant up-regulation event at only one timepoint. We observe a pulse pattern at each of the 3-minute (*Pulse-3*), 10-minute (*Pulse-10*) or 30-minute (*Pulse-30)* timepoints (Figure 5.2). We note that a fraction of up-regulated genes categorized as *Stag-60* may actually exhibit a 60-minute pulse if further data were gathered at later timepoints. However, extrapolating from the relative proportions of *Pulse-30* and *Stag-30* group sizes (17.1%; N=32), we expect a similar proportion of genes in *Stag-60* belong to a hypothetical *Pulse-60* expression profile. *UT* patterns make up 7.9% (N=110) of up-regulated genes (Table 5.2) and can be further sub-divided by the timepoint of the up-tick (Figure 5.2D). Finally, *hump* patterns make up 5.4% (N=76)(Table 5.2) and are profiles that are *pulse*-like but significant up-regulation is sustained over two consecutive timepoints (Figure 5.2E). Of the remaining fraction, 2.6% (N=36) have profiles that are equally similar to more than one principal expression profile and 0.8% (N=11) show profiles that are outliers in that their profiles are not adequately similar to any principal expression profile found in this analysis (Table 5.2; Section 5.5.2).

5.2.3 RNAseq further supports existing comparative genomics predictions

Comparative genomics methods organize genes by patterns of conservation (KWAN *et al.* 2010). Several groups have used different ciliated and non-ciliated species in comparative genomics studies with the aim of identifying new human ciliopathy genes in *Chlamydomonas* and other organisms (AVIDOR-REISS *et al.* 2004; EFIMENKO *et al.* 2005; KWAN *et al.* 2010; LI *et al.* 2004; MERCHANT *et al.* 2007; OSTROWSKI *et al.* 2002). Li *et al.* identifies *Chlamydomonas* cilia genes as those also present in human, but absent from the non-ciliated *Arabidopsis thaliana* (LI *et al.* 2004). Merchant *et al.* distinguish *Chlamydomonas* cilia genes as those also present in both human and a ciliated fungus (*Phytophthora*), but absent from non-ciliated organisms (MERCHANT *et al.* 2007). We also apply the method described previously in (KWAN *et al.* 2010) to predict *Chlamydomonas* cilia genes as those conserved in *Homo sapiens* (human), *Mus musculus* (mouse) and *Danio rerio* (zebrafish), but absent from *Arabidopsis thaliana*, *Saccharomyces cerevisiae* and *Oriza sativa* (rice) (KWAN *et al.* 2010). Combining the predictions from these three methods takes advantage of complementary strengths that exist between each method and the species they use. Taken together, the three comparative genomics methods identify 646 putative cilia genes. Existing qRT-PCR and proteomic results (LI *et al.* 2004; PAZOUR *et al.* 2005) provide experimental support for 149 of the 646 (23.1%) putative cilia genes and our early timepoint RNAseq data supports 123 of the 149 (82.6%).

Furthermore, RNAseq provides novel support for an additional 74 (11.5%) of the 646 *Chlamydomonas* genes with human homologs identified by comparative genomics, but lacked supporting qRT-PCR or proteomic evidence in *Chlamydomonas* cilia. We divide this set of genes into five broad categories based on their existing annotation in relation to cilia, human disease or mutant phenotypes (Table 5.3). The different groups consist of genes that are known from studies in other organisms to associate directly with cilia (N=12), genes that have prior evidence of associating with genes that are themselves cilia related or that are associated with recognized ciliary diseases (N=8), genes that have been implicated in diseases or mutant phenotypes with no known ciliary basis (N=7), genes that have no known function or cilia involvement (N=13) and genes that have known functions with no known connection to cilia (N=34) (Table 5.3). The genes that have been reported as cilia related in other organisms include ones that have been implicated in defective hedgehog signaling (ARL13B) (CASPARY *et al.* 2007), Bardet-Biedl Syndrome (BBS4 and BBS7)(NACHURY *et al.* 2007), structural genes (DNAH8, DNAL4, DYNLT1, TUBA1A, SEPT7 and TTLL9)(KUBO *et al.* 2010; TANNER *et al.* 2008), Primary Ciliary Dyskinesia (LRRC50)(DUQUESNOY *et al.* 2009; LOGES *et al.* 2009), defective murine spermatogenesis (PACRG)(LORENZETTI *et al.* 2004) and Sensenbrenner syndrome or cranioectodermal dysplasia (WDR35)(GILISSEN *et al.* 2010; MILL *et al.* 2011). Genes that have prior evidence of associating with genes that are themselves cilia related or that are associated with recognized ciliary diseases include genes that are involved in Joubert Syndrome (AASDHPPT), cilia formation (CEP164 and TTLL3)(GRASER *et al.* 2007; PATHAK *et al.* 2011; WLOGA

et al. 2009), basal body assembly (MORN1)(LORESTANI *et al.* 2010), spermatogenesis (PHPT1 and SPATA4)(KAMATH *et al.* 2010; WANG *et al.* 2011), cystic kidney disease (PIH1D1)(INOUE *et al.* 2010) and adult onset obesity and retinal degeneration in mice (TUB)(STOLC *et al.* 2005).

Genes that have been implicated in diseases or phenotypes with no known connection to cilia are involved in carotid artery calcified plaque or CarCP (ABCA3 and CACNA1H), autism (CACNA1H), dyslexia (DYX1C1), interacting with NUDT9 (GLOD4), lysosomal storage disease or Sanfillipo Syndrome (GNS), neuronal degeneration (INPP4A) and the short-limb dwarf mouse phenotype (NPR2). Two genes found to associate with carotid artery calcified plaque are up-regulated. ABCA3 is a lipid transporter and CACNA1H is a calcium channel T-type, which are also associated with surfactant transport in the lung, one role of cilia in that organ. DYX1C1 is associated with dyslexia (CURRIER *et al.* 2011) and up-regulated by about 8-fold during ciliogenesis, which may suggest a potential cilia role in the condition. INPP4A is up-regulated about 9-fold and has been associated with neuronal degeneration (SASAKI *et al.* 2010). The short limb dwarf mouse has mutations in the guanylate cyclase natriuretic peptide receptor 2 (NPR2)(TSUJI and KUNIEDA 2005), which traffic to chemosensory neurons in *C. elegans* for dauer formation (FUJIWARA *et al.* 2010; HALLEM *et al.* 2011; HUME *et al.* 2009; JENSEN *et al.* 2010). Genes with annotated functions but no obvious cilia connection include ankyrins (ANK1 and ANK2), calcium-channels (CACNA1G, CACNA1I), cystein conjugate-beta lyase 2 found to associate with mitochondria (CCBL2), potassium voltage-gated channels (KCNB2 and KCNJ1),

kinesins (KIF5B, KIF6 and KIF9), a Lamin B receptor (LBR), a guanylate cyclase (NPR1), Golgi association (SEC14L5 and SEC22A), ubiquitination or ubiquitin association (SKP1, UBFD1 and UBXN1) and also a tumor suppressor (VWA5A). Finally, of the 13 genes that have no known function and no known cilia connection, one is a tumor suppressor (ZMYND10) six have shown preferential expression in ciliate tissues (CXorf41, LRRC6, SLC25A32, TEX9, ZMYND10 and ZNF474).

Table 5.3 Annotation of 78 Human homologs of cilia gene candidates in *Chlamydomonas* previously identified by comparative genomics sorted on pattern

CILIARY GENES (N=12)				
GENE	**log_2(maxFC)**	**PATTERN**	**DESCRIPTION**	**DEFECTS**
BBS4	2.00	Arch1	Bardet-Biedl Syndrome	Obesity, retinal degeneration, kidney disease[1]
DNAL4	3.42	Arch1	Axonemal dynein light chain	*Chlamydomonas* Dynein Light Chain LC10/DLL3/*oda12*[2]
DYNLT1	2.78	Arch1	Axonemal dynein light chain TcTex type	*Chlamydomonas* Inner dynein arm I1; TcTex1/DLT3
PACRG	2.71	Arch1	Parkin coregulated; sperm morphogenesis	Mouse spermatogenesis defective[3]
TUBA1A	2.97	Arch1	Tubulin	
WDR35	3.01	Arch1	IFT121	Sensenbrenner syndrome: Cranioectodermal dysplasia[4,5]
ARL13B	3.37	Arch2	Ciliary / mouse mutant	Defective hedgehog signaling[6]
LRRC50	4.43	Arch2	Axonemal dynein chaperonin, ODA7	Primary Ciliary Dyskinesia[7,8]

TTLL9	3.52	Arch2	Tubulin tyrosine ligase like	Inner dynein arm motility in Chlamydomonas
BBS7	2.08	Stag-10	Bardet-Biedl Syndrome	Obesity, retinal degeneration, kidney disease[1]
DNAH8	2.34	Stag-10	Axonemal dynein heavy chain	*Chlamydomonas* Outer arm dynein heavy chain γ
SEPT7	1.96	Stag-30	Ciliary diffusion barrier	siRNA tissue culture[9]
ASSOCIATION WITH CILIARY DISEASE OR OTHER CILIARY PROTEINS (N=8)				
GENE	**log_2(maxFC)**	**PATTERN**	**DESCRIPTION**	**DEFECTS**
MORN1	2.90	Arch1		Defective in basal body assembly in *Toxoplasma*[10]
PIH1D1	2.54	Arch1	PIH1 domain	Part of prefoldin complex (R2TP) RuvB1 and RuvB2 Reptin (RuvB2) is implicated in cystic kidney disease in zebrafish. Assembly factor in sea urchin[11]
TTLL3	3.40	Arch1	Tubulin monoglycase TTLL3	Short cilia in zebrafish[12,13]
TUB	1.53	Arch1	Tubby	Adult onset obesity and retinal degeneration in mice[14]
CEP164	2.2	Pulse-30	Centrosomal protein	Primary cilia formation[15]
PHPT1	2.88	Stag-30	Phosphohistidine kinase	Highly expressed in testis Dephosphorylates ATP-Citrate lyase[16] (See Table 5.4)
SPATA4	3.2	UT3	Spermatogenesis associated	Spermatogenesis associated /osteoblast differentiation[17]
AASDHPPT	1.54	Ambiguous	LYS5	Joubert syndrome with pipecolic acidemia
IMPLICATED IN DISEASES OR MUTANTS NOT KNOWN TO HAVE CILIARY BASIS (N=7)				
GENE	**log_2(maxFC)**	**PATTERN**	**DESCRIPTION**	**DEFECTS**

GLOD4	1.97	Arch1		Down-regulated in RPGRORF15 and interacts w/ NUDT9
INPP4A	3.13	Hump1	Inositol polyphosphate-4-phosphatase, type I	Neuronal degeneration
DYX1C1	2.97	Hump2		Dyslexia
GNS	1.46	Hump2	Glucosamine (N-acetyl) 6 sulfatase	Lysosomal storage disease; Sanfillipo syndrome
NPR2	2.22	Pulse-3	Guanylate cyclase	Short limb dwarf mouse
CACNA1H	1.40	Pulse-10	Calcium channel; T type	Carotid artery calcified plaque (CarCP) Autism
ABCA3	2.13	Stag-30	Lipid Transporter	Carotid artery calcified plaque (CarCP)
KNOWN FUNCTION BUT NO KNOWN CILIARY CONNECTION (N=32)				
GENE	**log_2(maxFC)**	**PATTERN**	**DESCRIPTION**	**DEFECTS**
CACNA1G	3.01	Arch1	Calcium channel	
CYP4X1	3.31	Arch1	Cytochrome P450	
DNAJC27	3.11	Arch1	Chaperonin	
GYLTL1B	2.68	Arch1	Glycosyltransferase like 1B	Paralog of LARGE mouse mutant
KIF6	2.34	Arch1	Kinesin	
LBR	1.67	Arch1	Lamin B receptor	
RHBG	6.25	Arch1	Rh family glycoprotein	ammonium transport
VWA5A	5.11	Arch1	BCSC-1/tumor suppressor	

KCNB2	2.88	Arch2	Potassium voltage gated channel	
KIF9	3.36	Arch2	Kinesin	
SEC14L5	2.06	Arch2		Golgi/No function
NPR1	1.92	Pulse-3	Guanylate Cyclase	
PEF1	1.44	Pulse-3	Penta EF hand	ER folding
CACNA1I	2.26	Pulse-10	Calcium channel	
TMEM65	1.50	Pulse-30	Transmembrane protein	
KCNJ1	3.22	Stag-3	Potassium voltage gated channel K+ efflux pathway	
PDE4C	2.21	Stag-3	Phosphodiesterase	
PSMD10	4.00	Stag-3	Non-ATPase regulatory subunit	Proteosome
ANK1	3.12	Stag-10	Ankyrin1	
ANK2	2.99	Stag-10	Ankyrin2	
PLA2G7	3.14	Stag-10	Phospholipase A2	Arachidonic acid pathway
HCCS	2.61	Stag-30	Holocytochrome C synthase	Mitochondria
RBM45	1.87	Stag-30	RNA binding protein	Deubiquinitation
SEC22A	1.81	Stag-30	SNARE	Golgi
SKP1	1.52	Stag-30	S phase kinase ubiquitin ligases	Ubiquitination

TXNRD1	1.55	Stag-30	Thioredoxin reductase	
CCBL2	1.48	Stag-60	cysteine conjugate-beta lyase 2	Mitochondria
GDA	2.11	Stag-60		
KIF5B	2.08	Stag-60	Kinesin	
MTR	2.04	Stag-60	5-methyltetrahydrofolate-homocysteine methyltransferase	Folate metabolism
NUDT14	2.45	Stag-60	UDP-glucose pyrophosphatase	
UBFD1	1.84	Stag-60	Polyubiquitin binding protein	Ubiquitination
HSPBP1	2.49	UT5	Hsp70 binding protein	Chaperonin
UBXN11	4.49	Outlier	Ubiquitin associated	
NO KNOWN FUNCTION AND NO CILIA CONNECTON (N=13)				
GENE	**log_2(maxFC)**	**PATTERN**	**DESCRIPTION**	**DEFECTS**
ANKRD50	3.31	Arch1		
SVEP1	1.55	Arch1	Sushi	
TEX9	2.23	Arch1		Testis enriched
ZNF474	2.16	Arch1		Up-regulated in mice ciliated tissue; Highly expressed in testis
CXorf41	2.91	Arch2		Expressed highly in testis and trachea
KIAA0562	3.20	Arch2		

LRRC6	3.36	Arch2		Testis enriched
ZMYND10	4.57	Arch2	Zinc finger/tumor suppressor	
C1orf53	5.84	Stag-30		
HEPHL1	2.11	Stag-30		
KRTAP10-6	1.55	Stag-60	Keratin associated protein	
SLC25A32	1.97	Stag-60		Mitochondrial expressed in brain
TROVE2	2.99	UT1		

[1](NACHURY *et al.* 2007)
[2](TANNER *et al.* 2008)
[3](LORENZETTI *et al.* 2004)
[4](MILL *et al.* 2011)
[5](GILISSEN *et al.* 2010)
[6](CASPARY *et al.* 2007)
[7](LOGES *et al.* 2009)
[8](DUQUESNOY *et al.* 2009)
[9](KIM *et al.* 2010)
[10](LORESTANI *et al.* 2010)
[11](INOUE *et al.* 2010)
[12](PATHAK *et al.* 2011)
[13](WLOGA *et al.* 2009)
[14](STOLC *et al.* 2005)
[15](GRASER *et al.* 2007)
[16](KAMATH *et al.* 2010)
[17](WANG *et al.* 2011)
All human homologs have a BLASTP E-val of better than 1E-10 to the Chlamydomonas *gene. Blank cells indicate no available annotation.*

5.2.4 RNAseq identifies 188 novel ciliopathy gene candidates

There are 188 human genes that have *Chlamydomonas* homologs that are up-regulated during ciliogenesis. Our timeseries data indicates that these genes have an additional annotation of being up-regulated during ciliogenesis, which

suggests potential ciliary roles. These genes can be categorized into 14 annotation categories (Figure 5.3; Table 5.4; Table 5.5). In decreasing order of the number of genes found in each annotation group, they are solute carriers/facilitators/transporters (N=23), Golgi and trafficking (N=19), chaperonins (N=18), kinases and phosphatases (N=15), mitochondria (N=13), lipid and inositol metabolism (N=10), cilia proteins from non-*Chlamydomonas* studies (N=8), genes attributed to diseases or mutant phenotypes that have not been related to cilia (N=8), ubiquitin (N=7), cell cycle (N=7), disulfide bonds (N=5) and DNA repair (N=3). There are 16 additional genes that have no prior annotation (Table 5.4) and 36 genes that previously lacked any association to cilia or ciliogenesis (Figure 5.3; Table 5.5). The mitochondria annotation set is the only annotation group that is significantly enriched for an expression pattern. *Stag-30* is exhibited by 9 out of 13 genes (P=2.07E-6). Membrane transporters make up the largest of the coherent annotation groups (Table 5.4) and we note that SLC25A6 is a membrane transporter that was found by proteomic analysis (PAZOUR *et al.* 2005). The next largest annotation group is the Golgi/membrane trafficking proteins (Table 5.4). This group of proteins would not be found by proteomics or comparative genomics, but recent work has shown that IFT20 and GMAP210 are Golgi proteins (FOLLIT *et al.* 2008). The third largest annotation group is the chaperonins, including Hsp40, Hsp70 and Hsp90, which were observed previously to be up-regulated (STOLC *et al.* 2005). Since tubulin is the major protein of cilia and it requires tubulin folding cofactors such as chaperonin containing TCP1 and TCPCT-complex proteins, it is reasonable that this should be a major class. Mutations in the *ASQ2* gene show a role for the tubulin folding

co-chaperone, TBCCD1, in mother-daughter basal body linkage (FELDMAN and MARSHALL 2009) and in centrosome and Golgi positioning (GONCALVES *et al.* 2010). Fourth most common are the kinesins and phosphatases that perform post-translational modifications (Table 5.4). This is a diverse class and is likely to have a wide range of functions. It includes two cyclin dependent kinases, two aurora kinases and three aarF domain kinases. Aurora A kinase has been implicated in ciliary disassembly (LANDER *et al.* 2001; PUGACHEVA *et al.* 2007). ADCK3 is a mitochondrial protein that acts on Q10 biosynthesis (LAGIER-TOURENNE *et al.* 2008) and NEK4 has been implicated in altering sensitivity to the action of Taxol in *Chlamydomonas* (DOLES and HEMANN 2010). There is a group of genes that affect inositol and lipid function and biosynthesis (Table 5.4). Several inositol biosynthetic genes have been implicated in ciliary function. Morpholinos to inositol kinase (Ipk1) in zebrafish and patients with Joubert Syndrome that have mutations in INPP5E show ciliary defects (BIELAS *et al.* 2009; SARMAH *et al.* 2007). Furthermore, PICALM (phosphatidylinositol binding clathrin assembly protein) has been recently implicated as having a role in late-onset Alzheimer's disease (JUN *et al.* 2010; KOK *et al.* 2011). Seven genes affect ubiquitin based processes. Ubiquitin conjugation has been implicated in ciliary disassembly (HUANG *et al.* 2009). STAMBPL1 may serve as an interesting link as a STAM binding protein. STAM is an ESCRT-O protein that interacts with USP8 ubiquitin pathway for movement to the membrane (BERLIN *et al.* 2010). There are also a number of genes in cell cycle control, disulfide bond reduction and DNA repair (Table 5.4). A number of genes have been previously implicated in

cilia or ciliopathies. Two are likely to be missed by comparative genomics methods

