ANALYSIS OF THE POLYMORPHIC *SER* GENE FAMILY IN *TETRAHYMENA THERMOPHILA*

PATRATH PONSUWANNA

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY (BIOCHEMISTRY)
FACULTY OF GRADUATE STUDIES
MAHIDOL UNIVERSITY
2015

Thesis
entitled

ANALYSIS OF THE POLYMORPHIC *SER* GENE FAMILY IN *TETRAHYMENA THERMOPHILA*

……..……...
Miss Patrath Ponsuwanna
Candidate

……..……...
Asst. Prof. Thanat Chookajorn,
Ph.D. (Biochemistry, Molecular and Cell Biology)
Major advisor

……..……...
Prof. Sumalee Tungpradabkul,
Ph.D. (Molecular Biology)
Co-advisor

……...
Assoc. Prof. Wilai Noonpakdee,
Ph.D. (Pharmacology)
Co-advisor

……..…….....
Prof. Patcharee Lertrit,
M.D., Ph.D. (Biochemistry)
Dean
Faculty of Graduate Studies
Mahidol University

……..……...
Asst. Prof. Kittisak Yokthongwattana,
Ph.D. (Agricultural & Environmental Chemistry)
Program Director
Doctor of Philosophy Program in Biochemistry
Faculty of Science, Mahidol University

Thesis
entitled

ANALYSIS OF THE POLYMORPHIC *SER* GENE FAMILY IN *TETRAHYMENA THERMOPHILA*

was submitted to the Faculty of Graduate Studies, Mahidol University
for the degree of Doctor of Philosophy (Biochemistry)
on
July 13, 2015

……....
Miss Patrath Ponsuwanna
Candidate

……..……...	…….....
Asst. Prof. Thanat Chookajorn, Ph.D. (Biochemistry, Molecular and Cell Biology) Member	Prof. Prapon Wilairat, Ph.D. (Biochemistry) Chair
……..…...	……..……...
Prof. Prasit Palittapongarnpim, M.D. (Pediatrics) Member	Prof. Sumalee Tungpradabkul, Ph.D. (Molecular Biology) Member
……..…...	……..……...
Sittiporn Pattaradilokrat, Ph.D. (Genetics) Member	Assoc. Prof. Wilai Noonpakdee, Ph.D. (Pharmacology) Member
…….....	…….....
Prof. Patcharee Lertrit, M.D., Ph.D. (Biochemistry) Dean Faculty of Graduate Studies Mahidol University	Prof. Skorn Mongkolsuk, Ph.D. (Biological Science) Dean Faculty of Science Mahidol University

ACKNOWLEDGEMENTS

I wish to express my sincere thanks to my advisor, Asst. Prof. Thanat Chookajorn, for his expertise, guidance and patience. I would never be able to finish my dissertation without his perseverance with good science and mentorship. I also would like to thank all member of Chookajorn's lab for their help and support. They have taught me an invaluable lesson.

I would like to express my gratitude to my Thesis Committee Members for their guidance and suggestion. In particular, I wish to thank Prof. Prapon Wilairat for his valuable advices and Nitsara Karoonuthaisiri for her assistance with DNA microarray data analysis.

I take this opportunity to express my gratitude to all Department faculty members for their help and support. I am also grateful to my family, friends and my cat for keeping me alive during the hard times.

Patrath Ponsuwanna

ANALYSIS OF THE POLYMORPHIC *SER* GENE FAMILY IN *TETRAHYMENA THERMOPHILA*

PATRATH PONSUWANNA 5037194 SCBC / D

Ph.D.(BIOCHEMISTRY)

THESIS ADVISORY COMMITTEE: THANAT CHOOKAJORN, Ph.D., SUMALEE TUNGPRADABKUL, Ph.D., WILAI NOONPAKDEE, Ph.D.

ABSTRACT

Tetrahymena thermophila, a free-living ciliate, displays immobilization antigens (i-ag) on its surface. These proteins are cysteine-rich and are linked to the membrane by a glycosylphosphatidylinositol (GPI) anchor. They are encoded by a family of polymorphic *Ser* genes. Thirteen *Ser* genes were characterized so far, representing only a portion of the reported i-ag. Characterization of *Ser* gene family is impeded by its sequence polymorphism. In this study an algorithm to select *Ser* candidate from *T. thermophila* MAC genome was developed. The *Ser* prediction algorithm exploited two characteristic features of *Ser*—repetitive cysteine pattern and GPI anchor signal. The algorithm successfully selected Ser candidates including known *Ser*. *Ser* candidates were compared to known *Ser* and classified into subtypes by phylogenetic analysis. Ser candidates were found to be located as gene tandem array, suggesting that *Ser* gene family expands via gene duplication. *Ser* genes located on the same tandem array may have similar gene expression profile. The *Ser* prediction algorithm can select *Ser* candidates that did not belong to previously characterized subtype. Because the algorithm does not rely on sequence similarity, it can perform *Ser* identification more extensively than the sequence homology approach.

KEY WORDS: IMMOBILIZATION ANTIGEN / *TETRAHYMENA THERMOPHILA* / POLYMORPHIC GENE FAMILY

133 pages

การวิเคราะห์กลุ่มยีน *SER* ซึ่งมีความหลากหลายทางพันธุกรรมใน *TETRAHYMENA THERMOPHILA*

ANALYSIS OF THE POLYMORPHIC *SER* GENE FAMILY IN *TETRAHYMENA THERMOPHILA*

พัทธรัต พลสุวรรณา 5037194 SCBC / D

ปร.ด. (ชีวเคมี)

คณะกรรมการที่ปรึกษาวิทยานิพนธ์: ธนรรถ ชูขจร, Ph.D., สุมาลี ตั้งประดับกุล, Ph.D., วิไล หนุนภักดี, Ph.D.

บทคัดย่อ

Tetrahymena thermophila เป็นสิ่งมีชีวิตเซลล์เดียวประเภทดำรงชีวิตอิสระ (free-living) ชนิดหนึ่งซึ่งแสดงโปรตีน immobilization antigen (i-ag) บนผิวของมัน โปรตีนที่ผิวเหล่านี้มีกรดอะมิโน cysteine เป็นส่วนประกอบมากและยึดโยงอยู่กับเยื่อหุ้มเซลล์ด้วย glycosylphosphatidyl-inositol (GPI) โปรตีนเหล่านี้ถูก encode โดยกลุ่มยีน *Ser* ซึ่งเป็นกลุ่มยีนที่มีความหลากหลายทางพันธุกรรม จวบปัจจุบันนี้มีการศึกษาและอธิบายลำดับนิวคลีโอไทด์ (characterize) ของยีน *Ser* ไว้ทั้งสิ้น 13 ยีนซึ่งนับเป็นเพียงส่วนหนึ่งของ i-ag ทั้งหมดที่เคยมีการรายงานไว้ การศึกษาลักษณะยีน *Ser* ประสบปัญหาจากความหลากหลายทางพันธุกรรมในลำดับนิวคลีโอไทด์และกรดอะมิโน การศึกษานี้ได้พัฒนาอัลกอริทึมเพื่อค้นหาและคัดเลือกยีน *Ser* จากมาโครนิวเคลียร์จีโนมของ *T. thermo-phila* อัลกอริทึมสำหรับทำนายยีน *Ser* อาศัยลักษณะเด่นสองประการของโปรตีนที่ยีนนี้ encode ได้แก่ลำดับกรดอะมิโนที่ประกอบด้วย cysteine ซ้ำ (repetitive) และ GPI anchor signal อัลกอริทึมนี้สามารถคัดยีนกลุ่มหนึ่งเป็นตัวเลือกยีน *Ser* ซึ่งในกลุ่มนี้มียีน *Ser* ที่เคยมีการ characterize เอาไว้รวมอยู่ด้วย ตัวเลือกยีน *Ser* ถูกนำมาศึกษาเปรียบเทียบและจัดหมวดหมู่โดยการวิเคราะห์ด้วย phylogenetic การศึกษานี้พบยีน *Ser* ซึ่งไม่ถูกจัดอยู่ในกลุ่ม (subtype) เดียวกับยีน *Ser* ซึ่งเคยมีการ characterize เอาไว้ และพบว่ายีนในกลุ่มตัวเลือกยีน *Ser* อยู่เป็นกลุ่มของยีนบนโครโมโซมในลักษณะเป็น tandem array ซึ่งบ่งชี้ว่ากลุ่มยีน *Ser* อาจขยายขนาดกลุ่มโดยการเพิ่มชุดของยีน (gene duplication) ยีน Ser ใน tandem array เดียวกันมักมีรูปแบบของการแสดงออกคล้ายคลึงกัน อัลกอริทึมนี้สามารถค้นหาและระบุยีน *Ser* ได้อย่างครอบคลุมกว่าวิธีการค้นหาโดยอาศัยความคล้ายคลึงของลำดับนิวคลีโอไทด์ หรือกรดอะมิโน (sequence homology)

133 หน้า

CONTENTS

CONTENTS (cont.)

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES (cont.)

LIST OF ABBREVIATIONS

aa	amino acid
APase	alkaline phosphatase
Ca	calcium
Cbs	chromosome breakage site
cDNA	complementary DNA
Cys	cysteine
DBL	Duffy-binding like
DNA	deoxyribonucleic acid
ER	endoplasmic reticulum
EST	expressed sequence tag
Fe	ferrum
FRET	fluorescence resonance energy transfer
GFP	green fluorescent protein
GO	gene ontology
GPI	glycosylphosphatidylinositol
HMM	Hidden Markov model
i-ag	immobilization antigen
IES	internal eliminated sequence
kb	kilobase
kDa	kiloDalton
LSU rRNA	large subunit ribosomal ribonucleic acid
MAC	macronucleus
Mb	megabase
MIC	micronucleus
ml	mililiter
mtDNA	mitocondrial deoxyribonucleic acid
NJ	neighbour-joining

LIST OF ABBREVIATIONS (cont.)

NN	neural network
PfEMP1	Plasmodium falciparum erythrocyte membrane protein 1
PI-PLC	phosphatidylinositol phospholipase C
PLAP	placental alkaline phosphatase
rDNA	ribosomal deoxyribonucleic acid
RFLP	restriction fragment length polymorphism
RNA	ribonucleic acid
rRNA	ribonucleic acid interference
RT-PCR	reverse transcription polymerase chain reaction
SOM	self organizing map
tRNA	transfer ribonucleic acid
VSG	variable surface glycoprotein
μm	micrometer

CHAPTER I
INTRODUCTION AND LITERATURE REVIEW

1.1 Biology of Tetrahymena thermophila

Tetrahymena thermophila is a single-celled ciliate found in temperate freshwater (Collins and Gorovsky 2005; Simon et al. 2007). It is often found in decaying plants at the mud-water interface together with other *Tetrahymena* species (Doerder et al. 1995). The freshwater sources where *T. thermophila* is found range from still water, running water, and ponds; at the elevation from sea level to 10,000 feet (Elliott 1959). *T. thermophila* is a member of phylum Ciliophora, which includes organisms such as *Paramecium*, *Oxytricha*, and *Ichthyophthirius* (fish parasite). Based on phylogenetic analysis of small subunit ribosomal RNA (Sogin et al. 1986; Struder-Kypke et al. 2001) and mitochondrial cytochrome c oxidase subunit 1 (*cox*1) sequences (Chantangsi and Lynn 2008), genus *Tetrahymena* appears to be monophyletic. There are reports of tetrahymenas from various regions worldwide; twenty-nine strains of *T. thermophila* have been isolated exclusively from North America region (Simon et al. 2007). Due to lack of knowledge on specific food preference, presumably *T. thermophila* feeds on bacteria (Simon et al. 2007); however, it can also grow in a variety of axenic media under laboratory condition (Elliott 1959; Cassidy-Hanley 2012). In general, tetrahymenas are preyed on by other ciliates, invertebrate larvae such as mosquito larvae, invertebrates, and small vertebrates (Simon et al. 2007). In common with other ciliates, it exhibits nuclear dualism with micronucleus (MIC) as a germline nucleus and macronucleus (MAC) as a nucleus specific for gene expression (Karrer 2000). MIC is transcriptionally inert in general (Mayo and Orias 1981), and MAC is transcriptionally active. *Tetrahymena* strains collected from natural habitat are mostly amicronucleate, suggesting that MIC is not essential for vegetative growth (Karrer 2000). Natural amicronucleate Tetrahymena may have multiple origins (Doerder 2014). However, amicronucleate *Tetrahymena*

generated in laboratory are usually not viable (Karrer 2000). The diploid MIC contains five pairs of chromosome (Karrer 2000). The polyploid MAC contains about 200-300 acentromeric chromosomes (Karrer 2000; Altschuler and Yao 1985; Conover and Brunk 1986).

1.1.1 Morphology of *T. thermophila*

T. thermophila has a pyriform shape (pear-shaped), but the lack of food may induce it to form an elongated shape known as 'rapid swimmer' (Frankel 2000). Its size is about 50 μm in length and about 60 μm in maximum width (Karrer 2000), but the dimension may vary even within the same strain under different conditions (Elliott 1959). Cell growth occurs longitudinally while cell division occurs in transverse. *T. thermophila* has almost all the basic structural components of animal cell except intermediate filament and characteristic features of plant cell such as cell wall and chloroplast. The complex arrangement of structure on the cell surface is one of the distinctions of *T. thermophila*. *T. thermophila*'s plasma membrane covers the whole cell surface with cilia membrane. Underneath plasma membrane of *T. thermophila* lie the networks of interconnected flattened vesicles called cortical alveoli. Two layers of alveolar membranes, outer alveolar membrane and inner alveolar membrane, surround cortical alveoli. Cortical alveoli extend longitudinally along anterior-posterior axis and laterally from ciliary row to the midline between two adjacent rows. One possible function of cortical alveoli is the storage of plasma membrane's precursor; which is based on an observation of close positioning of cortical alveoli and plasma membrane, and ruthenium staining of cortical alveoli that normally are not stained unless the cortical alveoli are exposed to the environment (Williams 1983). Another possible function of cortical alveoli is the Ca^{2+} storage site, which is based on preliminary observation of peripheral calcium band (Stelly et al. 1995). Presence of cortical alveoli is a conserved feature among alveolate evolutionary lineage, which includes apicomplexan (e.g. malarial parasites) and dinoflagellates. Further underneath the cortical alveoli are layers of microtubule bands and ciliary units including basal bodies. Dense core secretory granules (mucocysts) arranged in longitudinal rows also lie beneath cortical alveoli. Arrays of mitochondria are in parallel to rows of basal

bodies and secretory granules. Flattened endoplasmic reticulum and arrays of small Golgi elements are beneath mitochondria arrays. (Frankel 2000). The surface of *T. thermophila* is covered with rows of cilia that originate from basal bodies and streak along anteroposterior axis (Karrer 2000). Generally, *T. thermophila* has 17 – 21 rows of cilia (Elliott 1959). There are three distinct surface 'landmarks' on *T. thermophila* cell surface: oral apparatus, cytoproct, and contractile vacuole pore. Oral apparatus is made up of four compound ciliary elements (hence the name 'tetrahymena') called undulating membranes. It is located at the anterior end of cell. Cytostome (i.e. cell mouth) is located at the posterior end of the oral apparatus and is a site where food vacuole is formed. Contractile vacuole pores, which are the site where a single contractile vacuole removes its content, are located at the posterior end of cell. There are generally two contractile vacuole pores, but *T. thermophila* with one or three pores is rarely found (Frankel 2000). Cytoproct (i.e. cell anus), which is the site where food vacuoles excrete, is also located at the posterior end. One MAC at the diameter of 10 μm is visible in the middle of the organism's body. Smaller MIC (1 – 2 μm in diameter) lies adjacent to the MAC in the sexually active strains.

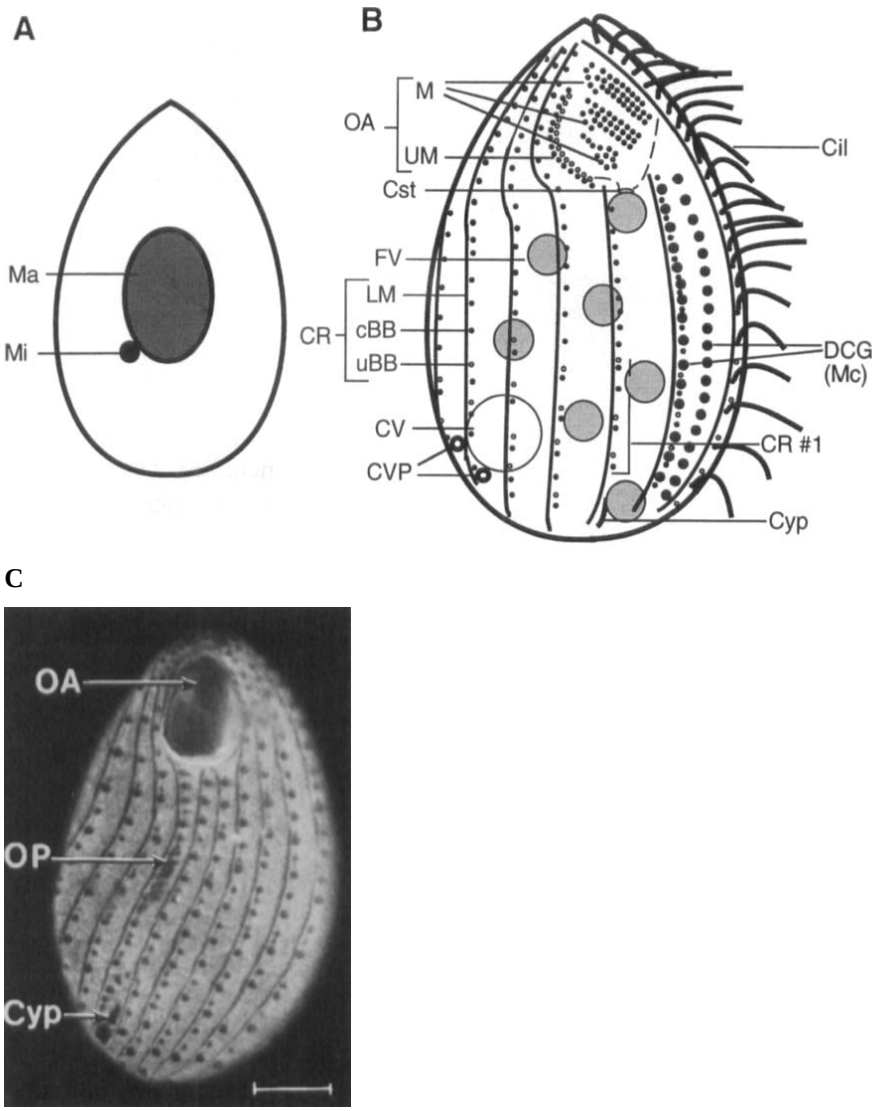


Figure 1.1 Illustration and microscopic image of *Tetrahymena*. (A) Simplified figure shows pear-shaped cell containing large macronucleus (Ma) and small MIC (mi). (B) Detailed figure depicts structural components: OA, oral apparatus; M, membranelle; UM, undulating membrane; Cst, cytostome; Cil, cilia; FV, food vacuole; CR, ciliary row; LM, longitudinal microtubule band; cBB, ciliated basal body; uBB, unciliated basal body; DCG (Mc), dense core secretory granule (Mucocyst); CV, contractile vacuole; CVP, contractile vacuole pore; Cyp, cytoproct. (C) Fluorescence image depicts oral apparatus at the anterior and pytoproct at the posterior. Oral primordium (OP) will undergoes further development to form ciliary structure. (Frankel 2000).

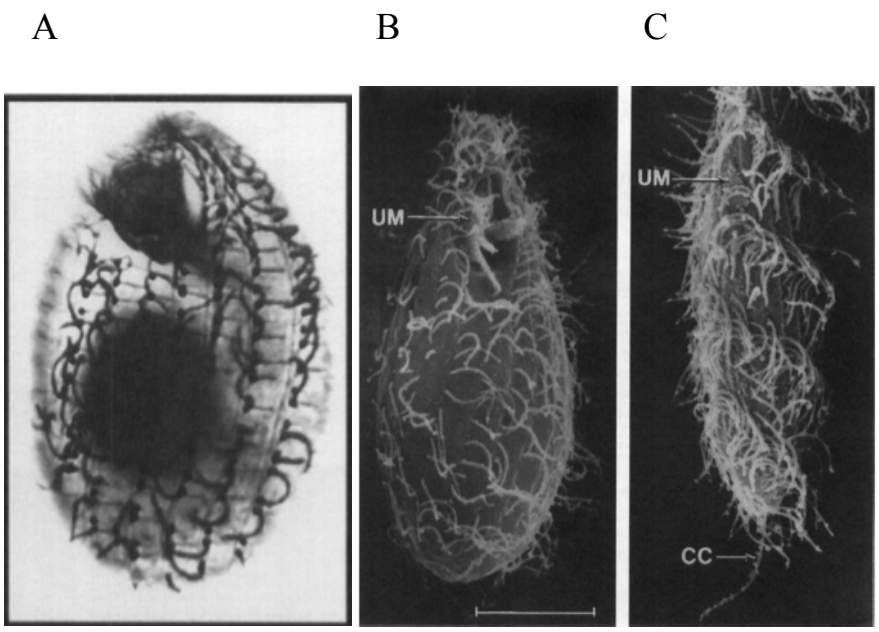


Figure 1.2 Light microscopic (LM) and electron microscopic (EM) image of *Tetrahymena*. (A) LM image of *Tetrahymena* that was stained with protein-silver staining. Ciliary row and oral apparatus were visualized. Macronucleus is also visible in the middle of the cell. (B) EM of *Tetrahymena* that was starved for 2 hours. Its overall shape resembles growing *Tetrahymena*. Undulating membrane (UM) was depicted. (C) EM of *Tetrahymena* that was starved for 5 hours. Its shape was elongated. Long caudal cilium (CC) is visible. (Frankel 2000).

1.1.2 Nutrients for *T. thermophila*

T. thermophila has two modes of nutrient intake: by phatocygosis via its oral apparatus and via direct transportation of nutrients in solution (Basmussen and Orias 1975). Availability of the latter renders food vacuole dispensable, allowing a mutant with defective phagocytosis to be grown in specific media (Cassidy-Hanley 2012). *T. thermophila* may uptake the nutrient from decaying plant in natural freshwater habitat (Doerder et al. 1995) by extracellular digestion using secreted acid hydrolase. The mutant with defective acid hydrolase secretion shows impaired growth in media with wheat grain as the sole nutrient source (Florin-Christensen et al. 1990). In laboratory condition, *T. thermophila* can be grown in various media including proteose peptone-based media, bacterized media, chemically defined media and skimmed milk media originally developed for inexpensive large-scaled culture (Cassidy-Hanley 2012). *T. thermophila* can grow quickly under laboratory condition with the doubling time within 2-3 hours (Frankel 2000) and as fast as 1.4 hours in optimal media (Kiy and Tiedtke 1992). *T. thermophila* can be grown indefinitely given large media volume (Orias et al. 2000). *T. thermophila* requires eleven essential amino acids, six essential B-complex vitamins, Fe^{3+}, a few trace metals, guanine and pyrimidine (preferably uracil) for growth (Elliott 1959). Lipids and carbohydrates are stimulatory for *T. thermophila*'s growth but are not necessary (Elliott 1959). Its lack of ability to synthesize purine and pyrimidine ring results in requirement in preformed purine and pyrimidine (Frankel 2000). No *de novo* purine synthesizing genes were found in *T. thermophila* genome by genomic analysis using known apicomplexan purine salvage enzymes as TBLASTN query (Chaudhary et al. 2004). Lack of *de novo* purine synthesis is common in other alveolates (Chaudhary et al. 2004).

1.1.3 Reproduction of *T. thermophila*

Like other single-celled organisms, *T. thermophila* can reproduce asexually and sexually. *T. thermophila* reproduces asexually by binary fission when food is abundant (Karrer 2000). During asexually reproduction, MIC undergoes mitotic division, and acentromeric MAC undergoes amitotic division. MAC division in *T. thermophila* occurs without chromosome condensation and MAC membrane

dissolution. MAC is elongated and constricted to form two new MACs for progeny cells instead. Lack of centromere and inability to equally assort genetic content in MAC underlies a phenomenon known as phenotypic assortment. Through the phenotypic assortment process, heterozygous loci in the cell eventually become homozygous in the progeny cells after several rounds of vegetative fission. *T. thermophila* can undergo indefinite vegetative growth, but they show aging and reduced fertility. Sexual reproduction in natural population of *T. thermophila* is frequent and consistent with change in food supply (Doerder et al. 1995). Under laboratory condition, sexual mating of *T. thermophila* can be induced by starvation in inorganic media (Cassidy-Hanley 2012). Starvation drives *T. thermophila* either to transform into an elongated fast-moving 'rapid swimmer' form or to conjugate with another mating type; the transformations are not mutually exclusive (Frankel 2000). *T. thermophila* have seven mating types, assigned by roman numeral, which are specified by *mat* alleles capable of being expressed in an array of mating types (Simon et al. 2007). Mating type of *T. thermophila* is determined during macronuclear development. Four progeny nuclei have independent mating type determination. This means the progeny cells may not have the same mating type, and that they are capable of mating among themselves. When two *T. thermophila* cells with different mating types are paired to conjugate, each of their MIC undergoes meiosis. MIC moves away from MAC and elongates approximately 50-folds into a 'crescent MIC', which is visible as a thin line curving around MAC (Karrer 2000). Crescent MIC condenses and undergoes two meiotic divisions. Out of the four micronuclei that arise from two meiotic divisions, three degenerates, and only one MIC is selected. The selected MIC then undergoes mitotic division and eventually produces two pronuclei that are genetically identical. One pronucleus (stationary pronucleus) remains within the parental cell, and the other one (migratory pronucleus) is exchanged between cells of mating pair. Pronuclei received from the mating pair and parental pronuclei then fuse together to form a synkaryon, which is a progenitor of daughter MIC and MAC. Synkaryon immediately undergoes two postzygotic mitotic divisions that result in two nuclei located at the anterior end and another two nuclei at the posterior end. The nuclei at the anterior then decondense and differentiate into new MAC, and nuclei at the posterior are maintained as MIC. While both new MACs undergo differentiation,

parental MAC partially supports the process. New MACs and MICs migrate to the center of the cell. Parental MAC migrates to the posterior end and eventually becomes degraded. Parental MAC DNA is not transmitted to the progeny (Harrison and Karrer 1985; Karrer 2000). Following MAC differentiation, the paired cells are separated. One MIC degrades, and another divides mitotically. Finally, each cell undergoes cytokinesis, producing total four progeny cells. These progenies cannot immediately undergo sexual reproduction and require time to mature. The maturing time is estimated at 65 asexual productions for *T. thermophil*a inbred strain B. (Karrer 2000).

The nuclei destined to be new MAC undergo MAC differentiation—the process that rearranges five germline chromosomes into polyploid somatic chromosomes. DNA rearrangement includes elimination of DNA at specific sequences and amplification of MAC copies (Karrer 2000). There are two types of DNA elimination process. The first type of DNA elimination is deletion-ligation. The deletion of specific DNA sequence, which is referred to as 'internal eliminated sequence (IES)', is followed by ligation of flanking sequences. IES includes transponsable elements and repeated sequences unique to MIC chromosome (Orias et al. 2011). The second type of DNA elimination is chromosome breakage. The deletion of highly conserved 15-bp sequence, which is referred to as 'chromosome breakage sequence (Cbs)', results in fragmentation of chromosome into smaller pieces (Hamilton et al. 2006). Each of five MIC undergoes about 50 chromosome breakage events, producing about 200-300 MAC chromosomes of size from 20 kb to over 1500 kb (Karrer 2000; Conover and Brunk 1986). Telomere is added to the fragmented chromosome pieces. Removal of IES occurs via programmed RNAi-mediated comparison between parental MAC genome and progeny MIC genome (Orias et al. 2011). By DNA rearrangement, IES and Cbs accounted for approximately 15% of DNA content (Yao and Gorovsky 1974), are specifically removed from MAC. DNA rearrangement can occur at the alternative junction and is suggested to contribute to phenotypic diversity (Howard and Blackburn 1985). Each somatic MAC chromosomes is amplified to approximately 45 copies. Phenotypic assortment of 45 copies of each chromosome causes higher phenotypic variation compared to diploid organism (Orias et al. 2011). Interestingly, rRNA-encoding gene is amplified to

approximately 9,000 copies in MAC (Gall 1974) from only one copy of Cbs-flanked rDNA-encoding gene in MIC (Karrer 2000). The MAC developmental process described above produces difference in genetic complexity between MAC and MIC genome.