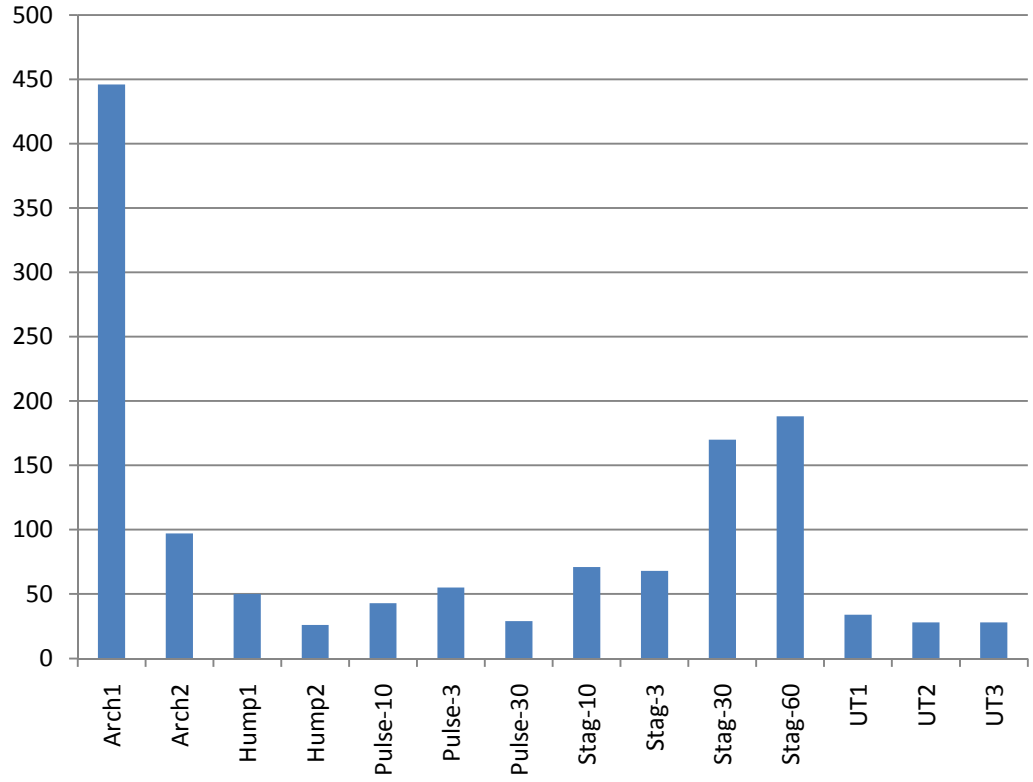

Figure 5.3 Annotation group distribution of 188 Human cilia genes

because they are small proteins and fail the 1E-10 similarity cutoff (TXNDC3 and DYNLT1). Tubulin, actin, katanin and KIF21A are found in non-ciliated organisms and would fail the negative cutoff. KIF21A has been implicated in movement of dendrites and its role in cilia would be interesting (MARSZALEK *et al.* 1999). Sixteen genes are annotated as open reading frames or by motifs. Leucine rich repeats (LRR) and WD repeats (WDR) are motifs found in other ciliary proteins (COLE 2003; DUQUESNOY *et al.* 2009). Thirty-six genes have existing annotations but were not associated with cilia, but it is interesting to note that 22

of them have enzymatic roles. Finally, nine of the 188 human homologs have been associated with human disease or mutant phenotypes in model organisms (Table 5.5). TULP1 is a tubby-like protein that is up-regulated about 14-fold during ciliogenesis; mouse mutants show retinal degeneration (XI *et al.* 2003). Furthermore, TULP3 has been implicated in retrograde transport (MUKHOPADHYAY *et al.* 2010). Cut-like homeobox 1 (CUX1) is a transcription factor and a homolog of the protein cut in *Drosophila melanogaster*. CUX1 has been implicated in transcriptional regulation of the RPGRIPIL and FTO (fat mass and obesity associated) genes (STRATIGOPOULOS *et al.* 2011) and follows the *Pulse-30* expression profile, suggesting that its role may be transient though essential. PRMT1 is an arginine methyltransferase. Sloboda and co-workers found that several proteins in flagella are methylated (SLOBODA and HOWARD 2009). A deletion of PRMT1 in *Chlamydomonas* results in altered motility (ESPARZA, GIDDINGS and DUTCHER, *in preparation*). CALN5 encodes a calpain, which is a calcium-dependent cysteine protease. CALN5 has been associated with polycystic ovary syndrome (GONZALEZ *et al.* 2006). Two other calpains (FAP135 and FAP226) were found in cilia by proteomic analysis (PAZOUR *et al.* 2005). NXN is a nucleoredoxin. In a whole exome sequencing project of a mouse that shows a recessive perilethal phenotype, a splice site variant was found (BOLES *et al.* 2009). The mouse has cleft palate defects, which has been observed in mice with defects in sonic hedgehog signaling (LIPINSKI *et al.* 2010). CLPTM1 is predicted to be a transmembrane protein that has been implicated in cleft lip and palate (BARONI *et al.* 2010), but there has not been validation of this role via mutational analysis. However, its expression is altered by exposure to nicotine, which has

been implicated in cleft lip and palate (BARONI *et al.* 2010). SHOC2 has been implicated in Noonan Syndrome, which is characterized by dysmorphic features, webbed neck, cardiac anomalies, short stature and cryptorchidism. It is generally an autosomal dominant trait and is often caused by defects in the MAPK pathway (KOMATSUZAKI *et al.* 2010). This method may provide a number of new genes that will serve as candidates for ciliopathy disease genes.

Table 5.4 Annotation of 144 of 188 Human homologs of novel *Chlamydomonas* cilia gene candidates that can be assigned an annotation group sorted on pattern

SOLUTE CARRIERS/FACILITATORS /TRANSPORTERS (N=23)		
GENE	**log_2(maxFC)**	**PATTERN**
ATP2C1	1.48	Arch1
ATP8B2	2.28	Arch1
SLC25A45	2.23	Arch1
NIPA2	3.15	Arch1
SLC10A7	3.79	Arch1
ATP8B4	2.63	Arch2
SLC25A6	1.43	Pulse-3
ABCA2	1.37	Pulse-30

ABCC4	1.67	Stag-10
ABCF2	2.41	Stag-10
SLC26A5	3.63	Stag-10
SLC11A1	3.64	Stag-10
ATP6V1C2	1.45	Stag-30
SLC17A6	1.51	Stag-30
SLC35B1	1.72	Stag-30
ABCB9	2.83	Stag-30
SLC7A14	2.95	Stag-30
MFSD5	4.89	Stag-30
ABCG2	2.03	Stag-60
HIAT1	2.84	Stag-60
ABCB6	3.43	Stag-60
ATP2A3	1.64	UT2
ABCG2	2.01	Ambiguous

GOLGI AND TRAFFICKING (N=19)		
GENE	**log_2(maxFC)**	**PATTERN**
VPS28	1.6	Arch1
SEC61A2	1.94	Arch1
YKT6	2.16	Arch1
GOLT1B	2.66	Arch1
GOSR2	3.01	Arch1
TBC1D17	3.95	Arch1
ARL6	4.65	Arch2
CPNE9	3.22	Hump1
STAMBPL1	1.37	Pulse-10
RABAC1	1.42	Pulse-10
KIFC3	1.91	Stag-3
GBF1	1.4	Stag-30
YOD1	1.58	Stag-30
ARF3	1.81	Stag-30

ERP29	1.87	Stag-30
SEC31B	1.95	Stag-30
KDELR1	1.98	Stag-30
DERL3	1.39	UT1
VAMP7	1.74	UT1
CHAPERONINS (N=18)		
GENE	**log_2(maxFC)**	**PATTERN**
DNAJB5	1.94	Arch1
CCT6A	2.05	Arch1
CCT8	2.13	Arch1
ERO1l	2.14	Arch1
CCT3	2.16	Arch1
TCP1	2.21	Arch1
CCT2	2.25	Arch1
CCT7	2.31	Arch1
DNAJC10	2.97	Arch1

DNAJB11	2.19	Hump1
DNAJA3	1.8	Stag-10
HSPA5	3.64	Stag-10
DNAJC7	1.91	Stag-30
DNAJC2	1.77	Stag-60
HSPA9	1.45	Stag-60
BBS9	1.6	Pulse-30
HSP4L	1.62	UT2
HSP90AB1	3.82	
KINASES AND PHOSPHATASES (N=15)		
GENE	**log_2(maxFC)**	**PATTERN**
MAP3K11	1.45	Arch1
PPP4C	1.53	Arch1
NEK4	1.76	Arch1
AURKA	2.12	Arch1
CDKL1	Stag-3	Arch2

AURKC	2.21	Hump1
PRKACB	2.41	Hump1
DUSP4	1.7	Pulse-10
RPS6KA1	1.45	Stag-3
ADCK2	1.97	Stag-30
ADCK1	2.31	Stag-30
AK2	2.54	Stag-30
MLK4	1.42	Stag-60
STK38L	1.61	UT3
BRSK2	1.95	Ambiguous
MITOCHONDRIA (N=13)		
GENE	**$\log_2$(maxFC)**	**PATTERN**
MRPL24	1.44	Stag-30
TIMM23	1.58	Stag-30
HTRA2	1.66	Stag-30
COQ9	1.72	Stag-30

TIMM17A	1.73	Stag-30
COQ5	1.75	Stag-30
ALDH2	2	Stag-30
NDUFS4	2.29	Stag-30
ADCK3	2.66	Stag-30
COQ3	1.49	Stag-60
TAZ	1.9	Stag-60
GFER	2.45	Stag-60
SOD2	1.9	UT5
LIPID AND INOSITOL METABOLISM (N=10)		
GENE	**log_2(maxFC)**	**PATTERN**
IMPA	1.47	Stag-60
PICALM	1.48	Arch1
CDIPT	1.54	Arch1
ISYNA1	1.76	Arch1
SC5DL	2.91	Arch1

SPTLC1	3.39	Arch2
ACLY	2.18	Hump1
SACM1L	2.01	Hump1
LPCAT1	1.57	Stag-30
LSS	1.84	Stag-30
UBIQUITIN (N=7)		
GENE	**log_2(maxFC)**	**PATTERN**
UBTD2	1.55	Arch1
DCUN1D1	1.7	Arch1
WDR5	2.6	Arch2
USP2	2	Hump1
UBA5	2.51	Stag-10
ARIH2	1.47	Stag-30
WUSUB1	2.75	Stag-60
CILIA PROTEINS (N=8)		
GENE	**log_2(maxFC)**	**PATTERN**
KIF21A	2.36	Arch1

TXNDC3	2.68	Arch1
RUVBL1	2.86	Arch1
DYNLT1	3.29	Arch1
TUBB	3.3	Arch1
RUVBL2	3.49	Arch1
ACT	6.73	Stag-10
KAT	1.54	Stag-30
CELL CYCLE (N=7)		
GENE	**log_2(maxFC)**	**PATTERN**
CNNM2	1.44	Arch1
CCNB2	1.65	Stag-30
CDKL2	1.46	Stag-60
MCM5	1.52	Stag-60
GINS	1.96	Stag-60
CDC42BPG	2.29	UT3
MAD1L1	1.41	UT4

DNA REPAIR (N=3)		
GENE	**log_2(maxFC)**	**PATTERN**
ERCC8	1.82	UT1
NUDT6	1.88	Stag-30
NUDT5	2.19	Arch1
DISULFIDE BONDS (N=5)		
GENE	**log_2(maxFC)**	**PATTERN**
GPX7	1.75	Hump2
PDIA6	1.37	Stag-60
PDIA3	1.41	Stag-60
HAGH	1.61	Stag-60
GPX4	1.75	UT2
UNDEFINED FUNCTIONS (N=16)		
GENE	**log_2(maxFC)**	**PATTERN**
RCBTB2	1.79	Arch1
ZFAND5	1.83	Arch1
LRRC61	2.55	Arch1
C22orf25	3.04	Arch1

GPATCH8	1.64	Hump1
ZDHHC11	2.38	Hump1
WDR49	2.4	Hump2
C4orf29	1.42	Pulse-10
MACROD2	1.68	Pulse-10
FAM119B	2.34	Pulse-10
KLHDC3	2.06	Pulse-3
C8orf38	3.68	Stag-3
GLIPRIL1	1.37	Stag-60
ACO68499.1	2.15	Stag-60
ZCCHC24	2.92	Stag-60
C22orf13	1.5	UT3

All human homologs have a BLASTP E-val of better than 1E-10 to the Chlamydomonas *gene.*

Table 5.5 Annotation of 44 of 188 Human homologs of novel *Chlamydomonas* cilia gene candidates involved in other processes or diseases sorted on pattern

OTHER CELLULAR PROCESSES (N=35)			
GENE	**log_2(maxFC)**	**PATTERN**	**DESCRIPTION**
GLRX5	1.66	Arch1	Glutaredoxin
TMEM19	1.76	Arch1	Transmembrane protein
CYB5A	2.06	Arch1	Cytochrome 5
CYP26A1	2.69	Arch1	Cytochrome P450
L2HGDH	2.82	Arch1	L-2-hydroglutamate dehydrogenase
CYP51A1	3.93	Arch1	Cytochrome P450
CWC27	3.93	Arch1	Peptidyl-prolyl isomerase
ECE2	2.71	Arch2	Endothelin converting enzyme
CYP4V2	3.58	Arch2	Cytochrome P450
TXNDC9	4.58	Arch2	Translation initiation factor
GPD1L	1.89	Pulse-3	Glycerol-3-phosphate dehydrogenase 1-like
DGAT2	2.23	Pulse-3	Diaycl glycerol-O-transferase
RPS15	1.53	Pulse-30	Ribosomal protein

TM9SF4	1.6	Pulse-30	Transmembrane protein
HIST4H4	2.56	Pulse-30	Histone H4
NLRC5	3.3	Pulse-30	NOD family/Receptor
ALDH18A1	1.6	Stag-10	Aldehyde dehydrogenase 18 family member A1
H3F3C	1.57	Stag-30	Histone H3
ENTPD6	1.67	Stag-30	Ectonucleoside triphosphate diphosphohydrolase
NCOA7	1.75	Stag-30	Coactivator
GGH	1.96	Stag-30	gamma-glutamyl hydrolase
P4HB	2.23	Stag-30	prolyl 4-hydroxylase
GCH1	2.2	Stag-30	GTP cyclohydralase
MSI2	2.64	Stag-30	RNA binding
EXT2	1.37	Stag-60	Exotensis
EIF2B1	1.39	Stag-60	Translational Initiation factor
DDX49	1.47	Stag-60	
ECD	1.54	Stag-60	Ecdysoneless homolog

ATG5	1.54	Stag-60	
LIPF	1.7	Stag-60	Lipase
ABHD3	1.77	Stag-60	Membrane bound hydrolase
MTHFR	2.8	Stag-60	Methylene tetrahydrofolae reductase
CPVL	3.04	Stag-60	carboxypeptidase
EPHX4	3.42	Stag-60	Epoxide hydrolase
PHGDH	4.02	Stag-60	Phosphoglycerate dehydrogenase
GENES WITH ASSOCIATED DISEASES OR MUTANT PHENOTYPES (N=9)			
GENE	**log_2(maxFC)**	**PATTERN**	**DESCRIPTION**
VPS4	1.72	Arch1	Cytokinesis
CHMP4B	1.57	Hump1	ESCRT III/Cytokinesis; interacts with VPS4
CALN5	2.5	Arch2	Polycystic ovary syndrome
NXN	5.94	Outlier	Perinatal lethal in mice
TULP1	3.75	Pulse-3	Retinitis pigmentosa
CUX1	1.46	Pulse-30	Transcription factor that regulates RPGRIP1L that results in COACH syndrome
SHOC2	1.7	Pulse-30	Noonan Syndrome

CLPTM1	1.78	Stag-30	Cleft Lip and palate gene
PRMT1	1.53	Stag-60	Arginine protein methyltransferase

All human homologs have a BLASTP E-val of better than 1E-10 to the Chlamydomonas *gene.*

5.3 Discussion

High-throughput transcriptome sequencing over a timeseries of interest facilitates the annotation of genes involved in a given biological process of interest and helps to further elucidate the regulation dynamics that underlie the biological process in question. We have leveraged the inducible ciliogenesis response in *Chlamydomonas* to support the involvement and better describe the regulation dynamics of known and novel *Chlamydomonas* homologs of human ciliopathy disease genes during ciliogenesis.

We performed high-throughput transcriptome sequencing of *Chlamydomonas* before pH-shock and at 3, 10, 30 and 60 minutes into ciliogenesis and find that 83% of the 99.4 million resulting sequencing reads can be aligned to the v4 assembly by TopHat (TRAPNELL *et al.* 2009). This fraction is similar to the fraction of genomic sequence that remains unresolved (MERCHANT *et al.* 2007) and the fraction of high-throughput genome sequencing reads that align to the same assembly (DUTCHER, *in preparation*). We quantified the transcript abundance of 11,315 gene models on the v4 assembly from each sample using Cufflinks (TRAPNELL *et al.* 2010) and find that 98% of the GreenGenie2

predicted models are supported by some detectable expression evidence, which further supports the genome-wide prediction accuracy claims originally made by Kwan *et al.* (KWAN *et al.* 2009).

We find that 1400 gene models satisfy our minimum fold-change cutoff of 2.5-fold up-regulation and that our sequencing, assembly, abundance and fold-change computation pipeline is both highly sensitive (89.0%) and very specific (96.0%). This represents a 30.9% increase in reliability over previous genome-wide transcriptional analysis of flagellar regeneration in *Chlamydomonas* (STOLC *et al.* 2005). This unprecedented degree of reliability is likely a result of combining the technological advances in next-generation whole-transcriptome sequencing, a new genome assembly and updated gene models predicted by a *Chlamydomonas* specific genefinder than any other potential factors. The fact that 10.2% (N=1163) of the entire gene catalog shows moderate up-regulation of between four and sixteen fold may indicate that ciliogenesis requires the cooperation of a large group of genes in a complicated, distributed process, which is not entirely unexpected given the inherent complexity, diverse functionality and evolutionary age of the organelle.

Many forms of experimental expression validation of cilia gene predictions measure candidate gene expression levels for evidence of up-regulation at 30 minutes into ciliogenesis because it is thought to be the time that most cilia genes exhibit peak expression in general (STOLC *et al.* 2005). Surprisingly, our whole-transcriptome data indicates that, while a substantial proportion of up-regulated genes do reach peak measured values at 30 minutes (N=362; 25.9%), there is a similar proportion of up-regulated genes with peak expression values at 60

minutes (N=385; 27.5%). Moreover, the greatest proportion of up-regulated genes reach peak measured expression values at 10 minutes (N=576; 41.1%). The remaining fraction (N=77; 5.5%) of up-regulated genes peak at 3 minutes (Table 5.1). Perhaps the most striking result is that over one-third of up-regulated genes (N=475; 33.9%) are at or below basal expression levels at 30 minutes and likely to be miscategorized as not up-regulated if transcript abundance is measured only at that timepoint. Given the reliability demonstrated by our data, these peak expression results indicate that, should a candidate gene not show up-regulation at the 30-minute timepoint, the 10-minute and 60-minute timepoints ought to be considered prime alternative timepoints for expression-based validation.

We detect a small number of principal regulation patterns from the 1400 expression profiles by adapting the correlation-based profile clustering method described earlier (BRADY *et al.* 2007) and determine 16 principal expression profiles (Figure 5.2). These profiles can be further grouped into five regulation pattern groups: *Arch*, *Staggered*, *Pulse*, *Up-tick* and *Hump* (Figure 5.2). The six most common principal expression profiles represent 74.3% (N=1040) of all up-regulated genes (Table 5.2). While 38.8% (N=543) of all up-regulated genes follow either *Arch1* or *Arch2* (Table 5.2), we also find that a similar proportion (35.5%; N=497) exhibit a staggered temporal signature reminiscent of expression profiles found in regulatory cascades in profiles *Stag-3*, *Stag-10*, *Stag-30* and *Stag-60* (Figure 5.2B) (KIM *et al.* 2008; ORLANDO *et al.* 2008; QIAN *et al.* 2001). *Pulse* patterns account for 9.1% (N=127) of up-regulated genes (Table 5.2). This expression profile pattern group is characterized by significant up-regulation at a single measured timepoint (Figure 5.2C). Presumably, these genes are required

transiently at key times during ciliogenesis and these regulation pattern gene sets may be particularly enriched in genes that are essential for ciliogenesis.

We find 78 human genes with *Chlamydomonas* homologs that are up-regulated during ciliogenesis, which were previously predicted to be involved in cilia by comparative genomics studies but lacked any qRT-PCR or proteomic evidence. In this group, there are several interesting genes that have mutant phenotypes or that are associated with human disease (Table 5.3). The up-regulation of these genes suggests additional diseases or mutant phenotypes that may be further studied in *Chlamydomonas*, including autism, dyslexia, carotid artery calcified plaque (CarCP), lysosomal storage disease (Sanfillipo syndrome), episodic ataxia, neuronal degeneration and the short limb dwarf mouse phenotype (Table 5.3).

Our data identifies an additional 188 human genes with *Chlamydomonas* homologs that are up-regulated during ciliogenesis. This set includes eight genes that have existing evidence of their involvement in cilia or ciliogenesis (KAT, KIF21A, TXNDC3, RUVBL1, DYNLT1, TUBB, RUVBL2, ACT). There are another nine genes with existing non-ciliopathy disease genes or mutant phenotypes to which our data assigns novel cilia annotations, thereby indicating some cilia involvement in the associated diseases or mutant phenotypes, including COACH syndrome, Noonan syndrome, cleft lip and palate gene, polycycstic ovary syndrome and perinatal lethality in mice (Table 5.5). Figure 5.3 illustrates how the largest category in the remainder of the set contains 23 solute carriers/transporters out of the 300 that have been found in human, followed by Golgi/membrane trafficking proteins (N=19). The next groups are chaperonins

(N=18), followed by kinases and phosphatases (N=15). The remainder of the coherent protein annotation groups include proteins that are involved in or that are associated with mitochondria (N=13), lipid and inositol metabolism (N=10), cell cycle (N=8), ubiquitin (N=7), proteins that act on disulfide bonds (N=5) and DNA repair (N=3). The mitochondrial annotation group may be of particular interest as there is a strong enrichment for the pattern *Stag-30* (P=2.07E-6). One could hypothesize that, aside from the recently reported roles of IFT20 and GMAP210 involvement at the Golgi (FOLLIT *et al.* 2008), there are genes associated to the mitochondria that are also essential for ciliogenesis that are subject to precise temporal regulation. These results suggest new areas where there may be cilia involvement and indicate the potential of using *Chlamydomonas* as a model organism for the study of these diseases and phenotypes.

5.4 Summary

In this chapter, we leverage the fact that transcript abundance of most genes that encode known cilia components are greatly increased in *Chlamydomonas* during ciliogenesis and perform high-throughput sequencing to measure the *Chlamydomonas* transcriptome at various points during ciliogenesis. Our results lend further support of a ciliary role for 372 genes that have existing *Chlamydomonas* evidence of cilia association and provide novel evidence of ciliogenesis involvement for 289 *Chlamydomonas* homologs of human genes.

Most importantly, our analysis has identified 254 novel ciliopathy human disease genes and many new diseases and mutant phenotypes that may be cilia based or involve cilia in some way. These data provide the necessary evidence to demonstrate, for the first time, a regulation program that hints at the elaborate and carefully tuned process of cilia biosynthesis. Our results reinforce the advantages of using *Chlamydomonas* as a model organism for ciliopathies and exemplify the importance of emerging model organisms in furthering our understanding of human disease.

5.5 Methods

5.5.1 RNAseq of *Chlamydomonas* transcriptome during ciliogenesis

Chlamydomonas cell cultures were grown in 150mL Rich medium (R) to a concentration of 7.2E6 cells per mL and proportion flagellated was 87.5% averaged over two samples. Cells were spun down in 50mL conical tubes in a Sorvall RT6000 for 10 minutes at 3500 RPM in room temperature. Cells were resuspended in 25mL HEPES buffer and a 5mL aliquot was taken and diluted to 50mL in R as “pre-treatment” sample. Acetic acid (0.5N) was added to the remaining 20mL with constant stirring to a pH of 4.1 as measured by a Corning pH meter 240 at 24C. After 45 seconds, pH was restored to 7.1 with 0.5N KOH. Deflagellation was confirmed under 40X magnification with a phase microscope. Deflagellated cells were diluted to 100mL with R then poured into a 600mL beaker and further diluted to 200mL R (22C). Aliquots of 50mL were taken at 0,

7, 27 and 57 minutes, spun in Sorvall RT6000 for 3 minutes at 3500 RPM, bringing the total number of timepoints to five.

RNA is isolated with a Qiagen RNeasy kit using a modified animal tissue protocol. Briefly, homogenization of the lysate was achieved by the needle method but passing the lysate through 20 times, instead of 5 as instructed for animal cells. Isolated RNA was prepared for Illumina sequencing per Illumina protocols and sequenced on the Illumina Gene Sequencing Machine.

Reads were aligned to the v4 Chlamydomonas genome assembly (MERCHANT *et al.* 2007) using the TopHat alignment software suite (TRAPNELL *et al.* 2009). Transcript abundance for 11,315 final GreenGenie2 gene models predicted on the v4 assembly (KWAN *et al.* 2009) were computed in FPKMs using the Cufflinks software suite (TRAPNELL *et al.* 2010).

5.5.2 Expression profile clustering

In this study, we compute the principal regulation profiles of every gene with an FC greater than 0.95 at some timepoint using the clustering method adapted from previous work (BRADY *et al.* 2007). Briefly, genes are first sorted by profile decreasing variance. The top 75% of genes are then grouped by fuzzy c-means clustering, which aims to assign, for each profile a probability membership to a given cluster. In contrast to regular k-means clustering, fuzzy c-means clustering allows multi-cluster membership for a given gene. Once membership is determined, the method determines the appropriate membership probability cut-off such that the average gene is assigned to one cluster (BRADY *et al.* 2007). After c-means clustering, similar clusters are collapsed by combining

clusters that have median profiles with a 1-Pearson correlation distance and cutting the single-linkage hierarchical clustering tree at a similarity cut-off (default: 0.1). Finally, all original input genes are assigned to the resultant "principal profiles" by evaluating the Pearson correlation between each input profile and each principal profile. Every gene is assigned to every profile for which the Pearson correlation coefficient satisfies the Pearson cutoff (default: 0.85). At this step, the method also determines the appropriate membership probability cut-off such that the average gene is assigned to one cluster (BRADY *et al.* 2007).

Chapter 6

Closing remarks and future directions

This work has describes the development of two new computational methods and a computational pipeline for accelerating the annotation of emerging model organisms that can greatly inform our understanding of human pathologies and their underlying biology and genetics. In Chapter 3, we described an effective genefinder training method that does not require a model organism to have any experimentally determined gene models and demonstrated this method on the *Chlamydomonas* genome to define a more accurate gene catalog in GreenGenie2. When we compared the prediction accuracy of GreenGenie2 to the latest *Chlamydomonas* specific genefinder GeneMark.hmm-3.0, we found that our EST-based training method outperforms the competition by a significant margin, making most performance gains on bounding and single exons. Five independent experiments of high-throughput transcriptome sequencing in Chapter 5 lends further support to 98% of the GreenGenie2 v4 gene models.

Chapter 4 described APACE: a novel, scalable comparative genomic method that automatically adjusts alignment scores to correct for biases introduced by differences in evolutionary distance without a predetermined phylogenetic tree. Furthermore, this method determines the appropriate alignment cutoff for each individual protein instead of using an arbitrary constant cutoff E-value for every protein alignment. We demonstrated the predictive performance of APACE on co-crystallization data in yeast and show that it is substantially more sensitive than existing methods. We also demonstrate its efficacy in predicting genes that are essential in the lifecycle of the malaria-causing parasite *Plasmodium falciparum* and found that APACE has a success rate similar to human gene-target selection. Most importantly, we demonstrated that APACE is able to identify a more accurate set of cilia genes than Procom and is thus an effective tool for automated protein characterization in our example emerging ciliopathy model organism, *Chlamydomonas*.

Finally, Chapter 5 uses high-throughput sequencing to detect genes that are up-regulated in *Chlamydomonas* during ciliogenesis. This component annotates 372 human genes with a role in ciliogenesis, identifies new diseases and phenotypes that may have cilia underpinnings and provides the first description of the early ciliogenesis gene regulation program in *Chlamydomonas*.

This dissertation has presented computational methods for accelerating the accurate annotation of the green alga *Chlamydomonas reinhardtii*. However, the methods presented here can be applied to any emerging model organism from gene model, through protein characterization to analyzing how genes are regulated in any process of interest. For example, aside from ciliopathies,

Chlamydomonas is only one of many emerging algal model organisms that have recently been recognized as essential for biofuels research and development. The list of emerging model organisms continues to grow. Some examples include the choanoflagellate *Monosiga brevicolis* for studying animal development and the origins of multicellularity (KING *et al.* 2008), the urochordate *Ciona intestinalis* for the study of the origins of vertebrate life, the zygomycete fungus *Phycomyces blakesleeanus* for the study of signal transduction pathways in response to environmental cues, the placozoan *Trichoplax adhaerens*, which is the simplest known animal with the smallest known animal genome that will be an important model organism for the study of how animal life evolved (SRIVASTAVA *et al.* 2008), and the gastropod snail *Lottia gigantean* as an emerging model in the studies of ecology and conservation.

This work complements any existing methods and resources for a given model organism. The methods developed in this dissertation are designed to help inform the biologist on the bench direct her experiments and also help the computational biologist at the terminal in his efforts to process genomes of multiple emerging model organisms to be as informative for downstream computational analyses as possible.

Appendix A

Colorfy: a heat-map visualization method for CLUSTALW multiple sequence alignments

Multiple sequence alignments (MSA) were color-coded using the online MSA column percentage composition coloring tool, Colorfy (http://bifrost.wustl.edu/colorfy). Colorfy takes as input any standard ALN format MSA such as default CLUSTALW output (LARKIN *et al.* 2007) and outputs the corresponding color-coded MSA. Colorfy groups the twenty amino acids into eight separate conservation groups ([G, A], [V, L, I], [F, Y, W], [C, M], [K, R, H], [D, E, N, Q], [S, T], [P]). Percentage composition is defined on a per column basis and categorized as Majority Identity, Conserved Minority or Insufficient Conservation. A column is Majority Identity when at least 61% of the amino acids in that column are identical. A column is Conserved Minority when at least 61% of the amino acids in that column belong to the same conservation group and no

amino acid makes up more than 60% of that column. A column is Insufficient Conservation when its composition fails to satisfy any of the prior two conditions. Columns are colored based on percentage composition (Blue: 61 to 70; Green: 71 to 80; Gold: 81–90; Red: 91 to 100). Colors codes are divided into two shades, dark and light. A Majority Identity column can have up to two colors in the column: dark to indicate the positions of the identity amino acid and light to indicate positions of amino acids belonging to the same group as the identity amino acid. A Conserved Minority is colored the light color of the corresponding percentage composed of the majority amino acid group. Columns categorized as Insufficient Conservation are left uncolored. If a column satisfies Majority Identity at a lower percentage and Conserved Minority at a higher percentage, the Majority Identity categorization takes precedence and the column is colored per the Majority Identity percentage.

Appendix B

Data table of 1400 *Chlamydomonas* genes that are up-regulated during ciliogenesis

Gene	Profile	MaxTP	Log_2 (MaxFC)
c1_t1000	Arch1	t30	2.957583
c1_t1005	Arch1	t30	2.995049
c1_t1006	UT2	t60	1.639738
c1_t1011	Stag-60	t60	2.061193
c1_t1028	Stag-3	t10	4.273609
c1_t1041	UT3	t10	2.46975
c1_t1065	UT2	t60	1.698629
c1_t1068	Stag-3	t10	4.185475
c1_t1072	Arch1	t30	2.916105
c1_t1075	Stag-30	t60	1.98499
c1_t1088	Stag-30	t30	1.428164
c1_t1090	Stag-30	t30	1.914067
c1_t1092	Arch1	t10	4.101091
c1_t1105	UT4	t30	1.929734
c1_t1107	Stag-60	t60	1.46396
c1_t1110	Pulse-3	t3	1.889931
c1_t114	Arch1	t30	3.201701
c1_t1140	Stag-30	t30	1.642047

Gene	Profile	MaxTP	Log_2 (MaxFC)
c1_t1141	Stag-30	t30	1.913461
c1_t1166	Hump1	t10	2.726867
c1_t1171	Arch2	t10	3.66671
c1_t1185	Pulse-3	t3	1.381461
c1_t12	Stag-60	t60	2.159599
c1_t120	Stag-60	t60	1.546627
c1_t121	Arch1	t30	2.49429
c1_t1216	Stag-30	t60	1.674781
c1_t122	Stag-60	t60	4.13159
c1_t1222	Stag-60	t60	1.472054
c1_t1227	Stag-3	t10	1.917082
c1_t1228	Arch1	t10	2.917072
c1_t1229	Stag-3	t10	2.221433
c1_t123	Stag-30	t60	4.524522
c1_t1235	Pulse-3	t3	1.473121
c1_t1240	Arch2	t10	3.717335
c1_t1243	Arch1	t10	3.401731
c1_t1249	Stag-30	t30	1.879572
c1_t1256	Stag-60	t60	1.546627
c1_t1272	Stag-3	t10	2.986638
c1_t1273	Arch1	t10	3.638549
c1_t1278	UT3	t10	1.632164
c1_t1293	Arch2	t10	3.387405
c1_t1294	Stag-10	t30	1.601002
c1_t1309	Arch1	t10	2.9973
c1_t1343	UT3	t10	3.206274
c1_t1349	Arch1	t30	1.56166
c1_t1376	Stag-10	t30	2.695762
c1_t1383	Arch1	t30	3.388285
c1_t1401	Stag-10	t30	2.067772
c1_t1428	Arch1	t10	3.058081
c1_t1448	Hump1	t10	2.692834
c1_t1452	Arch1	t10	2.230767
c1_t1459	Arch1	t10	2.600645
c1_t1460	Stag-3	t10	4.005455
c1_t1468	Stag-60	t60	1.618776
c1_t1478	Arch1	t10	2.127658
c1_t1484	Stag-60	t60	1.39705
c1_t1508	Arch2	t30	1.879572
c1_t1509	outlier	t30	2.292803

Gene	Profile	MaxTP	Log_2 (MaxFC)
c1_t1513	Stag-3	t10	3.979876
c1_t1518	Arch1	t30	1.784513
c1_t1523	Arch1	t10	1.942993
c1_t183	Pulse-3	t3	3.673058
c1_t184	Pulse-3	t3	3.349866
c1_t194	Stag-3	t10	2.925973
c1_t207	UT1	t60	3.745929
c1_t210	ambiguous	t10	1.873801
c1_t216	Pulse-3	t3	4.522921
c1_t230	Stag-60	t60	1.453559
c1_t231	Stag-10	t30	1.983908
c1_t245	Stag-30	t30	1.842394
c1_t255	Pulse-3	t3	2.058091
c1_t278	Pulse-10	t10	1.426889
c1_t282	UT3	t10	4.22134
c1_t284	Arch1	t10	3.047405
c1_t292	Stag-30	t30	1.61654
c1_t306	Arch1	t10	1.851367
c1_t316	Stag-60	t60	1.3767
c1_t317	UT1	t10	2.374056
c1_t342	Hump1	t10	2.941583
c1_t349	Arch1	t10	3.565174
c1_t36	Arch1	t10	1.450803
c1_t37	Arch1	t10	1.424969
c1_t371	Arch1	t10	2.695257
c1_t380	Arch1	t10	3.220211
c1_t39	Arch1	t30	1.479729
c1_t401	Stag-30	t30	1.378212
c1_t404	ambiguous	t10	3.119554
c1_t435	Stag-30	t60	1.589691
c1_t443	Pulse-3	t10	3.190246
c1_t445	Arch2	t10	3.676247
c1_t452	Stag-30	t30	1.582593
c1_t455	Stag-60	t60	1.698629
c1_t472	Hump1	t10	3.021566
c1_t492	Stag-60	t60	1.524337
c1_t50	Hump1	t10	2.417596
c1_t51	UT3	t10	1.870483
c1_t517	Pulse-10	t10	1.371139
c1_t520	Arch1	t30	7.195485