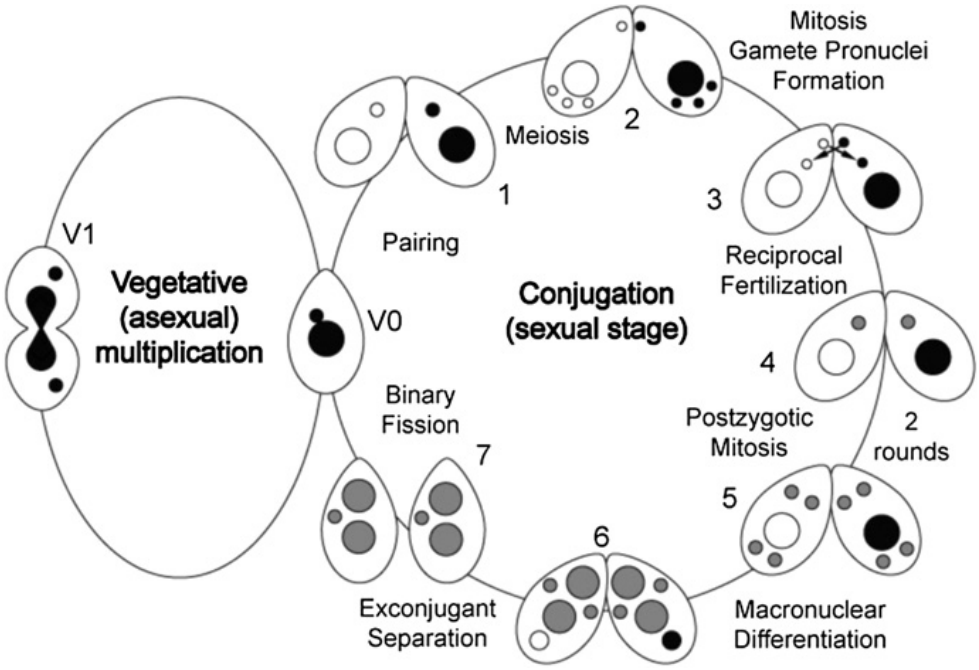


Figure 1.3 Asexual and sexual stage in *T. thermophila*. During asexual stage, vegetative *T. thermophila* (V0) undergoes binary fission (V1) to generate progenies. MIC divides mitotically, but MAC divides amitotically. During sexual stage, *T. thermophila* under starvation pairs with another cell with different mating type (1). Meiosis of MIC and degradation of three posterior micronuclei (2) is followed by exchange of pronuclei between the mating pair (3). Stationary and migratory pronuclei fertilize to form diploid synkaryon, which undergoes two rounds of mitosis (4) and results in four nuclei (5). Nuclei at the anterior differentiate into new MAC, and nuclei at the posterior become new MIC (6). Parental MAC undergoes degradation during progeny MAC differentiation process. Finally, the conjugated pair separates, and one MIC is degraded (7). Remaining MIC undergoes one round of mitosis to produce two MIC. Eventually each conjugant undergoes cytokinesis to produce four progenies in total. Each progeny has one MIC and one MAC. (Orias et al. 2011)

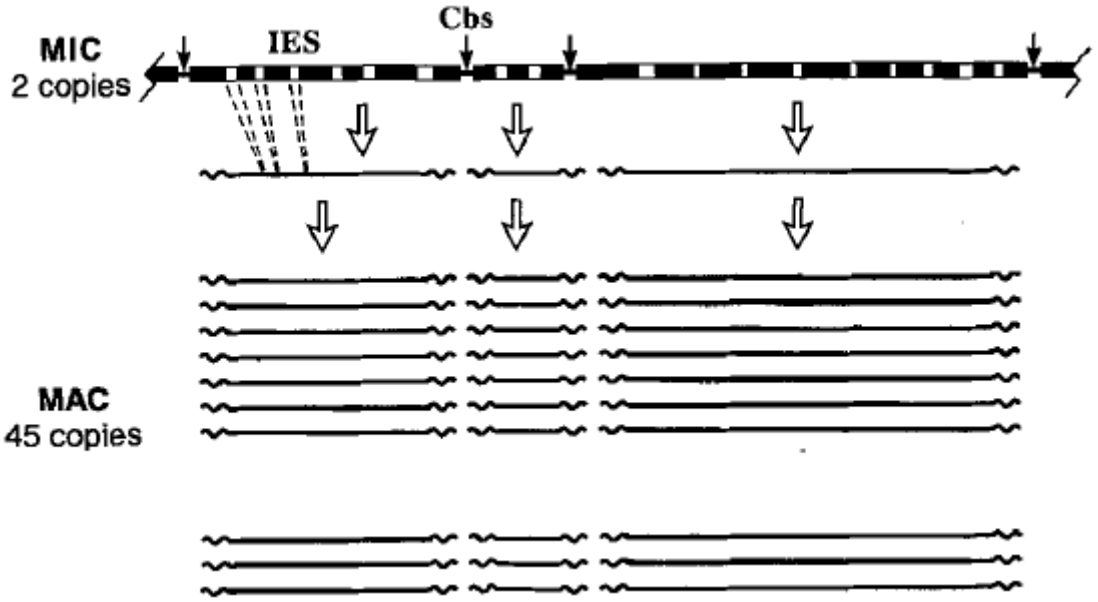


Figure 1.4 Illustration of MAC development in *T. thermophila*. The process that MAC genome is generated from MIC genome is called MAC development. MIC-specific sequences namely internal eliminated sequences (IES) are removed. MIC breaks at the specific sites namely chromosome breakage site (Cbs) and creates several MAC chromosomes from five MIC chromosomes. Each piece undergoes amplification and addition of telomere (shown as wavy line) at both ends. The length and the copy number vary for each MAC chromosome species. (Orias 2000).

1.1.4 *T. thermophila* strains

Until 1976, *T. thermophila* was designated as "*T. pyriformis* variety 1" and subsequently as "*T. pyriformis* syngen 1" due to the original designation of a group of morphologically indistinguishable tetrahymena species as *T. pyriformis* (Nanney and Simon 2000), making *T. pyriformis* a cryptic species complex. It was found in 1955 that there are eight 'varieties' of '*T. pyriformis*' that were unable to crossbreed each other, hence the denoted numeral of these eight genetically distinct '*T. pyriformis*' variety (Gruchy 1955). The term 'syngen' was coined later in 1957 to refer to genetically distinct morphospecies, i.e. species with indistinguishable morphology, when identification method not based on morphology was not available (Hall and Katz 2011). A number of inbred strains of *T. thermophila* were derived from natural isolates. *T. thermophila* strains were usually named after the site of collection; strains derived from natural isolates collected from Wood Hole, Massachusetts were named WH. Some strains were developed as the laboratory crosses and were named after their laboratory of origin, e.g. CU strain was developed by a laboratory from Cornell University (Cassidy-Hanley 2012). Inbred strain A and B were derived from strain WH. Inbred strain C was derived from natural isolates collected from Vermont. Inbred strain D was derived from natural isolates collected from Michigan (Nanney and Simon 2000). Each inbred strain carries different homologous mating allele, e.g. inbred strains A are homologous for *matA* allele. Inbred strain A, C, and D are capable of expressing all mating types except mating type IV and mating type VII, and inbred strain B can express all mating types except mating type I (Nanney and Simon 2000). Inbred strain B is the most genetically stable and is assigned as the reference strain (Allen et al. 1998). Progenies of inbred strain B and strain C3 were used for mapping genes on MIC and MAC genome (Orias et al. 2000). Strain SB210 that was used for genome sequencing was also derived from inbred strain B.

MIC-defective strains are called star strains. The name of these strains are designated by their originated inbred strains followed by an asterisk and the Roman numeral indicating mating type, e.g. A*III. Due to defective MIC, star strains can mate with normal *T. thermophila* cell but cannot contribute genetic materials to the progeny. Conjugation with star strains occurs through an alternative conjugation process called

‘genetic exclusion’ (Allen 1967). Star strains are useful genetic tool and are employed for constructing homozygous strain through two rounds of mating, creating functional heterokaryons strain, i.e. strain with non-expressed mutant allele in MIC and expressed non-mutant allele in MAC. Star strains are also used for conjugation rescue (Satir et al. 1986), short-circuit genomic exclusion (Bruns et al. 1976), and uniparental cytogamy (Cole and Bruns 1992; Cassidy-Hanley 2012) (details of each process are described below). Several *T. thermophila* mutant strains are available from the *Tetrahymena* Stock Center (http://tetrahymena.vet.cornell.edu). These strains include drug resistant mutants, temperature sensitive mutants, exocytosis mutants, and conjugation mutants (Cassidy-Hanley 2012). Drug resistant strains are heterokaryons homozygous for resistance in drug such as cyclohexamide (protein synthesis inhibitor), 6-methylpurine (adenine analog) (Byrne et al. 1978), and paromomycin (protein synthesis inhibitor) (Bruns et al. 1985). They carry drug resistance alleles in germinal MIC but confer drug sensitive phenotype from their MAC. The mutants show drug resistance phenotype only after another round of conjugation, which brings drug resistant allele to transcriptionally active MAC. These mutants can be used to select whole-genome homozygotes via conjugation with induced self-fertilization (namely cytogamy). Temperature-sensitive mutations may occur from the defects at various biological processes such as cell division, phagocytosis, and morphological development (Cassidy-Hanley 2012). Exocytosis mutants have defective mucocyst secretion, and are useful for purification of cell organelles such as cilia or cortex preparation for *in vitro* assay of motor proteins since the secreted substance can interfere with the purification (Orias et al. 2000). These mutants were used in a study of secretory granule formation and exocytosis (Turkewitz 2004). Conjugation mutants are useful in the studies of conjugation process. Mutant strains with defect in early, middle, and late stage conjugation are also available (Cassidy-Hanley 2012).

There are conjugation variants in *T. thermophila* that can be induced, such as genomic exclusion and cytogamy, and are useful in research application. In genomic exclusion process, normal *T. thermophila* cell conjugates with a star strain cell and transfers its migratory pronucleus to the conjugant. It receives no pronucleus in return from its conjugant pair because the star strain carries defective MIC. Exconjugants become homozygotes for their MIC genome and retain their old MACs.

Their mating types remain the same, and they can undergo a second round of conjugation immediately if left without food. By the end of the second round of conjugation, exconjugants are both MIC and MAC homozygous (Allen 1967). When C* strain is used for conjugation, there is about 10% of first-round exconjugants that haploid pronuclei diploidize. These small fractions of exconjugants then undergo postzygotic development and MAC differentiation, resulting in whole-genome homozygotes by only one round of conjugation. This method is called 'short-circuit genomic exclusion' (Bruns et al. 1976). Cytogamy is an alternative conjugation process that can be employed to generate whole-genome homozygotes as well. In cytogamy, exchange of migratory pronuclei between the conjugants is blocked by administration of hypoosmotic shock, and self-fertilization of each conjugant is induced (Orias and Hamilton 1979). Each exconjugants become a whole-genome homozygote. Usually the two exconjugants are different genetically. Cytogamy can be employed for isolating recessive mutant. Uniparental cytogamy is yet another method that is a combination of genomic exclusion and cytogamy to create whole-genome homozygotes (Cole and Bruns 1992). A heterokaryon carrying drug resistance marker in MIC conjugates with a star strain cell and is induced to self-fertilize by pronuclei exchange blockage (hence the naming 'uniparental'). The star strain conjugant will eventually perish due to degradation of old MAC without new MAC replacement, leaving its whole-genome homozygous exconjugant to be selectable by the drug treatment. The unique biology of *T. thermophila* makes it a useful model organism leading to many seminal discoveries in the field of molecular biology (Simon et al. 2007). The nuclear dimorphism of inactive germinal MIC and active somatic MAC together with phenotypic assortment allows lethal mutation and loss of the whole chromosome to be maintained as heterokaryons that can produce viable progeny.

1.1.5 Other *Tetrahymena* species

T. thermophila is a member of *Tetrahymena* group that can be found in various part of the world. *Tetrahymena* genus belongs to phylum Ciliophora, class Oligohymenophorea, order Hymenostomatida, family Tetrahymenidae. Tetrahymenas were collected from freshwater habitat in North America, Central America, Columbia,

Mexico, Europe, Australia, New Zealand, Pacific island, and Asia (Simon et al. 2007). Some tetrahymenas are found at specific geographical areas; *T. pigmentosa* is found only in North America, and *T. capricornis* is found only in Australia (Simon et al. 2007). On the other hand, tetrahymenas such as *T. elliotti* and *T. australis* were found from freshwater source worldwide. Even though many tetrahymenas including *T. thermophila* are freshwater free-living ciliates, some tetrahymenas are facultative parasites which were isolated from larval and adult invertebrates. These parasitic tetrahymenas include *T. limacis*, *T. rostrata*, and *T. chiromoni* which parasitize slugs, snails, and larvae of midge flies, respectively. There is a report of one tetrahymena species (*T. corlissi*) that was isolated from skin, muscle, and viscera of freshwater fish and can cause disease in fish (Hoffman et al. 1975). Before the advent of genetic-based identification method, tetrahymenas were classified into three complexes based on their morphologies and life traits: *pyriformis* complex, *patula* complex, and *rostrata* complex. This classification as complexes was shown later to be paraphyletic. The *pyriformis* complex consists of microstomatous, bacterivorous tetrahymenas with smaller cell size and fewer ciliary rows. The *rostrata* complex consists of parasitic, histophagous tetrahymenas with larger cell and more ciliary rows. Some tetrahymenas from *rostrata* complex can form resting cyst. The *patula* complex consists of tetrahymenas that can transform between microstome and macrostome. Tetrahymena species are morphologically similar, or even indistinguishable and caused an uncertain identity problem of '*T. pyriformis*' complex. Yet on the other hand, tetrahymenas possess high variety in molecular components (Simon et al. 2007). Based on an analysis of D2 LSU rRNA sequences, tetrahymenas can be divided into a few groups, with *T. elliotti* and *T. malaccensis* in the same group with *T. thermophila* (Nanney et al. 1998).

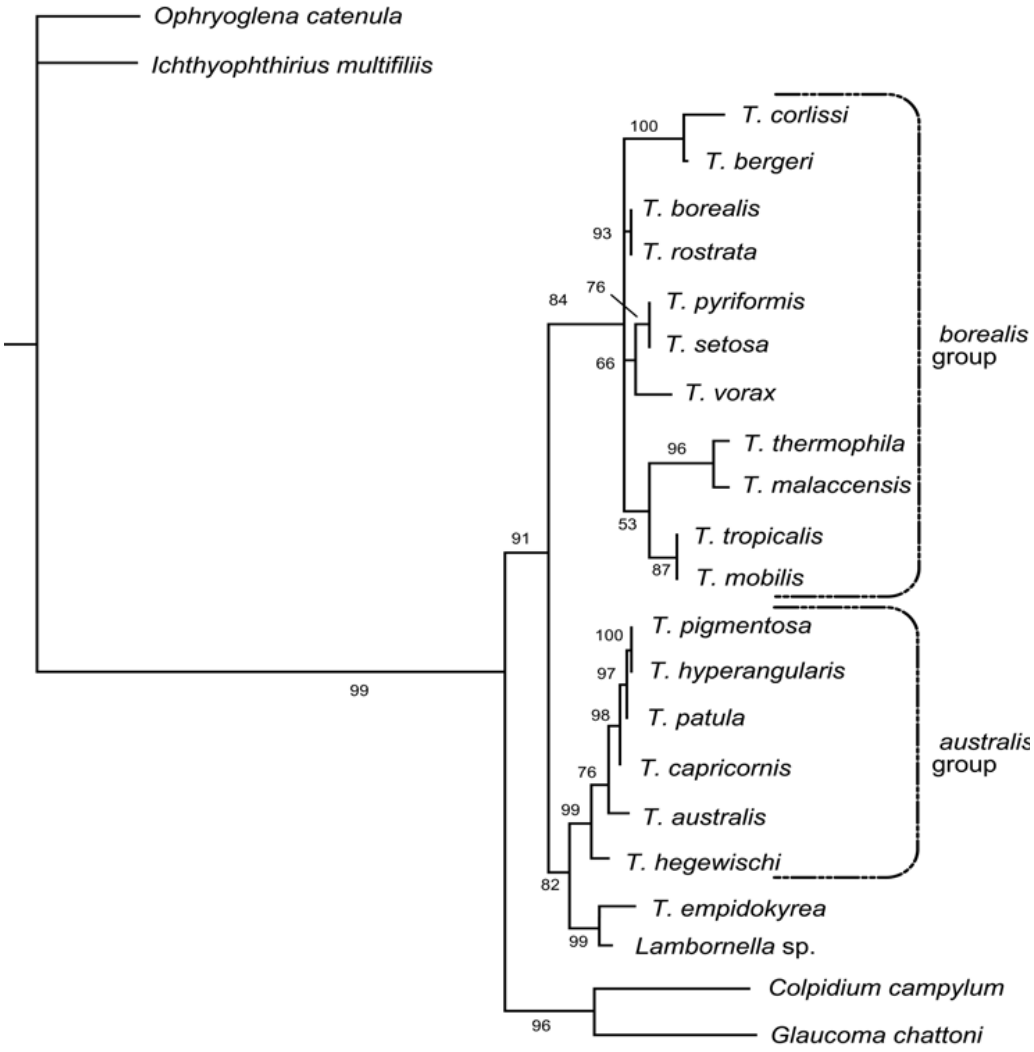


Figure 1.5 Tetrahymena phylogeny tree inferred from small subunit ribosomal DNA sequences. Tree was estimated by maximum likelihood method. Branch support value was showed on the branch. *Glaucoma chattoni* and *Colpidium campylum* are two tetrahymenid species that were chosen as outgroup. Two ophryoglenid, *Ophryoglena catenula* and *Ichthyophthirius multifiliis*, were also chosen as outgroup. (Struder-Kypke et al. 2001).

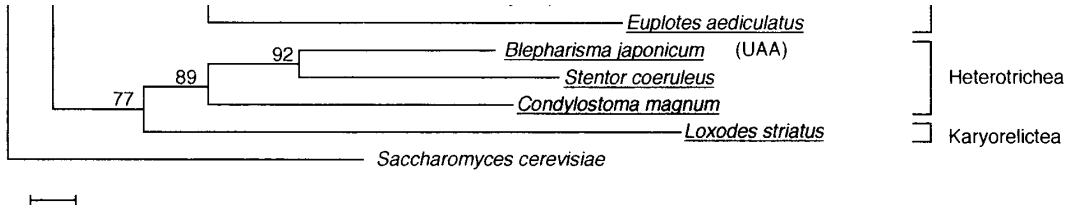


Fig. 1. Phylogenetic pattern of the genetic code in ciliates. Phylogenetic tree of the ciliates generated by the Neighbor-Joining method. The analys is restricted to the 5′ end of the 28S rRNA molecule (see Materials and methods for details). Among the 250 unambiguously aligned sites, 136 ar variable. Representatives from seven classes are included in the tree. *Saccharomyces cerevisiae* is taken as an outgroup. The 28S rRNA sequence *Oxytricha sp.* is not available. Its position (marked by a broken line) is indicated in reference to its position in the 18S rRNA tree (Greenwood *et* 1991; Leipe *et al.*, 1994). All substitutions are taken into account in the matrix data. The few gaps postulated in the selected domains that correspond to a single nucleotide were treated as a fifth base. Bootstrap values (>55%) that corroborate topological elements in the figure are indicated above the branches. Species underlined in colour are species in which the glutamine residues are encoded by universal codons and by th additional codons UAA and/or UAG. Black underlined species use, as so far determined, the universal glutamine codons. Termination codons whe known are indicated. Scale bar, 1% of nucleotide substitutions.

Heterotrichea and Karyorelictea). These data show that deviations are extensively present in the phylum. Deviation from the universal code is by far the most common for glutamine, since deviant cysteine codons have only been observed in two species of the genus *Euplotes*. Second, we have superimposed the new and previous stop codon

Current 18S and 28S ribosomal RNA phylogenies of ciliates provide a very congruent pattern of diversificati consisting of several well-resolved major branches whi correspond to most of the morphologically defined clas (Greenwood *et al.*, 1991; Baroin Tourancheau *et al.*, 19 Leipe *et al.*, 1994). The solid topological features of

Figure 1.6 Phylogeny tree of Tetrahymena and other ciliates. The tree was inferred from sequence of 28S rRNA. Bootstrap value is shown on the selected branch. The position of *Oxytricha sp.* in the tree was inferred from its position in 18S rRNA tree due to unavailability of 28S rRNA sequence. *Saccharomyces cerevisiae* was included as outgroup. (Tourancheau et al. 1995)

1.1.6 Genome sequence of *T. thermophila*

T. thermophila is the first tetrahymenas to have its genome sequenced. The MAC genome of *T. thermophila* strain SB210, which was derived from inbred strain B, has been sequenced by shotgun sequencing method to 9X coverage (Eisen et al. 2006). The assembly of bulk MAC (rDNA and mtDNA excluded) contains 1,177 scaffolds with the estimated MAC genome size of about 104 Mb (Eisen et al. 2006; Coyne et al. 2008). Its genome is AT-rich; and codon usage is biased toward AT-rich codon (Eisen et al. 2006). Gene prediction was initially performed by gene prediction algorithm TIGRscan that was trained several rounds with *P. falciparum* gene model, which is a relatively well-annotated single-celled eukaryote genome, and the initial set of putative *T. thermophila* genes (Eisen et al. 2006). Gene model was gradually improved using combined results from gene prediction algorithm, expressed sequence tag (EST), and microarray data (Coyne et al. 2008). Majority of gene model predicted by *ab initio* approach agreed with gene model predicted by EST analysis (Coyne et al. 2008). However, subsequent RNA-seq analysis reported the correction of predicted gene model including incorrect intron boundary, two genes that were predicted as one gene and vice versa, and incorrect splice site (Xiong et al. 2011). It shows that gene prediction should be performed by multiple approaches for verification. Currently gene annotation of *T. thermophila* MAC genome is being updated. The latest update is June 2014 version. There are 24,725 predicted protein-coding genes in MAC genome, which is a high number for unicellular eukaryote when compared to approximately 5,000 - 6,500 predicted genes in malaria parasite (*Plasmodium falciparum*) and yeast (*Saccharomyces cerevisae* and *S. pombe*) (Eisen et al. 2006; Coyne et al. 2008). Eisen and colleagues (Eisen et al. 2006) suggested that high number of genes in *T. thermophila* was due to retention of ancestral functions or was because other eukaryotic lineages had already lost the genes, and might be contributed by the expansion of gene families. Genomic resources for *T. thermophila* and other *Tetrahymena* (presently, *T. borealis*, *T. elliotti*, and *T. malaccensis*) can be accessed online via the community-maintained *Tetrahymena* Genome Database (TGD; http://www.ciliate.org).

MAC genome analysis does not find whole genome or segmental duplication, but extensive numbers of duplicated gene tandem was found (Eisen et al.

2006). 1,603 tandem cluster of 2-15 genes were found; 67% were simple gene pairs and 96% contains <5 genes. It appears that many of the paralogus genes in *T. thermophila* are the results of small duplication events (Eisen et al. 2006).

Table 1.1 Statistics of *T. thermophila* genome compared to other single-celled eukaryotes. For *T. thermophila* and *I. multifiliis*, the data is shown for MAC genome. Genome sequencing of *P. falciparum* was performed separately for each chromosome.

	T. thermophila	*I. multifiliis*	*P. falciparum*	*S. cerevisae*
Predicted genome size	104 Mb	49 Mb	23 Mb	12 Mb
Coverage	9X	19X	14.5X	15X,18X
Average GC content	22%	15.9%	19.4%	38.3%
Number of predicted genes	27,424	8,096	5,268	6,561

1.1.7 Transcriptome and proteome of *T. thermophila*

Gene expression profile of *T. thermophila* was initially demonstrated by expressed sequence tag (EST) analysis, which was applied for re-annotation of *T. thermophila* MAC genomic sequence (Coyne et al. 2008). Data was obtained from *T. thermophila* cultured in six different conditions (4 growth conditions, 1 starvation, and 1 conjugation). Transcriptomic of *T. thermophila* during growth, starvation, and conjugation were analyzed over time in each condition by microarray (Miao et al. 2009) and RNA-seq approach (Xiong et al. 2011). Gene expression data from both approaches were shown to have strong correlation (Xiong et al. 2011). The result indicates that majority of gene model from MAC genome annotation are indeed expressible genes. Deep sequencing technique, which has high sensitivity to low-abundance transcripts, lowers the percentage of predicted genes without detected gene expression to 3.8%. About a thousand transcribed regions that encode protein but were missing from gene model were described by RNA-seq. Genes that are specifically up-regulated during growth, starvation, and conjugation were identified, and their biological functions were annotated with gene ontology (GO) biological process information. Genes that are up-regulated during growth (1,002 genes determined by RNA-seq) are primarily involved in small molecule metabolic process, amine metabolic process, protein folding, and translation. Gene expression during conjugation is of particular interest due to DNA rearrangement that reorganizes MIC into MAC. Change, both up-regulation and down-regulation, in large number of genes can be observed during conjugation stage. Genes that are up-regulated during conjugation (1,894 genes determined by RNA-seq) are primarily involved in cell cycle, DNA replication, DNA repair, DNA recombination, and chromosome organization. Out of 213 genes that are up-regulated during starvation, only 62 genes are annotated with GO information and major biological function represented in this gene group cannot be determined. Genes that are adjacently located or are located on the same chromosome did not show clear tendency to share similar gene expression (Miao et al. 2009).

Analysis of gene network in *T. thermophila* from microarrays was reported (Xiong et al. 2011). Genes that are linked by similar expression profiles are clustered together as a module of genes by unsupervised clustering methods. Few genes are

highly connected in the network and indicate that the gene network of *T. thermophila* is scale-free. Highly-connected genes (i.e. 'hubs') were shown to have specifically high expression during growth or early conjugation, suggesting their possible pivotal role during specific stages.

Transcriptomic and gene network data of *T. thermophila* is available on *Tetrahymena* Functional Genomics Database (TetraFGD) (Xiong et al. 2013). The database provides gene expression value from microarray with two normalization methods performed separately by two laboratories. RNA-seq reads coverage plot and microarray probe location are shown for each gene model and transcript.

Availability of complete MAC genome sequence assisted the proteomic analysis by mass spectrometry in *T. thermophila*. Proteome of *T. thermophila* was analyzed for protein group of interest such as phagosome proteins, mitochondrial proteins, ciliate pellicular proteins, and phosphoprotein (Jacobs et al. 2006; Smith et al. 2007; Gould et al. 2011; Tian et al. 2014). Analysis of phagosomal proteins from purified phagosome identified 28 proteins that have functions related to phagocytosis in other organisms and 12 novel phagosome-involved candidates (Jacobs et al. 2006). Candidate phagosomal proteins were confirmed by their association with phagosome using localization of GFP-tagged candidate proteins in phagosome. However, the proteome analysis suffered from contamination during phagosome preparation. Jacobs et al reported the presence of SerH3, which is a surface immobilization antigen, in their phagosomal proteome study. The analysis of mitochondrial proteome in *T. thermophila* was the first among single-celled eukaryotes (Smith et al. 2007). About 45% of *T. thermophila* mitochondrial proteins have no homologues outside ciliate lineage, indicating that mitochondrial genome of ciliates gained additional genes during the course of evolution (Smith et al. 2007). The phosphoproteome of mixed population of *T. thermophila* from growth, starvation, and conjugation stage was determined by enrichment of phosphopeptides followed by mass spectrometry analysis, which resulted in identification of 1008 phosphoproteins (Tian et al. 2014). Identified phosphoproteins were subjected to functional annotation with GO term, providing additional information for *T. thermophila* genome annotation. Phosphoproteins may have roles in regulation, and the data gives clue for study of

signaling pathway in *T. thermophila*. Proteomic analysis of *T. thermophila* during temperature shift is not available yet.