Gene	Profile	MaxTP	Log_2 (MaxFC)
c1_t538	Arch2	t10	2.158301
c1_t541	Stag-3	t10	3.931532
c1_t559	Arch1	t30	2.30824
c1_t581	Stag-30	t60	1.546627
c1_t586	Arch1	t10	2.82849
c1_t587	Arch1	t10	2.931931
c1_t598	Arch1	t10	1.376416
c1_t606	Arch2	t10	4.434138
c1_t610	Arch1	t30	2.703798
c1_t616	Arch1	t30	2.005692
c1_t617	Pulse-3	t10	3.486893
c1_t618	UT3	t10	1.607956
c1_t620	Stag-30	t60	2.786464
c1_t628	Arch2	t10	3.990826
c1_t631	Pulse-3	t3	2.222125
c1_t65	Stag-60	t60	1.772452
c1_t659	Stag-30	t30	1.974732
c1_t662	Stag-30	t30	1.726372
c1_t663	UT1	t60	1.815819
c1_t671	Stag-30	t60	1.91727
c1_t688	Stag-30	t30	1.999128
c1_t691	Arch1	t10	2.606315
c1_t693	Stag-60	t60	1.407223
c1_t695	Hump2	t30	1.661393
c1_t701	Stag-60	t60	1.482903
c1_t712	Arch1	t30	3.576268
c1_t721	Pulse-10	t10	2.677426
c1_t736	Stag-3	t10	3.221437
c1_t738	Stag-30	t60	2.638733
c1_t743	Stag-3	t10	3.245891
c1_t749	UT1	t10	3.569357
c1_t755	Arch1	t10	2.338479
c1_t758	Stag-30	t60	5.519174
c1_t759	Stag-60	t60	3.041778
c1_t760	Stag-60	t60	3.236282
c1_t773	Stag-30	t60	1.600324
c1_t775	Arch2	t10	4.582028
c1_t795	Arch1	t10	3.896431
c1_t804	Arch1	t10	2.548391
c1_t807	Hump1	t10	2.194094

Gene	Profile	MaxTP	Log_2 (MaxFC)
c1_t810	Stag-10	t30	3.040033
c1_t813	Hump2	t10	1.414398
c1_t830	Arch1	t30	1.384047
c1_t832	Pulse-3	t3	1.637964
c1_t838	UT3	t10	2.977016
c1_t843	Stag-30	t60	2.951754
c1_t85	Stag-60	t60	1.546627
c1_t863	Arch1	t10	1.721712
c1_t89	Arch2	t10	3.693732
c1_t90	Arch1	t30	2.776813
c1_t901	Stag-10	t30	2.077394
c1_t902	Arch1	t30	2.520388
c1_t910	Arch2	t10	3.977352
c1_t924	Pulse-10	t10	2.055004
c1_t926	Stag-60	t60	1.961661
c1_t927	Stag-60	t60	2.283599
c1_t933	Arch1	t10	3.447031
c1_t935	ambiguous	t10	1.664711
c1_t975	UT1	t60	2.013497
c10_t100	UT2	t30	3.464531
c10_t101	UT2	t60	3.574912
c10_t102	Stag-60	t60	1.546627
c10_t104	Arch1	t30	2.432283
c10_t110	Stag-3	t10	4.004907
c10_t131	Stag-60	t60	1.485226
c10_t136	Stag-3	t10	4.114653
c10_t141	Arch1	t10	3.669783
c10_t143	UT3	t10	2.761996
c10_t16	Arch1	t10	3.857521
c10_t183	Stag-60	t60	1.961661
c10_t191	Arch1	t30	3.309672
c10_t196	Stag-30	t60	1.806687
c10_t197	Arch1	t10	3.992933
c10_t207	Arch2	t10	3.335944
c10_t216	Stag-30	t60	2.315295
c10_t224	Stag-60	t60	2.855454
c10_t227	Arch1	t30	2.144191
c10_t230	Arch1	t30	3.283963
c10_t245	Hump1	t10	2.376465
c10_t250	Stag-30	t60	2.655266

Gene	Profile	MaxTP	Log_2 (MaxFC)
c10_t266	Hump1	t10	2.055177
c10_t267	Hump2	t30	2.166928
c10_t279	Arch1	t30	2.821825
c10_t282	Arch1	t10	1.989779
c10_t295	Stag-3	t10	3.949421
c10_t310	Arch1	t30	3.620256
c10_t324	Stag-60	t60	2.808985
c10_t325	Stag-30	t60	6.781835
c10_t337	Stag-30	t30	1.650097
c10_t342	Hump1	t10	2.209415
c10_t344	Arch2	t10	2.878667
c10_t359	Stag-60	t60	2.148086
c10_t365	Stag-30	t60	1.615487
c10_t385	Pulse-3	t3	1.492497
c10_t424	Stag-60	t60	1.654699
c10_t425	Stag-30	t60	2.084781
c10_t426	Arch1	t30	2.313607
c10_t439	Arch1	t30	2.207713
c10_t447	Stag-30	t30	1.755414
c10_t454	Stag-10	t60	2.588238
c10_t46	Hump1	t10	1.838354
c10_t472	Hump2	t30	1.752196
c10_t48	Arch1	t10	1.536441
c10_t489	ambiguous	t3	1.405241
c10_t491	Arch1	t10	2.509236
c10_t506	Arch1	t10	3.498795
c10_t517	Arch1	t30	2.163033
c10_t526	Arch1	t30	2.184428
c10_t530	Stag-60	t60	2.758895
c10_t533	Stag-60	t60	2.541509
c10_t539	Arch2	t10	3.197503
c10_t540	Pulse-30	t30	2.20288
c10_t548	Pulse-10	t10	2.261338
c10_t566	Pulse-30	t30	1.464537
c10_t579	Stag-60	t60	2.184053
c10_t589	Stag-30	t60	1.546627
c10_t602	UT2	t30	2.615736
c10_t604	Stag-60	t60	1.502235
c10_t607	Arch1	t10	1.401027
c10_t608	Arch1	t10	2.363639

Gene	Profile	MaxTP	Log_2 (MaxFC)
c10_t611	outlier	t10	1.909032
c10_t616	Stag-10	t30	3.607877
c10_t622	Stag-30	t60	1.812588
c10_t632	Arch1	t10	3.422938
c10_t633	Arch1	t10	1.787225
c10_t640	Arch1	t10	2.67278
c10_t650	Arch1	t10	3.292706
c10_t661	Arch1	t30	3.319685
c10_t663	Arch1	t30	3.355045
c10_t671	Stag-60	t60	1.3767
c10_t705	Stag-30	t60	1.729488
c10_t709	Arch1	t30	3.013429
c10_t711	Arch1	t10	1.62068
c10_t716	Pulse-3	t3	1.492497
c10_t717	Arch1	t30	2.250763
c10_t780	Stag-60	t60	3.194329
c10_t791	Stag-60	t60	1.835296
c10_t797	UT3	t10	1.498975
c10_t8	Arch2	t10	1.917948
c10_t805	Arch1	t30	2.164302
c10_t817	Stag-3	t10	3.730059
c10_t819	Arch1	t30	2.368963
c10_t82	UT2	t60	1.819642
c10_t827	Arch1	t10	1.475516
c10_t869	UT1	t60	3.746766
c10_t894	Arch1	t10	2.748709
c10_t919	Stag-60	t60	1.403095
c10_t92	Hump1	t10	3.200128
c10_t940	Arch1	t30	3.671399
c10_t961	Stag-10	t30	3.899475
c10_t973	Arch2	t10	3.524504
c10_t974	Arch1	t10	3.616389
c10_t980	Stag-30	t30	1.444859
c10_t981	Stag-3	t10	3.04514
c10_t982	Arch1	t10	2.536878
c10_t99	Stag-60	t60	1.406449
c11_t101	Stag-60	t60	1.9904
c11_t11	Stag-60	t60	1.894547
c11_t128	Arch1	t10	2.718066
c11_t13	Arch1	t10	3.45322

Gene	Profile	MaxTP	Log_2 (MaxFC)
c11_t143	Arch1	t10	3.922067
c11_t15	Stag-10	t30	2.623728
c11_t154	Arch1	t10	4.977947
c11_t158	Arch1	t10	3.465498
c11_t170	Arch1	t30	3.014901
c11_t180	Hump1	t10	1.974862
c11_t191	Stag-10	t10	2.341307
c11_t196	Arch1	t30	3.53518
c11_t205	Arch1	t10	1.417787
c11_t206	Arch1	t10	3.831654
c11_t212	Arch1	t30	3.117246
c11_t220	UT2	t60	1.469457
c11_t222	Arch1	t30	3.189092
c11_t23	Stag-10	t30	1.387274
c11_t249	Stag-10	t30	1.994136
c11_t25	Arch1	t30	1.739501
c11_t271	Stag-60	t60	1.815834
c11_t283	Arch1	t10	2.595711
c11_t315	UT3	t10	1.437164
c11_t317	Arch2	t10	4.071026
c11_t319	Arch1	t10	3.255917
c11_t321	Arch1	t10	2.897639
c11_t329	Hump1	t10	1.762007
c11_t33	Arch1	t10	4.086059
c11_t34	Pulse-3	t3	2.991298
c11_t347	Pulse-3	t3	2.299843
c11_t35	Arch1	t10	3.13124
c11_t51	UT1	t60	1.482759
c11_t79	ambiguous	t60	1.832179
c11_t82	Pulse-10	t10	1.714355
c11_t93	Arch1	t30	2.942579
c11_t94	Arch1	t10	1.505063
c11_t98	Arch1	t10	1.727757
c12_t1017	Arch1	t10	1.96068
c12_t1019	Stag-60	t60	1.6621
c12_t1020	Arch1	t30	3.139131
c12_t1032	Arch1	t30	1.422917
c12_t1048	Arch1	t10	2.20868
c12_t1059	Arch1	t10	2.277799
c12_t106	Stag-30	t30	1.520673

Gene	Profile	MaxTP	Log_2 (MaxFC)
c12_t1060	Arch2	t10	2.625258
c12_t1061	Arch1	t10	3.542278
c12_t1063	Stag-30	t30	3.103396
c12_t107	Stag-30	t60	3.090844
c12_t1072	Arch1	t30	2.952216
c12_t1086	outlier	t60	5.371053
c12_t1100	Arch1	t30	1.924469
c12_t1105	Stag-3	t10	3.638001
c12_t1117	Arch1	t30	1.554562
c12_t1159	Arch1	t10	3.50924
c12_t1174	UT1	t60	2.546631
c12_t1181	Stag-60	t60	1.799632
c12_t1184	ambiguous	t10	2.973914
c12_t1208	Arch1	t30	2.986436
c12_t1216	Arch1	t10	1.785133
c12_t1217	UT1	t60	1.809515
c12_t1224	Arch1	t10	2.424968
c12_t1234	Arch1	t10	4.238494
c12_t1243	Arch1	t30	3.388732
c12_t1244	Stag-30	t60	1.871882
c12_t1254	Stag-30	t60	2.426786
c12_t1261	Stag-60	t60	1.836132
c12_t1279	Arch1	t30	3.387592
c12_t1284	Arch1	t10	3.886721
c12_t1288	Arch1	t30	3.15402
c12_t1291	Stag-30	t30	4.170341
c12_t1292	Stag-60	t60	1.515465
c12_t1293	Stag-60	t60	2.028775
c12_t130	Stag-30	t60	2.131582
c12_t1300	Pulse-3	t3	2.145201
c12_t1325	Stag-30	t60	2.008967
c12_t1327	Hump1	t10	1.808462
c12_t133	Arch1	t10	3.609291
c12_t1336	Stag-10	t60	4.050035
c12_t1337	UT4	t30	1.879572
c12_t1338	Arch1	t10	2.009357
c12_t134	UT3	t10	1.526242
c12_t1355	Arch1	t30	3.302387
c12_t138	Arch1	t10	3.697858
c12_t1385	Stag-60	t60	1.462157

Gene	Profile	MaxTP	Log_2 (MaxFC)
c12_t1388	Arch1	t10	1.481504
c12_t1389	Hump1	t10	2.009934
c12_t1393	Arch2	t10	3.731617
c12_t142	Arch2	t30	1.899928
c12_t1429	Arch2	t10	2.60356
c12_t1437	Arch1	t10	3.29852
c12_t1456	Arch1	t30	3.01203
c12_t1459	Stag-10	t30	2.114789
c12_t1462	Stag-30	t30	1.814478
c12_t1463	UT1	t60	1.584599
c12_t1487	UT4	t30	2.421881
c12_t149	UT5	t3	1.909248
c12_t1491	Stag-30	t60	1.934178
c12_t1492	Arch2	t10	4.581206
c12_t1507	Arch1	t30	3.015651
c12_t1516	Pulse-3	t3	2.229454
c12_t1518	Stag-10	t30	1.807019
c12_t1530	Arch1	t10	3.90236
c12_t1531	Arch1	t10	3.45876
c12_t1532	Arch2	t10	3.074701
c12_t1533	Stag-60	t60	3.421092
c12_t1536	Arch1	t30	3.209001
c12_t156	Stag-30	t30	1.397422
c12_t1561	Arch1	t30	2.04642
c12_t1570	Stag-30	t60	2.643537
c12_t1571	Stag-3	t10	4.012741
c12_t17	Stag-10	t30	2.885924
c12_t196	Arch1	t10	2.346962
c12_t201	Pulse-10	t10	1.679542
c12_t226	Arch1	t30	3.161161
c12_t230	Pulse-30	t30	1.5964
c12_t239	Arch1	t30	2.696714
c12_t241	Arch1	t10	1.526718
c12_t242	Pulse-10	t10	1.407027
c12_t243	Arch1	t10	3.898133
c12_t264	Arch1	t10	4.853544
c12_t281	Arch1	t10	2.160681
c12_t283	Stag-3	t10	4.473018
c12_t288	Arch1	t10	3.723279
c12_t298	Arch2	t10	5.032842

Gene	Profile	MaxTP	Log_2 (MaxFC)
c12_t299	Stag-10	t60	2.916668
c12_t303	Arch1	t10	3.789556
c12_t309	Stag-30	t60	1.6621
c12_t32	Arch2	t10	3.905462
c12_t33	Hump1	t10	2.334252
c12_t338	Arch1	t10	3.706904
c12_t34	Stag-3	t10	3.96783
c12_t345	Arch1	t30	2.804585
c12_t35	Hump1	t10	2.69934
c12_t357	Stag-30	t30	1.519345
c12_t36	Stag-3	t10	4.017935
c12_t374	Stag-3	t10	3.907727
c12_t390	UT2	t3	2.330328
c12_t408	UT1	t60	2.890584
c12_t409	ambiguous	t10	3.426487
c12_t42	Stag-60	t60	2.131582
c12_t429	Stag-3	t10	3.448503
c12_t43	Stag-3	t3	1.756438
c12_t435	Arch1	t10	3.49653
c12_t437	Hump1	t10	2.122911
c12_t438	Stag-10	t30	3.142608
c12_t450	Arch1	t30	3.238201
c12_t467	Arch2	t10	4.652778
c12_t473	Stag-60	t60	1.836132
c12_t508	Stag-60	t60	1.374673
c12_t517	Arch2	t10	2.268638
c12_t519	Stag-30	t60	1.394049
c12_t521	Arch2	t10	3.332409
c12_t529	Arch1	t10	3.262006
c12_t543	Stag-10	t10	2.855483
c12_t552	Arch1	t10	1.622599
c12_t553	Arch1	t10	3.771479
c12_t557	Stag-3	t10	3.968551
c12_t558	Pulse-3	t3	4.074142
c12_t566	Arch1	t10	2.234071
c12_t605	Stag-3	t10	3.518575
c12_t606	Arch1	t10	3.744255
c12_t625	Arch1	t10	1.598189
c12_t637	Arch1	t10	2.870184
c12_t64	Stag-3	t10	3.9458

Gene	Profile	MaxTP	Log_2 (MaxFC)
c12_t672	Stag-30	t30	1.670569
c12_t679	Arch2	t10	3.358551
c12_t684	Stag-60	t60	3.206173
c12_t693	Stag-3	t10	2.08393
c12_t696	Arch1	t30	2.974924
c12_t701	Stag-60	t60	1.961661
c12_t742	Arch1	t30	1.537163
c12_t747	Arch1	t30	3.56073
c12_t748	Arch1	t30	3.469148
c12_t75	Arch1	t10	4.118635
c12_t760	Stag-30	t60	2.472621
c12_t767	Arch1	t30	3.786469
c12_t770	Arch1	t10	2.469057
c12_t776	Stag-30	t30	1.948706
c12_t778	Arch1	t30	2.525481
c12_t780	Hump1	t10	2.125667
c12_t798	UT1	t60	1.389525
c12_t822	UT1	t60	2.039668
c12_t831	Stag-30	t30	1.872748
c12_t834	Arch1	t10	2.639324
c12_t87	Hump1	t10	2.10531
c12_t89	Arch1	t30	1.702914
c12_t893	Arch1	t30	3.139694
c12_t894	Arch1	t30	3.421755
c12_t910	ambiguous	t3	2.366958
c12_t915	Arch1	t30	3.480473
c12_t921	Arch1	t30	2.155213
c12_t931	Arch2	t10	3.605497
c12_t932	outlier	t3	2.068839
c12_t943	Stag-60	t60	1.421874
c12_t963	Pulse-30	t30	1.440289
c12_t964	Arch1	t10	3.318458
c12_t970	Arch1	t10	3.921548
c12_t978	Stag-60	t60	1.737885
c13_t117	Arch1	t10	3.506629
c13_t122	Arch1	t10	1.760708
c13_t143	Stag-60	t60	2.365659
c13_t150	Stag-60	t60	2.296352
c13_t151	Hump2	t30	1.645105
c13_t157	Hump2	t30	1.714758

Gene	Profile	MaxTP	Log_2 (MaxFC)
c13_t158	Pulse-10	t10	3.171044
c13_t160	Pulse-10	t10	2.628028
c13_t166	Stag-60	t60	3.605526
c13_t167	Stag-30	t30	2.464542
c13_t168	Arch1	t10	6.253852
c13_t175	Arch1	t30	3.130793
c13_t180	Hump2	t30	1.511843
c13_t187	Stag-60	t60	2.187371
c13_t221	Hump2	t10	2.19496
c13_t222	Arch1	t30	2.951047
c13_t235	Arch1	t10	3.773888
c13_t248	Arch1	t10	2.826196
c13_t291	Hump2	t10	1.460729
c13_t306	Arch1	t30	2.857301
c13_t31	Stag-10	t30	2.747959
c13_t33	Arch1	t30	3.295895
c13_t337	ambiguous	t60	2.965099
c13_t347	Stag-30	t60	2.196128
c13_t374	Stag-30	t60	4.889741
c13_t377	Stag-30	t60	1.824158
c13_t39	Hump1	t10	2.804888
c13_t390	Arch1	t10	3.450941
c13_t392	Arch1	t30	2.497002
c13_t393	Pulse-30	t30	1.439704
c13_t397	Stag-60	t60	2.097477
c13_t4	Pulse-10	t10	2.346962
c13_t411	Arch1	t30	2.982469
c13_t420	Stag-10	t30	2.512641
c13_t421	Stag-60	t60	2.925136
c13_t423	Stag-30	t30	2.879576
c13_t425	Arch1	t10	3.221437
c13_t43	Arch2	t10	3.083934
c13_t432	UT1	t60	1.639738
c13_t448	Stag-10	t60	1.818459
c13_t462	Arch1	t10	3.867562
c13_t464	Arch1	t10	3.629345
c13_t466	Stag-10	t30	1.548372
c13_t509	outlier	t30	1.912913
c13_t519	UT2	t60	3.931431
c13_t52	Arch1	t30	2.957409

Gene	Profile	MaxTP	Log_2 (MaxFC)
c13_t54	Arch1	t30	1.409353
c13_t560	Stag-30	t30	2.567435
c13_t572	Arch1	t60	4.821617
c13_t573	Stag-30	t30	1.900635
c13_t574	Arch1	t60	5.119836
c13_t591	Stag-60	t60	1.377099
c13_t630	Stag-30	t30	1.49557
c13_t662	UT2	t60	2.882231
c13_t665	UT5	t60	1.546627
c13_t672	Arch1	t10	3.057908
c13_t682	Hump1	t10	1.931192
c13_t686	Stag-3	t3	1.525174
c13_t688	Pulse-3	t3	4.564211
c13_t716	Arch1	t10	4.315014
c13_t727	Stag-30	t60	5.843175
c13_t728	Stag-60	t60	2.197455
c13_t753	Stag-3	t10	4.378897
c13_t760	Stag-60	t60	1.474117
c13_t762	Stag-30	t30	2.41562
c13_t786	Arch1	t10	2.391887
c13_t8	ambiguous	t3	3.176396
c13_t807	Stag-60	t60	2.283599
c13_t808	UT2	t60	4.909635
c13_t81	Stag-60	t60	1.46862
c13_t823	Stag-60	t60	1.769018
c13_t852	Arch1	t10	2.38678
c13_t865	Hump1	t10	2.543053
c13_t879	UT1	t60	2.406098
c13_t887	Stag-60	t60	3.427714
c13_t911	Arch2	t10	1.457108
c13_t916	Stag-60	t60	1.384087
c13_t932	UT3	t60	2.168573
c14_t113	Arch1	t10	2.796347
c14_t116	Arch1	t10	2.759688
c14_t117	Arch1	t10	2.146673
c14_t123	Pulse-30	t30	1.37392
c14_t138	Stag-60	t60	1.387929
c14_t163	Stag-30	t60	2.176738
c14_t180	Stag-60	t60	1.705395
c14_t190	Arch1	t30	3.089026

Gene	Profile	MaxTP	Log_2 (MaxFC)
c14_t193	Stag-10	t60	2.441747
c14_t194	Stag-30	t60	1.971299
c14_t195	Arch1	t10	3.62858
c14_t196	Arch1	t10	3.788128
c14_t20	Arch1	t10	1.47983
c14_t211	Stag-3	t10	3.521445
c14_t212	Arch1	t10	2.066156
c14_t224	UT2	t60	1.517513
c14_t254	Stag-30	t60	1.769018
c14_t258	Pulse-3	t3	2.196835
c14_t26	UT2	t60	1.961661
c14_t270	Stag-60	t60	2.611797
c14_t318	Pulse-30	t30	1.702698
c14_t319	Pulse-10	t10	1.763724
c14_t326	Arch2	t10	3.978058
c14_t345	Arch2	t10	3.645243
c14_t346	Arch2	t10	3.575763
c14_t356	Stag-10	t30	2.836512
c14_t359	Stag-30	t60	2.378586
c14_t36	Stag-60	t60	2.108657
c14_t394	Stag-30	t60	3.06476
c14_t408	Stag-10	t30	3.110422
c14_t409	Arch1	t10	4.017704
c14_t442	Arch2	t10	4.434585
c14_t448	Arch2	t10	1.498513
c14_t472	Stag-60	t60	1.546627
c14_t475	Pulse-3	t3	1.690204
c14_t480	Arch1	t30	4.08023
c14_t489	Arch2	t10	3.732887
c14_t505	Arch1	t10	3.581577
c14_t521	Stag-60	t60	2.447085
c14_t54	Arch2	t10	3.102631
c14_t62	UT2	t60	1.41872
c14_t63	Arch1	t30	2.520792
c14_t86	Stag-3	t10	4.561095
c14_t89	Arch1	t10	2.20249
c14_t96	Arch1	t30	3.257317
c14_t97	Arch1	t10	3.702417
c15_t107	Pulse-10	t10	1.812948
c15_t108	Pulse-10	t10	1.437484

Gene	Profile	MaxTP	Log_2 (MaxFC)
c15_t109	Pulse-10	t10	1.487808
c15_t115	Arch1	t10	4.276408
c15_t128	Stag-60	t60	1.525563
c15_t194	ambiguous	t60	1.961661
c15_t246	Hump1	t10	2.579582
c15_t268	Pulse-3	t3	4.177785
c15_t273	Stag-10	t30	1.673223
c15_t293	Arch1	t10	5.073107
c15_t305	Arch1	t30	3.190621
c15_t307	Stag-3	t10	3.372054
c15_t323	Pulse-30	t30	2.467961
c15_t338	Stag-10	t30	3.777957
c15_t340	Stag-30	t30	3.155419
c15_t41	Stag-60	t60	1.620709
c15_t48	Arch2	t10	3.105416
c16_t11	Hump2	t10	2.091821
c16_t124	Arch1	t10	3.095259
c16_t127	Pulse-3	t3	4.493447
c16_t130	Arch1	t10	4.252704
c16_t146	UT2	t60	2.283599
c16_t152	Stag-60	t60	1.591971
c16_t162	Arch1	t10	2.992395
c16_t166	Hump1	t10	1.589432
c16_t17	Pulse-30	t30	2.161157
c16_t171	Stag-60	t60	2.835819
c16_t176	Stag-10	t30	1.397924
c16_t18	Pulse-30	t30	2.559918
c16_t194	Arch1	t30	2.149428
c16_t198	Stag-60	t60	1.972554
c16_t199	Stag-60	t60	2.464166
c16_t208	Arch1	t10	2.190704
c16_t209	Arch1	t10	2.398913
c16_t215	Pulse-3	t3	1.492497
c16_t219	Arch1	t10	2.366337
c16_t220	Arch1	t10	2.156079
c16_t222	UT1	t60	3.54662
c16_t231	Arch1	t30	2.066228
c16_t265	Arch2	t10	3.248906
c16_t274	Pulse-10	t10	1.480696
c16_t28	Stag-3	t10	2.952071

Gene	Profile	MaxTP	Log_2 (MaxFC)
c16_t295	Stag-60	t60	1.86855
c16_t30	Stag-60	t60	2.342086
c16_t302	Stag-10	t30	2.989495
c16_t307	Stag-60	t60	1.605518
c16_t324	Arch1	t10	2.919164
c16_t333	Stag-30	t30	1.424459
c16_t335	Arch1	t10	2.596173
c16_t345	Arch1	t10	3.013631
c16_t346	Pulse-10	t10	1.979782
c16_t353	Arch1	t30	3.414383
c16_t359	Arch1	t30	5.53301
c16_t364	Arch1	t10	3.455341
c16_t370	UT4	t30	1.410089
c16_t372	Arch1	t10	1.825557
c16_t4	Pulse-3	t3	2.229454
c16_t404	Pulse-3	t3	3.110739
c16_t405	Pulse-3	t3	3.750762
c16_t439	Pulse-3	t3	3.156862
c16_t44	Stag-3	t3	1.720096
c16_t457	Arch1	t10	1.736731
c16_t485	Stag-60	t60	1.402339
c16_t49	Pulse-10	t10	1.751028
c16_t50	Stag-60	t60	1.974487
c16_t54	Arch1	t10	2.065348
c16_t546	Arch1	t10	2.827163
c16_t547	Stag-3	t10	4.07645
c16_t549	Stag-60	t60	1.410534
c16_t572	Stag-60	t60	1.769018
c16_t58	Stag-3	t10	3.438519
c16_t590	Stag-10	t30	2.245641
c16_t596	UT5	t60	4.13159
c16_t597	UT2	t60	1.618776
c16_t61	Arch1	t30	1.444354
c16_t610	UT5	t60	2.283599
c16_t613	Stag-60	t60	1.546627
c16_t620	Stag-60	t60	1.597972
c16_t625	Arch2	t10	4.426087
c16_t63	outlier	t3	1.492497
c16_t642	Stag-60	t60	2.10671
c16_t651	Arch1	t10	3.765838

Gene	Profile	MaxTP	Log_2 (MaxFC)
c16_t659	Stag-30	t30	3.045356
c16_t666	Arch1	t10	2.219283
c16_t677	Pulse-10	t10	1.604262
c16_t696	Stag-30	t60	2.283599
c16_t705	Stag-30	t60	1.475891
c16_t715	ambiguous	t3	3.979833
c16_t728	Stag-60	t60	1.466961
c16_t768	Arch1	t30	2.87897
c16_t77	Stag-30	t30	1.670468
c16_t824	Arch2	t10	3.308071
c16_t829	Arch1	t10	2.907117
c16_t846	UT1	t60	2.107431
c16_t849	Arch1	t10	3.366991
c16_t860	Pulse-30	t30	1.420143
c16_t861	Stag-30	t60	2.605522
c16_t862	Hump1	t10	1.984398
c16_t869	Arch1	t10	3.507754
c16_t87	Arch2	t10	2.757409
c16_t885	Stag-30	t60	2.305614
c16_t903	Pulse-10	t10	1.370624
c16_t92	Arch1	t30	3.271888
c16_t926	Stag-10	t30	3.295115
c16_t927	Stag-60	t60	2.955447
c16_t928	Stag-10	t30	3.351554
c16_t929	Stag-60	t60	1.395002
c16_t930	Stag-60	t60	3.436702
c16_t931	Stag-60	t60	2.421102
c16_t945	UT4	t30	2.800141
c16_t953	Stag-60	t60	3.527721
c16_t954	Stag-60	t60	2.977968
c16_t961	Stag-60	t60	2.546631
c16_t962	ambiguous	t10	2.488086
c16_t970	Stag-3	t3	1.644499
c16_t976	Stag-30	t30	1.880149
c16_t978	Stag-60	t60	2.927776
c16_t987	Arch2	t10	1.762007
c16_t995	Arch1	t10	2.560856
c16_t996	Arch1	t10	2.219009
c17_t140	Pulse-3	t3	1.492497
c17_t157	Arch1	t10	3.521734

Gene	Profile	MaxTP	Log_2 (MaxFC)
c17_t158	Arch1	t10	3.609637
c17_t162	Arch1	t10	1.993328
c17_t167	Arch1	t10	3.476881
c17_t168	Arch1	t30	1.870757
c17_t178	Arch1	t30	2.290423
c17_t215	UT4	t30	1.783791
c17_t224	Arch1	t10	2.372425
c17_t235	Stag-10	t30	1.620695
c17_t239	Stag-30	t60	1.862837
c17_t280	Arch1	t10	1.472761
c17_t300	Stag-30	t60	2.8543
c17_t303	Arch2	t10	3.415768
c17_t324	Stag-60	t60	2.217999
c17_t33	Arch1	t10	4.137043
c17_t332	ambiguous	t3	3.515213
c17_t333	Pulse-3	t3	3.570814
c17_t34	UT3	t10	2.346962
c17_t341	Stag-60	t60	1.900116
c17_t363	Pulse-3	t3	1.372166
c17_t365	Arch1	t30	2.678666
c17_t366	Stag-10	t60	4.13159
c17_t367	Stag-10	t30	2.045164
c17_t368	Pulse-30	t30	1.60295
c17_t369	Pulse-3	t3	1.54126
c17_t381	Arch2	t10	2.888333
c17_t388	Arch2	t10	3.951931
c17_t389	Hump1	t10	1.886569
c17_t422	UT3	t10	2.700148
c17_t426	Stag-60	t60	1.447095
c17_t435	Arch2	t10	3.523364
c17_t45	Arch1	t30	2.419587
c17_t455	Arch1	t30	3.633024
c17_t458	Arch1	t10	3.275538
c17_t491	Pulse-10	t10	1.930831
c17_t492	Pulse-10	t10	1.408017
c17_t497	Stag-30	t30	2.044587
c17_t499	Stag-30	t30	2.130269
c17_t502	Arch1	t10	3.122006
c17_t518	Arch1	t10	2.057312
c17_t521	Stag-30	t60	1.55146

Gene	Profile	MaxTP	Log_2 (MaxFC)
c17_t537	Arch1	t10	2.140512
c17_t538	Stag-3	t10	3.50497
c17_t541	Stag-10	t30	3.636803
c17_t542	Stag-10	t30	3.002508
c17_t545	Arch1	t10	2.696714
c17_t551	Arch2	t10	3.27297
c17_t552	Arch2	t10	3.600404
c17_t553	Pulse-10	t10	1.462446
c17_t561	Stag-3	t10	3.657665
c17_t564	Arch1	t30	2.552257
c17_t57	Stag-60	t60	1.3767
c17_t588	UT1	t10	2.990159
c17_t593	Stag-30	t60	1.597424
c17_t598	UT2	t60	5.090953
c17_t6	Stag-30	t30	1.481922
c17_t600	Arch1	t10	1.76042
c17_t610	Arch1	t10	3.636948
c17_t627	Hump1	t10	2.116852
c17_t634	Arch1	t30	2.161042
c17_t638	Arch1	t30	2.466576
c17_t642	UT5	t3	1.907531
c17_t666	Stag-60	t60	2.693742
c17_t670	Arch1	t30	2.675146
c17_t679	Stag-30	t30	1.772654
c17_t684	ambiguous	t60	1.961661
c17_t69	Arch1	t10	1.887536
c17_t697	Arch1	t30	2.273731
c17_t708	Stag-10	t30	2.021172
c17_t71	Arch1	t30	1.530122
c17_t721	Stag-3	t10	3.45739
c17_t729	Stag-60	t60	1.769018
c17_t731	Stag-30	t60	2.168154
c17_t739	Stag-60	t60	1.599098
c17_t787	Stag-60	t60	2.077135
c17_t789	Arch1	t30	2.604858
c17_t798	Arch1	t10	5.526575
c17_t799	Arch1	t10	4.904961
c17_t800	Arch1	t10	4.253858
c17_t828	Arch1	t10	2.823441
c17_t829	Pulse-10	t10	1.550493

Gene	Profile	MaxTP	Log_2 (MaxFC)
c17_t832	Pulse-10	t10	1.762007
c17_t839	UT2	t60	1.624633
c17_t856	Arch1	t30	2.740861
c17_t870	Arch2	t10	4.743091
c17_t871	Arch1	t10	2.499036
c17_t873	Arch2	t10	2.289817
c17_t890	Arch2	t10	3.345119
c17_t897	Arch1	t30	2.767681
c17_t905	Stag-60	t60	2.661844
c17_t912	Stag-10	t30	1.863414
c17_t935	Hump1	t10	3.126998
c2_t1001	Pulse-3	t3	1.787687
c2_t1005	Pulse-3	t3	2.121454
c2_t1015	Stag-10	t30	2.280107
c2_t1016	Hump1	t10	2.252638
c2_t1026	Arch1	t30	2.846221
c2_t1028	Arch1	t10	3.849471
c2_t105	UT2	t60	1.755212
c2_t1060	Stag-3	t10	4.133971
c2_t1061	Arch2	t10	4.267319
c2_t1062	Arch1	t10	4.117134
c2_t1065	Stag-3	t10	3.550501
c2_t1076	Arch1	t10	1.492771
c2_t1096	Stag-10	t30	3.417283
c2_t1114	Stag-60	t60	1.802546
c2_t1117	Arch1	t30	1.484807
c2_t112	Arch2	t10	3.791518
c2_t113	Arch1	t30	3.067328
c2_t1135	Stag-30	t60	3.735513
c2_t1136	Stag-30	t60	2.196503
c2_t1137	Arch1	t10	3.953143
c2_t1144	Arch1	t10	3.235027
c2_t1145	Arch1	t30	3.490875
c2_t118	Arch1	t10	3.671183
c2_t1192	ambiguous	t30	1.394147
c2_t1202	Stag-30	t30	1.643576
c2_t1205	Arch1	t10	1.429067
c2_t1206	Pulse-3	t3	4.511812
c2_t1219	UT1	t60	1.735259
c2_t1229	UT5	t60	2.248714

Gene	Profile	MaxTP	Log_2 (MaxFC)
c2_t1231	UT5	t60	2.223943
c2_t1242	Arch1	t30	3.303858
c2_t1251	Stag-60	t60	2.243247
c2_t1256	ambiguous	t60	3.243251
c2_t1259	ambiguous	t3	1.907531
c2_t126	Arch1	t30	1.721958
c2_t1260	Arch1	t10	2.585252
c2_t1276	Arch2	t10	3.907987
c2_t1281	Arch1	t10	2.669722
c2_t13	Arch1	t10	3.327345
c2_t131	Arch1	t30	1.443907
c2_t1321	Arch2	t10	4.578696
c2_t1330	ambiguous	t60	2.965099
c2_t1393	Arch1	t10	3.603246
c2_t1395	Stag-30	t30	2.110735
c2_t1396	Stag-10	t30	2.290163
c2_t1400	Arch2	t10	2.060111
c2_t1425	Stag-3	t10	2.213109
c2_t1428	UT1	t60	4.049126
c2_t1436	Pulse-10	t10	2.899514
c2_t1437	Pulse-10	t10	1.67507
c2_t1445	Arch1	t10	3.108286
c2_t1447	Arch1	t30	3.571363
c2_t1454	Stag-30	t60	1.542905
c2_t1461	Arch1	t10	1.440108
c2_t1476	Arch2	t10	3.634827
c2_t1477	Arch1	t10	3.124574
c2_t150	Arch1	t30	3.642747
c2_t1505	Arch1	t30	1.511843
c2_t151	outlier	t60	3.828682
c2_t152	Arch1	t30	3.79423
c2_t158	Stag-60	t60	1.819642
c2_t159	Stag-60	t60	2.562544
c2_t160	Stag-60	t60	1.671405
c2_t233	Arch1	t10	2.753585
c2_t239	Stag-60	t60	1.546627
c2_t273	Stag-30	t60	1.396292
c2_t295	Stag-60	t60	1.421095
c2_t300	Stag-30	t60	2.236552
c2_t314	Arch1	t30	3.482464