1.2 Antigenic variation

The surface of single-celled eukaryotes is covered with variable surface proteins. Regulated switching between members of protein family is called 'antigenic variation'. Free-living and parasitic single-celled eukaryotes both exhibit antigenic variation. Parasitic protozoa use antigenic variation to escape immune response from their hosts and to interact with ligands on host tissue. *Plasmodium falciparum*, a human malaria parasite, has the *var* gene family of about 60 members encoded for *Plasmodium falciparum* Erythrocyte Membrane Protein 1 (PfEMP1). PfEMP1 is transported to infected erythrocyte membrane. Trypanosomes, the human and animal blood parasite, use antigenic variation of their VSG protein for evasion from host immunity. For parasitic protozoa, their surface antigen is regulated in a manner that only one type is expressed at a time. Expression controls in *P. falciparum* and *Trypanosoma* are epigenetics. For parasitic single-celled eukaryotes, it is clear that antigenic variation plays role in parasite's survival; varying surface antigens helps the parasite to escape the immune response. However, the role of antigenic variation in free-living organism is less obvious.

1.2.1 History of antigenic variation discovery in *Plasmodium*

Antigenic variation in *Plasmodium* was suggested by an early experiment in *P. knowlesi* (Brown et al. 1968). Sera collected from cured monkeys showed specific agglutination reaction to erythrocytes infected by parasite from the same stabilate i.e. parasite from the blood that was drawn from the same source at the same time point. Cross-reaction was not observed from sera of monkeys infected by parasites from two different stabilate groups, showing that the antigens displayed by parasite of two stabilates were of different variants. Immunized monkeys showed protective immunity against the reinfection by the parasite from the same stabilate that

was used for immunization. However, some monkeys were infected leading to chronic relapsing or even death despite being subjected to reinfection by parasite from the same stabilate. Additional experiment by agglutination test of schizont-infected erythrocyte proved that there was change in antigenic character expressed by the parasite. Sera taken at several time points from monkey with chronic malaria infection have different antibody specificities in response to variant antigen, suggesting that antigenic variation occurred in *P. knowlesi*. Stabilate that was retained by serial blood transfer from infected monkey to normal monkey also showed antigenic change. Brown and colleagues speculated that such antigenic variation helps the parasite to survive the immune response (Brown et al. 1968). Antigenic variation in *P. falciparum* was similarly demonstrated. One population of *P. falciparum* bearing new antigenic determinant can be detected from the monkeys that were infected with another antigenically distinct parasite population (Hommel et al. 1983). The reinfection with the same parasite stabilate may follow by a slight increase in parasitaemia and detection of an antigenically different parasite population. In addition, Hommel and colleagues speculated that waves of parasitaemia might come from the recurrence of previous infection after antigenic variation. Proof of 'antigenic variation' rather than 'antigenic diversity' created by the mixed population was shown by an experiment in cloned *P. falciparum* (Biggs et al. 1991). Two parasite clones, derived from the same parental line and maintained *in vitro* for six months, expressed different erythrocyte surface antigens. Antisera raised against each clones do not cross-react and cytoadherence inhibition is clonal specific. *P. falciparum* erythrocyte membrane protein (PfEMP1), known erythrocyte surface antigen and probably responsible for agglutination and cytoadherence properties in the parasite, was shown to be antigenically different between these two parasite clones by immunoprecipitation. The result hence confirmed antigenic variation in *P. falciparum* and suggested that PfEMP1 is an antigen of antigenic variation.

1.2.2 Cloning of antigenic variant genes in *P. falciparum*

The gene encoding PfEMP1, a protein responsible for *P. falciparum* antigenic variation, was successfully cloned and characterized (Baruch et al. 1995;

Smith et al. 1995; Su et al. 1995). Cloning cDNA library of *P. falciparum* was created and was screened for transcripts that appeared only in parasites that produced knob protrusion from surface of the infected erythrocytes. Cloned cDNAs from knob-protruding parasite were proven to be PfEMP1 transcript by immunoprecipitation and adherence inhibition test (Baruch et al. 1995). The genes encoding PfEMP1 were named '*var* genes' and were characterized as a gene family (Su et al. 1995). DNA segments on chromosome 7 that are polymorphic were sequenced, and several genes containing multiple domains similar to erythrocyte binding proteins (namely Duffy antigen-binding protein) from *P. vivax* and *P. knowlesi* were identified. These genes belong to a large variant gene family (hence the name '*var*'); they form clusters on chromosome with the head-to-tail arrangement. The complete sequence of *var* genes contains two exons separated by 0.8 – 0.12 kb intron (Su et al. 1995). Predicted protein sequence showed that the proteins are highly diverged and contain cysteine-rich conserved motifs with consistent structural features including putative transmembrane and acidic terminal regions. Another accompanying study showed that the switch of *var* gene expression is correlated with change in antigenic determinant of cloned parasites (Smith et al. 1995). The *var* transcripts of *P. falciparum* clones were amplified using primers that bind to Duffy Binding-Like (DBL) domain in *var* genes and were subjected to Southern blot analysis using *var*-specific DNA probe. The antigenicity was shown to be correlated to *var* gene expression. The *var* transcripts of antigenically distinct *P. falciparum* clones were sequenced and were shown to be distinct. Despite the sequence variation among *var* genes, DBLα domain (i.e. the DBL domain at the most N-terminal) of *var* gene is conserved enough to allow the design of the 'universal primers' that can amplify majority of *var* genes in laboratory isolates (45 different DBLα fragments amplified out of 60 *var* genes in *P. falciparum* 3D7) (Taylor et al. 2000).

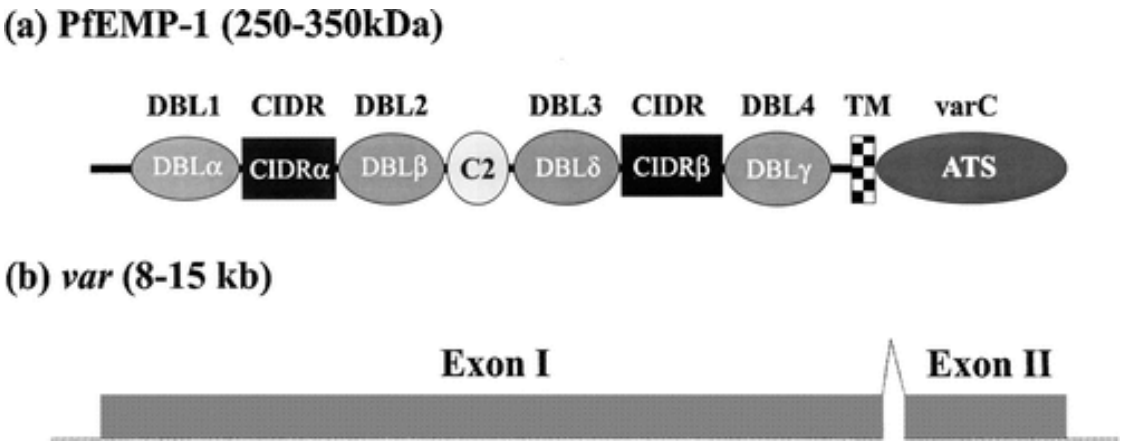


Figure 1.7 Structure of PfEMP1 protein and *var* gene in *P. falciparum*. (A) PfEMP1 contains Duffy-binding like (DBL) domain (grey oval), cysteine-rich interdomain region (CIDR) domain (black box), C2 domain (light grey oval), transmembrane (TM) domain (checked box), and acidic terminal sequence (ATS) domain. The first DBL and CIDR domain (DBLα and CIDRα) are the most conserved domain and are common in all *var* genes. The other variant of these two domains may vary in order and presence, causing sequence variation in *var* gene family. (B) *var* gene has two exons. Exon I encodes the variable portion that protrudes outside erythrocyte, and exon II encodes the ATS domain that tethers the structure within erythrocyte cytoplasm. (Kyes et al. 2001).

1.2.3 Regulation mechanism of antigenic variation in *P. falciparum*

P. falciparum transcribes multiple *var* transcripts simultaneously, but gradually only one variant of *var* transcript is detected and only one type of PfEMP1 protein is displayed on the red blood cell membrane. Early studies using reverse-transcription PCR (RT-PCR) with DBLα –specific primers (Chen et al. 1998; Scherf et al. 1998) and hybridization of amplified *var* transcripts from panned parasites (i.e. selected by defined surface adhesive phenotype by receptor-binding panning assay) using *var* probes (Scherf et al. 1998) showed that multiple *var* transcripts can be detected during the ring stage but only one transcript species was detected during the trophozoite stage. A nuclear run-on analysis using nuclei from two trophozoite lines expressing two different *var* variants showed that nascent mRNA from each parasite lines hybridized with their respective *var* probes, suggesting that *var* expression is regulated at the transcription level (Scherf et al. 1998). Subsequent studies (Noviyanti et al. 2001; Duffy et al. 2002) showed that a population of parasite, including parasites extracted from a single infected erythrocyte, expressed various *var* transcripts during the trophozoite stage as well, but the result agreed with previous studies that there is a single dominant PfEMP1 molecule displayed on the infected erythrocyte surface.

Intron of *var* genes has crucial role in *var* expression regulation. Expression vector containing *luc* reporter gene, 5' flanking region of *var* gene, and *var* intron showed expression repression after entering the S phase, suggesting the assemble of silent chromatin structure as *var* regulatory mechanism and *var* intron as regulatory element (Deitsch et al. 2001). Subsequent study (Calderwood et al. 2003) proved that *var* intron is necessary for silencing the associated *var* promoter. The silencing of reporter gene in a plasmid construct with *var* intron coincides with the detection of sterile transcript. The *var* intron has highly conserved structure with AT-rich region and possesses its own promoter activity that may explain the origin of sterile transcript. Once the promoter activity of *var* intron is removed by gene deletion, *var* gene silencing effect is abolished. Silencing effect occurs whether *var* intron is inserted in either orientation, suggesting bidirectional transcriptional activity (Calderwood et al. 2003). The bidirectional promoter activity of *var* intron was directly shown by the expression of two different luciferase reporter genes that have *var* intron placed between them and by detection of both sense and antisense mRNA

transcript from *var* intron (Epp et al. 2009). Similar plasmid construct with *var* intron replaced by other *P. falciparum* gene (*hrp*, *hsp86*, and *rRNA*) promoters showed the same S phase-dependent silencing effect, and that episomal *var* promoter without promoter activity coupling from *var* intron or other promoters (i.e. no promoter 'pairing') was not recognized by *var* expression control mechanism (Dzikowski et al. 2007). The antisense *var* intron transcript was detected from either active or inactive *var* genes of parasite in late trophozoite - schizont stage, suggesting that *var* intron transcript alone is not direct signal for *var* gene silencing (Epp et al. 2009). The study speculated that *var* intron transcript associates with chromatin, as it was found precipitated with antibody against histone H3 in higher degree than actively transcribed mRNA (Epp et al. 2009). TG-rich region, which may act as insulator-like pairing element, was found in both 5'UTR and intron of *var* genes (Avraham et al. 2012). Disruption of this pairing element causes the gene under *var* promoter and intron pair to express constitutively instead of being regulated by *var* control mechanism (Avraham et al. 2012).

1.3 Immobilization antigen of *T. thermophila*

In the early studies of ciliates including *Tetrahymena*, variation of surface antigen was utilized to determine strains, namely serotyping. The surface protein called immobilization antigen (i-ag) covers the cell surface of *T. thermophila* (Margolin et al. 1959; Williams et al. 1985). Incubation with the antibody against i-ag causes *T. thermophila* to cease its movement, hence the name. Various subtypes of *T. thermophila* i-ag were described based on an immobilzation assay with specific antibodies. Subtypes H, L, and T were found to be expressed in different culture temperatures from the observation that *T. thermophila* cultured at different temperatures showed different serotypes. (Margolin et al. 1959; Phillips 1967). Subtype H is expressed at "high" temperature (20-35ºC), while subtype L is expressed at "low" temperature (<20 ºC) (Margolin et al. 1959). Subtype T is expressed at temperature above 36ºC and was named "torrid" subtype (Phillips 1967; Smith et al. 1992). Yet culture temperature is not the sole factor for i-ag expression. When *T.*

thermophila is cultured in media with 10 – 200 mM NaCl, expression of subtype S (originally named subtype St, probably after "salt" condition) i-ag can be detected (Grass 1972). Subtype I was described from *T. thermophila* cultured in presence of diluted H-antisera (Margolin et al. 1959). Subtype M and subtype P were described from *T. thermophila* mutants, and are expressed at the same temperature range as subtype H (Doerder and Berkowitz 1987; Smith et al. 1992). Subtype J and subtype K were described from *T. thermophila* natural isolates from ponds in Allegheny National Forest (ANF), northwestern Pennsylvania; both i-ags were not detected from the inbred strains of *T. thermophila* commonly used in laboratory experiments (Saad and Paul Doerder 1995). Characterization of i-ag protein was performed by purification of antigen using column chromatography, followed by SDS electrophoresis and immunoblotting with subtype-specific antisera that were raised by injecting whole *T. thermophila* cell or soluble extract into rabbit (Smith et al. 1992). These i-ag proteins were characterized on protein gel and shown to be distinct subtype (Smith et al. 1992; Saad and Paul Doerder 1995). Based on the result from protein gel electrophoresis, *T. thermophila* i-ags have molecular weight of 25-60 kDa (Doerder and Berkowitz 1987; Smith et al. 1992). Presence of i-ags with varying molecular weight suggested that each i-ag subtype is a family of proteins (Smith et al. 1992; Doerder and Berkowitz 1987).

Following characterization of i-ag proteins, gene sequences of three subtypes of i-ag were identified. Sublinages of *T. thermophila* inbred strains (A1, C2, C3, B2, B3, D1) were selected for homozygosity at H loci (Nanney and Dubert 1960; Nanney and Simon 2000); thus the gene encoding for subtype H i-ag was the first to be identified. The gene encoding for i-ag was named *Ser* after the word "serotype" (Doerder et al. 1985). Up to now, there are six subtype H genes (*Ser*H1, *Ser*H3, temperature-sensitive *Ser*H3, *Ser*H4, *Ser*H5 and *Ser*H6) (Tondravi et al. 1990; Deak and Doerder 1995; Gerber et al. 2002), one subtype J genes (*Ser*J) (Doerder 2000), and six subtype L paralogs (*Ser*LA, *Ser*LB, *Ser*LC, *Ser*LD, *Ser*LE1, *Ser*LE2) (Doerder and Gerber 2000) identified. *Ser*H3 is the first *Ser* gene to be molecularly characterized by sequencing of *Ser*H3-homozygous, inbred strain B's MAC genomic library that had shown to hybridize with *Ser*H3 probe (Tondravi et al. 1990). *Ser*H1 was sequenced a few years later (Deak and Doerder 1995), followed by *Ser*H4, *Ser*H5, and *Ser*H6

(Gerber et al. 2002). Despite reports of H2 i-ag (Margolin et al. 1959; Saad and Paul Doerder 1995), its gene has never been successfully characterized. Partial *Ser*J sequence was sequenced using *Ser*H primer before the full-length *Ser*J sequence can be obtained (Doerder 2000). No additional *Ser*J genes have been reported in literatures despite the observation from southern blot experiment that *Ser*J has paralogs in *T. thermophila* inbred strain B (Doerder 2000). Characterization of *Ser*L genes, which express in low temperature condition, was aided by the availability of *T. thermophila* mutants that constitutively express subtype L i-ag due to an inability to express subtype H i-ag (Doerder and Gerber 2000). Sequencing of *Ser*L was also initiated with partial sequencing using *Ser*H primer (Doerder and Gerber 2000). The study of *Ser*L genes reported the presence of pseudogene (Genbank accession: AF312774) (Doerder and Gerber 2000).

Sequence data of the thirteen *Ser* genes (see Table 1.3) show that these intron-free genes are, like *T. thermophila* MAC genome, AT-rich and biased toward AT-rich codon (Doerder and Gerber 2000; Gerber et al. 2002; Doerder 2000). The translated polypeptides are about 140-450 amino acids in length and are rich in alanine, cysteine, serine, and threonine. *Ser* sequence can be divided roughly into 3 regions: N-terminus region of about 100 amino acids, region of tandem repeats, and C-terminus region of about 30 amino acids. N-terminus region contains hydrophobic motif that is predicted to be ER translocation signal (Gerber et al. 2002). C-terminus region is similar to GPI signal sequence, and GPI attachment site is predicted at this region (Gerber et al. 2002). The region of tandem repeats in each Ser subtypes differ by the number of tandem repeat and number of cysteine residues in each repeat. Further details about Ser tandem repeat are discussed below.

Despite the characterization of various i-ag in *T. thermophila*, only a portion of their genes was identified. The successful gene identification was based on sequence similarity that allowed the gene of three different subtypes to be amplified. However, this approach was obviously inadequate for identification of *Ser* genes. *Ser*H2 gene was not successfully characterized. Only one *Ser*J gene was characterized despite the indication of its paralogs. Various i-ag subtypes such as SerT, SerI, SerM, SerP, and SerK had no reports of gene characterization. The polymorphic nature of *Ser* gene, both within and among subtypes, renders the identification process by sequence

similarity inadequate. In order to gain the full grasp of antigenic variation in *T. thermophila*, another identification approach is required.

Table 1.2 Summary of subtypes and putative genes for *T. thermophila* i-ag

Subtype	Expression condition	Putative gene
H	Cultured at temperature 20-35ºC	SerH
L	Cultured at temperature <20ºC	SerL
T	Cultured at temperature >36ºC	n/a
I	Cultured with anti-H sera	n/a
S	Cultured with 10 – 200 mM NaCl	n/a
J	n/a	SerJ
K	n/a	n/a
M	Cultured at temperature 20-35ºC	n/a
P	Cultured at temperature 20-35ºC	n/a

Table 1.3 Thirteen known Ser. The NCBI accession number and *T. thermophila* strain that each subtype was characterized were given. All genes do not contrain intron. Number of tandem repeats in this table is shown as reported in the reference.

Subtype	Accession	Strain	Protein length (aa)	No. of tandem repeats	Reference
SerH1	AAA91970	A	436	3.5	Deak and Doerder 1995
SerH3	AAF06326	B	423	3.5	Tondravi 1990
SerH4	AAL23952	B3	397	3.5	Gerber 2002
SerH5	AAL23953	ANF5906, ANF6707	395	3.5	Gerber 2002
SerH6	AAL23954	ANF18021	421	3.5	Gerber 2002
SerJ	AAF44706	ANF814-1	438	4	Doerder 2000
SerLA	AAG38116	B	316	5	Doerder and Gerber 2000
SerLB	AAG38117	B	316	5	Doerder and Gerber 2000
SerLC	AAG38118	B	305	5	Doerder and Gerber 2000
SerLD	AAG38107	ANF18211	371	6	Doerder and Gerber 2000
SerLE1	AAG38119	B	148	2	Doerder and Gerber 2000
SerLE2	AAG38108	ANF18027-4, BJ1-1	151	2	Doerder and Gerber 2000

Gradual i-ag subtype switching was observed when temperature was shifted (Margolin et al. 1959; Phillips 1967) and *T. thermophila* can retain its original i-ag subtype for a while (Juergensmeyer 1969). One study of *T. thermophila* in natural habitat shows that there is seasonal variation in expression of i-ag; the frequency of expression of subtype H and J is higher in warmer months, while expression of subtype L is more frequent during colder months (Saad and Paul Doerder 1995). There may be no active molecular mechanism to remove old i-ag from the cell surface as nongrowing *T. thermophila* was shown to retain its i-ag as long as 48 hours (Williams et al. 1985).

The mechanism behind *Ser* expression control is not well understood. It was found that half-life of *Ser*H3 (normally expressed at 20-35ºC) mRNA was rapidly decreased when temperature was shifted to 40ºC (Hallberg et al. 1984; Love et al. 1988). Similar result was obtained from a study of *Ser*H1 (Deak and Doerder 1995) and *Ser*H4 (Gerber et al. 2002); however, *Ser*H5 transcribes and expresses as H5 i-ag even at 40ºC (Gerber et al. 2002). An *in vitro* mRNA assay suggested that stability of *Ser* transcript is affected rather by cellular component than culture temperature, as transcripts of SerH3 grown in 30ºC and 40ºC have the same half-life when incubated in the cell extract from the same source (McMillan et al. 1995). Treatment of protein synthesis inhibitor and protein kinase inhibitor can prolong *Ser*H3 mRNA half-life during temperature shift from 30ºC to 40ºC, suggesting that rapid *Ser* mRNA destabilization during temperature shift is not spontaneous but instead requires protein synthesis (McMillan et al. 1995). These observations may explain temperature-dependent expression in a certain *Ser* subtypes. Expression of *Ser*L was suggested to be regulated by both transcriptional control (no *Ser*L transcript detected in cells expressing H and T i-ag) and by mRNA stability (*Ser*L transcript detected in cells expressing S and K i-ag) (Doerder and Gerber 2000).

The role of *T. thermophila* i-ag remains unclear, though it may involve sensing the environment or prey-predator recognition like *Paramecium* surface antigen (Simon and Schmidt 2007).

1.3.1 Immobilization antigen in other ciliates

Similar i-ag proteins were described in other ciliates as well. *Paramecium* i-ag shows high resemblance to i-ag of *T. thermophila* and was often mentioned in comparison when discussing *T. thermophila* i-ag. I-ags of *Paramecium* have a high molecular weight of 250-300kDa (Hansma and Kung 1975; Forney et al. 1983), and are encoded by genes of 6.7-8.5 kb open reading frame, which are about five times larger than those of *T. thermophila*. The periodic cysteines and repeats were observed in the sequence of Paramecium i-ag (Prat et al. 1986).

Ichthyophthirius multifiliis, a fish parasite, has similar i-ag to *T. thermophila* repetitive periodic Cys feature and GPI anchor. Five serotypes of *I. multifiliis* were described, namely serotype A, B, C, D, and E. Their protein sequences are 40-57% identical to each other. One *I. multifiliss* i-ag (designated IAG52A) has high sequence similarity to SerL as resulted from BLAST search on SwissProt+Trembl database (Lin et al. 2002). A recent analysis of *I. multifiliis* MAC genome identifies 17 candidate i-ag genes and 4 pseudogenes (Coyne et al. 2011). These candidate i-ag genes encode 303-340 amino acid residue proteins. Most of them were found to be in tandem array alignment. *I. multifiliis* genome has fewer number of protein-coding genes compared to free-living *Tetrahymena* and *Paramecium*, which Coyne and colleagues suggested that parasitic lifestyle causes *I. multifiliis* to loss its ciliate-specific gene families (Coyne et al. 2011).

Table 1.4 Comparison between immobilization antigen (i-ag) of *T. thermophila* and *Ichthyophthirius multifiliis,* and surface antigen (SA) of *Paramecium* sp. (Doerder and Gerber 2000).

Organism	Protein length (aa)	No. of tandem repeat	No. of Cys residue/repeat	Intron	GPI anchor signal
T. thermophila H1 i-ag	436	3.5	8	No	Yes
T. thermophila J i-ag	438	4	10	No	Yes
T. thermophila LA i-ag	316	5	6	No	Yes
Paramecium SA	2233-2717	31-38	8	No	Yes
I. multifiliis i-ag	455-468	4	6	No	Yes

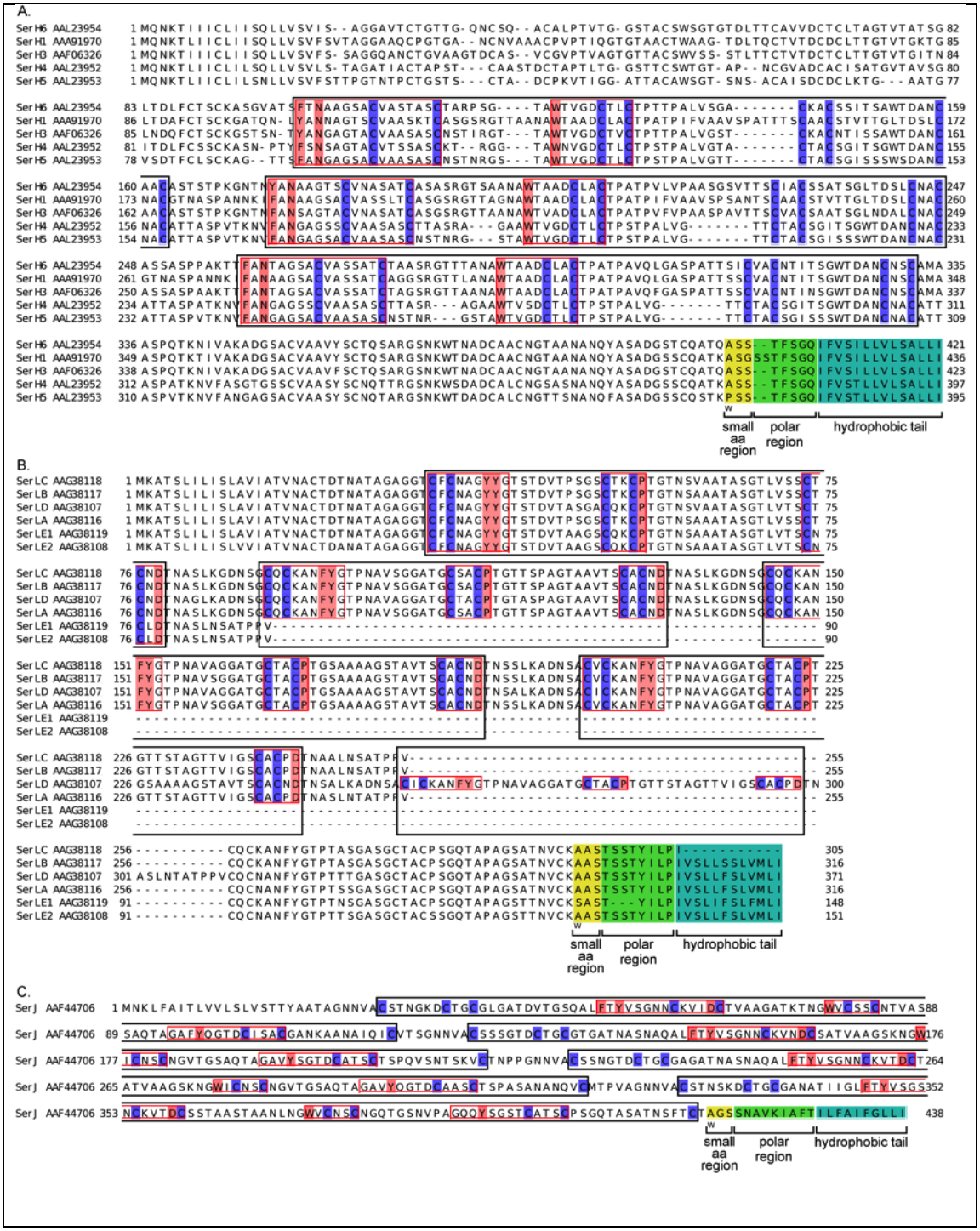
1.3.2 Cysteine-rich motif in *T. thermophila* i-ag

One common characteristic of these Ser proteins is the tandem cysteine-rich repeat (Doerder and Gerber 2000; Gerber et al. 2002). Tandem repeat was reported in the previous studies of *Ser* gene identifications. These studies described the pattern of repetitive sequence, number of repeat, and number of cysteine residues per repeat of *T. thermophila* i-ag subtype H, J, and L (Doerder and Gerber 2000). According to these studies, SerL, SerH and SerJ contain 6, 8, and 10 cysteine residues per repeat respectively (Doerder and Gerber 2000; Gerber et al. 2002; Doerder 2000). The number of amino acid residues between two cysteine residues shows consistent trend of alternating between short (about 2-4 residues) and long (up to 21 residues) stretches. The length of each repeat is varied approximately from 55 to 100 amino acid residues (see Table 1.5) (Doerder 2000; Doerder and Gerber 2000; Gerber et al. 2002). This information is critical for generalized repeat pattern that is required in this study. Such features are also common in surface protein of other unicellular eukaryotes (Kusch and Schmidt 2001).