Gene	Profile	MaxTP	Log_2 (MaxFC)
c2_t334	Arch1	t30	2.801569
c2_t337	Arch2	t10	4.327782
c2_t34	Arch1	t30	1.545155
c2_t366	Stag-30	t60	1.511223
c2_t371	Arch1	t10	3.838579
c2_t374	Arch1	t10	3.139968
c2_t381	UT2	t60	1.518379
c2_t383	Arch1	t30	3.932989
c2_t389	Arch1	t10	3.737489
c2_t391	Arch1	t10	3.923885
c2_t392	Arch1	t30	3.64569
c2_t419	outlier	t60	5.94373
c2_t420	outlier	t60	5.735809
c2_t426	Pulse-30	t30	2.559918
c2_t432	Arch2	t10	3.318011
c2_t438	UT2	t60	3.662108
c2_t447	Stag-10	t30	2.414841
c2_t504	Hump1	t10	2.165052
c2_t510	Stag-30	t30	1.869877
c2_t519	Arch1	t30	1.706146
c2_t538	Stag-30	t60	1.425944
c2_t549	Hump2	t10	2.318296
c2_t57	ambiguous	t10	1.945445
c2_t60	Pulse-10	t10	1.422927
c2_t609	Arch2	t10	2.037417
c2_t614	Stag-10	t30	2.230984
c2_t619	Arch1	t10	3.778
c2_t636	Stag-60	t60	1.667251
c2_t648	Arch1	t10	2.766498
c2_t659	Stag-30	t60	1.553205
c2_t662	Arch1	t30	3.180926
c2_t671	UT1	t30	3.402409
c2_t684	Stag-30	t30	2.130327
c2_t69	Arch1	t10	3.04211
c2_t7	Stag-3	t10	4.283304
c2_t711	Stag-30	t60	1.633491
c2_t715	Stag-60	t60	1.548906
c2_t731	ambiguous	t10	1.931927
c2_t74	Arch1	t30	1.669155
c2_t76	Arch1	t30	1.856878

Gene	Profile	MaxTP	Log_2 (MaxFC)
c2_t763	Arch1	t10	2.85052
c2_t789	Arch1	t30	2.59046
c2_t792	Arch2	t10	2.495026
c2_t805	Arch1	t10	3.305041
c2_t817	Arch1	t10	3.582457
c2_t849	Stag-60	t60	2.155675
c2_t857	Stag-60	t60	2.068334
c2_t858	UT1	t60	2.488909
c2_t867	UT3	t10	1.762007
c2_t874	Arch1	t10	3.456769
c2_t880	Stag-10	t30	2.549098
c2_t893	Stag-60	t60	2.05359
c2_t904	Stag-30	t60	2.284969
c2_t914	Arch1	t10	2.773235
c2_t918	Pulse-3	t3	2.292197
c2_t930	ambiguous	t10	1.668895
c2_t932	Stag-60	t60	2.464166
c2_t954	Stag-60	t60	1.464162
c2_t963	ambiguous	t10	2.420222
c3_t1019	Arch2	t10	4.101495
c3_t1029	Pulse-3	t3	1.438045
c3_t1032	Arch2	t10	4.855188
c3_t1044	Arch1	t10	3.099861
c3_t1049	Arch1	t10	2.142142
c3_t1054	Arch2	t10	3.656655
c3_t1055	Hump1	t10	1.647832
c3_t1062	Arch1	t10	3.184533
c3_t1112	Arch1	t30	3.116943
c3_t1114	Stag-30	t60	3.316958
c3_t1115	Stag-60	t60	1.391651
c3_t1119	Stag-30	t60	1.897534
c3_t1120	Stag-60	t60	2.257327
c3_t1126	Stag-30	t60	2.546631
c3_t1160	Pulse-3	t3	3.452326
c3_t118	Arch1	t10	3.110725
c3_t1182	Stag-60	t60	1.961661
c3_t1196	Pulse-3	t3	2.091057
c3_t1209	Stag-30	t60	1.928494
c3_t122	Hump2	t30	1.428736
c3_t1222	Stag-3	t10	3.724202

Gene	Profile	MaxTP	Log_2 (MaxFC)
c3_t1232	UT3	t10	1.785739
c3_t1249	UT5	t60	2.103204
c3_t1260	Arch1	t30	3.800102
c3_t1269	Arch2	t10	3.649528
c3_t150	Hump1	t10	1.872748
c3_t157	Stag-10	t60	2.086916
c3_t170	Arch2	t10	3.644017
c3_t179	Arch1	t30	2.221678
c3_t192	Pulse-3	t3	1.422103
c3_t201	Arch1	t10	5.046432
c3_t224	Pulse-10	t10	2.005303
c3_t243	Arch1	t10	3.76542
c3_t26	Stag-10	t30	1.841413
c3_t264	Stag-60	t60	1.961661
c3_t297	Stag-3	t10	3.62431
c3_t299	Arch2	t10	2.642657
c3_t306	Arch1	t30	2.533574
c3_t32	Pulse-3	t3	1.747493
c3_t361	Stag-60	t60	1.686063
c3_t362	Stag-3	t10	3.257692
c3_t373	Stag-60	t60	2.427089
c3_t377	UT3	t10	2.849467
c3_t391	Arch1	t10	4.364888
c3_t399	Pulse-3	t3	2.129923
c3_t403	Arch1	t30	2.131091
c3_t422	Pulse-10	t10	1.984398
c3_t424	Hump1	t10	2.800141
c3_t425	Arch2	t10	2.998483
c3_t430	Stag-3	t10	2.82259
c3_t440	Stag-30	t60	2.182307
c3_t458	Arch1	t30	1.944537
c3_t461	Stag-30	t60	1.961661
c3_t462	Pulse-30	t30	1.496955
c3_t466	Arch1	t10	2.221433
c3_t468	Arch2	t10	4.095277
c3_t477	Stag-60	t60	2.925136
c3_t478	Stag-30	t60	2.564506
c3_t482	Stag-60	t60	1.528838
c3_t496	UT2	t60	1.618776
c3_t502	Arch1	t30	1.779478

Gene	Profile	MaxTP	Log_2 (MaxFC)
c3_t516	Pulse-30	t30	2.559918
c3_t518	Arch1	t10	3.202581
c3_t522	Arch1	t10	3.376397
c3_t530	Stag-3	t10	3.856425
c3_t552	Stag-60	t60	3.929743
c3_t557	ambiguous	t10	2.857575
c3_t558	Arch2	t10	4.888529
c3_t567	Arch1	t10	2.617987
c3_t572	Stag-10	t60	6.734255
c3_t577	Stag-60	t60	1.98535
c3_t594	Arch1	t10	3.619982
c3_t597	Arch1	t30	1.454742
c3_t598	Stag-60	t60	1.961661
c3_t603	Arch1	t30	2.810687
c3_t606	Arch2	t10	3.622146
c3_t615	UT2	t60	2.853059
c3_t637	Arch1	t10	1.758256
c3_t638	Arch1	t10	2.488937
c3_t646	Stag-30	t60	4.785304
c3_t647	Stag-10	t60	6.614771
c3_t648	Stag-30	t60	4.918768
c3_t655	Stag-3	t10	3.761366
c3_t656	Arch1	t30	2.895417
c3_t661	Stag-30	t60	3.31804
c3_t662	UT1	t60	2.533199
c3_t665	Stag-10	t10	1.677436
c3_t725	Stag-60	t60	1.401287
c3_t728	Arch1	t30	2.619934
c3_t729	Arch1	t10	1.50287
c3_t730	Stag-3	t10	3.058773
c3_t740	Arch1	t30	3.557744
c3_t741	Stag-3	t10	1.40586
c3_t743	Arch1	t10	1.659907
c3_t745	UT2	t60	1.546627
c3_t746	Arch1	t10	2.161893
c3_t752	Arch1	t30	2.912267
c3_t769	Arch1	t30	3.767093
c3_t78	Pulse-10	t10	2.346962
c3_t780	UT1	t10	1.679715
c3_t783	Hump2	t10	1.808779

Gene	Profile	MaxTP	Log_2 (MaxFC)
c3_t819	Stag-60	t60	1.447095
c3_t838	Arch1	t30	3.082102
c3_t855	Stag-10	t30	2.754047
c3_t857	Arch1	t10	2.668899
c3_t858	Arch1	t30	2.548189
c3_t863	Stag-30	t30	3.594042
c3_t88	Arch1	t30	2.29201
c3_t881	Stag-60	t60	2.065997
c3_t886	Arch1	t10	1.980647
c3_t887	Pulse-3	t3	1.400055
c3_t89	Arch1	t30	2.111831
c3_t907	Stag-10	t30	3.678973
c3_t908	Stag-10	t30	1.844774
c3_t916	Pulse-30	t30	1.379215
c3_t938	Stag-3	t10	1.569393
c3_t959	Stag-30	t60	1.572653
c3_t961	UT1	t60	1.791741
c3_t983	Hump2	t10	2.161734
c3_t984	Arch1	t10	1.873109
c4_t116	Stag-30	t60	2.428402
c4_t117	Stag-30	t60	2.740443
c4_t122	Arch1	t30	2.970394
c4_t132	Pulse-30	t30	2.442512
c4_t133	Pulse-3	t3	1.422114
c4_t175	Arch1	t10	2.848688
c4_t180	Hump2	t10	3.792903
c4_t182	Pulse-30	t30	1.608172
c4_t183	Arch1	t10	4.806396
c4_t186	UT3	t10	1.624503
c4_t187	Pulse-3	t3	2.218273
c4_t211	Stag-3	t10	3.285565
c4_t212	Stag-60	t60	2.027001
c4_t23	Arch2	t10	3.522225
c4_t234	Stag-60	t60	2.158055
c4_t239	Stag-60	t60	1.403278
c4_t240	Stag-60	t60	1.546627
c4_t244	Stag-60	t60	1.95426
c4_t272	Pulse-10	t10	1.464768
c4_t292	Arch1	t10	4.199858
c4_t296	UT1	t60	1.926589

Gene	Profile	MaxTP	Log_2 (MaxFC)
c4_t314	ambiguous	t3	1.396915
c4_t319	Arch1	t10	1.486697
c4_t320	Hump2	t10	1.738188
c4_t335	Hump2	t30	1.464537
c4_t340	Stag-60	t60	1.692844
c4_t347	Stag-30	t30	1.754043
c4_t359	Stag-60	t60	1.769018
c4_t362	Arch1	t30	3.664676
c4_t363	Pulse-10	t10	1.630173
c4_t384	Stag-10	t60	2.909541
c4_t398	Stag-60	t60	1.768239
c4_t406	Stag-60	t60	2.283599
c4_t65	Pulse-3	t3	1.907531
c4_t77	Stag-30	t30	2.113058
c4_t83	Arch1	t10	2.570378
c4_t95	ambiguous	t10	1.755313
c4_t96	Hump1	t10	1.832468
c5_t104	Arch1	t10	2.122479
c5_t106	Stag-60	t60	1.698629
c5_t141	Arch2	t10	3.079202
c5_t147	Stag-30	t30	1.619901
c5_t151	Arch1	t10	2.232628
c5_t170	Pulse-3	t3	1.928595
c5_t178	Stag-60	t60	2.033998
c5_t198	Stag-60	t60	1.546627
c5_t205	Stag-3	t10	2.351492
c5_t206	Arch1	t10	2.713753
c5_t210	Pulse-30	t30	2.677757
c5_t242	Hump2	t30	1.845135
c5_t276	UT3	t10	2.671135
c5_t28	Arch1	t10	3.154323
c5_t280	Hump1	t10	2.188597
c5_t290	UT2	t60	1.640157
c5_t292	Arch1	t30	3.222303
c5_t293	Arch1	t30	3.540835
c5_t294	UT3	t10	3.330692
c5_t296	Stag-60	t60	1.433985
c5_t310	UT3	t10	2.08393
c5_t322	Hump2	t30	2.789206
c5_t337	Stag-30	t60	2.209588

Gene	Profile	MaxTP	Log_2 (MaxFC)
c5_t347	Arch1	t30	1.692974
c5_t354	Stag-30	t30	1.747464
c5_t358	Stag-10	t30	3.12218
c5_t360	Stag-10	t30	3.48043
c5_t366	Stag-30	t30	1.84802
c5_t38	Arch1	t10	1.812631
c5_t397	ambiguous	t10	1.906464
c5_t401	Stag-30	t30	2.95223
c5_t407	Stag-30	t60	1.56694
c5_t409	Pulse-30	t30	1.53031
c5_t410	Stag-30	t30	3.647869
c5_t55	UT1	t60	3.54662
c5_t91	Arch2	t10	3.578922
c5_t92	UT1	t60	2.50599
c6_t1010	Arch2	t10	3.796452
c6_t1015	Arch1	t30	2.127499
c6_t1022	Arch1	t10	1.650746
c6_t1061	Arch2	t10	3.367957
c6_t1085	Pulse-3	t3	2.522552
c6_t1092	Stag-30	t30	1.405748
c6_t1111	Arch1	t10	4.029851
c6_t1123	UT3	t10	2.285561
c6_t1149	Stag-3	t10	3.959592
c6_t1161	UT4	t30	1.450486
c6_t1168	Stag-60	t60	2.394629
c6_t1180	Arch2	t10	3.367726
c6_t1190	Stag-10	t60	3.045976
c6_t1203	UT2	t60	1.518061
c6_t1208	Stag-60	t60	1.385599
c6_t1234	Arch1	t10	1.549858
c6_t1240	Stag-10	t30	2.921212
c6_t1244	Arch1	t10	1.955313
c6_t1245	Stag-3	t3	3.977914
c6_t1255	Stag-30	t60	2.836136
c6_t139	Stag-30	t60	1.905714
c6_t154	UT3	t10	3.465613
c6_t161	Arch1	t10	3.995977
c6_t166	UT4	t30	1.797699
c6_t20	Hump1	t10	1.810986
c6_t235	Stag-10	t30	1.800628

Gene	Profile	MaxTP	Log_2 (MaxFC)
c6_t236	Arch1	t30	1.782132
c6_t265	Arch1	t10	3.20538
c6_t267	Stag-3	t10	2.729045
c6_t276	Stag-30	t60	2.047487
c6_t277	Stag-60	t60	1.740467
c6_t282	Arch1	t30	1.921742
c6_t29	Stag-30	t60	1.453155
c6_t306	Pulse-10	t10	2.525423
c6_t32	Stag-10	t30	2.080265
c6_t323	Arch1	t10	1.686828
c6_t324	Arch1	t10	1.830722
c6_t332	Hump1	t10	2.001364
c6_t375	Arch2	t10	2.644965
c6_t38	Stag-30	t60	2.868554
c6_t380	Stag-30	t30	1.388326
c6_t389	Pulse-30	t30	3.161161
c6_t403	Stag-3	t10	4.181536
c6_t405	Arch1	t30	2.55898
c6_t409	Hump1	t10	1.774385
c6_t421	Stag-30	t60	1.722693
c6_t426	Stag-30	t60	1.57108
c6_t428	Stag-60	t60	3.387924
c6_t429	Stag-30	t60	1.500273
c6_t44	Stag-60	t60	2.037576
c6_t444	Arch1	t10	1.955833
c6_t445	Arch1	t10	4.604289
c6_t454	Pulse-10	t10	1.492165
c6_t456	UT1	t10	1.407594
c6_t477	Stag-3	t10	3.775807
c6_t49	Arch1	t10	3.015521
c6_t502	Stag-30	t60	1.676931
c6_t535	Arch1	t10	2.590604
c6_t556	Stag-60	t60	1.953755
c6_t557	Stag-60	t60	2.453202
c6_t559	Stag-60	t60	1.616049
c6_t611	Arch1	t30	1.902626
c6_t612	Arch1	t10	1.837157
c6_t618	Stag-3	t10	1.920501
c6_t627	Arch1	t30	2.405636
c6_t654	Arch1	t30	1.912956

Gene	Profile	MaxTP	Log_2 (MaxFC)
c6_t680	ambiguous	t10	1.832396
c6_t691	Hump1	t10	1.571008
c6_t708	Arch1	t30	3.40603
c6_t734	Arch1	t30	2.71319
c6_t755	UT1	t60	2.769022
c6_t765	Stag-60	t60	2.402232
c6_t766	UT3	t10	2.176233
c6_t77	UT4	t30	1.879572
c6_t773	Stag-60	t60	2.108499
c6_t779	UT2	t60	2.176103
c6_t827	Stag-60	t60	1.511858
c6_t875	UT2	t30	1.641412
c6_t876	Stag-3	t10	3.410618
c6_t879	Arch1	t10	1.496926
c6_t900	Arch1	t30	2.68047
c6_t902	Arch1	t30	2.828
c6_t903	Arch1	t30	3.352506
c6_t921	outlier	t60	4.488282
c6_t922	Stag-10	t60	5.313013
c6_t926	Stag-60	t60	1.836132
c6_t953	Stag-30	t30	1.61654
c6_t956	Stag-60	t60	1.546627
c6_t985	Arch1	t10	3.71142
c7_t102	UT3	t10	2.7638
c7_t104	ambiguous	t60	1.454525
c7_t117	Arch1	t10	3.049785
c7_t12	Hump1	t10	2.079962
c7_t127	Arch2	t10	3.386496
c7_t128	Pulse-30	t30	1.47403
c7_t155	Pulse-10	t10	1.715595
c7_t158	Hump1	t10	2.820541
c7_t164	Stag-10	t30	2.393503
c7_t169	Stag-30	t60	3.421092
c7_t170	Stag-30	t60	3.421092
c7_t172	Stag-60	t60	2.37785
c7_t21	Arch1	t10	2.126922
c7_t212	Stag-30	t30	1.427879
c7_t213	Arch1	t10	3.33968
c7_t225	ambiguous	t10	1.895485
c7_t230	Arch1	t30	2.62911

Gene	Profile	MaxTP	Log_2 (MaxFC)
c7_t235	ambiguous	t60	1.546627
c7_t243	Stag-30	t30	2.411335
c7_t247	Arch1	t10	2.478997
c7_t273	Arch1	t10	1.659503
c7_t275	Stag-10	t30	2.231359
c7_t291	Stag-60	t60	2.103017
c7_t307	Arch1	t30	2.548636
c7_t325	Stag-60	t60	1.377458
c7_t34	Hump2	t30	2.409532
c7_t369	Arch1	t30	3.704192
c7_t373	Arch2	t10	2.884149
c7_t374	Stag-30	t60	1.438101
c7_t378	Pulse-10	t10	1.700606
c7_t385	Arch1	t10	2.947109
c7_t397	Stag-30	t30	2.923968
c7_t400	Hump1	t10	2.073946
c7_t414	Arch1	t10	2.388786
c7_t434	Stag-30	t30	2.464542
c7_t456	Stag-30	t60	1.404606
c7_t486	Stag-30	t60	1.891272
c7_t488	Arch1	t30	3.106238
c7_t489	Arch2	t10	2.907059
c7_t504	Arch1	t10	2.347726
c7_t532	Pulse-30	t30	3.855747
c7_t57	Stag-30	t60	1.961661
c7_t587	Arch1	t10	5.134855
c7_t620	Arch1	t30	3.236903
c7_t632	Arch1	t10	3.245227
c7_t651	Pulse-3	t3	1.745748
c7_t655	Arch1	t30	4.414878
c7_t656	Stag-10	t30	4.171495
c7_t663	Stag-30	t60	1.935981
c7_t665	Stag-30	t60	2.159599
c7_t672	Stag-30	t60	4.015612
c7_t674	Stag-60	t60	1.698629
c7_t675	Stag-60	t60	2.178498
c7_t688	Arch2	t10	3.534184
c7_t695	UT1	t60	2.510246
c7_t706	Stag-30	t60	1.75035
c7_t725	Arch1	t30	3.238591