Table 1.5 Repeat block reported in previous studies of Ser.

Subtype	Repeat block	Reference
H1, H3	$CX_6CX_{17}CX_2CX_{18}CX_2CX_{11}CX_2CX_{19}$	Deak and Doerder 1995
J	$CX_6CX_2CX_{21}CX_4CX_{13\text{-}15}CX_2CX_{18}CX_3CX_{11}CX_{9\text{-}10}$	Doerder 2000
LA	$CX_{12\text{-}15}CXCX_{15\text{-}17}CX_2CX_{14\text{-}18}CX$	Doerder and Gerber 2000

Figure 1.8 Sequence feature of *T. thermophila* known i-ag subtype-H (A), subtype-L (B), and subtype-J (C). Selected known *Ser* protein sequences are aligned to show their cysteine pair, repetitive block and GPI anchor site. Cys residues are highlighted (dark blue). Each repetitive block is indicated by black box. Red box indicates sequence feature that appears in each repetitive block. The region predicted as GPI anchor signal by FragAnchor is color-shaded. Predicted GPI attachment site is marked by letter ‘‘w’’. According to data from known *Ser*, number of Cys residues per repetitive block is unique for each subtype (SerL: 6 Cys per block; SerH: 8 Cys per block). The length of each repetitive block differs among various subtypes and is also varied between 55–100 aa. GPI anchor signal predicted by FragAnchor exhibits region of small amino acids (Ala, Gly, Ser) where GPI is attached (yellow), followed by polar region (green) and hydrophobic tail (light blue).


A.
small aa region
polar region
hydrophobic tail
B.
small aa region
polar region
hydrophobic tail
C.
small aa region
polar region
hydrophobic tail

Even though sequence variation is a hallmark of these highly diverged surface proteins, they often contain repetitive cysteine-rich motifs. The periodic cysteine residues form disulfide bonds in a consistent pattern among proteins in the same family. The formation of the disulfide bond might introduce an extremely hydrophobic moiety into a protein (Chookajorn et al. 2004). It was suggested that 'hydrophobic collapse' might play a crucial role in protein folding via hydrophobic core nucleus that drives the folding process (Chang 2011). Disulfide bond formation could allow protein to become highly diverged while maintaining the overall fold. An example can be found in the SCR gene family in sporophytic self-incompatible plants that contains periodic cysteine repeats in highly diverged protein sequences. Despite their high degree of diversity, SCR proteins are malleable to the same overall fold, which makes it flexible to the change in order to avoid inbreeding (Chookajorn et al. 2004). *T. thermophila* i-ag also has alternating short and long stretches of amino acid sequences linked by cysteine pairs. Different number of cysteine pair per one repeating sequence block in different i-ag subtype was documented (Doerder and Gerber 2000). Despite high sequence variation among *Ser* genes, the pattern of cysteine rich motif is a common feature among them, which might suggest a similar role like in the case of SCR.

This study took advantage of cysteine-rich motif in i-ag and availability of *T. thermophila* MAC genome sequence. Cysteine-rich motif cannot be identified by sequence homology search due to high sequence variation. Amino acid residues between cysteine pair vary in term of residue species and length (Figure 1.7), rendering the homology search to yield poor result. Therefore, the strategy to identify *T. thermophila* i-ag is not the direct search for similar sequence but instead the detection of the periodic cysteine pairs regardless of the variation of residues in-between.

1.4 GPI anchor

1.4.1 General information about GPI anchor

Glycosylphosphatidylinositol (GPI) anchor provides a mean to display

proteins on the cell surface besides transmembrane domain and hydrophobic moiety such as lipid. GPI anchor is assembled from phosphatidylinositol, which is embedded in cell membrane, and chains of glycan that attach to the protein. GPI anchor exists in various eukaryotic organisms, plants, fungi, all major vertebrate cell types and tissues, and particularly abundant in protists (Ferguson 2009; Paulick and Bertozzi 2008). However, there are no reports of GPI anchor in prokaryotic organisms yet. Beside the function of GPI anchor as the surface protein display, it is not known whether GPI anchor has any other biological role. GPI anchor itself may affect the conformation and structure of the attached protein. Putative roles of GPI anchor include lipid raft partitioning, signal transduction, cellular communication, apical membrane targeting, and prion disease pathogenesis (Paulick and Bertozzi 2008). GPI-anchored proteins are assumed to be associated with lipid raft due to similar insolubility profile, but the evidence from fluorescence resonance energy transfer (FRET) microscopy is ambiguous (Paulick and Bertozzi 2008). Another speculation is that GPI-anchored proteins are not restricted by cytoskeletal structure; therefore they can move freely in cell membrane and to interact with ligands. However, there are no strong supports for such speculation that GPI-anchored proteins have advantage in mobility (Ferguson 2009) . The loose association with cell membrane may allow GPI-anchored protein to spontaneously transfer and incorporate into exogenous cell membrane without altering its characteristic and function. Trypanosomal VSG was found to be able to integrate into human red cell membrane and was suspected to be associated with pathogenic condition of trypanosomiasis (Rifkin and Landsberger 1990). The transfer mechanism is not yet known.

The existence of GPI anchor was speculated from the release of alkaline phosphase (APase) enzyme from the cell surface by the treatment with purified phosphatidylinositol phospholipase C (PI-PLC), which reacts on phosphatidylinositol, from *Bacillus cereus*. Manipulation of ionic environment cannot achieve the same result, suggesting that the APase is covalently attached to the cell surface via membrane-embedded component (Low 1987). Enzymes with similar characteristic were also purified from other bacteria such as *Staphylococcus aureus* and *Clostridium novyi* (Paulick and Bertozzi 2008). These results led to the conclusion that APase is anchored to the membrane by phosphatidylinositol. Chemical structure of GPI was

elucidated from a study of *T. brucei* VSG via combination of nuclear magnetic resonance (NMR) spectroscopy, mass spectrometry, chemical modification, and exoglycosidase digestion (Ferguson et al. 1988).

GPI anchors, though having structural diversity depending on organism and cell types, share a common core structure. The core structure of GPI consists of phosphoethanoalamine, chain of mannoses, nonacetylated glucosamine, and phosphatidylinositol. Phosphoethanoalamine links to the C-terminus of displayed protein by the amide linkage through its amino group. Phosphate group at the other terminus of phosphoethanoalamine links to the carbon 6 of the first mannosyl residue by a phosphodiester link. Three mannosyl residues are linked via α1-2, α1-6, and α1-4 glycosidic bond, respectively. The third mannosyl residue is linked to a nonacetylated glucosamine. Phosphatidylinositol is glycosidically attaced to glucosamine at the reducing end via its inositol ring. Lastly, the phosphatidylinositol anchors the entire structure to the membrane via its hydrophobic part consisting of fatty acid (Ferguson 2009). GPI anchor may be varied in modification at the side chains of core mannosyl residues by various glycan, and at phosphatidylinositol by variation of length and saturation of fatty acid chain (Ferguson 2009). Inositol ring may be modified by palmitate. In higher eukaryotes, core mannosyl residues may be modified by additional residue of phosphoethanoalamine. Core mannosyl residues may also have modification of glycan substitutions e.g. galactose, mannose and acetylgalactosamine (Ferguson 2009). For example, at the core mannose residues, *Trypanosma* VSG has α-linked galactosyl modification, and mammalian cell surface antigen Thy-1 has β-linked N-acetylglucosamine modification. Complex side chain such as N-acetylgalactosamine-containing polysaccharide,can also be found in mammalian cells and protists (Paulick and Bertozzi 2008). The variation of fatty acid chain attached to the phosphatidylinositol includes diacylglycerol, akyl-acylglycerol, stearoyl-*lyso*glycerol, and ceramide. The length of fatty acid chain is also varied between 14 – 28 carbons. Fatty acid component of GPI anchor can be either saturated or unsaturated. The significance of such variation in GPI anchor modification is unknown (Ferguson 2009). GPI-anchored protein can be cleaved off from the anchor by treatment with phosphatidylinositol-specific phospholipase C (PI-PLC). However, treatment with PI-PLC is not a clear-cut method to determine GPI-anchored protein

because modification at GPI anchor may result in resistance from PI-PLC treatment. For example, GPI anchor with palmitate modification at its inositol ring is resistant to PI-PLC treatment (Roberts et al. 1988).

Both the biosynthesis and attachment of the GPI anchor moiety occur in endoplasmic reticulum (ER) (Ferguson 2009). The rapid attachment of GPI moiety to the protein was shown from a study in the variant surface glycoprotein (VSG) of the parasitic ciliate *Trypanosoma brucei* (Bangs et al. 1985). GPI precursor is assembled on the cytoplasmic side of ER membrane and is attached to the C-terminus of the target protein. GPI-anchored protein contains the N-terminal signal sequence that targets the protein to ER, and GPI anchor signal sequence at the C-terminus that can be recognized by membrane-embedded transamidase in its propeptide form. Both N- and C-terminal signal sequences are eventually removed and no longer appear in the final product. The GPI anchor biosynthesis was elucidated by the studies using cell-free systems in *Trypanosoma*, *Toxoplasma*, yeast, and mammalian cell. The studies show that the basic GPI anchor biosynthesis is generally conserved with some variations. First, phosphatidylinositol is attached with acetylglucosamine, which is rapidly deacetylated. Then, mannosyl residues and phostatidylethanolamine are added sequentially to the deacetylated glucosamine.

Another distinct feature among *Ser* genes is the consensus sequence at the C-terminus specific for Glycosylphosphatidylinositol (GPI) anchor modification. *T. thermophila* i-ag subtype H was shown to be GPI-anchored protein by radiolabelling at GPI anchor components (Ko and Thompson 1992; Ron et al. 1992). Putative GPI anchor site was predicted to be located at the C-terminus of the Ser proteins (Doerder 2000; Doerder and Gerber 2000; Gerber et al. 2002). Variable surface antigen in free-living cililate such as *Paramecium primaurelia* was also shown to be GPI-anchored (Capdeville 2000). This study exploited this information in combination with periodic cysteine pairs to develop a *T. thermophila Ser* detection strategy based on translated protein sequences.

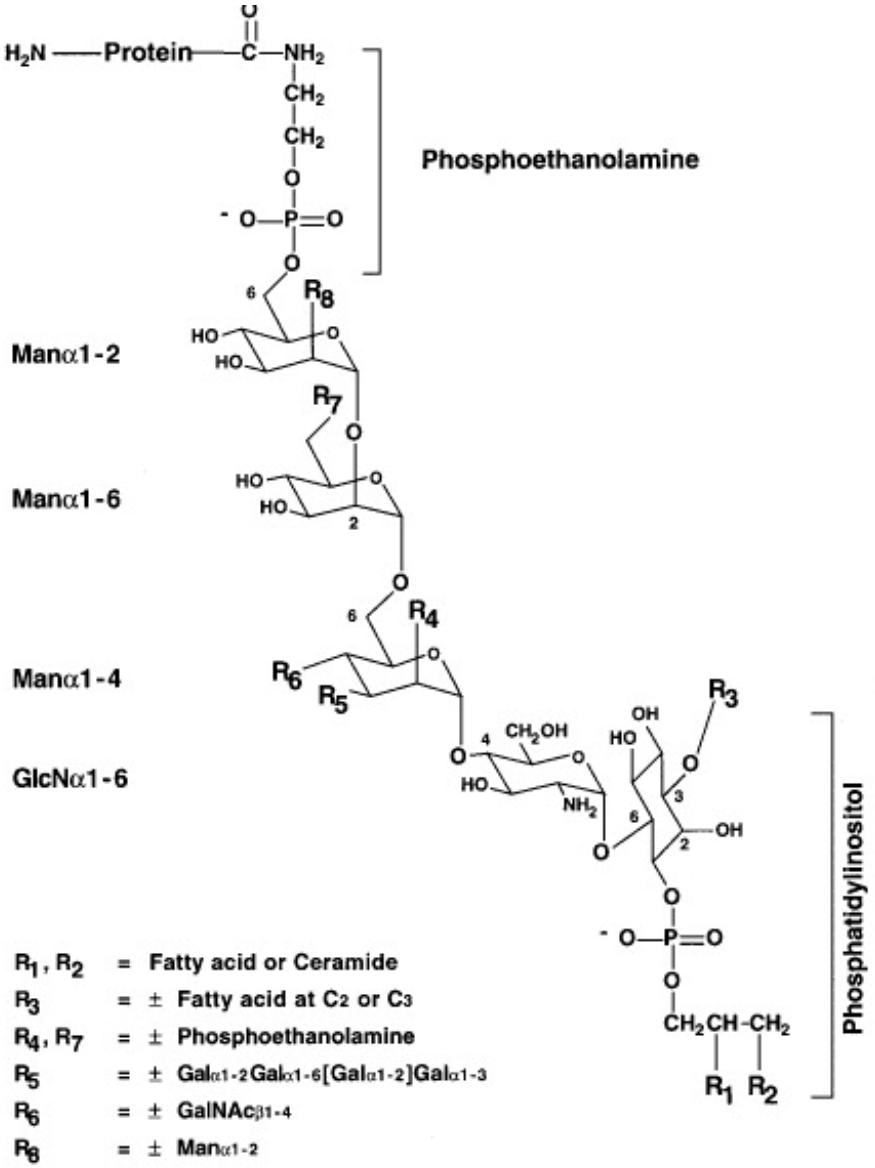


Figure 1.9 Chemical structure of GPI anchor. GPI anchor consists of phospho -ethanolamine, chain of mannose, glucosamine, and phosphatidylinositol. Phospho -ethanolamine is linked to the C-terminus of the attached protein by phosphodiester bond. Phosphatidylinositol is the part that anchor the whole structure to the membrane. Glycosidic bond linking each mannose residue is indicated. GPI anchor may be modified by at various position shown here as R group. (Ferguson 2009)

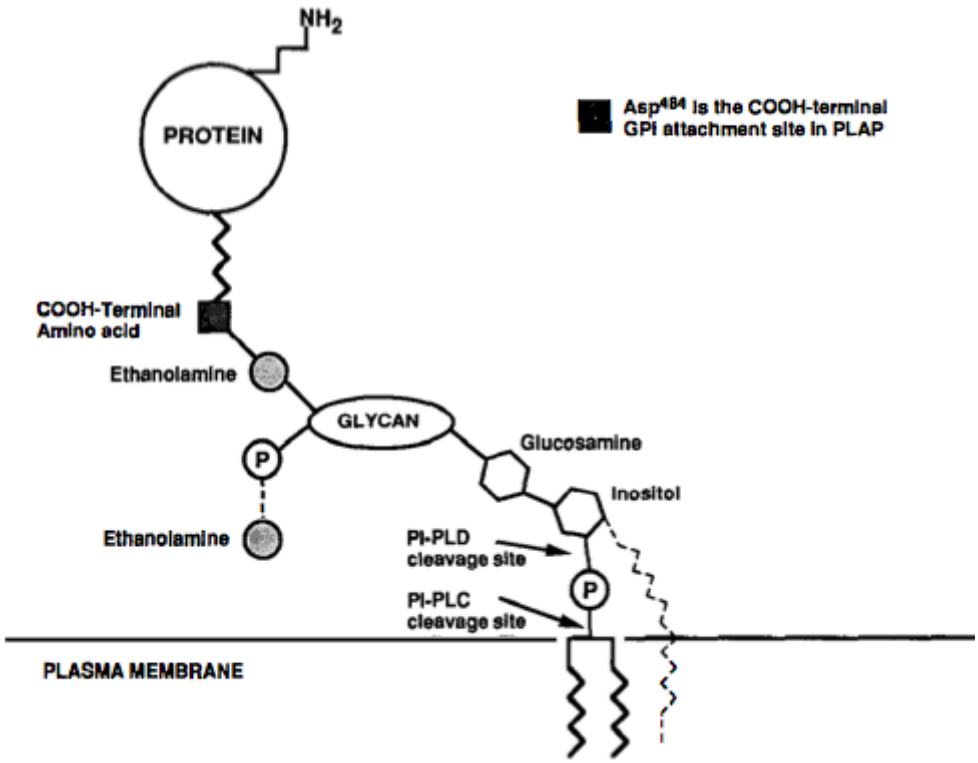


Figure 1.10 Attachment of GPI anchor on membrane and protein. Protein is attached to the GPI anchor via its C-terminus. GPI anchor attaches to the membrane by its phosphatidylinositol (shown here as wavy line). The site where phosphatidylinositol-specific phospholipase C (PI-PLC) and D (PI-PLD) cleaves GPI anchor is indicated. GPI-anchored protein carries GPI anchor signal at its C-terminus which is cleaved off during GPI anchoring process. Placental alkaline phosphatase (PLAP), one of well-studied GPI-anchored protein, is attached to the GPI anchor at Asp484 residue as indicated here by black box. (Udenfriend and Kodukula 1995).

1.4.2 Prediction of GPI-anchored protein

Presence of GPI signal sequence at C-terminus indicates that the protein is GPI-anchored. Artificial addition of GPI signal sequence to the C-terminus of a protein causes it to become GPI-anchored (Ferguson 2009). However, GPI signal has no detectable conserved sequence and hence cannot be identified by sequence similarity approach. Characteristics of GPI signal such as optimal length and hydrophobicity were determined by generating a mutant protein (Udenfriend and Kodukula 1995). The signal sequences surrounding the GPI attachment site can be defined as the regions of amino acid residues with different physical properties such as size and hydrophobicity (Eisenhaber et al. 1998). GPI anchor signal can be divided into three regions—GPI attachment site (ω site), spacer region of polar residues (ω+3 to ω+8), and hydrophobic region (ω+9 to C-terminus) (Eisenhaber et al. 1998). The ω site was experimentally determined in relatively few GPI-anchored proteins (Micanovic et al. 1990; Gerber et al. 1992; Kodukula et al. 1993; Moran et al. 1991), but these data provide information for prediction of GPI anchor signal. The data from site-directed mutation experiment of placental APase, a model GPI-anchored protein, was used to generate a table of allowable amino acid substitution around GPI attachment site (ω, ω+1 and ω+2 site) (Kodukula et al. 1993). By the amino acid preference value given in the table, the probability of given sequence being ω site can be calculated (Kodukula et al. 1993). A refined model based on the physical properties of the GPI signal sequence of experimentally verified GPI-anchored protein data from Swiss-Prot database has further improved prediction power (Eisenhaber et al. 1998).

Such physical properties of signal sequence appear to be required for recognition by GPI transamidase enzyme (Eisenhaber et al. 1998). The attachment site and following spacer region are composed of small, polar amino acid residues that may fit into the active site of GPI transamidase enzyme. The hydrophobic region of GPI signal is the region that spans through ER membrane. There may be the minimum length for hydrophobic region require for GPI attachment, as altering length of this region affects the attachment of GPI moiety (Udenfriend and Kodukula 1995) These features allow the prediction of GPI-anchored protein based on set of rules (Eisenhaber et al. 1999). Several GPI anchor protein prediction programs are freely available e.g. GPI-SOM (Fankhauser and Maser 2005), big-PI Predictor (Eisenhaber et

al. 1999) and FragAnchor (Poisson et al. 2007). These programs exploited the availability of protein database to create a verified GPI-anchored protein training set, but used different algorithm to calculate probability of given input sequence being GPI-anchored protein. Big-PI uses scoring function based on position-weighted sequence alignment and physical properties of amino acid sequence around ω site. GPI-SOM employs self-organizing map (SOM), which is a tool for unsupervised data classification to distinguish between GPI-anchored and non GPI-anchored proteins. FragAnchor employs a tandem system of neural network (NN) and Hidden Marcov Model (HMM) to calculate likelihood and classify the predicted GPI-anchored proteins into score-based classes, giving an additional flexibility in prediction. Big-PI and GPI-SOM integrate ER localization signal prediction into their GPI-anchored protein prediction algorithm, but it is not implemented in FragAnchor.

In this study, GPI anchor signal is one of the two criteria for identifying *Ser* candidates. The performance of GPI-anchored protein prediction program would greatly impact the quality of *Ser* identification algorithm developed here. Besides precision of prediction, flexibility is also crucial. As GPI anchor prediction programs are developed using a set of verified GPI-anchored proteins as training set, programs that do not allow flexible prediction for deviated GPI anchor signal would yield poor result. Some GPI-anchored protein predictors may also require ER signal sequence at N-terminus as input, which may affect result in the case of fragmented or partial input sequences. In addition, the scope of work that took translated protein sequences from whole *T. thermophila* genome (24,725 sequences) requires GPI anchor prediction program to handle large data volume.

Among all available GPI-anchored protein prediction programs, FragAnchor was selected in this study due to its advantages that is suitable for the study. One advantage is its accepted input format: FragAnchor can take multiple protein sequences as a plain Fasta-formatted file as input. Another advantage of FragAnchor is its flexible degree of precision. Rather than return a list of predicted GPI-anchored proteins, FragAnchor classifies the submitted proteins into four categories, ranking from highest to lowest probability of being GPI-anchored protein. This allows users to select the degree of precision that fits their needs. By the validation test using 593 GPI-anchored protein sequence available from Swiss-Prot

database, NN component of FragAnchor can predict 91% of manually curated GPI-anchored proteins (Poisson et al. 2007). HMM component was also tested and yielded similar performance (Poisson et al. 2007). The capability to predict GPI-anchored protein based on the signal at C-terminus alone is also advantageous that the sequence with complete GPI-anchor signal but lacks its N-terminal is not excluded from the analysis. Finally, the comparative prediction efficiency test using positive set of GPI-anchored proteins showed that FragAnchor has higher sensitivity than Big-PI and GPI-SOM (Poisson et al. 2007).

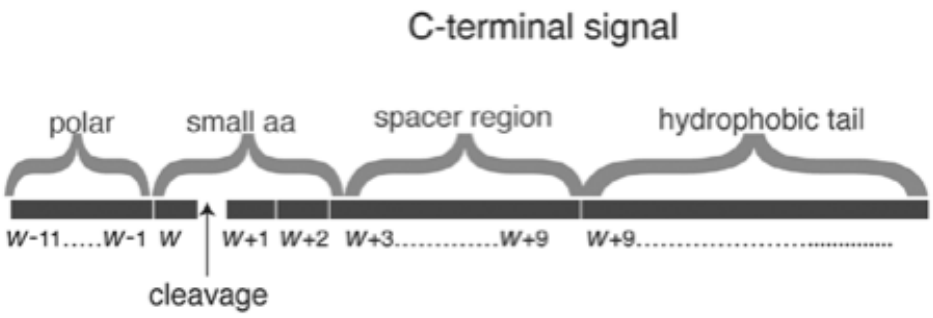


Figure 1.11 General composition of GPI anchor signal. GPI anchor signal locates at the C-terminus of the GPI-anchored protein and is cleaved off during the GPI attachment process. The signal sequence is not conserved, but has regions of definable physical characteristic. Around the GPI attachment (ω) are small amino acid residues. This region follows by spacer region that contains polar residues, and hydrophobic tail that contains hydrophobic residues. The region preceding cleavage site contains polar residues. (Poisson et al. 2007).

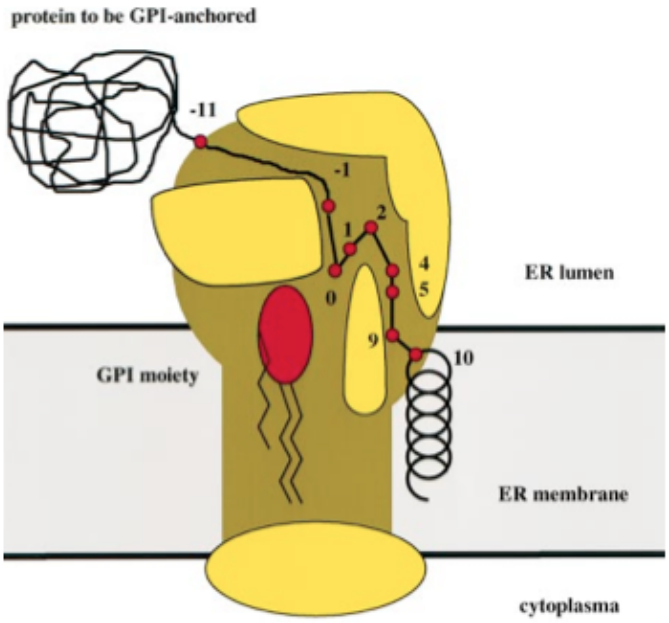


Figure 1.12 Scheme of putative GPI transamidase. The restriction of physical characteristic of GPI signal may be due to interaction of the signal with the transamidase enzyme. The GPI attachment site (ω) is denoted by number 0. Region of small and polar amino acid residues in GPI signal may interact with enzyme cavity. Hydrophobic tail of GPI signal may span through ER membrane. The region after GPI attachment site is cleaved off. (Eisenhaber et al. 1998).

1.5 Summary of *Ser* candidate selection strategy

The objective of this study is to determine *Ser* candidates from *T. thermophila* MAC genome using two selected sequence features as criteria. The *Ser* candidate selection algorithm was developed as two consecutive steps: the first step selects *T. thermophila* proteins that contains periodic cysteine motif, the second step selects proteins that contains GPI anchor signal. On the first step, custom Perl script was developed to search each *T. thermophila* translated sequences for the defined *Ser* cysteine motif. Cysteine motif was determined from thirteen known i-ag sequences and was written as a general search pattern (namely, regular expression). Perl, a computer scripting language, was selected as a tool in this study due to being suitable for manipulating text-based data such as protein sequence, and availability of biological data-oriented module (namely BioPerl (Stajich et al. 2002)) that facilitate the work with large number of amino acid sequences. A list of *T. thermophila* proteins with matching cysteine pattern was created, and was then subjected to the GPI anchor signal prediction by FragAnchor on the second step. *T. thermophila* proteins that contains matching cysteine pattern and receives high probability score as GPI-anchored protein were finally selected as *Ser* candidates. *Ser* candidates were then subjected to further analyses to determine subtypes, chromosomal location, and associated expression pattern. To determine subtype of Ser candidates, phylogenetic analysis using two different tree estimation methods was employed with known Ser data as reference. Gene expressions of *Ser* candidates were analyzed using available data of gene expression from microarray.

Ser candidates identified from this study will serve as a starting point for a detailed study of *T. thermophila* antigenic variation. Further experiment that shows the binding with specific antibodies will be required to prove that the candidates are indeed Ser. Additional identification of Ser candidates on other *T. thermophila* isolates may allow a comparative study of Ser that will provide clue whether these genes are under positive selection. The developed Ser candidate identification algorithm may be applied for identification of i-ag in other ciliates that exhibit similar sequence features (periodic cysteine pairs and GPI anchor signal) as *T. thermophila*. The algorithm requires only genome data of organism of interest and may be further customized to work in pipeline with genome annotation process.

CHAPTER II
MATERIALS AND METHODS

2.1 Sequence and genome data

The whole predicted protein sequences of *T. thermophila* version October 2008 were downloaded from Tetrahymena Genome Database (TGD) accessible via http://ciliate.org/index.php/home/downloads. MAC genome of *T. thermophila* strain SB210 was sequenced to 9X coverage by whole genome shotgun sequencing method (Eisen et al. 2006). Sequencing reads were divided into separate bin of mitochondrial DNA (mtDNA), ribosomal DNA (rDNA), and MAC DNA, and assembly was performed separately for each bin. Low coverage of MIC-specific sequence in the assembled reads showed that MIC contamination was low. MAC DNA bin contained 1,971 scaffolds after the sequencing reads assembly. The number of scaffolds was reduced to 1,177 after an additional comparative analysis between MAC and MIC genome (Coyne et al. 2008). The size of MAC genome is about 104 Mb. Forty-four percent of sequenced genome (accounted for 45.9 Mb from 125 scaffolds) are telomere-capped at both ends, and 31% (accounted for 31.8 Mb from 120 scaffolds) are telomere-capped at one end. Sequencing gaps were closed by primer walking (Coyne et al. 2008). Finally, 60% of *T. thermophila* genome was finished with assembled chromosomes that are telomere-capped at both ends.

T. thermophila MAC genome contains 24,725 protein-coding genes as predicted by *ab initio* approach (Eisen et al. 2006). In brief, gene prediction algorithm, namely TIGRscan, underwent two rounds of training. In the first round of training, a set of *P. falciparum* cDNA sequences was used as training set and gene prediction was performed on *T. thermophila* genomic sequence. Predictions from the first round were used to improve parameters for the second round of training. Sixty curated *T. thermophila* genes resulted from prediction algorithm were used in combination with original training set to adjust parameters for splice site and start/stop codon. Predicted

gene models from *ab initio* approach were compared with gene prediction by alignment of 9,122 expressed sequence tag (EST) clusters to *T. thermophila* genome assembly. Only 4.4% of EST clusters were aligned at the intergenic region as predicted by the gene model. Intron size distribution, GC content, and splice site are similar between the two gene prediction approaches. Large number of predicted genes has matched genes from other species via Blastp search. Gene model was refined and re-annotated using data from EST generated from another microarray experiment, data of sequence similarity to other organisms, and *ab initio* gene prediction (Coyne et al. 2008). Manual gene model examination and correction were also performed. The additional refinement described above resulted in 14% of gene models to be updated. Final estimation of protein-coding genes in *T. thermophila* MAC genome is 24,725. Data was submitted to Genbank and TGD. *T. thermophila* genome annotation version October 2008 was the latest version during the study period.

T. thermophila predicted protein sequences were available as a plain text file in the Fasta format. Each sequence was annotated in Fasta header section with an isoform ID and a gene description. An ID translation table was created using the gene attribute table supplemented with *T. thermophila* genome data (available as final_gene_attributes.txt file in *T. thermophila* genome version October 2008). ID translation table was composed of a column of *T. thermophila* isoform ID and a column of corresponding Gene Model Identifier, which begins with 'TTHERM' and follows by 8-digit numbers, e.g. TTHERM_00000010. Gene Model Identifier is used as identifier in this study because it is readily available from genome data attribute and is already used in public *T. thermophila* database. Chromosomal location of each Ser candidate was retrieved from Tetrahymena Functional Genome Database (TetraFGD) (Xiong et al. 2013). Although majority of scaffold are telomere-capped (Eisen et al. 2006), the term 'scaffold' is used when referring to the physical location of Ser candidate instead of 'chromosome' for consistency with the original data from the database. Thirteen known *Ser* gene data (listed in Table 1.3) from NCBI database were used in this study as reference.

2.2 Search pattern determination

Cysteine (Cys) pattern of *Ser* i-ag which will be employed in Ser search algorithm was initially determined based on data from previous studies of subtype H, subtype J, and subtype L *Ser* genes (Table 1.2). Previous study suggested that the common features of repetitive blocks, which occupy the region of tandem repeat, in *Ser* genes are:

Feature 1. Each block contains 6-10 Cys residues, depending on subtypes.

Feature 2. Each block contains approximately 55-100 amino acid residues.

Feature 3. The number of amino acid residues spanning between two Cys residues alternate between 'short (2-4 residues)' and 'long (from 6 to up to 21 residues)'.

Based on these features, search pattern was set accordingly as '$CX_{(\geq 6)}CX_{(\geq 1)}CX_{(\geq 1)}CX_{(\geq 1)}CX_{(\geq 1)}C$' (C = cysteine; X = any amino acid except C; number in parenthesis = number of X). The number of Cys residues in the search pattern is set to six which is the lowest number of Cys residues in one repeat block as reported in previous studies (Feature 1). By fixing the number of X between the first pair of C at the minimum of six residues, a sequence matched with this pattern is expected to contain at least one 'long' span (Feature 3). To ensure that the search returns as much candidates as possible, the search pattern was given more weight toward Feature 1.

The length limit for the search pattern was determined by performing the search with *T. thermophila* predicted protein sequences while varying the set length in each search in order to choose the optimum number. The reported length of *Ser* repeat block (Feature 2) was used initially as a guideline but was not strictly obeyed. The length was eventually varied between 30 – 300 residues, and the numbers of candidates picked by the search were recorded. Judging from the result, the length of 120 amino acid residues was selected for *Ser* search pattern. Detail of length determination process is discussed in Result section.

2.3 Pattern search script

The cysteine pattern search part was performed by custom developed script. Perl was chosen as programming language for the pattern search script development due to its powerful text processing tools which is suitable for text-based data and its availability of modules specific for parsing biological data (namely Bio-perl). The developed pattern search script reads Fasta-formatted plain text file as input, and returns the result as two plain text files. One is an output report file which is automatically named 'testresult_newscript_6C_$[interval]aa' (i.e. if the length limit was set to 100, the result output file name will be 'testresult_newscript_6C_100aa'). Another is the Fasta-formatted file of the sequences that matches the search pattern, which is automatically named as the output report file with 'fasta' extension (e.g. testresult_newscript_6C_100aa.fasta). The script requires Bio-perl module to parse Fasta-formatted sequence. The script does not provide input interface, and the change in any search setting (e.g. input file name and length limit) must be done by manually editing the script. The script must be placed in the same directory as the input file. To perform the search, the script was simply executed via the command line terminal.

The algorithm of pattern search script was shown in Figure 2.1. In brief, each *T. thermophila* predicted protein sequence in Fasta format was parsed from the input text file as a string. Next, the script searches for the letter 'C' (Cys residue) and selects the portion of sequence from the matched 'C' until the end as a substring. The script then searches the substring for the defined pattern (i.e. $CX_{(\geq 6)}CX_{(\geq 1)}CX_{(\geq 1)}CX_{(\geq 1)}CX_{(\geq 1)}C$) within the defined length limit. The number of matched along the whole substring is counted. In the next step, the script goes back to the parsed sequence, looks for another substring started with the next 'C' and repeats the search again. After all the substrings started with 'C' were searched, the script determined the sum of pattern match counts. If pattern match is found (count is not zero), the parsed protein sequence will be marked as matched. The process is repeated for the next protein sequence. After all protein sequences in the input file are parsed and searched , the script print out the sequences which were marked as matched into the output file. The output report shows the search pattern and length limit setting, number of matched sequences, number of sequences not matched, the Fasta header of matched sequences,

the position of Cys residues, and the matching string. The example of the output report is shown in Figure 2.2.

```
#!/usr/bin/perl -w

# 2012/02/20
#for Tth Ser gene identification. I use Bioperl module so the
script handles fasta sequence better.

use Bio::SeqIO;
use Bio::Seq;

$pattern = "C[^C]{6,}C[^C]+C[^C]+C[^C]+C[^C]+C"; $patternname =
'6C'; #CX6CXCXC
$interval = 300;
$numpatmatch = 0; $numNotmatch = 0; $numMatch = 0;
@outputseq = (); @outputpat = ();

my $filename = "tta1_oct2008_finalrelease.txt"; #Name the file to
be opened.
my $format = "Fasta"; #Must be fasta format.
my $outfile = "testresult_newscript_${patternname}_${interval}aa";

$inseq = Bio::SeqIO->new( #SeqIO object brings in the input file
                          -file   => "<$filename",
                          -format => $format,
                          );
$seq_out = Bio::SeqIO->new(
                            -file   => ">$outfile.fasta",
                            -format => $format,
                            );

while (my $seq = $inseq->next_seq) {
#  print $seq->seq, "\n";
  $numpatmatch = 0; #reset count of the times the pattern matches.
  $oneseqtemp = $seq->seq;
  $subseq = $oneseqtemp;
  $oneidtemp = $seq->display_id;
  $seqlength = $seq->length;
  $keepPat = "";
  print ">$oneidtemp\t$seqlength\n$oneseqtemp\n";

#while-loop checking for all possible C
  while ($oneseqtemp =~ /C/g) {
    $Cpos = length($`);
    $subseq = substr $oneseqtemp, $Cpos; #Select substring
afterward the matched C
    print "substring: $subseq\n";
    if (($subseq =~ /$pattern/) && (length($&) <= $interval)) {
      print $Cpos,"\t",$&,"\n";
      $keepPat .= "$Cpos\t$&\n";
      $numpatmatch++; #increase count of number of times the
pattern matches.
    }
  } #end of while-loop checking for all possible pattern matches

#now the judgement...
```

```
  if ($numpatmatch == 0) { #if not a single pattern is found...
    $numNotmatch++;
    next;
  } elsif ($numpatmatch > 0) { #but if at least one pattern is
found in sequence...
    $numMatch++;
    $seq_out->write_seq($seq); #export matched seq in fasta format
for later use.
    push(@outputseq,$oneidtemp);
    push(@outputpat,$keepPat);
  }
} #end of while-$seq

print "Matched: $numMatch\tNot Matched: $numNotmatch\n"; #Report


# Write result report to a file
unless (open(TESTRESULT, ">$outfile")) {
     print "Cannot open file \"$outfile\" to write to! \n";
     exit;
}
print TESTRESULT "pattern: $pattern $patternname\t interval:
$interval\t\n";
print TESTRESULT "total match: $numMatch\tnot match:
$numNotmatch\n\n";
for ($i = 0; $i < $numMatch; $i++) {
     print TESTRESULT ">$outputseq[$i]\n";
     print TESTRESULT "|Cys found position|\t|Pattern matched|\n";
     print TESTRESULT "$outputpat[$i]\n";
}
close (TESTRESULT);
```

Figure 2.1 Pattern search script.

```
pattern: C[^C][6,}C[^C]+C[^C]+C[^C]+C[^C]+C 6C  interval: 120
total match: 4925  not match: 19801

>127.m00065
|Cys found position|    |Pattern matched|
234
     CMDKIHIYPFLTRICKKLAQEQKNDTSVLECGRRICINCLVVYLIKKTSYNYVQKLITKQNQQD
KTYLQYLDDFKYFKIKC
248
     CKKLAQEQKNDTSVLECGRRICINCLVVYLIKKTSYNYVQKLITKQNQQDKTYLQYLDDFKYFK
IKCPSQDC

>127.m00073
|Cys found position|    |Pattern matched|
1412
     CPPEAIQKRLNKDQKLNGNQSQIGKSNLFKAKSEYELKEFGVFRSTYIKIERELLIQRMKQIES
LNVGTIMICNECGKKLSVFECIECDLTFC

>127.m00092
|Cys found position|    |Pattern matched|
18   CTNFSDSFSSCVYTYCQQDLNCYNQAIWQLECGVNKC
28   CYNQAIWQLECGVNKCQTINDATLWYNCLVQSFSQTCQVSYGVANDARDYC
33   CYNQAIWQLECGVNKCQTINDATLWYNCLVQSFSQTCQVSYGVANDARDYC
39   CYNQAIWQLECGVNKCQTINDATLWYNCLVQSFSQTCQVSYGVANDARDYC

>127.m00058
|Cys found position|    |Pattern matched|
28   CRDVYSYFDSNGNQCLYCSKNCALCSSNNNC
42   CQQDYYLNEQGQCLSDCGTSLLKGLNSFQCYNFSDINCLKYNQQNVC
45   CQQDYYLNEQGQCLSDCGTSLLKGLNSFQCYNFSDINCLKYNQQNVC
49   CQQDYYLNEQGQCLSDCGTSLLKGLNSFQCYNFSDINCLKYNQQNVC
52   CQQDYYLNEQGQCLSDCGTSLLKGLNSFQCYNFSDINCLKYNQQNVC
58   CQQDYYLNEQGQCLSDCGTSLLKGLNSFQCYNFSDINCLKYNQQNVC
61   CQQDYYLNEQGQCLSDCGTSLLKGLNSFQCYNFSDINCLKYNQQNVC
73   CGTSLLKGLNSFQCYNFSDINCLKYNQQNVCIVCRDGYILNDDGIC
77   CGTSLLKGLNSFQCYNFSDINCLKYNQQNVCIVCRDGYILNDDGIC
90   CYNFSDINCLKYNQQNVCIVCRDGYILNDDGICKNVLC
98   CLKYNQQNVCIVCRDGYILNDDGICKNVLCSQFNGIYDNTIQKC
```

Figure 2.2 An excerpt from output report generated by pattern search script.

2.4 *Ser* candidate refinement by GPI-anchored protein prediction

To refine *Ser* candidate set, *T. thermophila* protein sequences that passed the pattern search script were subsequently subjected to GPI anchor prediction algorithm. A web-based FragAnchor program was chosen for GPI prediction due to its reported high specificity, ability to process multiple input sequences, and flexibility in scoring system as classes. FragAnchor is accessible via http://navet.ics.hawaii.edu /~fraganchor/NNHMM/NNHMM.html. Fasta-formatted file created by the pattern search script was uploaded into FragAnchor as an input. The program did not provide any manual parameter setting and thus required no adjustment.

Detail of GPI prediction algorithm followed a published protocol (Poisson et al. 2007). In short, FragAnchor uses a tandem system of neural network (NN) and Hidden Marcov Model (HMM) to predict and classify GPI-anchored proteins. The program selected the last 50 aa from the C-terminus as input based on the analysis of GPI-anchored protein from Swiss-Prot database, which showed that they have maximum length of 45 residues. Each amino acid in the input sequence was encoded with hydropathy value and molecular weight. NN was trained with annotated 79 GPI-anchored protein sequences and 79 non GPI-anchored protein sequences from Swiss-Prot database. Prediction performance of neural network was asserted by the control set of 134 GPI sequences and 134 non-GPI sequences from Swiss-Prot database. The method was shown to have 93% precision when the selection threshold was set to 0.90. Additional validation test showed that NN had 95% specificity (Poisson et al. 2007). Input sequence with NN score less than 0.90 is rejected, and those with NN score at least 0.90 is ‘accepted’ by NN, which is passed further to HMM. To create model for GPI-anchored protein prediction, HMM graph (i.e. the network that each amino acid residue is treated as node that is linked to another residue) was divided into three regions (anchor site, spacer region, and hydrophobic region) according to the structure of GPI anchor signal. Parameters were estimated by iteration of Baum-Welch algorithm that determines the unknown parameters for HMM. HMM was trained with another set of 66 annotated GPI-anchored proteins from Swiss-Prot database, which was different from the set that was used for training and validation of NN. Model was validated by 500 bootstrapping on training set. HMM calculates score based on the likelihood that the input sequence is similar to the sequences in the training set, and

normalizes by the length of the sequence. HMM refines the prediction by assigning probability score to each input sequence and classifying them into four classes based on score—'highly probable' class (HMM score >= 5.40), 'probable' class (2.20 <= HMM score < 5.40), 'weakly probable' class (0.20 <= HMM score < 2.20), and 'potential false positive' class (HMM score < 0.20). Each class was defined by Receiver Operating Characteristic (ROC) curve that visualizes the performance of prediction by plotting specificity (reciprocal of rate of false positive) against sensitivity (rate of true positive) while the discrimination threshold was varied. Validation test showed that the specificity of HMM varied from 93.89% to 99.89% depending on classes (Poisson et al. 2007). In this study, only the sequences that are classified into 'highly probable' class by FragAnchor were selected as *Ser* genes candidates. The result report has detail of the input sequence in the Fasta format, NN score, HMM classification and score, three predicted GPI signals with highest score, and predicted GPI attachment sites (ω site) (Figure 2.3).

```
Time_stamp: Mon Feb 18 01:40:07 2013
Input_count: 4925
Time_used: 22.685422 seconds

Statistics:

Invalid_sequence_count: 110 (2.23%)

NN:
NN_accept: 591 (12%)
NN_reject: 4224 (85.77%)

HMM:
HMM_highly_probable: 222 (4.51%)
HMM_probable: 71 (1.44%)
HMM_weakly_probable: 35 (0.71%)
HMM_potential_false_positive: 263 (5.34%)
HMM_not_scored: 0 (0%)



>127.m00065 hypothetical protein
MNQISQNKQYQSQINYNFKQGVQNNRFQTDFNQNAYQANQFNFQQQNQNMMDYNQNVQYQ
QPDQQMYQQNQQYYSQQNQNQQLGNNGMENYQMQPNDMQVNSMNMQYQNQYDYNQMQQYQ
QNGEFYNQQNPNNLNNPNGQMIDNNQNQQYNQQAYYNQNMQQFDQQQMQYINQVQQQKKG
IIDMTNYQIPSFQKAQQIDPDLTQEQYNLYTQIIQKEEALNSTFQKLVSKDLIFCMDKIH
IYPFLTRICKKLAQEQKNDTSVLECGRRICINCLVVYLIKKTSYNYVQKLITKQNQQDKT
YLQYLDDFKYFKIKCPSQDCQEEQGNCYLQRKQVEDMINCRQGFINK

NEURAL NET
Score: 0.000047
REJECT


>127.m00092 hypothetical protein
MRTYTLISLLIFVSSVLSCTNFSDSFSSCVYTYCQQDLNCYNQAIWQLECGVNKCQTIND
ATLWYNCLVQSFSQTCQVSYGVANDARDYCLNSYDFSQDCTNVTDGGSQPGSGRNVVVIT
TGSNKLFMSLTLMTILIFLFA

NEURAL NET
Score: 0.999962
ACCEPT

HMM
Score 1: 7.68679 *** 115
Structure: [NVV] VITTGSNKLFMSLTLMTILIFLFA
Classification: Highly probably
```

Figure 2.3 An excerpt from FragAnchor combined result report. The proportion of input classified into each class, and details including the score and predicted GPI attachment site were given.

2.5 ID translation script

The FragAnchor result report is not readily usable for further analysis because the text is not structured in the way that allows convenient data import/export. Sequences of *Ser* candidates were not directly available in the Fasta format. Moreover, isoform ID which is the gene identifier used in *T. thermophila* MAC genome data needs to be translated into Gene Model Identifier which is another identifier that is commonly used in *T. thermophila* resource database. To solve this problem, a Perl script was developed to translate gene ID from the FragAnchor result report and to create a Fasta-formatted sequence file for further use. This Perl script requires three input files, which should be placed in the same directory as the script:

File 1. An ID translation table

File 2. A file contains Fasta-formatted protein sequences of *T. thermophila*

File 3. A file contains Fasta header of *Ser* candidates with isoform ID

All files are plain text file. A translation table was prepared as described in "Sequence and genome data" section. A file containing Fasta-formatted protein sequences of *T. thermophila* was already created by pattern search script as described in "Pattern search script" section. A file containing Fasta header of *Ser* candidates is created from FragAnchor result report by retrieving all lines in the result report file that start with '>' symbol using the UNIX grep command.

ID translation script requires Bio-perl module to run. It does not provide user interface, so any desired change in setting must be done by manual script editing. The detail of ID translation script is shown in Figure 2.4. In brief, the script reads File 1, selects isoform ID, looks up File 2 to retrieve its sequence, and looks up File 3 to find the matching Gene Model ID. Finally, the script creates two output files. One is the plain text file containing *Ser* candidate sequences in the Fasta format with Gene Model ID as identifier ('TTHERM_' followed by 8 digits). Another is a tab-delimited file containing both isoform ID and matching Gene Model ID. The first output file is used for further analysis, and the second file serves as translation log file.

```
#!/usr/bin/perl -w

use Bio::SeqIO;
use Bio::Seq;

# 2012.02.21 Retrieve selected sequence for phylogenetic anal
# - Read from list of id and pick fasta sequence.
# - Add TTHERM and write another file.
#This will be useful when alignment will be saved in phylip format,
because
#phylip format allows only 10 characters in sequence header.
# - And generate a table & write another file.

my $idfilename = "fraganchor.6120.invalid"; #Name the file containing id
of sequence you want to select.
my $filename = "testresult_newscript_6C_120aa.fasta"; #Name the fasta
file to be opened.
my $format = "Fasta"; #Must be fasta format.
@seqArray = ();
$counter = 0;

open IDFILE, "$idfilename.id" or die "$!";
open TRANSL, "Tth_IDtranslation_table" or die "$!";
@translationTable = <TRANSL>;
close TRANSL;

$inseq = Bio::SeqIO->new( #SeqIO object brings in the input file
                          -file   => "<$filename",
                          -format => $format,
                          );
$seq_out = Bio::SeqIO->new(
                            -file   => ">$idfilename.fasta",
                            -format => $format,
                            );

while (my $seq = $inseq->next_seq) { push(@seqArray,$seq); }
@TTHERMarray = ();
@TABLEarray = ();

foreach $id (<IDFILE>) {
  $id =~ s/>(\d{1,4}\.m\d{5}) .+\n$/$1/;
  print $id,"\n";
  foreach my $seq (@seqArray) {
    $seqid = $seq->display_id;
    $seqdesc = $seq->desc;
    $seqAA = $seq->seq;
    if ($id eq $seqid) {
      $seq_out->write_seq($seq);
      foreach my $TLtable (@translationTable) {
     if ($TLtable =~ /$id/) { $TTHERM = $TLtable;
                     $TTHERM =~ s/^TTHERM_(\d{8})\t\d{1,4}.m\d{5}\n$/$1/;
                     print "$TTHERM\n";
                     last; }
      }
      push(@TTHERMarray,">$TTHERM\n$seqAA");
      push(@TABLEarray,"$id\tTTHERM_$TTHERM\t$seqdesc");
      $counter++;
      next;
    }#if-$id
  }
```

```
}#foreach-$id loop
# Write result report to a file
unless (open(TTHERMfile, ">$idfilename.TTHERM.fasta")) {
     print "Cannot open file \"$outfile\" to write to! \n";
     exit;
}
foreach my $printtemp (@TTHERMarray) { print TTHERMfile "$printtemp\n"; }
close (TTHERMfile);

unless (open(TABLEfile, ">$idfilename.csv")) {
     print "Cannot open file \"$outfile\" to write to! \n";
     exit;
}
foreach my $printtemp (@TABLEarray) { print TABLEfile "$printtemp\n"; }
close (TABLEfile);
close IDFILE;

print "\nRetreive $counter sequences listed in $idfilename.id from
$filename.\n";
```

Figure 2.4 ID translation script in Perl.

2.6 Additional Ser homolog search

To extend the Ser candidate search, the homolog search approach was employed. The initial set of 216 Ser candidates was used as blastp query against NCBI non-redundant (nr) protein database. Blastp parameters were set as default (gap penalty: 11; gap extension 1; conditional compositional score matrix adjustment). Blast hit table was downloaded as tab-delimited plain text file. All sequences that have 100% identity match to the query, and all sequences that pass the initial pattern search step but were not classified as 'highly probable' by FragAnchor ('probable', 'weakly probable', 'potential false positive', 'rejected', and 'invalid') were removed manually or using custom Pearl script (Figure 2.5). To retrieve additional information such as protein name and taxa ID for candidate list, another custom Perl script was developed to retrieve information from online database using NCBI accession as query (Figure 2.6). The candidate list was then sorted by Taxa ID, and was separated into two lists of candidates from *T. thermophila* and from other species. The homolog hit table for each species was generated by combination of blast hit table and retrieved information from Script 4 using another custom-developed Perl script (Figure 2.7). The candidate list from *T. thermophila* was manually inspected for cysteine pattern and was subjected to GPI anchor signal prediction by FragAnchor.

```
#!/usr/bin/perl -w
#DATE: 2014/02/28
# This script reads any string of text (A) as query, matches query to another
text (B),
#then removes any rows which contains query-matched text.
# I expect the query to contains gene/protein ID, so it will include symbol
character like ._,|;
# And target text is expected to be in tab-delimited format.
#
#Enter file name here:
#queryfilename = (A), targetfilename = (B)
#Don't forget to edit file name before each use.
#(A) is just a simple index of GI you want to remove from file (B). It cannot
contain anything else.
#$queryfilename = "test_query";
#$targetfilename = "test_hittable";
$queryfilename = "probableGI";
$targetfilename = "merge-hittable-sort-nocomment-sort_MINUS_SerGI";
#Do not change anything below this line.
#Declare result file name.
$resultfilename = "${targetfilename}_MINUS_${queryfilename}";
#Open file and load data so we will be ready to go.
open QUERY, "$queryfilename" or die "\nCan't open file: $!";
@querylist = <QUERY>;
close QUERY;
open TARGET, "$targetfilename" or die "\nCan't open file: $!";
@targetfile = <TARGET>;
close TARGET;

#Done loading file. Now let's use foreach loop to check query-by-query.
#We will use Refseq GI as query. I hope it's universal; If not, I'll be
doomed.
foreach $onequery(@querylist) {
  $i = 0;
  $onequery =~ s/\n//; #clean it
  $onequery =~ s/gi\|(\d*)\|\w*\|.*/$1/; #clean it. only GI is left.
#  print "$onequery\n"; #for debugging
#Go through every lines in target file. (Ouch. This will be long.)
#If matched, push this row to @rmvRows. If not, just ignore this row.
#I'm assuming there's no duplicate in query list.
  foreach $onetarget(@targetfile) {
    if ("$onetarget" =~ m/$onequery/g) {
      print "Matched: $onetarget";
      $targetfile[$i] = "";
    }
    $i = ++$i;
  } }
#Now there should be only not-matched target in @targetfile.
#Let's print the result file.
unless (open(RESULT, ">$resultfilename")) {
       print "Cannot open file \"$resultfilename\" to write to! \n";
       exit; }
foreach (@targetfile) {
       print RESULT "$_"; } close (RESULT);
```

Figure 2.5 Perl script for search and removal of Ser candidates using ID.

```
#!/usr/bin/perl -w
#http://www.bioperl.org/wiki/HOWTO:EUtilities_Cookbook#Get_GIs_for_a_list_of_
accessions
#http://www.bioperl.org/wiki/HOWTO:EUtilities_Web_Service
#http://www.ncbi.nlm.nih.gov.ejournal.mahidol.ac.th/books/NBK25500/
#Bio::DB::EUtilities

#edit this file name before each use.
$accfilename = "acc_0041-0060";

use Bio::DB::EUtilities;

open ACCESSION, "$accfilename";
my @accs = <ACCESSION>;
close ACCESSION;

my $factory = Bio::DB::EUtilities->new(-eutil => 'esearch',
                                       -email => 'user@email.com',
                                       -db    => 'protein',
                                       -term  => join(',',@accs) );

my @uids    = $factory->get_ids;

$factory->reset_parameters(-eutil => 'esummary',
                           -db    => 'protein',
                       -email => 'user@email.com',
                           -id    => \@uids);

print "ID\tCaption\tStatus\tTaxId\tTitle\n";
while (my $ds = $factory->next_DocSum) {
    print $ds->get_id,"\t";
#    while (my $item = $ds->next_Item())  {
       # not all Items have content, so need to check...
#      my $title = $item->get_contents_by_name('CreateDate');
#        printf("%-20s:%s\n",$item->get_name,$item->get_content) if $item-
>get_content;
#    }
    print $ds->get_contents_by_name('Caption'),"\t";
    print $ds->get_contents_by_name('Status'),"\t";
    print $ds->get_contents_by_name('TaxId'),"\t";
    print $ds->get_contents_by_name('Title');
    print "\n";
}
```

Figure 2.6 Perl script for data retrieval from protein database. The script takes accession number as input and returns a tab-delimited table of GI, accession, data status, taxa ID, and protein description. The script requires BioPerl module.

```
#!/usr/bin/perl -w
#
#Pro tip: Keep looking at the terminal. If something which is not acc turns
up, there's something wrong. Check regex as recommended below.
#
#Enter file name here:
#Don't forget to edit file name before each use.
#$queryfilename = "test_query";
#$targetfilename = "test_hittable";
$queryfilename = "homsearch.Tcr";
$targetfilename = "merge-hittable-sort-nocomment-sort";

#DATE: 2014/04/03
# The objective of this script is to combine list of organism (made by
GI2Info.pl script)
#to Blast table.
# OrgList is used as query. The file is tab-delimited and should look like
this (no header included):
#GI             acc            status  TaxId    Title
#115252917 CAJ66787  live 1311 putative cell-wall anchored surface adhesin
[Streptococcus agalactiae]
#115252925 CAJ66791  live 1311 putative cell-wall anchored surface adhesin
[Streptococcus agalactiae]
#....
#The script can detect acc if it looks like this: YP_001967238, CAJ66787
#If acc doesn't look like this, then the script won't work.
#And if it doesn't look like neatly clean acc on your terminal, regex can't
match your query.
#You will have to modify regex at this line:
#$onequery =~ s/\d*\t([A-Z0-9_]*)\t\w*\t\d*\t.*\n/$1/
#
# Blast table is used as target. The file is exported from NCBI Blast and
should look like this:
#00411420 gi|115252917|emb|CAJ66787.1|   56.8672.5551   20   1    118    168
     400  448  8e-0551.6
#00411430 gi|115252925|emb|CAJ66791.1|   34.9152.83106  49   3    104    192
     669  771  1e-0656.6
#....
#
# The script will select acc row from query (row 2), look up through target,
and then create a combined table
#by simply add data from target after data from query at the same line. User
can remove unwanted row later manually.
#The script will create new file as output. Output file is automatically
named "hittable.(queryfilename)".
#
#END DESCRIPTION.
#
#Do not change anything below this line.
#
#Declare result file name.
$resultfilename = "hittable.${queryfilename}";
@printout = ();

#Open file and load data so we will be ready to go.
open QUERY, "$queryfilename" or die "\nCan't open file: $!";
@querylist = <QUERY>;
close QUERY;
open TARGET, "$targetfilename" or die "\nCan't open file: $!";
@targetfile = <TARGET>;
close TARGET;
print "Matched:\n";
#Done loading file. Now let's use foreach loop to check query-by-query.
#We will use acc as query. I hope it's universal; If not, I'll be doomed.
```

```
foreach $onequery(@querylist) {
  $fullquery = $onequery;
  $fullquery =~ s/\n//; #clean it
  $onequery =~ s/\d*\t([A-Z0-9_]*)\t\w*\t\d*\t.*/$1/; #clean it. only
accession is left.
  $onequery =~ s/\n//; #clean it. I don't rmv newline at first because the
last line won't have newline character, so the last line will not be matched
and acc will not be read.
  print "$onequery\n"; #for debugging
#Go through every lines in target file. (Ouch. This will be long.)
  foreach $onetarget(@targetfile) {
    if ("$onetarget" =~ m/$onequery/) {
push(@printout,"$fullquery\t$onetarget");
    }
  }
}

#Let's print the result file.
unless (open(RESULT, ">$resultfilename")) {
     print "Cannot open file \"$resultfilename\" to write to! \n";
     exit;
}
foreach (@printout) {
     print RESULT "$_";
}
close (RESULT);
```

Figure 2.7 Perl script for generating homolog hit table. Output generated from Perl script in Figure 2.5 and Figure 2.6 are required as input.