Gene	Profile	MaxTP	Log_2 (MaxFC)
c7_t739	Stag-60	t60	1.688646
c7_t74	UT1	t60	3.662108
c7_t752	Stag-3	t3	1.4253
c7_t771	Stag-10	t30	2.879576
c7_t776	Stag-60	t60	1.631515
c7_t780	Pulse-3	t3	1.492497
c7_t798	UT3	t3	1.741549
c7_t807	Hump1	t10	2.324946
c7_t808	Hump1	t10	1.597511
c7_t817	Arch1	t30	3.366197
c7_t82	ambiguous	t60	1.6621
c7_t821	Stag-60	t60	2.020552
c7_t823	Pulse-3	t3	2.481522
c7_t865	Arch1	t30	2.623728
c8_t109	Hump1	t10	3.123767
c8_t12	Arch2	t10	3.00939
c8_t131	Arch2	t10	2.876792
c8_t136	UT3	t10	2.009934
c8_t139	Pulse-10	t10	2.849467
c8_t16	Arch2	t10	4.566014
c8_t168	Stag-60	t60	1.51721
c8_t170	Stag-30	t60	1.836132
c8_t18	Arch1	t10	4.323093
c8_t188	Pulse-10	t10	3.975115
c8_t190	Stag-60	t60	1.86855
c8_t191	Hump1	t3	1.621229
c8_t193	Arch2	t10	1.636478
c8_t196	Stag-3	t10	3.029486
c8_t208	Hump2	t10	1.567272
c8_t212	Stag-30	t60	1.879197
c8_t214	Stag-60	t60	1.783085
c8_t224	Stag-30	t60	1.553725
c8_t241	Stag-3	t10	3.092229
c8_t245	Stag-30	t30	2.559918
c8_t246	ambiguous	t60	2.827062
c8_t262	Stag-30	t60	1.645423
c8_t263	Stag-30	t60	2.088114
c8_t274	Stag-60	t60	1.482499
c8_t275	outlier	t3	2.456102
c8_t279	Arch2	t10	2.345663

Gene	Profile	MaxTP	Log_2 (MaxFC)
c8_t281	Pulse-10	t10	3.210328
c8_t282	Pulse-3	t3	2.053301
c8_t305	Arch1	t10	3.191573
c8_t310	Stag-60	t60	1.385423
c8_t323	UT3	t3	1.656474
c8_t335	Hump1	t10	2.120459
c8_t36	Arch2	t10	3.682248
c8_t360	Arch1	t30	3.400086
c8_t362	Hump2	t30	3.096803
c8_t363	Arch1	t30	3.146013
c8_t372	Arch1	t10	3.459814
c8_t376	Pulse-30	t30	1.527497
c8_t400	Arch1	t30	2.614856
c8_t411	Stag-30	t30	2.542534
c8_t419	Arch1	t10	4.19465
c8_t420	Arch1	t10	4.307714
c8_t421	Arch1	t10	3.831336
c8_t483	Arch2	t10	2.792163
c8_t491	UT5	t60	2.284969
c8_t499	Stag-60	t60	1.77489
c8_t500	ambiguous	t60	3.313135
c8_t507	Stag-30	t30	2.031055
c8_t510	Stag-10	t60	2.626816
c8_t527	Pulse-30	t30	1.436467
c8_t528	Pulse-30	t30	1.458666
c8_t54	Pulse-30	t30	2.822965
c8_t567	Stag-30	t60	1.621142
c8_t569	Stag-60	t60	1.815819
c8_t579	Arch1	t30	1.375269
c8_t64	Arch2	t10	3.523105
c8_t74	Stag-60	t60	2.610759
c8_t77	Stag-30	t60	2.167765
c8_t95	Arch1	t30	3.116871
c8_t97	Arch2	t10	2.712252
c9_t103	UT4	t30	1.821273
c9_t137	Stag-30	t60	2.11124
c9_t138	Stag-60	t60	1.454352
c9_t159	ambiguous	t3	2.492487
c9_t169	Stag-30	t30	1.515162
c9_t189	ambiguous	t3	2.011016

Gene	Profile	MaxTP	Log_2 (MaxFC)
c9_t191	Arch2	t10	3.668889
c9_t197	Arch1	t30	2.850664
c9_t204	Arch1	t10	1.547723
c9_t206	Arch1	t10	3.50924
c9_t216	Pulse-3	t3	1.434105
c9_t219	Arch1	t10	2.249623
c9_t222	Arch1	t30	2.05196
c9_t234	Arch1	t30	5.0943
c9_t235	Stag-30	t60	1.783301
c9_t275	Hump1	t10	3.215738
c9_t279	Hump1	t10	1.93223
c9_t284	Arch1	t10	1.836854
c9_t285	Stag-30	t60	2.179061
c9_t296	UT4	t30	1.879572
c9_t335	Stag-60	t60	2.861138
c9_t345	Arch1	t10	3.636255
c9_t358	UT1	t10	2.627162
c9_t360	Arch1	t10	1.540495
c9_t444	Arch1	t10	4.032087
c9_t474	Stag-30	t60	1.582694
c9_t484	Stag-3	t10	3.688798
c9_t501	Stag-60	t60	2.14057
c9_t509	Stag-10	t30	3.714767
c9_t51	Stag-30	t30	2.008607
c9_t518	Stag-30	t60	1.483624
c9_t54	Pulse-3	t3	2.229454
c9_t541	Stag-30	t60	1.532402
c9_t543	Pulse-10	t10	1.770547
c9_t561	Pulse-30	t30	1.464537
c9_t582	Arch1	t30	2.860143
c9_t59	Arch1	t10	2.655093
c9_t600	Hump2	t10	1.677436
c9_t605	Hump2	t10	1.677436
c9_t608	Stag-60	t60	2.243247
c9_t614	Stag-3	t10	3.401759
c9_t63	Arch2	t10	2.095904
c9_t632	Arch1	t30	3.209722
c9_t636	Arch1	t30	2.25121
c9_t640	Pulse-10	t10	1.698182
c9_t650	Stag-30	t30	1.616511

Gene	Profile	MaxTP	Log_2 (MaxFC)
c9_t73	Stag-10	t60	2.136747
c9_t86	Stag-60	t60	1.698629
c9_t99	Stag-10	t60	3.036945
s18_t110	Pulse-3	t3	1.875936
s18_t126	Stag-30	t60	2.260443
s18_t151	Stag-60	t60	1.431149
s18_t153	Arch1	t10	3.556835
s18_t154	Hump2	t10	2.972327
s18_t156	Arch2	t10	2.344033
s18_t174	Arch1	t30	2.693497
s18_t177	Stag-60	t60	2.447085
s18_t21	Arch1	t30	2.806619
s18_t27	Stag-60	t60	1.898183
s18_t41	Stag-60	t60	1.3767
s18_t44	Pulse-30	t30	1.776477
s18_t45	Hump2	t30	1.762223
s18_t5	Arch1	t10	2.692156
s18_t60	Stag-60	t60	2.376696
s18_t80	Arch1	t30	1.726675
s18_t84	UT1	t10	2.083583
s18_t90	Pulse-10	t10	1.578842
s18_t95	Arch1	t30	3.094379

.

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:** 403-410.

ANANTHARAMAN, V., and L. ARAVIND, 2003 Application of comparative genomics in the identification and analysis of novel families of membrane-associated receptors in bacteria. BMC Genomics **4:** 34.

ARAVIND, L., H. WATANABE, D. J. LIPMAN and E. V. KOONIN, 2000 Lineage-specific loss and divergence of functionally linked genes in eukaryotes. Proceedings of the National Academy of Sciences of the United States of America **97:** 11319-11324.

AVIDOR-REISS, T., A. M. MAER, E. KOUNDAKJIAN, A. POLYANOVSKY, T. KEIL *et al.*, 2004 Decoding cilia function: defining specialized genes required for compartmentalized cilia biogenesis. Cell **117:** 527-539.

BALL, S., and P. DESCHAMPS, 2009 Starch Metabolism. The Chlamydomonas Sourcebook **2**.

BARKER, D., A. MEADE and M. PAGEL, 2007 Constrained models of evolution lead to improved prediction of functional linkage from correlated gain and loss of genes. Bioinformatics **23:** 14-20.

BARONI, T., C. BELLUCCI, C. LILLI, F. PEZZETTI, F. CARINCI *et al.*, 2010 Human cleft lip and palate fibroblasts and normal nicotine-treated fibroblasts show altered in vitro expressions of genes related to molecular signaling pathways and extracellular matrix metabolism. J Cell Physiol **222:** 748-756.

BERLIN, I., H. SCHWARTZ and P. D. NASH, 2010 Regulation of epidermal growth factor receptor ubiquitination and trafficking by the USP8.STAM complex. J Biol Chem **285:** 34909-34921.

BIELAS, S. L., J. L. SILHAVY, F. BRANCATI, M. V. KISSELEVA, L. AL-GAZALI *et al.*, 2009 Mutations in INPP5E, encoding inositol polyphosphate-5-phosphatase E, link phosphatidyl inositol signaling to the ciliopathies. Nat Genet **41:** 1032-1036.

BOLES, M. K., B. M. WILKINSON, L. G. WILMING, B. LIU, F. J. PROBST *et al.*, 2009 Discovery of candidate disease genes in ENU-induced mouse mutants by large-scale sequencing, including a splice-site mutation in nucleoredoxin. PLoS Genet **5:** e1000759.

BRADY, S. M., D. A. ORLANDO, J. Y. LEE, J. Y. WANG, J. KOCH *et al.*, 2007 A high-resolution root spatiotemporal map reveals dominant expression patterns. Science **318:** 801-806.

BRENT, M., 2007 How does eukaryotic gene prediction work? Nat Biotechnol **25:** 883 - 885.

CASPARY, T., C. E. LARKINS and K. V. ANDERSON, 2007 The graded response to Sonic Hedgehog depends on cilia architecture. Dev Cell **12:** 767-778.

COKUS, S., S. MIZUTANI and M. PELLEGRINI, 2007 An improved method for identifying functionally linked proteins using phylogenetic profiles. BMC Bioinformatics **8 Suppl 4:** S7.

COLE, D. G., 2003 The intraflagellar transport machinery of Chlamydomonas reinhardtii. Traffic **4:** 435-442.

CURRIER, T. A., M. A. ETCHEGARAY, J. L. HAIGHT, A. M. GALABURDA and G. D. ROSEN, 2011 The effects of embryonic knockdown of the candidate dyslexia susceptibility gene homologue Dyx1c1 on the distribution of GABAergic neurons in the cerebral cortex. Neuroscience **172:** 535-546.

DESPER, R., and O. GASCUEL, 2002 Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. J Comput Biol **9:** 687-705.

DOLES, J., and M. T. HEMANN, 2010 Nek4 status differentially alters sensitivity to distinct microtubule poisons. Cancer Res **70:** 1033-1041.

DUQUESNOY, P., E. ESCUDIER, L. VINCENSINI, J. FRESHOUR, A. M. BRIDOUX *et al.*, 2009 Loss-of-function mutations in the human ortholog of Chlamydomonas reinhardtii ODA7 disrupt dynein arm assembly and cause primary ciliary dyskinesia. Am J Hum Genet **85:** 890-896.

DUTCHER, S., 2009 Basal bodies and associated structures. The Chlamydomonas Sourcebook **3**.

DYMEK, E., P. LEFEBVRE and E. SMITH, 2004 PF15p is the Chlamydomonas homologue of the Katanin p80 subunit and is required for assembly of flagellar central microtubules. Eukaryot Cell **3:** 870 - 879.

EFIMENKO, E., O. E. BLACQUE, G. OU, C. J. HAYCRAFT, B. K. YODER *et al.*, 2006 Caenorhabditis elegans DYF-2, an orthologue of human WDR19, is a component of the intraflagellar transport machinery in sensory cilia. Mol Biol Cell **17:** 4801-4811.

EFIMENKO, E., K. BUBB, H. Y. MAK, T. HOLZMAN, M. R. LEROUX *et al.*, 2005 Analysis of xbx genes in C. elegans. Development **132:** 1923-1934.

FELDMAN, J. L., and W. F. MARSHALL, 2009 ASQ2 encodes a TBCC-like protein required for mother-daughter centriole linkage and mitotic spindle orientation. Curr Biol **19:** 1238-1243.

FENG, D. F., and R. F. DOOLITTLE, 1997 Converting amino acid alignment scores into measures of evolutionary time: a simulation study of various relationships. J Mol Evol **44:** 361-370.

FLIEGAUF, M., T. BENZING and H. OMRAN, 2007 When cilia go bad: cilia defects and ciliopathies. Nat Rev Mol Cell Biol **8:** 880-893.

FOLLIT, J. A., J. T. SAN AGUSTIN, F. XU, J. A. JONASSEN, R. SAMTANI *et al.*, 2008 The Golgin GMAP210/TRIP11 anchors IFT20 to the Golgi complex. PLoS Genet **4:** e1000315.

FUJIWARA, M., T. TERAMOTO, T. ISHIHARA, Y. OHSHIMA and S. L. MCINTIRE, 2010 A novel zf-MYND protein, CHB-3, mediates guanylyl cyclase localization to sensory cilia and controls body size of Caenorhabditis elegans. PLoS Genet **6:** e1001211.

GABISON, E. E., T. HOANG-XUAN, A. MAUVIEL and S. MENASHI, 2005 EMMPRIN/CD147, an MMP modulator in cancer, development and tissue repair. Biochimie **87:** 361-368.

GILISSEN, C., H. H. ARTS, A. HOISCHEN, L. SPRUIJT, D. A. MANS *et al.*, 2010 Exome sequencing identifies WDR35 variants involved in Sensenbrenner syndrome. Am J Hum Genet **87:** 418-423.

GONCALVES, J., S. NOLASCO, R. NASCIMENTO, M. LOPEZ FANARRAGA, J. C. ZABALA *et al.*, 2010 TBCCD1, a new centrosomal protein, is required for centrosome and Golgi apparatus positioning. EMBO Rep **11:** 194-200.

GONZALEZ, A., M. E. SAEZ, M. J. ARAGON, J. J. GALAN, P. VETTORI *et al.*, 2006 Specific haplotypes of the CALPAIN-5 gene are associated with polycystic ovary syndrome. Hum Reprod **21:** 943-951.

GRASER, S., Y. D. STIERHOF, S. B. LAVOIE, O. S. GASSNER, S. LAMLA *et al.*, 2007 Cep164, a novel centriole appendage protein required for primary cilium formation. J Cell Biol **179:** 321-330.

HAAS, B., A. DELCHER, S. MOUNT, J. WORTMAN, R. SMITH *et al.*, 2003 Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res **31:** 5654 - 5666.

HALLEM, E. A., W. C. SPENCER, R. D. MCWHIRTER, G. ZELLER, S. R. HENZ *et al.*, 2011 Receptor-type guanylate cyclase is required for carbon dioxide sensation by Caenorhabditis elegans. Proc Natl Acad Sci U S A **108:** 254-259.

HARRIS, E., 2009 The Chlamydomonas Sourcebook. **1**.

HEGEMANN, P., and P. BERTHOD, 2009 Sensory photoreceptors and light control of flagellar activity. The Chlamydomonas Sourcebook **3**.

HUANG, K., D. R. DIENER and J. L. ROSENBAUM, 2009 The ubiquitin conjugation system is involved in the disassembly of cilia and flagella. J Cell Biol **186:** 601-613.

HUME, A. N., J. BUTTGEREIT, A. M. AL-AWADHI, S. S. AL-SUWAIDI, A. JOHN *et al.*, 2009 Defective cellular trafficking of missense NPR-B mutants is the major mechanism underlying acromesomelic dysplasia-type Maroteaux. Hum Mol Genet **18:** 267-277.

INOUE, M., M. SAEKI, H. EGUSA, H. NIWA and Y. KAMISAKI, 2010 PIH1D1, a subunit of R2TP complex, inhibits doxorubicin-induced apoptosis. Biochem Biophys Res Commun **403:** 340-344.

JANSE, C. J., H. KROEZE, A. VAN WIGCHEREN, S. MEDEDOVIC, J. FONAGER *et al.*, 2011 A genotype and phenotype database of genetically modified malaria-parasites. Trends Parasitol **27:** 31-39.

JAROSZEWSKI, L., Z. LI, S. S. KRISHNA, C. BAKOLITSA, J. WOOLEY *et al.*, 2009 Exploration of uncharted regions of the protein universe. PLoS Biol **7:** e1000205.

JENSEN, V. L., N. J. BIALAS, S. L. BISHOP-HURLEY, L. L. MOLDAY, K. KIDA *et al.*, 2010 Localization of a guanylyl cyclase to chemosensory cilia requires the novel ciliary MYND domain protein DAF-25. PLoS Genet **6:** e1001199.

JIANG, Z., 2008 Protein function predictions based on the phylogenetic profile method. Crit Rev Biotechnol **28:** 233-238.

JOTHI, R., T. M. PRZYTYCKA and L. ARAVIND, 2007 Discovering functional linkages and uncharacterized cellular pathways using phylogenetic profile comparisons: a comprehensive assessment. BMC Bioinformatics **8:** 173.

JUN, G., A. C. NAJ, G. W. BEECHAM, L. S. WANG, J. BUROS *et al.*, 2010 Meta-analysis confirms CR1, CLU, and PICALM as alzheimer disease risk loci and reveals interactions with APOE genotypes. Arch Neurol **67:** 1473-1484.

KAMATH, V., C. N. KYATHANAHALLI, B. JAYARAM, I. SYED, L. K. OLSON *et al.*, 2010 Regulation of glucose- and mitochondrial fuel-induced insulin secretion by a cytosolic protein histidine phosphatase in pancreatic beta-cells. Am J Physiol Endocrinol Metab **299:** E276-286.

KARIMPOUR-FARD, A., L. HUNTER and R. T. GILL, 2007 Investigation of factors affecting prediction of protein-protein interaction networks by phylogenetic profiling. BMC Genomics **8:** 393.

KASTENMAYER, J., L. NI, A. CHU, L. KITCHEN, W. AU *et al.*, 2006 Functional genomics of genes with small open reading frames (sORFs) in S. cerevisiae. Genome Res **16:** 365 - 373.

KEELING, P. J., G. BURGER, D. G. DURNFORD, B. F. LANG, R. W. LEE *et al.*, 2005 The tree of eukaryotes. Trends Ecol Evol **20:** 670-676.

KENT, W., 2002 BLAT - the BLAST-like alignment tool. Genome Res **12:** 656 - 664.