2.7 Phylogenetic analysis

The translated sequences of *Ser* candidates were aligned using ClustalX 2.0.12 (Larkin et al. 2007) with default multiple alignment parameters (gap opening penalty = 10; gap extension penalty = 0.2; Gonnet series weight matrix). Thirteen known Ser genes (Table 1.3) sequences were included in the analysis. Alignment was then adjusted manually. TTHERM_01098980 was excluded from phylogenetic analysis because its sequence has an unusual sequence length of 3751 amino acid residues, preventing it from being aligned. Neighbor-joining (NJ) tree was calculated with 1000 bootstrap replicates using ClustalX (Larkin et al. 2007). To determine the best-fit model for Maximum likelihood (ML) tree estimation of *Ser* candidates, protein evolutionary model was selected using ProtTest program (Abascal et al. 2005). ProtTest employs Phyml to compute likelihood score of each candidate model, and then select the best fitted model. The selected candidate models were WAG, Dayhoff, JTT, VT, Blosum62, and LG with invariable sites (+I), gamma distribution of rate of amino acid change (+G), and observed amino acid frequency (+F) parameters. ProtTest then measure model fit using Akaike Information Criterion (AIC), corrected Akaike Information Criterion (AICc), and Bayesian Information Criterion (BIC). For our dataset, VT+G model yielded the highest likelihood score as determined by ProtTest, and was determined as the best-fitted model by all three measurements. ML tree was estimated with 100 bootstrap replicates using RAxML 7.2.8 (Stamatakis 2006) as implemented on the CIPRES Science Gateway (Miller 2010). Parameters were set accordingly to best-fitted model determined by ProtTest. ML tree estimation was executed under GAMMA model of rate heterogeneity (+G parameter) on the *Ser* candidate dataset using empirical base frequency and the VT substitution model. RAxML returned program execution info (RAxML_info.result), the best-scoring ML tree (RAxML_bestTree.result), all 100 bootstrapped trees (RAxML_bootstrap.result), and the best-scoring ML tree with bootstrap support value written (RAxML_bipartitions.result) as output that can be downloaded as batch file from CIPRES system. Phylogenetic trees were then created from the best-scoring ML tree using Dendroscope (Huson et al. 2007).

2.8 *Ser* gene expression analysis

T. thermophila genome-wide gene expression (growth, starvation, and conjugation) microarray data was retrieved from *Tetrahymena* Functional Genomics Database (TetraFGD) (http://tfgd.ihb.ac.cn/) (Xiong et al. 2013). Details of microarray experiment and data analysis were previously described (Miao et al. 2009). In brief, *T. thermophila* strain B2086 and CU428 that were derived from inbred strain B were used for expression analysis in growth, starvation, and conjugation condition. MAC genomic sequence version 2006 obtained from strain SB210, which also was derived from inbred strain B, was used for microarray probe design. Microarray probes were designed as 60-mer oligonucleotides, covering predicted protein-coding genes, non-coding RNA and tRNA. Each of *T. thermophila* sequence was covered by 13-14 unique probes that were evenly distributed over the length of the sequence. Random oligonucleotide probes that were not corresponding to *T. thermophila* genome sequence were used for measuring background signal from non-specific binding. Total 4,308 random probes were generated with comparable length and GC content to the actual gene probes. Expression data during growth phase were collected from the culture of strain CU428 in triplicates at three time points—low growth (10^5 cell/ml, logarithmic growth phase), medium growth (3.5×10^5 cell/ml, deceleratory growth phase), and high growth (10^6 cell/ml, stationary phase). For starvation condition, strain CU428 was cultured to the concentration of 2×10^5 cell/ml, then the cells were collected, washed, and starved in Tris (pH 7.5). Data during starvation phase were collected in triplicate every 3 hours for seven time points (0^{th}, 3^{rd}, 6^{th}, 9^{th}, 12^{th}, 15^{th}, 24^{th} hour). For conjugation condition, strain B2086 and CU428 at the concentration of 10^5 cell/ml were starved in Tris (pH 7.5) for 18 hours and then mixed to initiate conjugation. Data during conjugation phase were collected in duplicate every 2 hours for ten time points (0^{th}, 2^{nd}, 4^{th}, 6^{th}, 8^{th}, 10^{th}, 12^{th}, 14^{th}, 16^{th}, 18^{th} hour). Determined background signal was subtracted from signal intensity of all probes. Array was normalized by quantile normalization method, which transforms the distribution of expression in each array into identical distribution that can be compared. Expression value for any open reading frame was determined from normalized value from probes by Robust Multiarray Average (RMA) method. The constant expression level of two known constitutively expressed genes confirmed data normalization.

MultiExperiment Viewer (MeV) program (available from www.tm4.org) was employed to evaluate gene expression clusters and to classify the expression pattern into subgroups. Expression data of Ser candidates were adjusted by log-2 transformation. Because no prior knowledge of the number of cluster for Ser candidate data, the number of cluster was estimated based on the data itself. To determine the number of cluster that fits Ser candidate dataset, K-Means clustering Support module (KMS) available in MeV was used. This module repetitively clusters the dataset given the percentage of co-occurrence, the number of cluster, and the number of iteration. The number of iteration was set to 200 because the clusters appear to be saturated by 200 rounds of iteration. The percentage of co-occurrence in each cluster was set to 50, which means data were assigned to the same cluster for at least 50% of iteration rounds. The pre-set number of clusters was varied between 10, 20, 30, and 40. KMS module returns the percentage of data that was not persistently assigned to any certain cluster. The best clustering parameters are 30 clusters using median as central value, which resulted in 24% of data not being persistently assigned to any cluster. Clustering was then performed with K-Means Clustering module (KMC) for 200 iterations. Distance was calculated by Pearson correlation because it gives expression profiles with similar expression trend (e.g. genes that are up-regulated at the same time point but at different degree) a smaller distance score.

CHAPTER III
RESULTS

3.1 Identification of *Ser* gene candidates

Due to the high degree of polymorphism, identification of *T. thermophila Ser* genes by sequence homology alone is limited by low sequence conservation. In order to systematically search for *Ser* candidates, two criteria were applied based on the common features found in existing i-ag proteins (six *Ser*H, one *Ser*J and six *Ser*L), namely, the presence of Cys residue pattern block $CX_{long}CX_{short}C$ and the GPI-anchor signal located at the C-terminus. For the first criterion, a Perl script was set up to search for the *T. thermophila* proteins which, for any window frame containing 6 Cys residues, the number of amino acid between the first Cys pair was equal or more than 6 and the number of amino acid between the other Cys pairs was at least 1 ($CX_{(\geq 6)}CX_{(\geq 1)}CX_{(\geq 1)}CX_{(\geq 1)}CX_{(\geq 1)}C$). The number of Cys residues in each block was set to 6 because it is the lowest number of Cys residues observed in the known Ser sequences that was reported (Table 1.5). However, the length of each block was not consistently determined in previous study of *Ser* genes due to indefinite periphery of block. Approximately the length of each repeat was reported between 55 – 100 residues (Doerder 2000; Gerber et al. 2002; Doerder and Gerber 2000). Therefore, to determine the value that will be fixed as the length of each block for the search pattern, the length was varied from 50 to 300 residues. For the second selection criterion, FragAnchor classified Ser candidates as being “highly probable” GPI-anchored proteins including all known *Ser* genes. Sequences classified as “probable” or “weakly-probable” GPI-anchored proteins were excluded in order to avoid false positives.

By varying the length of window frame, increasing the interval frame length also increases the number of output returned from pattern search and subsequently the number of *Ser* candidates that passed the GPI-anchored protein filter

(Figure 3.1) The interval was finally fixed to 120 residues because it is the data point where the number of the output started to show sign of saturation. Further increase of interval length did not yield relatively higher number of *Ser* candidates and may otherwise decrease the specificity of *Ser* detection. Therefore, after determining the number of hits versus search pattern for saturation in gene numbers and change in phylogenetic pattern, the search criteria was limited to the minimum of 6 Cys residues within a 120 amino acid interval. Any proteins containing $CX_{(\geq 6)}CX_{(\geq 1)}CX_{(\geq 1)}$ $CX_{(\geq 1)}CX_{(\geq 1)}C$ sequence within 120 amino acid residues was selected, resulting in 4,925 hits out of 24,725 predicted proteins in *T. thermophila* genome.

4,925 sequences matched by the pattern search script were further submitted to FragAnchor. 110 sequences (2.23%) were initially discarded as invalid sequence by FragAnchor. Remaining sequences were selected for candidate GPI-anchored protein by neural network (NN). 4,224 sequences (85.77%) were subsequently rejected by FragAnchor by due to having scores lower than 0.9, leaving 591 sequences (12%) as potential GPI-anchored protein candidates to be passed further into Hidden Marcov Model (HMM) for probability score assignment. FragAnchor assigned 222 sequences (4.51%) into 'highly probable' class, 71 sequences (1.44%) into 'probable' class, 35 sequences (0.71%) into 'weakly probable' class, and 263 sequences (5.34%) into 'potential false positive' class. To avoid false positive, only sequences in 'highly probable' class were selected as *Ser* candidates. Six candidates were removed due to unusual sequence length that obstruct phylogenetic analysis and due to lack of expression data. Therefore, the final set of *Ser* candidate consists of 216 sequences. Summary of *Ser* prediction algorithm is shown in Figure 3.2.

The *Ser* prediction algorithm described above has successfully identified SerH3, SerLA, SerLB, SerLC, and SerLE1 as *Ser* candidates. These known *Ser* genes were experimentally identified from the genome of *T. thermophila* strain B, which shares the same genetic background as the reference strain (strain SB210) that was used for macronuclear genome sequencing. The rest of experimentally identified *Ser* genes that were not identified by the described *Ser* prediction algorithm above were identified from *T. thermophila* of different strains and were not expected to be found the genome data used in this study (Table 1.3). Size of known i-ag is approximately

300-400 amino acid residues with three repetitive blocks. The smallest i-ag identified was about 150 amino acid residues with one repetitive block (SerLE1 and Ser LE2) (Table 1.3). *Ser* candidates found from search algorithm have the same range of length variation. Larger *Ser* candidates (600 – 1000 residues) are also found. Inspection on sequence alignment reveals that these large *Ser* candidates carry more repetitive blocks, causing them to be larger. Extremely large *Ser* candidates (>1500 aa) were discarded to avoid complication in sequence alignment.

To compare the *Ser* prediction algorithm with conventional sequence homology search approach, homology search was performed with NCBI blastp. The conventional homology search only resulted in *Ser* genes from the same subtype. For examples, the search with SerH3 (AAF06326) resulted in seventy-one sequences of *T. thermophila* proteins. Twenty-nine of them belong to the *SerH* subtype. The rest showed either weak homology or without clear GPI anchor signal.

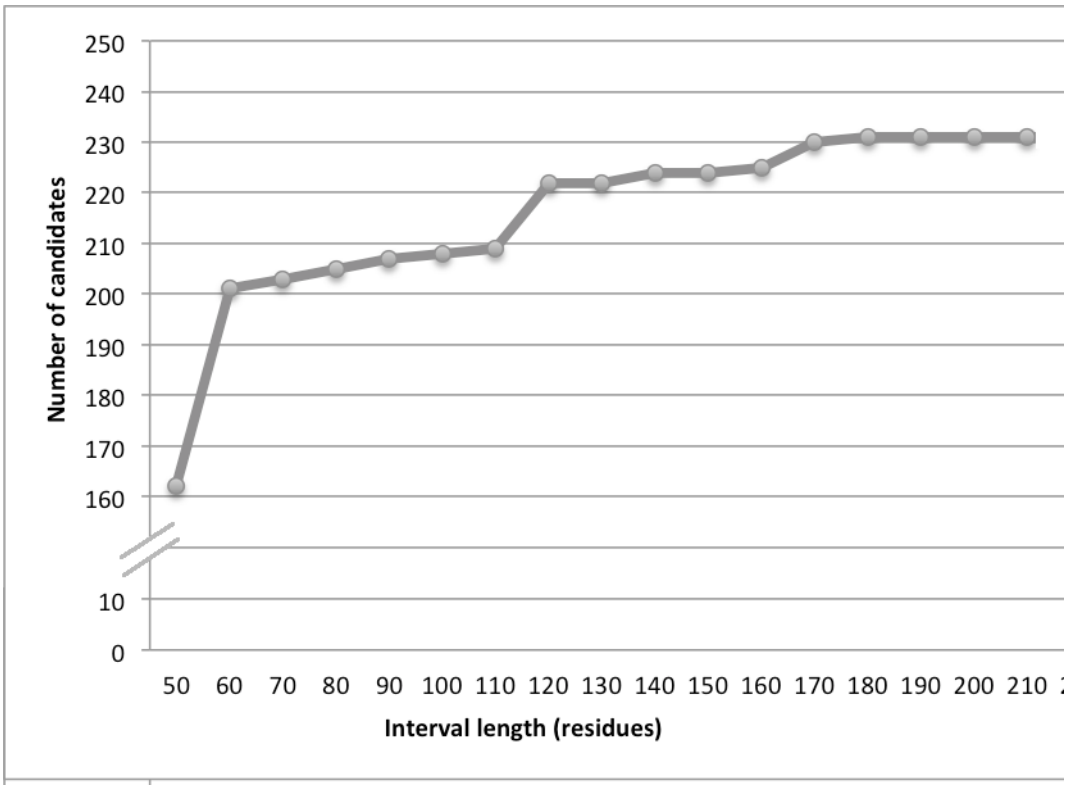


Figure 3.1 Number of Ser candidates that pass both Ser pattern search and GPI-anchored protein prediction when the length of each repeat block is varied. The number of output increases dramatically when the length is increased from 50 to 60. The increase of output quantity gradually slows down and eventually reaches saturation. Increasing interval length from 110 to 120 results in the last sharp increase from 209 candidates to 222 candidates. Therefore, the interval length of 120 residues was chosen.

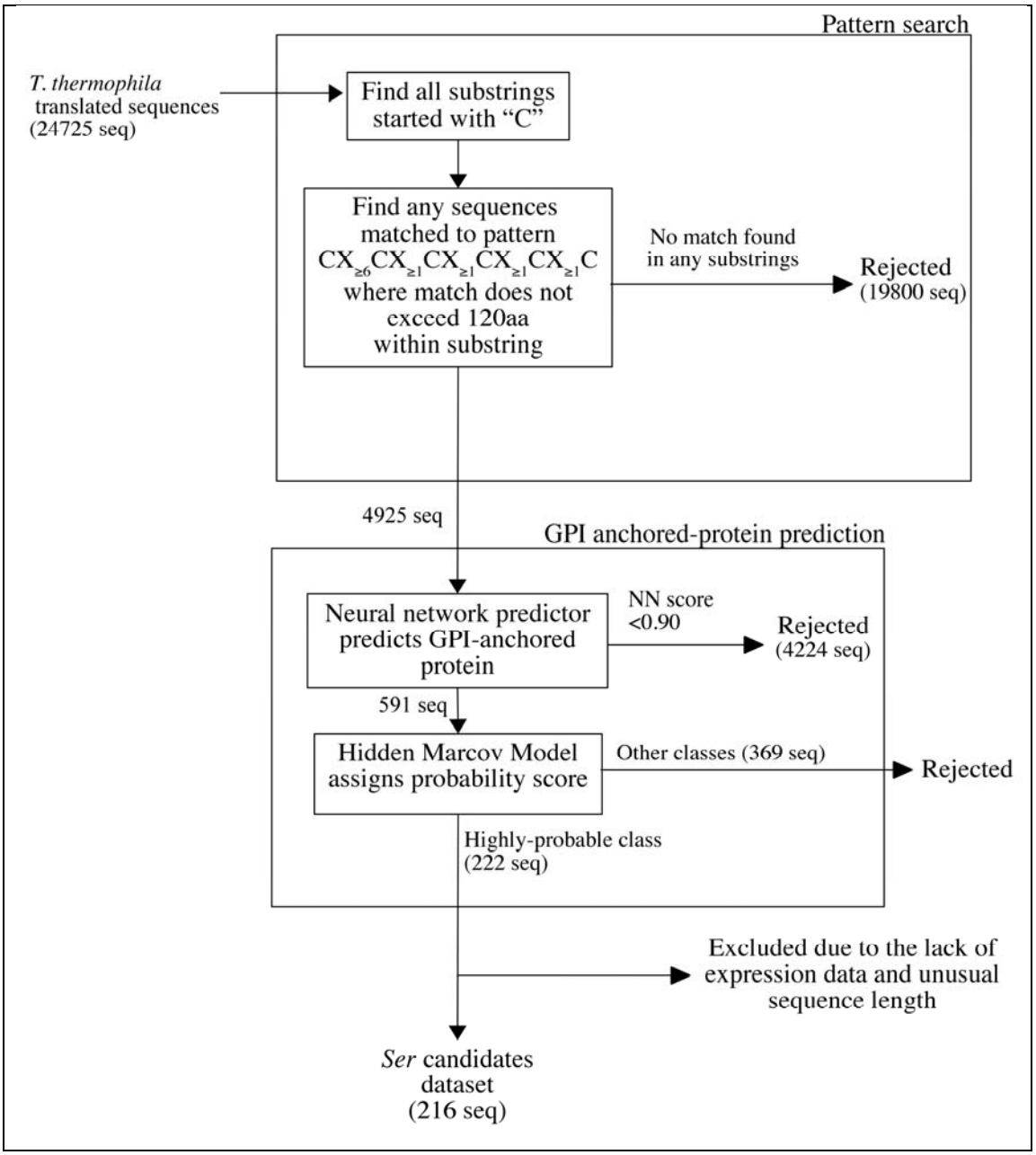


Figure 3.2 Ser prediction algorithm diagram. The process is divided into two parts —pattern search (top box) and GPI-anchored protein prediction using FragAnchor (bottom box). The number of input and output from each step is denoted.

3.2 Classification of *Ser* genes candidates by phylogenetic analysis

Ser candidate dataset was classified into subtype by phylogenetic analysis. All thirteen known *Ser* sequences of three known subtype, subtype-H, subtype-L, and subtype-J, were included in phylogenetic construction as the reference. *Ser* candidates that were found clustered in the same clade with known *Ser* subtype were then identified as that certain subtype. Two different methods of phylogenetic tree estimation, neighbor-joining (NJ) and maximum likelihood (ML) were employed in order to confirm the tree topology.

Ninety-eight (45%) of the *Ser* candidates could be grouped with three identified *Ser* subtypes, *Ser*H, *Ser*L and *Ser*J. For NJ tree, each subtypes form their separated clade with good bootstrap support (*Ser*H: 58.7% NJ bootstrap support; *Ser*L: 75.9% NJ bootstrap support; *Ser*J: 99.8% NJ bootstrap support). Two distinct branches of *Ser* candidates were found as the closest branch of *Ser*J and *Ser*L branch, respectively. Thus, these two branches were named J* (74.6% NJ bootstrap support) and L* (42.2% NJ bootstrap support) to reflect their phylogenetic association with *Ser*J and *Ser*L group (Figure 3.3). ML tree yields similar result with good bootstrap support (*Ser*H: 61% ML bootstrap support; *Ser*J: 86% bootstrap support; J*: 79% bootstrap support; L*: 94% bootstrap support) (Figure 3.4). One *Ser* candidate (TTHERM_00263370) has inconsistency between NJ tree and ML tree. This candidate was grouped with L* group in NJ tree, but was grouped with *Ser*L group in ML tree. Such inconsistency might cause low (17%) bootstrap support of *Ser*L branch in ML tree. When TTHERM_00263370 is excluded, all *Ser* candidates that were grouped with known *Ser*L in NJ tree were also found grouped with known *Ser*L in ML tree. Moreover, the topology of *Ser*L sub-branches in ML tree is comparable to NJ tree. *Ser*L branch consists of three sub-branches with good bootstrap support in both NJ tree (96.1%, 100%, 96.3%) and ML tree (82%, 82%, 42%). Overall, ML tree confirms the tree topology observed in NJ tree. Therefore, *Ser* candidates grouped with known *Ser*H, *Ser*L, and *Ser*J by phylogenetic analysis were classified into their corresponding subtypes.

The number of *Ser* candidates classified into each subtypes are as follows: *Ser*H, 30; *Ser*L, 34; *Ser*J, 10; L*, 17; J*, 7. For the larger group such as *Ser*H and *Ser*L, phylogenetic analysis shows that these experimentally identified i-ags are

grouped together rather than disperse throughout the whole subtype clades. Among all molecularly characterized subtype, subtype-J is the only subtype that was not identified from *T. thermophila* inbred strain B (Table 1.3). However, the *Ser* candidate identification algorithm described above is able to pick up candidates similar to known subtype-J, suggesting that *Ser* gene repertoire in *T. thermophila* is extensive and may not be strain-specific.

Multiple sequence alignment of *Ser* candidates of each subtype were inspected for the presence of repetitive block, number of cysteine residues in each repetitive block, and sequence features that are unique for each subtypes and are present in every repetitive block (Figure 1.7). Based on such features of known *Ser*, subtype classification by phylogenetic analysis was confirmed for each *Ser* candidate. Grouping of J* and L* was also confirmed by presence of repetitive block and unique sequence feature that is different from subtype-J and subtype-L, respectively.

One hundred and eighteen (55%) of the *Ser* candidates cannot be classified as any subtypes due to lack of phylogenetic association with known *Ser* subtypes. Because known *Ser* subtypes could only be a portion of *Ser* subtypes, the possibility that these candidates are actual *Ser* genes are not excluded. Instead, the candidates were annotated as ‘unclassified’ subtype. One inconsistency between NJ and ML trees was found in *Ser* candidate in ‘unclassified’ category. TTHERM_00329880 is grouped with unclassified *Ser* candidates in NJ tree, but it is grouped with subtype-H *Ser* candidates in ML tree (Figure 3.3, Figure 3.4).

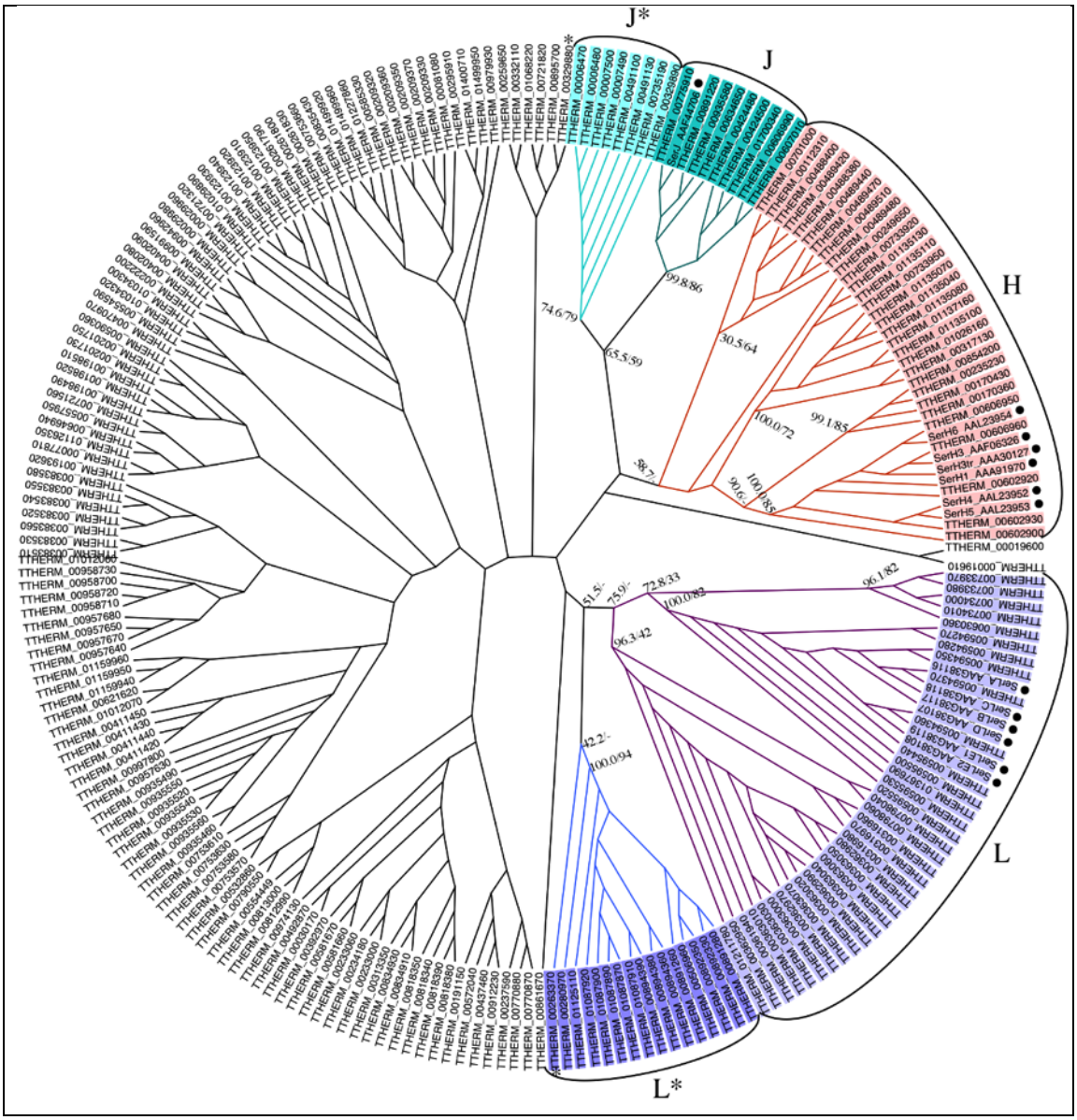


Figure 3.3 Neighbor-Joinging (NJ) cladogram of Ser candidates. Known Ser proteins are also included in this figure (black dots). 100-replicate bootstrapping was performed. Group-assigned candidates are highlighted. Bootstrap support percentages for the Neighbor- Joining (NJ) tree and the Maximum Likelihood (ML) tree (dash indicates undetermined bootstrap support value) are respectively shown on selected branch nodes. Among group-assigned candidates, TTHERM_00263370 and TTHERM_00329880 (marked with asterisk) are inconsistent between NJ and ML trees.

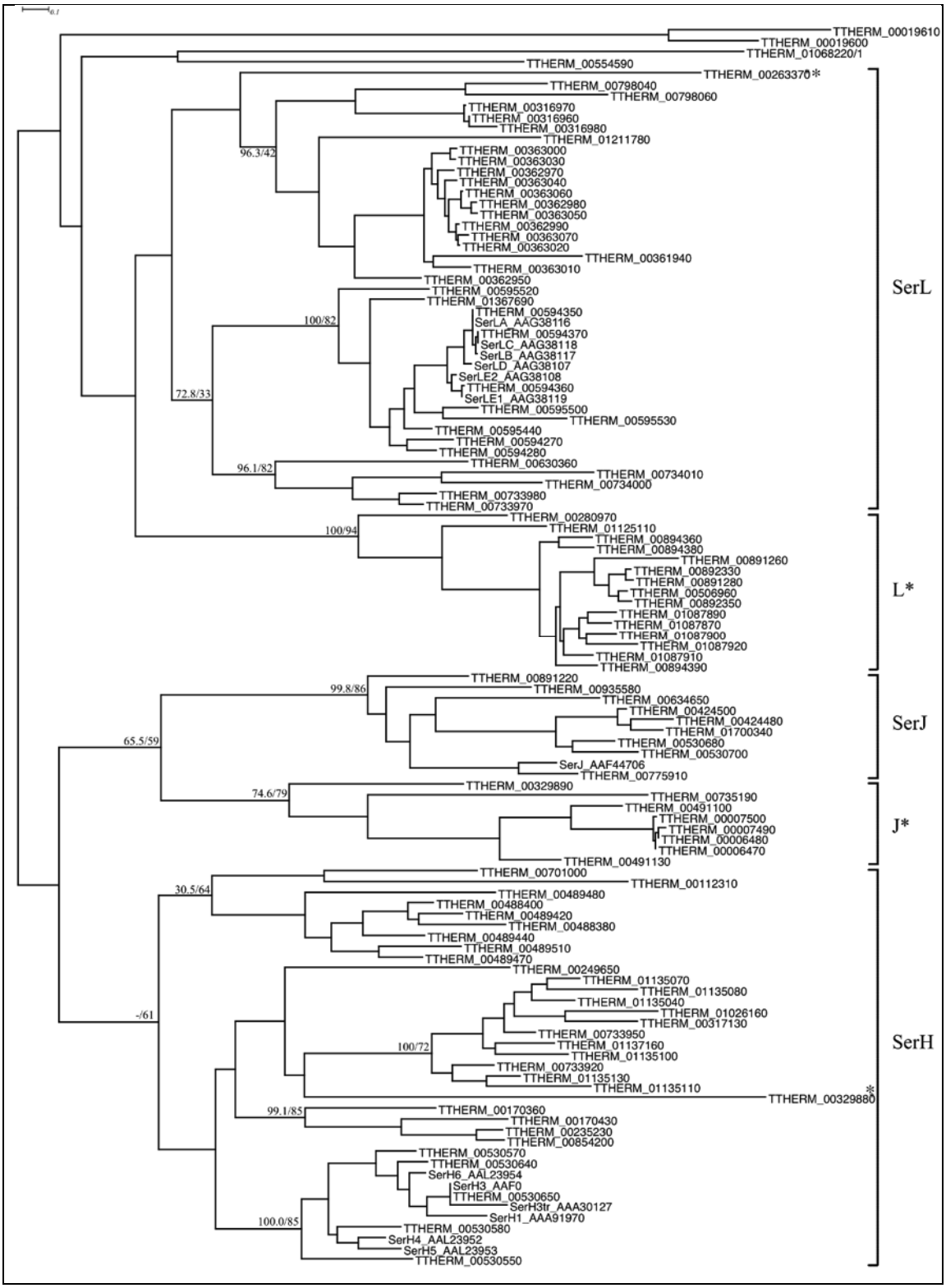


Figure 3.4 Unrooted Maximum Likelihood (ML) phylogram of subtype-classified *Ser* candidates. Known Ser proteins are also included in this figure. Percentage boot -strap support of Neighbor Joining (NJ) and ML tree were indicated on the selected

branches. Dash indicates the undetermined bootstrap support value. Inconsistency between ML and NJ tree is marked with asterisk.

3.3 *Ser* chromosomal location

Ser candidates were found distributed on 76 out of the estimated 250-300 MAC scaffold in *T. thermophila*. For subtype-classified *Ser* candidates, majority of them (84 candidates, 85.7%) are located in close proximity to one another. Few (14 candidates, 14.3%) could be found as a sole *Ser* gene on one scaffold (referred to as 'lone *Ser*'). Here, *Ser* gene tandem was defined as an array of *Ser* genes located within 10 kb from the adjacent *Ser* genes. *Ser* candidate genes tend to form a tandem array of the same subtype. In addition, *Ser* genes in close proximity on the same scaffold tend to have the same orientation. Two exceptions are observed: subtype-H TTHERM_00602920 on Scaffold 84 and subtype-L TTHERM_00595520 on Scaffold 3835, which lie in opposite orientation from surrounding *Ser* genes (Figure 3.5). There are six MAC scaffolds that contain tandem arrays with only one subtype of *Ser* gene (Scaffold 38, array of 13 subtype-L *Ser* genes; Scaffold 47, array of 2 subtype-J *Ser* genes; Scaffold 135, array of 2 subtype-L *Ser* genes; Scaffold 227, array of 5 subtype L* *Ser* genes; Scaffold 249, array of 7 subtype-H *Ser* genes; Scaffold 3835, array of 9 subtype-L *Ser* genes).

The number of genes in *Ser* tandem array is varied from two to thirteen genes. It is notable that *Ser* subtype with high number of candidates identified, such as subtype-H (30 candidates) and subtype-L (34 candidates), were found to form a tandem of multiple genes. In contrast, *Ser* subtype with fewer number of candidates, such as subtype-J (10 candidates) and subtype J* (7 candidates), were found as an array of two genes. Larger gene tandem (>2 genes) was not observed for subtype-J and group J*. There appears to be no preference for scaffold size or chromosomal region where *Ser* genes are located. *Ser* genes tandems were found at both the center and the end of MAC scaffold.

The sequence similarity of genes in the same tandem is higher than genes located in other tandem. Overall, *Ser* genes of the same subtype have approximately 20% - 30% sequence identity. *Ser* genes that are located in the same tandem array have sequence similarity above 60% with average about 70% to 80%, which is higher than the same *Ser* subtype located in other tandem. Percent sequence identity above 90% can also be found among *Ser* genes in the same tandem array. For example,

TTHERM_00363060 that is a subtype-L genes in a tandem of 13 genes located on Scaffold 38 have 58% sequence identity to TTHERM_00361940 and TTHERM_00362950 that are located in the same tandem, and have 80% - 94% sequence identity to other ten genes in the this tandem. Similar result is observed from other genes in this tandem, as well as in other *Ser* tandem regardless of *Ser* subtype. A few inter-tandem pair of highly similar *Ser* genes is also observed. Two subtype-H *Ser* genes located on Scaffold 120 (TTHERM_00733920 and TTHERM_00733950) have about 70% sequence similarity to subtype-H *Ser* genes on Scaffold 249. Similar observation was found with subtype-L* *Ser* genes on Scaffold 158 and 3723. Lone subtype-L* *Ser* gene located on Scaffold 3723 (TTHERM_00506960) has more than 64% similarity to *Ser* genes located as tandem on Scaffold 158 that also has the same degree of sequence similarity among them. The result suggests that gene duplication is the source of *Ser* gene expansion in *T. thermophila*.

The distribution of unclassified *Ser* candidates is comparable to the subtype-classified group. Out of 118 unclassified *Ser* candidates, 93 out of them (78.8%) were found to form a *Ser* gene tandem array, and 26 candidates (22%) were found as lone *Ser*. Similarity in physical gene distribution suggests that they are Ser candidates of unknown subtype.

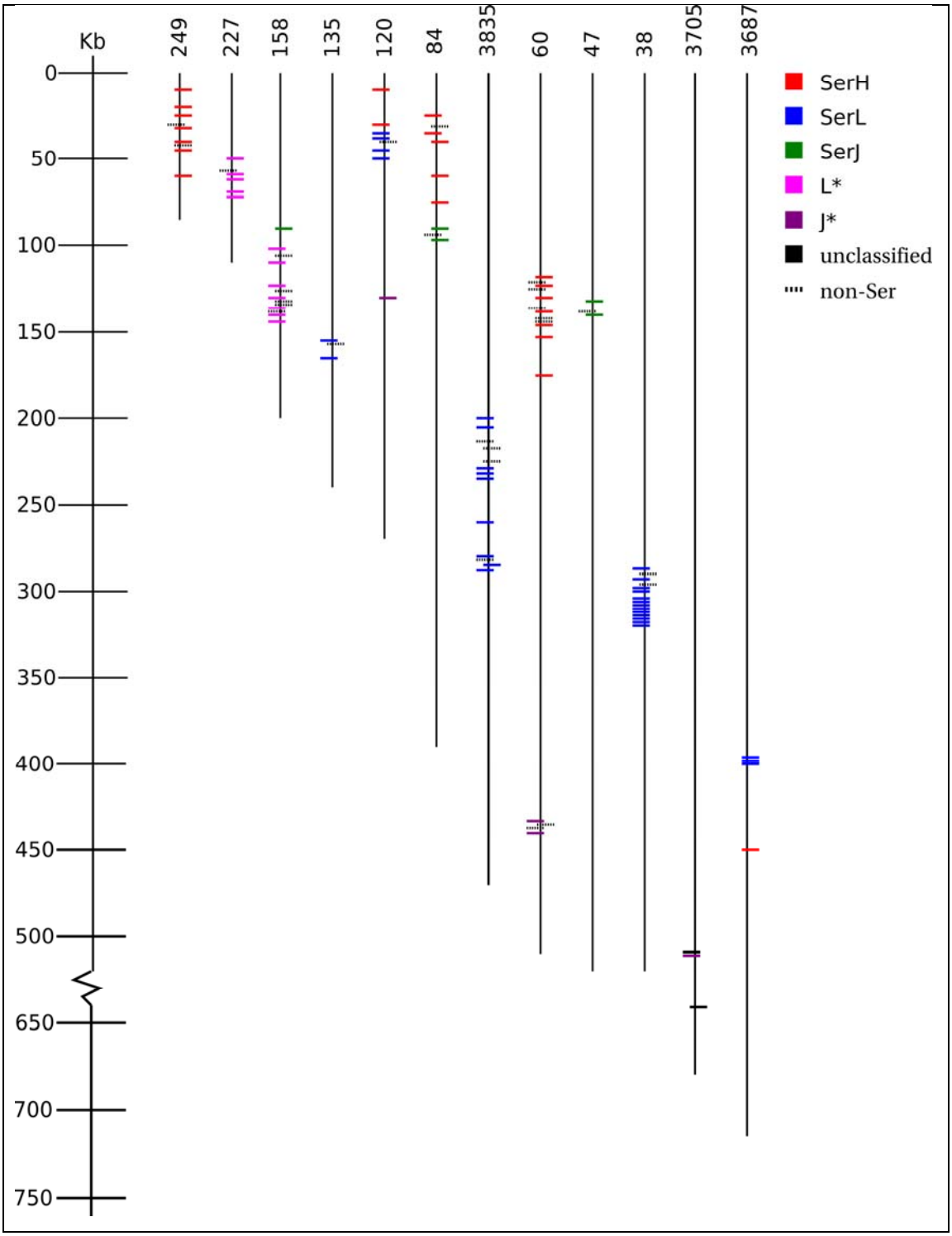


Figure 3.5 Distribution of subtype-classified *Ser* candidates on *T. thermophila* MAC scaffolds as gene tandem array. Each short horizontal line represents one *Ser* gene. *Ser* subtype is color-coded. Dash line represents non-*Ser* gene that locates within *Ser* tandem arrays. Gene orientation on plus or minus strand is depicted by left or right

alignment respectively. Scaffold number is shown on the top of each scaffold line (vertical line). Only scaffolds containing classified *Ser* tandem array are shown.

3.4 *Ser* gene expression analysis

Ser expression patterns during growth, starvation and conjugation were analyzed based on DNA microarray data. Because no biological data suggests how many groups should *Ser* genes candidate be clustered, the number of clusters was determined based on the data itself without prior information. Unsupervised data classification method was considered, but was not employed in this study due to difficulty in defining the expression cluster. *Ser* expression data was grouped into 30 expression clusters by their expression patterns (Figure 3.7, Figure 3.8). The number of *Ser* candidates assigned into cluster is varied. The largest cluster contains 18 *Ser* candidates. Fifteen clusters contain more than 5 *Ser* candidates, and only one cluster contains a single *Ser* candidate with distinct expression pattern (up-regulation during starvation and conjugation).

No clear correlation was found among specific *Ser* subtypes, chromosomal location, and expression patterns. For certain *Ser* tandems, majority of *Ser* genes located in the same tandem were associated with the same expression pattern. For example, ten out of thirteen subtype-L *Ser* genes on scaffold 38 were found in expression cluster 4 (Figure 3.6, upper panel), which shows two up-regulation peaks during early starvation (0^{th} hour) and during early conjugation (0^{th} hour). The other three *Ser* candidates in scaffold 38 exhibit different expression pattern. TTHERM_00362950, one subtype-L gene located in subtype-L genes tandem on scaffold 38, is up-regulated at late time points during conjugation (8^{th} to 12^{th} hour) and was assigned into expression cluster 2. The other two genes (TTHERM_00361940 and TTHERM_00363010) have prolonged up-regulation at different time points during conjugation (stable up-regulation during conjugation, and slight up-regulation at 2^{nd} to 8^{th} hour) and were assigned to expression cluster 21 and expression cluster 6, respectively. Such observation was found in tandem of all *Ser* subtypes, including the unclassified *Ser* candidates. For example, six out of seven unclassified *Ser* candidates located as a tandem on scaffold 169 were classified into expression cluster 6 (up-regulation during growth and conjugation), and another *Ser* candidate was classified into expression cluster 22 (higher up-regulation during growth compared to expression cluster 22). Investigation of 5'-UTR region of subtype-L candidates in this tandem (i.e.

tandem of 13 genes on scaffold 38) did not strongly suggest UTR region as sole factor of gene expression control. Percent identity of UTR sequence of *Ser* candidates in this tandem is approximately 40% - 50% regardless of the expression pattern.

For some *Ser* tandems, their *Ser* candidates were not in the same expression cluster, but their expression patterns appeared to be stage-specific. For example, scaffold 60 contains a tandem array of 6 subtype-H genes grouped into four different expression clusters (expression cluster 1: TTHERM_00488380, TTHERM_00489440; expression cluster 14: TTHERM_00489420, TTHERM_00489470; expression cluster 20: TTHERM_00489480; expression cluster 21: TTHERM_00488400) (Figure 3.6, lower panel). Expression cluster 1 has up-regulation peak during conjugation at the 8^{th} and 10^{th} hour time point. Expression cluster 14 has up-regulation peak during conjugation at the 10^{th} and 12^{th} hour time point. Expression cluster 20 has up-regulation peak during the 12^{th} to 18^{th} hour time point. Expression cluster 21 has slight up-regulation during the 2^{nd} and 12^{th} hour time point. Therefore, *Ser* candidates located in tandem on scaffold 60 are expressed during conjugation but at different time points.

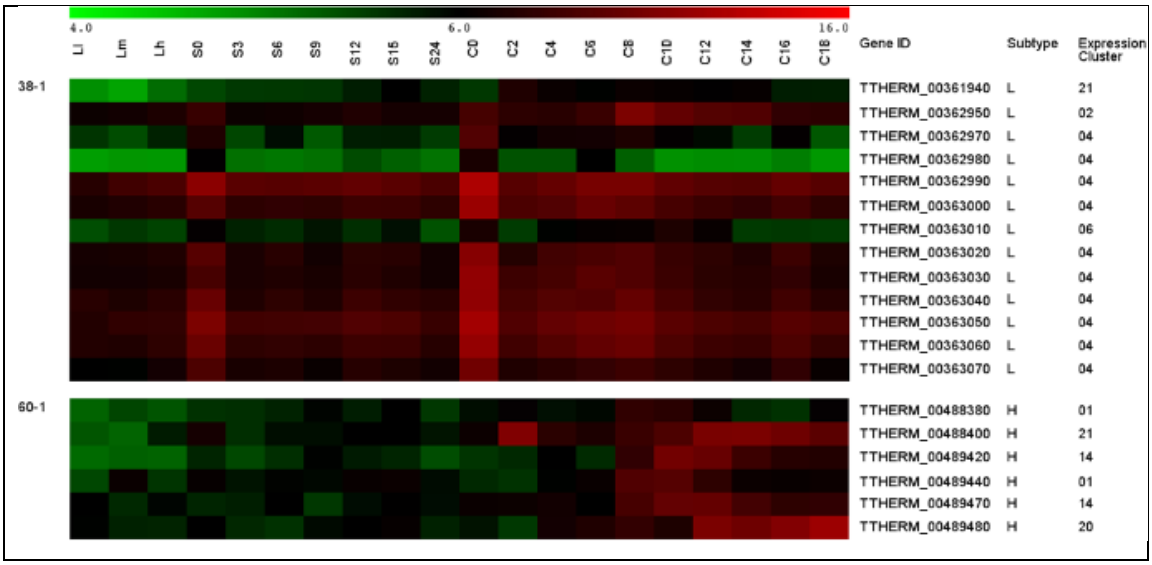


Figure 3.6 Heat map of gene expression of *Ser* candidates forming gene tandem on scaffold 38 and 60. *T. thermophila* gene expression data (Miao et al. 2009) was retrieved and only data of Ser candidates were selected for further analysis. Red indicates high level of expression, and green indicates low level of expression compared to median value. Scale bar represents log2-transformed gene expression value. Gene expression data was collected during growth at low (Ll), middle (Lm), and high (Lh) cell density, during starvation at seven time points (S0, S3, S6, S9, S12, S15, S24), and during conjugation at ten time points (C0, C2, C4, C6, C8, C10, C12, C14, C16, C18). Data is sorted by physical location on chromosome, not gene expression profile. Gene location is indicated at the leftmost column in ####-# format with the first four digits representing the scaffold ID and the last digit indicating the numerical code of the tandem array.

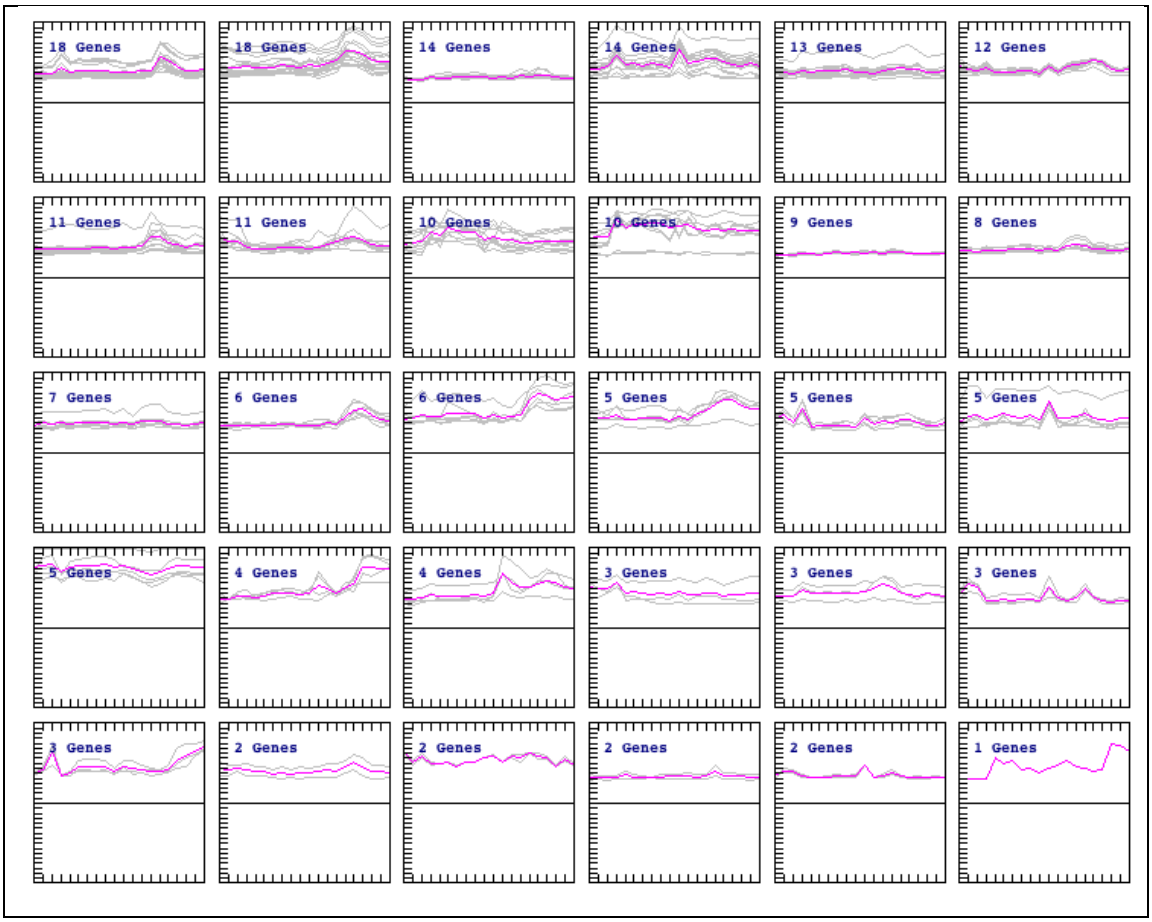


Figure 3.7 Expression profile of each *Ser* expression cluster. *T. thermophila* gene expression data (Miao et al. 2009) was retrieved and only data of Ser candidates were selected for further analysis. The expression profiles of Ser candidates were clustered and visualized using MeV software. Each grey line is the individual *Ser* gene expression profile. Pink line is the trend line of each expression cluster. The number of *Ser* candidates in each cluster is denoted.

Figure 3.8 *Ser* gene expression clusters. *T. thermophila* gene expression data (Miao et al. 2009) was retrieved and only data of Ser candidates were selected for further analysis. Each row represents one gene. Gene ID and subtype were listed. Unclassified *Ser* candidates are marked as X. Gene expression data was subjected to K-means clustering method using Pearson correlation to measure distance. Expression cluster ID is indicated as Arabic numeral on the left of expression heat map. Scale bar represents log2-transformed gene expression value. Red indicates expression value above median, and green indicates expression value below median.