KENT, W., C. SUGNET, T. FUREY, K. ROSKIN, T. PRINGLE *et al.*, 2002 The human genome browser at UCSC. Genome Res **12:** 996 - 1006.

KIM, C. S., P. RIIKONEN and T. SALAKOSKI, 2008 Detecting biological associations between genes based on the theory of phase synchronization. Biosystems **92:** 99-113.

KIM, S. K., A. SHINDO, T. J. PARK, E. C. OH, S. GHOSH *et al.*, 2010 Planar cell polarity acts through septins to control collective cell movement and ciliogenesis. Science **329:** 1337-1340.

KING, N., M. J. WESTBROOK, S. L. YOUNG, A. KUO, M. ABEDIN *et al.*, 2008 The genome of the choanoflagellate Monosiga brevicollis and the origin of metazoans. Nature **451:** 783-788.

KOK, E. H., T. LUOTO, S. HAIKONEN, S. GOEBELER, H. HAAPASALO *et al.*, 2011 CLU, CR1 and PICALM genes associate with Alzheimer's-related senile plaques. Alzheimers Res Ther **3:** 12.

KOMATSUZAKI, S., Y. AOKI, T. NIIHORI, N. OKAMOTO, R. C. HENNEKAM *et al.*, 2010 Mutation analysis of the SHOC2 gene in Noonan-like syndrome and in hematologic malignancies. J Hum Genet **55:** 801-809.

KONTTINEN, Y. T., M. AINOLA, H. VALLEALA, J. MA, H. IDA *et al.*, 1999 Analysis of 16 different matrix metalloproteinases (MMP-1 to MMP-20) in the synovial membrane: different profiles in trauma and rheumatoid arthritis. Ann Rheum Dis **58:** 691-697.

KUBO, T., H. A. YANAGISAWA, T. YAGI, M. HIRONO and R. KAMIYA, 2010 Tubulin polyglutamylation regulates axonemal motility by modulating activities of inner-arm dyneins. Curr Biol **20:** 441-445.

KULP, D., 2003 Protein coding gene structure prediction using generalized hidden Markov models.

KURAHARA, S., M. SHINOHARA, T. IKEBE, S. NAKAMURA, M. BEPPU *et al.*, 1999 Expression of MMPS, MT-MMP, and TIMPs in squamous cell carcinoma of the oral cavity: correlations with tumor invasion and metastasis. Head Neck **21:** 627-638.

KWAN, A., L. LI, D. KULP, S. DUTCHER and G. STORMO, 2009 Improving Gene-finding in Chlamydomonas reinhardtii:GreenGenie2. BMC Genomics **10:** 210.

KWAN, A. L., S. K. DUTCHER and G. D. STORMO, 2010 Detecting Coevolution of Functionally Related Proteins for Automated Protein Annotation, pp. 99-105 in *Proceedings of the 2010 IEEE International Conference on Bioinformatics and Bioengineering*. IEEE Computer Society.

LAGIER-TOURENNE, C., M. TAZIR, L. C. LOPEZ, C. M. QUINZII, M. ASSOUM *et al.*, 2008 ADCK3, an ancestral kinase, is mutated in a form of recessive ataxia associated with coenzyme Q10 deficiency. Am J Hum Genet **82:** 661-672.

LANDER, E. S., L. M. LINTON, B. BIRREN, C. NUSBAUM, M. C. ZODY *et al.*, 2001 Initial sequencing and analysis of the human genome. Nature **409:** 860-921.

LARKIN, M. A., G. BLACKSHIELDS, N. P. BROWN, R. CHENNA, P. A. MCGETTIGAN *et al.*, 2007 Clustal W and Clustal X version 2.0. Bioinformatics **23:** 2947-2948.

LEFEBVRE, P. A., and J. L. ROSENBAUM, 1986 Regulation of the synthesis and assembly of ciliary and flagellar proteins during regeneration. Annu Rev Cell Biol **2:** 517-546.

LI, J., S. LIN, H. JIA, H. WU, B. ROE *et al.*, 2003 Analysis of Chlamydomonas reinhardtii genome structure using large-scale sequencing of regions on linkage groups I and III. J Eukaryot Microbiol **50:** 145 - 155.

LI, J. B., J. M. GERDES, C. J. HAYCRAFT, Y. FAN, T. M. TESLOVICH *et al.*, 2004 Comparative genomics identifies a flagellar and basal body proteome that includes the BBS5 human disease gene. Cell **117:** 541-552.

LI, J. B., M. ZHANG, S. K. DUTCHER and G. D. STORMO, 2005 Procom: a web-based tool to compare multiple eukaryotic proteomes. Bioinformatics **21:** 1693-1694.

LICHTINGHAGEN, R., D. MICHELS, C. I. HABERKORN, B. ARNDT, M. BAHR *et al.*, 2001 Matrix metalloproteinase (MMP)-2, MMP-7, and tissue inhibitor of metalloproteinase-1 are closely related to the fibroproliferative process in the liver during chronic hepatitis C. J Hepatol **34:** 239-247.

LIN, H., and U. GOODENOUGH, 2007 Gametogenesis in the Chlamydomonas reinhardtii minus mating type is controlled by two genes, MID and MTD1. Genetics **176:** 913 - 925.

LIPINSKI, R. J., C. SONG, K. K. SULIK, J. L. EVERSON, J. J. GIPP *et al.*, 2010 Cleft lip and palate results from Hedgehog signaling antagonism in the mouse: Phenotypic characterization and clinical implications. Birth Defects Res A Clin Mol Teratol **88:** 232-240.

LOGES, N. T., H. OLBRICH, A. BECKER-HECK, K. HAFFNER, A. HEER *et al.*, 2009 Deletions and point mutations of LRRC50 cause primary ciliary dyskinesia due to dynein arm defects. Am J Hum Genet **85:** 883-889.

LOMSADZE, A., V. TER-HOVHANNISYAN, Y. CHERNOFF and M. BORODOVSKY, 2005 Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res **33:** 6494 - 6506.

LORENZETTI, D., C. E. BISHOP and M. J. JUSTICE, 2004 Deletion of the Parkin coregulated gene causes male sterility in the quaking(viable) mouse mutant. Proc Natl Acad Sci U S A **101:** 8402-8407.

LORESTANI, A., L. SHEINER, K. YANG, S. D. ROBERTSON, N. SAHOO *et al.*, 2010 A Toxoplasma MORN1 null mutant undergoes repeated divisions but is defective in basal assembly, apicoplast division and cytokinesis. PLoS One **5:** e12302.

MARSZALEK, J. R., J. A. WEINER, S. J. FARLOW, J. CHUN and L. S. GOLDSTEIN, 1999 Novel dendritic kinesin sorting identified by different process targeting of two related kinesins: KIF21A and KIF21B. J Cell Biol **145:** 469-479.

MERCHANT, S., S. PROCHNIK, O. VALLON, E. HARRIS, S. KARPOWICZ *et al.*, 2007 The Chlamydomonas genome reveals the evolution of key animal and plant functions. Science **318:** 245 - 250.

MILL, P., P. J. LOCKHART, E. FITZPATRICK, H. S. MOUNTFORD, E. A. HALL *et al.*, 2011 Human and Mouse Mutations in WDR35 Cause Short-Rib Polydactyly Syndromes Due to Abnormal Ciliogenesis. Am J Hum Genet **88:** 508-515.

MUKHOPADHYAY, S., X. WEN, B. CHIH, C. D. NELSON, W. S. LANE *et al.*, 2010 TULP3 bridges the IFT-A complex and membrane phosphoinositides to promote trafficking of G protein-coupled receptors into primary cilia. Genes Dev **24:** 2180-2193.

MURAKAMI, S., K. KUEHNLE and D. STERN, 2005 A spontaneous tRNA suppressor of a mutation in the Chlamydomonas reinhardtii nuclear MCD1 gene required for stability of the chloroplast petD mRNA. Nucleic Acids Res **33:** 3372 - 3380.

MURDOCH, J. N., and A. J. COPP, 2010 The relationship between sonic Hedgehog signaling, cilia, and neural tube defects. Birth Defects Res A Clin Mol Teratol **88:** 633-652.

NACHURY, M. V., A. V. LOKTEV, Q. ZHANG, C. J. WESTLAKE, J. PERANEN *et al.*, 2007 A core complex of BBS proteins cooperates with the GTPase Rab8 to promote ciliary membrane biogenesis. Cell **129:** 1201-1213.

NEI, M., Y. SUZUKI and M. NOZAWA, 2010 The neutral theory of molecular evolution in the genomic era. Annu Rev Genomics Hum Genet **11:** 265-289.

OMRAN, H., D. KOBAYASHI, H. OLBRICH, T. TSUKAHARA, N. T. LOGES *et al.*, 2008 Ktu/PF13 is required for cytoplasmic pre-assembly of axonemal dyneins. Nature **456:** 611-616.

ORLANDO, D. A., C. Y. LIN, A. BERNARD, J. Y. WANG, J. E. S. SOCOLAR *et al.*, 2008 Global control of cell-cycle transcription by coupled CDK and network oscillators. Nature **453:** 944-947.

OSTROWSKI, L. E., K. BLACKBURN, K. M. RADDE, M. B. MOYER, D. M. SCHLATZER *et al.*, 2002 A proteomic analysis of human cilia: identification of novel components. Mol Cell Proteomics **1:** 451-465.

PARKINSON, J., and M. BLAXTER, 2009 Expressed sequence tags: an overview. Methods Mol Biol **533:** 1-12.

PATHAK, N., C. A. AUSTIN and I. A. DRUMMOND, 2011 Tubulin Tyrosine Ligase-like Genes ttll3 and ttll6 Maintain Zebrafish Cilia Structure and Motility. J Biol Chem **286:** 11685-11695.

PAZOS, F., J. A. RANEA, D. JUAN and M. J. STERNBERG, 2005 Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome. J Mol Biol **352:** 1002-1015.

PAZOUR, G., and G. WITMAN, 2009 The Chlamydomonas Flagellum as a model for human ciliary disease. The Chlamydomonas Sourcebook **3**.

PAZOUR, G. J., N. AGRIN, J. LESZYK and G. B. WITMAN, 2005 Proteomic analysis of a eukaryotic cilium. J Cell Biol **170:** 103-113.

PELLEGRINI, M., E. M. MARCOTTE, M. J. THOMPSON, D. EISENBERG and T. O. YEATES, 1999 Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A **96:** 4285-4288.

PETTERSSON, E., J. LUNDEBERG and A. AHMADIAN, 2009 Generations of sequencing technologies. Genomics **93:** 105-111.

PUGACHEVA, E. N., S. A. JABLONSKI, T. R. HARTMAN, E. P. HENSKE and E. A. GOLEMIS, 2007 HEF1-dependent Aurora A activation induces disassembly of the primary cilium. Cell **129:** 1351-1363.

QIAN, J., M. DOLLED-FILHART, J. LIN, H. YU and M. GERSTEIN, 2001 Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions. J Mol Biol **314:** 1053-1066.

RAMBAUT, A., 2007 FigTree v1.2.2.

ROCHAIX, J.-D., 2009 State Transitions. The Chlamydomonas Sourcebook **2**.

ROGER, A. J., and L. A. HUG, 2006 The origin and diversification of eukaryotes: problems with molecular phylogenetics and molecular clock estimation. Philos Trans R Soc Lond B Biol Sci **361:** 1039-1054.

ROGIC, S., A. MACKWORTH and F. OUELLETTE, 2001 Evaluation of gene-finding programs on mammalian sequences. Genome Res **11:** 817 - 832.

ROZEN, S., and H. SKALETSKY, 2000 Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol **132:** 365-386.

RUNDHAUG, J. E., 2005 Matrix metalloproteinases and angiogenesis. J Cell Mol Med **9:** 267-285.

SARMAH, B., V. P. WINFREY, G. E. OLSON, B. APPEL and S. R. WENTE, 2007 A role for the inositol kinase Ipk1 in ciliary beating and length maintenance. Proc Natl Acad Sci U S A **104:** 19843-19848.

SASAKI, J., S. KOFUJI, R. ITOH, T. MOMIYAMA, K. TAKAYAMA *et al.*, 2010 The PtdIns(3,4)P(2) phosphatase INPP4A is a suppressor of excitotoxic neuronal death. Nature **465:** 497-501.

SLOBODA, R. D., and L. HOWARD, 2009 Protein methylation in full length Chlamydomonas flagella. Cell Motil Cytoskeleton **66:** 650-660.

SNELL, W., and U. GOODENOUGH, 2009 Flagellar Adhesion: Flagellar generated signaling and gamete fusion during mating. The Chlamydomonas Sourcebook **3**.

SRIVASTAVA, M., E. BEGOVIC, J. CHAPMAN, N. H. PUTNAM, U. HELLSTEN *et al.*, 2008 The Trichoplax genome and the nature of placozoans. Nature **454:** 955-960.

STARK, C., B. J. BREITKREUTZ, A. CHATR-ARYAMONTRI, L. BOUCHER, R. OUGHTRED *et al.*, 2011 The BioGRID Interaction Database: 2011 update. Nucleic Acids Res **39:** D698-704.

STARK, K., D. KIRK and R. SCHMITT, 2001 Two enhancers and one silencer located in the introns of regA control somatic cell differentiation in Volvox carteri. Genes Dev **15:** 1449 - 1460.

STOLC, V., M. P. SAMANTA, W. TONGPRASIT and W. F. MARSHALL, 2005 Genome-wide transcriptional analysis of flagellar regeneration in Chlamydomonas reinhardtii identifies orthologs of ciliary disease genes. Proc Natl Acad Sci U S A **102:** 3703-3707.

STORMO, G., 2000 Gene-finding approaches for eukaryotes. Genome Res **10:** 394 - 397.

STRATIGOPOULOS, G., C. A. LEDUC, M. L. CREMONA, W. K. CHUNG and R. L. LEIBEL, 2011 Cut-like homeobox 1 (CUX1) regulates expression of the fat mass and obesity-associated and retinitis pigmentosa GTPase regulator-interacting protein-1-like (RPGRIP1L) genes and coordinates leptin receptor signaling. J Biol Chem **286:** 2155-2170.

SUN, J., J. XU, Z. LIU, Q. LIU, A. ZHAO *et al.*, 2005 Refined phylogenetic profiles method for predicting protein-protein interactions. Bioinformatics **21:** 3409-3415.

TANNER, C. A., P. ROMPOLAS, R. S. PATEL-KING, O. GORBATYUK, K. WAKABAYASHI *et al.*, 2008 Three members of the LC8/DYNLL family are required for outer arm dynein motor function. Mol Biol Cell **19:** 3724-3734.

TATUSOV, R., N. FEDOROVA, J. JACKSON, A. JACOBS, B. KIRYUTIN *et al.*, 2003 The COG database: an updated version includes eukaryotes. BMC Bioinformatics **4:** 41.

TOBIN, J. L., and P. L. BEALES, 2009 The nonmotile ciliopathies. Genet Med **11:** 386-402.

TRAPNELL, C., L. PACHTER and S. L. SALZBERG, 2009 TopHat: discovering splice junctions with RNA-Seq. Bioinformatics **25:** 1105-1111.

TRAPNELL, C., B. A. WILLIAMS, G. PERTEA, A. MORTAZAVI, G. KWAN *et al.*, 2010 Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol **28:** 511-515.

TSUJI, T., and T. KUNIEDA, 2005 A loss-of-function mutation in natriuretic peptide receptor 2 (Npr2) gene is responsible for disproportionate dwarfism in cn/cn mouse. J Biol Chem **280:** 14288-14292.

VARSHNEY, R. K., J. C. GLASZMANN, H. LEUNG and J. M. RIBAUT, 2010 More genomic resources for less-studied crops. Trends Biotechnol **28:** 452-460.

WALLINGFORD, J. B., and B. MITCHELL, 2011 Strange as it may seem: the many links between Wnt signaling, planar cell polarity, and cilia. Genes Dev **25:** 201-213.

WANG, X., K. HARIMOTO, J. LIU, J. GUO, S. HINSHAW *et al.*, 2011 Spata4 promotes osteoblast differentiation through Erk-activated Runx2 pathway. J Bone Miner Res.

WEI, C., and M. BRENT, 2006 Using ESTs to improve the accuracy of de novo gene prediction. BMC Bioinformatics **7:** 327.

WIRSCHELL, M., G. PAZOUR, A. YODA, M. HIRONO, R. KAMIYA *et al.*, 2004 Oda5p, a novel axonemal protein required for assembly of the outer dynein arm and an associated adenylate kinase. Mol Biol Cell **15:** 2729 - 2741.

WISEMAN, B. S., M. D. STERNLICHT, L. R. LUND, C. M. ALEXANDER, J. MOTT *et al.*, 2003 Site-specific inductive and inhibitory activities of MMP-2 and MMP-3 orchestrate mammary gland branching morphogenesis. J Cell Biol **162:** 1123-1133.

WLOGA, D., D. M. WEBSTER, K. ROGOWSKI, M. H. BRE, N. LEVILLIERS *et al.*, 2009 TTLL3 Is a tubulin glycine ligase that regulates the assembly of cilia. Dev Cell **16:** 867-876.

WU, T., and C. WATANABE, 2005 GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics **21:** 1859 - 1875.

XI, Q., G. J. PAUER, K. A. WEST, J. W. CRABB and S. A. HAGSTROM, 2003 Retinal degeneration caused by mutations in TULP1. Adv Exp Med Biol **533:** 303-308.

YAU, W. L., H. L. LUNG, E. R. ZABAROVSKY, M. I. LERMAN, J. S. SHAM *et al.*, 2006 Functional studies of the chromosome 3p21.3 candidate tumor suppressor gene BLU/ZMYND10 in nasopharyngeal carcinoma. Int J Cancer **119:** 2821-2826.

Vita

Alan Lechuen Kwan

Date of Birth	December 24, 1982
Place of Birth	Vancouver, British Columbia, CANADA
Degree	B.S. Computer Science, May 2004 M.S. Computer Science, December 2006 Ph.D. Computer Science, May 2011
Professional Societies	IEEE Genetics Society of America
Publications	**Kwan** AL, Lin H and Dutcher SK. *Whole-transcriptome sequencing reveals an early ciliogenesis regulation program*. In preparation. Lin H, **Kwan** AL and Dutcher SK. *Synthesizing and salvaging NAD: lessons learned from Chlamydomonas reinhardtii*. PLoS Genet. 2010 Sep 9;6(9). pii: e10011005 **Kwan** AL, Dutcher SK and Stormo GD. *Detecting Coevolution of Functionally Related Proteins for Automated Protein Annotation*. Proceedings of the 10th International IEEE Conference on Bioinformatics and Bioengineering, pp. 99-105. **Kwan** AL, Li L, Kulp DC, Dutcher SK and Stormo GD. *Improving Gene-finding in Chlamydomonas reinhardtii: GreenGenie2*. BMC Genomics 2009, 10:210.

May 2011