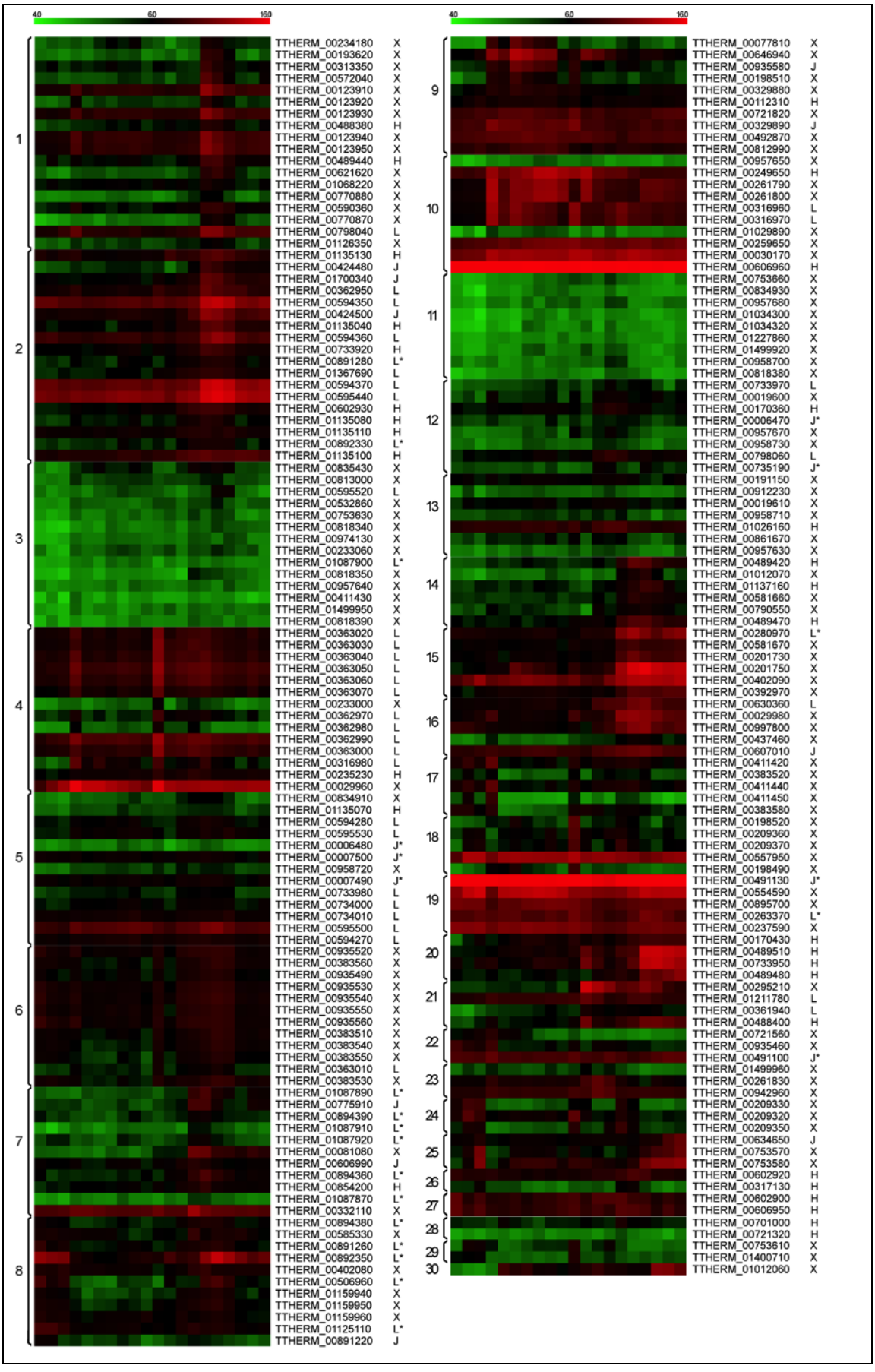

40
60
160
1
TTHERM_00234180 X
TTHERM_00193620 X
TTHERM_00313350 X
TTHERM_00572040 X
TTHERM_00123910 X
TTHERM_00123920 X
TTHERM_00123930 X
TTHERM_00488380 H
TTHERM_00123940 X
TTHERM_00123950 X
TTHERM_00489440 H
TTHERM_00621620 X
TTHERM_01068220 X
TTHERM_00770880 X
TTHERM_00590360 X
TTHERM_00770870 X
TTHERM_00798040 L
TTHERM_01126350 X
2
TTHERM_01135130 H
TTHERM_00424480 J
TTHERM_01700340 J
TTHERM_00362950 L
TTHERM_00594350 L
TTHERM_00424500 J
TTHERM_01135040 H
TTHERM_00594360 L
TTHERM_00733920 H
TTHERM_00891280 L*
TTHERM_01367690 L
TTHERM_00594370 L
TTHERM_00595440 L
TTHERM_00602930 H
TTHERM_01135080 H
TTHERM_01135110 H
TTHERM_00892330 L*
TTHERM_01135100 H
3
TTHERM_00835430 X
TTHERM_00813000 X
TTHERM_00595520 L
TTHERM_00532860 X
TTHERM_00753630 X
TTHERM_00818340 X
TTHERM_00974130 X
TTHERM_00233060 X
TTHERM_01087900 L*
TTHERM_00818350 X
TTHERM_00957640 X
TTHERM_00411430 X
TTHERM_01499950 X
TTHERM_00818390 X
4
TTHERM_00363020 L
TTHERM_00363030 L
TTHERM_00363040 L
TTHERM_00363050 L
TTHERM_00363060 L
TTHERM_00363070 L
TTHERM_00233000 X
TTHERM_00362970 L
TTHERM_00362980 L
TTHERM_00362990 L
TTHERM_00363000 L
TTHERM_00316980 L
TTHERM_00235230 H
TTHERM_00029960 X
5
TTHERM_00834910 X
TTHERM_01135070 H
TTHERM_00594280 L
TTHERM_00595530 L
TTHERM_00006480 J*
TTHERM_00007500 J*
TTHERM_00958720 X
TTHERM_00007490 J*
TTHERM_00733980 L
TTHERM_00734000 L
TTHERM_00734010 L
TTHERM_00595500 L
TTHERM_00594270 L
6
TTHERM_00935520 X
TTHERM_00383560 X
TTHERM_00935490 X
TTHERM_00935530 X
TTHERM_00935540 X
TTHERM_00935550 X
TTHERM_00935560 X
TTHERM_00383510 X
TTHERM_00383540 X
TTHERM_00383550 X
TTHERM_00363010 L
TTHERM_00383530 X
7
TTHERM_01087890 L*
TTHERM_00775910 J
TTHERM_00894390 L*
TTHERM_01087910 L*
TTHERM_01087920 L*
TTHERM_00081080 X
TTHERM_00606990 J
TTHERM_00894360 L*
TTHERM_00854200 H
TTHERM_01087870 L*
TTHERM_00332110 X
8
TTHERM_00894380 L*
TTHERM_00585330 X
TTHERM_00891260 L*
TTHERM_00892350 L*
TTHERM_00402080 X
TTHERM_00506960 L*
TTHERM_01159940 X
TTHERM_01159950 X
TTHERM_01159960 X
TTHERM_01125110 L*
TTHERM_00891220 J
40
60
160
9
TTHERM_00077810 X
TTHERM_00646940 X
TTHERM_00935580 J
TTHERM_00198510 X
TTHERM_00329880 X
TTHERM_00112310 H
TTHERM_00721820 X
TTHERM_00329890 J
TTHERM_00492870 X
TTHERM_00812990 X
10
TTHERM_00957650 X
TTHERM_00249650 H
TTHERM_00261790 X
TTHERM_00261800 X
TTHERM_00316960 L
TTHERM_00316970 L
TTHERM_01029890 X
TTHERM_00259650 X
TTHERM_00030170 X
TTHERM_00606960 H
11
TTHERM_00753660 X
TTHERM_00834930 X
TTHERM_00957680 X
TTHERM_01034300 X
TTHERM_01034320 X
TTHERM_01227860 X
TTHERM_01499920 X
TTHERM_00958700 X
TTHERM_00818380 X
12
TTHERM_00733970 L
TTHERM_00019600 X
TTHERM_00170360 H
TTHERM_00006470 J*
TTHERM_00957670 X
TTHERM_00958730 X
TTHERM_00798060 L
TTHERM_00735190 J*
13
TTHERM_00191150 X
TTHERM_00912230 X
TTHERM_00019610 X
TTHERM_00958710 X
TTHERM_01026160 H
TTHERM_00861670 X
TTHERM_00957630 X
14
TTHERM_00489420 H
TTHERM_01012070 X
TTHERM_01137160 H
TTHERM_00581660 X
TTHERM_00790550 X
TTHERM_00489470 H
15
TTHERM_00280970 L*
TTHERM_00581670 X
TTHERM_00201730 X
TTHERM_00201750 X
TTHERM_00402090 X
TTHERM_00392970 X
16
TTHERM_00630360 L
TTHERM_00029980 X
TTHERM_00997800 X
TTHERM_00437460 X
TTHERM_00607010 J
17
TTHERM_00411420 X
TTHERM_00383520 X
TTHERM_00411440 X
TTHERM_00411450 X
TTHERM_00383580 X
18
TTHERM_00198520 X
TTHERM_00209360 X
TTHERM_00209370 X
TTHERM_00557950 X
TTHERM_00198490 X
19
TTHERM_00491130 J*
TTHERM_00554590 X
TTHERM_00895700 X
TTHERM_00263370 L*
TTHERM_00237590 X
20
TTHERM_00170430 H
TTHERM_00489510 H
TTHERM_00733950 H
TTHERM_00489480 H
21
TTHERM_00295210 X
TTHERM_01211780 L
TTHERM_00361940 L
TTHERM_00488400 H
22
TTHERM_00721560 X
TTHERM_00935460 X
TTHERM_00491100 J*
23
TTHERM_01499960 X
TTHERM_00261830 X
TTHERM_00942960 X
24
TTHERM_00209330 X
TTHERM_00209320 X
TTHERM_00209350 X
25
TTHERM_00634650 J
TTHERM_00753570 X
TTHERM_00753580 X
26
TTHERM_00602920 H
TTHERM_00317130 H
27
TTHERM_00602900 H
TTHERM_00606950 H
28
TTHERM_00701000 H
TTHERM_00721320 H
29
TTHERM_00753610 X
TTHERM_01400710 X
30
TTHERM_01012060 X

Table 3.1 Gene ID of subtype-classified *Ser* candidates listed by their location and expression cluster. Unclassified *Ser* candidates located on the same MAC scaffold are also shown and marked as X. Location is indicated in ####-# format with the first four digits representing the scaffold ID and the last digit indicating the numerical code of the tandem array. Tandem array is defined as an array of *Ser* genes located no further than 10 kb from other *Ser* genes. Expression cluster as determined by K-mean clustering method (KMC) for each *Ser* candidate is also listed.

Gene ID	Location	Subtype	Expression Cluster
TTHERM_00006470	1-1	J*	12
TTHERM_00006480	1-1	J*	05
TTHERM_00007490	1-1	J*	05
TTHERM_00007500	1-1	J*	05
TTHERM_00019600	1-2	X	12
TTHERM_00019610	1-2	X	13
TTHERM_00733920	120-1	SerH	02
TTHERM_00733950	120-2	SerH	20
TTHERM_00733970	120-2	SerL	12
TTHERM_00733980	120-2	SerL	05
TTHERM_00734000	120-2	SerL	05
TTHERM_00734010	120-2	SerL	05
TTHERM_00735190	120-3	J*	12
TTHERM_00798040	135-1	SerL	01
TTHERM_00798060	135-1	SerL	12
TTHERM_00170360	14-1	SerH	12
TTHERM_00170430	14-2	SerH	20

Table 3.1 Gene ID of subtype-classified *Ser* candidates listed by their location and expression cluster. (cont.)

Gene ID	Location	Subtype	Expression Cluster
TTHERM_00891220	158-1	SerJ	08
TTHERM_00891260	158-2	L*	08
TTHERM_00891280	158-2	L*	02
TTHERM_00892330	158-3	L*	02
TTHERM_00892350	158-3	L*	08
TTHERM_00894360	158-3	L*	07
TTHERM_00894380	158-3	L*	08
TTHERM_00894390	158-3	L*	07
TTHERM_00935460	169-1	X	22
TTHERM_00935490	169-1	X	06
TTHERM_00935520	169-1	X	06
TTHERM_00935530	169-1	X	06
TTHERM_00935540	169-1	X	06
TTHERM_00935550	169-1	X	06
TTHERM_00935560	169-1	X	06
TTHERM_00935580	169-2	SerJ	09
TTHERM_00233000	19-1	X	04
TTHERM_00233060	19-2	X	03
TTHERM_00234180	19-3	X	01
TTHERM_00235230	19-4	SerH	04

Table 3.1 Gene ID of subtype-classified *Ser* candidates listed by their location and expression cluster. (cont.)

Gene ID	Location	Subtype	Expression Cluster
TTHERM_01087870	227-1	L*	07
TTHERM_01087890	227-1	L*	07
TTHERM_01087900	227-1	L*	03
TTHERM_01087910	227-1	L*	07
TTHERM_01087920	227-1	L*	07
TTHERM_00259650	23-1	X	10
TTHERM_00261790	23-2	X	10
TTHERM_00261800	23-2	X	10
TTHERM_00261830	23-2	X	23
TTHERM_00263370	23-3	L*	19
TTHERM_01135040	249-1	SerH	02
TTHERM_01135070	249-2	SerH	05
TTHERM_01135080	249-2	SerH	02
TTHERM_01135100	249-2	SerH	02
TTHERM_01135110	249-2	SerH	02
TTHERM_01135130	249-2	SerH	02
TTHERM_01137160	249-3	SerH	14
TTHERM_00316960	3687-1	SerL	10
TTHERM_00316970	3687-1	SerL	10
TTHERM_00316980	3687-1	SerL	04
TTHERM_00317130	3687-2	SerH	26

Table 3.1 Gene ID of subtype-classified *Ser* candidates listed by their location and expression cluster. (cont.)

Gene ID	Location	Subtype	Expression Cluster
TTHERM_00329880	3705-1	X	09
TTHERM_00329890	3705-1	SerJ	09
TTHERM_00332110	3705-2	X	07
TTHERM_00361940	38-1	SerL	21
TTHERM_00362950	38-1	SerL	02
TTHERM_00362970	38-1	SerL	04
TTHERM_00362980	38-1	SerL	04
TTHERM_00362990	38-1	SerL	04
TTHERM_00363000	38-1	SerL	04
TTHERM_00363010	38-1	SerL	06
TTHERM_00363020	38-1	SerL	04
TTHERM_00363030	38-1	SerL	04
TTHERM_00363040	38-1	SerL	04
TTHERM_00363050	38-1	SerL	04
TTHERM_00363060	38-1	SerL	04
TTHERM_00363070	38-1	SerL	04
TTHERM_00112310	3812-1	SerH	09
TTHERM_00123910	3812-2	X	01
TTHERM_00123920	3812-2	X	01
TTHERM_00123930	3812-2	X	01
TTHERM_00123940	3812-2	X	01

Table 3.1 Gene ID of subtype-classified *Ser* candidates listed by their location and expression cluster. (cont.)

Gene ID	Location	Subtype	Expression Cluster
TTHERM_00123930	3812-2	X	01
TTHERM_00123940	3812-2	X	01
TTHERM_00123950	3812-2	X	01
TTHERM_00594270	3835-1	SerL	05
TTHERM_00594280	3835-1	SerL	05
TTHERM_00594350	3835-1	SerLA	02
TTHERM_00594360	3835-2	SerLE1	02
TTHERM_00594370	3835-2	SerLC	02
TTHERM_00595440	3835-3	SerL	02
TTHERM_00595500	3835-4	SerL	05
TTHERM_00595520	3835-4	SerL	03
TTHERM_0059530	3835-4	SerL	05
TTHERM_00424480	47-1	SerJ	02
TTHERM_00424500	47-1	SerJ	02
TTHERM_00488380	60-1	SerH	01
TTHERM_00488400	60-1	SerH	21
TTHERM_00489420	60-1	SerH	14
TTHERM_00489440	60-1	SerH	01
TTHERM_00489470	60-1	SerH	14
TTHERM_00489480	60-1	SerH	20
TTHERM_00489510	60-2	SerH	20

Table 3.1 Gene ID of subtype-classified *Ser* candidates listed by their location and expression cluster. (cont.)

Gene ID	Location	Subtype	Expression Cluster
TTHERM_00489480	60-1	SerH	20
TTHERM_00489510	60-2	SerH	20
TTHERM_00491100	60-3	J*	22
TTHERM_00491130	60-3	J*	19
TTHERM_00602900	84-1	SerH	27
TTHERM_00602920	84-1	SerH	26
TTHERM_00602930	84-1	SerH	02
TTHERM_00606950	84-2	SerH	27
TTHERM_00606960	84-3	SerH	10
TTHERM_00606990	84-4	SerJ	07
TTHERM_00607010	84-4	SerJ	16
TTHERM_00701000	109	SerH	28
TTHERM_00775910	129	SerJ	07
TTHERM_00854200	151	SerH	07
TTHERM_00861670	153	L*	13
TTHERM_01026160	202	SerH	13
TTHERM_00249650	22	SerH	10
TTHERM_01125110	245	L*	08
TTHERM_01211780	275	SerL	21
TTHERM_01367690	350	SerL	02
TTHERM_00280970	3694	L*	15

Table 3.1 Gene ID of subtype-classified *Ser* candidates listed by their location and expression cluster. (cont.)

Gene ID	Location	Subtype	Expression Cluster
TTHERM_01367690	350	SerL	02
TTHERM_00280970	3694	L*	15
TTHERM_00506960	3723	L*	08
TTHERM_00630360	3831	SerL	16
TTHERM_01700340	573	SerJ	02
TTHERM_00634650	90	SerJ	25

Table 3.2 Gene ID of unclassified *Ser* candidates listed by their location and expression cluster. The table includes the rest of the unclassified *Ser* candidates with the same labeling system in Table 3.1

Gene ID	Location	Subtype	Expression Cluster
TTHERM_00721320	115-1	X	28
TTHERM_00721560	115-1	X	22
TTHERM_00721820	115-2	X	09
TTHERM_00753570	122-1	X	25
TTHERM_00753580	122-1	X	25
TTHERM_00753610	122-2	X	29
TTHERM_00753630	122-2	X	03
TTHERM_00753660	122-2	X	11
TTHERM_00770870	127-1	X	01
TTHERM_00770880	127-1	X	01
TTHERM_00812990	138-1	X	09
TTHERM_00813000	138-1	X	03
TTHERM_00818340	139-1	X	03
TTHERM_00818350	139-1	X	03
TTHERM_00818380	139-1	X	11
TTHERM_00818390	139-1	X	03
TTHERM_00834910	145-1	X	05
TTHERM_00834930	145-1	X	11
TTHERM_00835430	145-2	X	03

Table 3.2 Gene ID of unclassified *Ser* candidates listed by their location and expression cluster. (cont.)

Gene ID	Location	Subtype	Expression Cluster
TTHERM_00198490	17-1	X	18
TTHERM_00198510	17-1	X	09
TTHERM_00198520	17-1	X	18
TTHERM_00201730	17-2	X	15
TTHERM_00201750	17-3	X	15
TTHERM_00209320	17-4	X	24
TTHERM_00209330	17-4	X	24
TTHERM_00209350	17-4	X	24
TTHERM_00209360	17-4	X	18
TTHERM_00209370	17-4	X	18
TTHERM_00957630	177-1	X	13
TTHERM_00957640	177-1	X	03
TTHERM_00957650	177-1	X	10
TTHERM_00957670	177-2	X	12
TTHERM_00957680	177-2	X	11
TTHERM_00958700	177-3	X	11
TTHERM_00958710	177-3	X	13
TTHERM_00958720	177-3	X	05
TTHERM_00958730	177-3	X	12
TTHERM_01012060	196-1	X	30
TTHERM_01012070	196-1	X	14

Table 3.2 Gene ID of unclassified *Ser* candidates listed by their location and expression cluster. (cont.)

Gene ID	Location	Subtype	Expression Cluster
TTHERM_01034300	205-1	X	11
TTHERM_01034320	205-1	X	11
TTHERM_00402080	3713-1	X	08
TTHERM_00402090	3713-1	X	15
TTHERM_00411420	3714-1	X	17
TTHERM_00411430	3714-1	X	03
TTHERM_00411440	3714-1	X	17
TTHERM_00411450	3714-1	X	17
TTHERM_00313350	3810-1	X	01
TTHERM_00572040	3810-2	X	01
TTHERM_01159940	3823-1	X	08
TTHERM_01159950	3823-1	X	08
TTHERM_01159960	3823-1	X	08
TTHERM_00029960	3825-1	X	04
TTHERM_00029980	3825-1	X	16
TTHERM_00030170	3825-2	X	10
TTHERM_00581660	3827-1	X	14
TTHERM_00581670	3827-1	X	15
TTHERM_00077810	3828-1	X	09
TTHERM_00081080	3828-2	X	07

Table 3.2 Gene ID of unclassified *Ser* candidates listed by their location and expression cluster. (cont.)

Gene ID	Location	Subtype	Expression Cluster
TTHERM_00383510	40-1	X	06
TTHERM_00383520	40-1	X	17
TTHERM_00383530	40-1	X	06
TTHERM_00383540	40-1	X	06
TTHERM_00383550	40-1	X	06
TTHERM_00383560	40-1	X	06
TTHERM_00383580	40-1	X	17
TTHERM_01499920	425-1	X	11
TTHERM_01499950	425-2	X	03
TTHERM_01499960	425-2	X	23
TTHERM_00790550	133	X	14
TTHERM_00895700	159	X	19
TTHERM_00193620	16	X	01
TTHERM_00912230	162	X	13
TTHERM_00942960	172	X	23
TTHERM_00974130	181	X	03
TTHERM_00997800	188	X	16
TTHERM_01029890	204	X	10
TTHERM_01068220	217	X	01
TTHERM_01126350	246	X	01
TTHERM_01227860	282	X	11

Table 3.2 Gene ID of unclassified *Ser* candidates listed by their location and expression cluster. (cont.)

Gene ID	Location	Subtype	Expression Cluster
TTHERM_01126350	246	X	01
TTHERM_01227860	282	X	11
TTHERM_00585330	3675	X	08
TTHERM_01400710	368	X	29
TTHERM_00191150	3697	X	13
TTHERM_00237590	3706	X	19
TTHERM_00437460	3708	X	16
TTHERM_00492870	3711	X	09
TTHERM_00590360	3726	X	01
TTHERM_00646940	3727	X	09
TTHERM_00557950	3735	X	18
TTHERM_00392970	3813	X	15
TTHERM_00532860	3830	X	03
TTHERM_00554590	72	X	19
TTHERM_00621620	86	X	01
TTHERM_00295910	92	X	21

3.5 Additional Ser candidates identified by homolog search approach

To uncover as much *Ser* candidates as possible, additional search was performed by using 216 Ser candidates as blastp query against NCBI non-redundant (nr) protein database. The search returned 736 blastp hits. After remove all blast hits that were already picked up by pattern search script, there are sixteen hits left. However, they appear to have either low similarity score or show degree of similarity for only a small segment (Table 3.3). These sixteen genes were manually inspected for cysteine repeat pattern and were subjected to GPI-anchored protein classification by FragAnchor. Further investigation revealed that these sixteen genes lack matching cysteine pattern, causing them to fail the initial *Ser* candidate search. Six of them were rejected by FragAnchor as non GPI-anchored proteins, while four of them were classified as 'highly probable' GPI-anchored proteins (Figure 3.9). Phylogenetic analysis showed that these sixteen genes are not grouped with any subtype-classified gene in the initial candidate set, and they cannot be assigned with any known subtypes. Manual inspection of cysteine pattern in these proteins shows that they failed the initial pattern search due to 1) having Cys pair spanning an interval longer than 120 residues, or 2) having very short sequence length (<200 amino acid residues) and less than 6 Cys residues.

3.6 *Ser* homolog in other ciliate species identified by homology search

Ser homology search approach was also employed in other ciliates. The search was performed by using 216 Ser candidates as blastp query against NCBI non-redundant (nr) protein database. Blast hits were then categorized by organism, subjected to FragAnchor, and manually inspected. The search returned weak hits from ciliate such as *Ichthyophthirius multifiliis* (AAK94941, E-value: 0.008) and *Paramecium tetraurelia* (XP_001450224, E-value: 2.1).

For the fish parasite *I. multifiliis*, 104 unique blastp hits were returned using *T. thermophila Ser* candidates as queries. Blast hits from *I. multifiliis* include proteins annotated as immobilization antigen, putative leishmanolysin family protein, and zinc finger LSD1 subclass family protein. The latter two groups are cysteine-rich

protein, but they lack GPI anchor signal. They are unlikely to be *T. thermophila Ser* homolog. One protein from *I. multifiliis* (accession: AAK94941) annotated as immobilization antigen has about 22% identity to subtype LC i-ag and about 21% similarity to subtype LB and subtype LD. This protein was experimentally characterized to be i-ag of *I. multifiliis* (Lin et al. 2002). It contains GPI anchor signal, and periodic cysteine pair pattern. The length of this protein is 460 amino acid residues, which is comparable to majority of *T. thermophila Ser* candidates.

For free-living ciliate *P. tetraurelia*, 80 unique blastp hits were returned using *T. thermophila Ser* candidates as queries. The majority (78 hits) of blast hits are annotated as hypothetical protein. According to manual inspection, most of them are cysteine-rich protein without periodic cysteine pattern or GPI anchor signal. One protein from *P. tetraurelia* (accession: XP_001450224) has approximately 22% identity to confirmed subtype H i-ag (SerH1: 23.39% identity; SerH4: 19.65% identity; SerH5: 21.52% identity; SerH6: 22.57% identity). Its length is 2,693 amino acid residues, which is about six times larger than usual *Ser*. It contains multiple regions named Paramecium Surface Antigen (PSA) Repeat. This sequence is aligned well with known *P. tetraurelia* surface antigen (sequence identity: A surface antigen, 62.98%; A51 surface antigen, 63.42%; G surface antigen, 70.29%; 51B surface antigen, 57.37%), suggesting that it is *P. tetraurelia* surface antigen. However, FragAnchor assigned XP_001450224 with low probability score for GPI anchor signal (‘potential false positive’ class). Manual inspection of amino acid sequence at its C terminus reveals that, despite containing small hydrophobic amino acid residues that is characteristic of GPI anchor signal, it also contains arginine, aspartate, and histidine that do not fit into the hydrophobic criterion of GPI anchor signal.

Table 3.3 Sixteen *T. thermophila* proteins found by blastp against NCBI protein database using *Ser* candidates as query. Each protein was subjected to GPI-anchored protein prediction program (FragAnchor) and was classified accordingly to probability score using HMM ('Highly probable', 'Probable', 'Weakly probable', 'HMM score < 0.20'). Protein without GPI signal recognized by FragAnchor was rejected before probability scoring.

Accesssion	Gene ID	Predicted GPI anchor classification
XP_001009023	TTHERM_00260760	Highly probable
XP_001015544	TTHERM_00383590	Highly probable
XP_001015023	TTHERM_00675510	Highly probable
XP_001015019	TTHERM_00673490	Highly probable
XP_001008873	TTHERM_00187070	Highly probable
XP_001015026	TTHERM_00675540	Probable
XP_001015018	TTHERM_00673480	Probable
XP_001015017	TTHERM_00673470	Probable
XP_001012443	TTHERM_01031270	HMM score < 0.20
XP_001029448	TTHERM_01503030	HMM score < 0.20
XP_001024289	TTHERM_00994410	Rejected
XP_001028396	TTHERM_02444130	Rejected
XP_001007125	TTHERM_00209340	Rejected
XP_001016147	TTHERM_00818400	Rejected
XP_001012445	TTHERM_01032280	Rejected
XP_001033063	TTHERM_00471790	Rejected

>TTHERM_00260760
MKKAFILIFIGAIYSILSLVK**C**QTIN**C**NNHPIFIKSQNVIQNILKNDNVTSMNATGLYNV
LSPMYQSAIPKIAQDSDVGV**C**SDYVGDNS**CC**FKAFTQWIDNAALLKVLNIQSTNSAYQQL
VDNYVAYLYDGN**C**KQDFLPYNQVNNTAIYGNRNLKKFQSLRASLSAIKIGFAQNITSFTR
GVL**C**GM**C**AGVDVVQNYFNDKGLLKIQQSSLDYFIGNTTLNINYYLGNFTQQSISDVIDEF
NSGYIKISKPDGLT**C**ANYVKGKITQKFANLGISLQDSSGNLI**C**PGTIAFGDNSA**C**ENNLQ
GSPSLNPPSSSGRLLFQEIHDMVRMLQATVDAVGDTNIGINAVTNSTQTDSVIDEFGSNI
QPNFNTINP[NSS]NNLVFIVLFSLVALFM
>TTHERM_00383590
MNTYSIVLQISKRLINYKITVKK**C**YKEKYTTKNQKIKTALSNYIK**C**FLKFPPTEDNKNEG
DNKGEGENKDEGEMKDEEKKIEGNKNEEETKKDEGIKRNEENKK**C**KVEEGQKDEKRDN**C**S
KYQNGKTLK[SEQ]SVISSGNSLVFALTLSNLALLN
>TTHERM_00675510/1-168
MKSQLLVAFAIQLLLLVNADDN**C**NFTIDQLNNQVLKTKISSS**C**NN**C**FTSDVKLTIQSNTQQIISLKLTAIDA
SDPVQQTTLKPAEKSLTINPFFLIQYQSGGTYEAFNFMTT**C**QTKQINGNLVYNLVFNAKVNDNYQFFELNNG
YTTN[SLI]LQSLSYVALVLGVALLF
>TTHERM_00673490/1-169
MKLSFLITFTAYLLILTSAADNN**C**NFTIDQLNNKVLKTFKTQD**C**DN**C**FTSDVTLTIQSNSKQITSLKLTAVE
SSSPLQILTLQPFSETLILNTSFSIYYPLPNGFSTENLITS**C**QTKQINGKIAFDLTFMPKVNDHYYQIFTLN
NGYT[SSL]LLKSVASISMII**C**ILLFF
>TTHERM_00187070/1-164
MKKQILLTLAFYLLLIANASDNDDD**C**SFTVDQLNNQILKPDSTD**C**QN**C**FTSDVTMSIQADSQQITSITLTAN
DKSSPLQQLNLQPAEKSLTIDTSFRLQGQNMLYGYNFMNL**C**ATSKVNGKLTYVLGFQAKVQNTYQTFFLS[SE
G]VVLNLSYLILAL**C**SILLFV

Figure 3.9 Amino acid sequence of the four *T. thermophila* proteins that were identified by *Ser* homology search and were predicted to be 'highly probable' GPI-anchored protein by FragAnchor. The cysteine residues are in bold face. Predicted GPI anchor signal are underlined. Predicted GPI-attachment site is in bracket.

A.

```
AAK94941              MKFNILIILIISLFINELRAVNCPNGAAIANGQSDTGAADINTCTHCQKH 50
SerLC_AAG38118        MKATSLILISLAVIATVN---------ACTDTNATAGAGGTCFCN----- 36
                      ** . **:: :::: .           * :: :: :**..   *.

AAK94941              FYFNGGNPAGQAPGAVQFNPGVSQCIACQVHKADSQHRQGGDANLAAQCS 100
SerLC_AAG38           --------AGYYGTSTDVTP-SGSCTKCPTGTNSVAATASGTLVSSCTCN 77
                              **    :.:..*  ..*  * . . .     .*    :. *.

AAK94941              NLCPAGTAVEDGSPTFTQSLTQCVNCKPNFYFNGGNPTGQAPGAGQFDPT 150
SerLC_AAG38118        DTNASLKGDNSG-----------CQCKANFYG----------------- 98
                      :  .: .. :.*            :**.***

AAK94941              QLIANPDLANNPEVPNVSSPNGQCVACQVNKSDSQLRPGAQANLATQCNN 200
SerLC_AAG38118        ------------------------------------TPNAVSGGATGCS- 111
                                                       *.* :. ** *.

AAK94941              ECPTGTAIQDGAIFIYTQSISQCTFCKVDFYFNGGNPSAQNPGNGQFTPG 250
SerLC_AAG38118        ACPTGTTSPAG-----TAAVTSCACNDTNASLKGDNSGCQCKANFYGTP- 155
                       *****:   *     * :::.*:  ..:  ::*.*...*  .*   **

AAK94941              QLIANPDAATAAQIPMVPGPNSKCVACESKKTNSQSRSGLEANLAAQCGT 300
SerLC_AAG38118        --------------NAVAGGATGCTACP---TGSAAAAGSTAVTSCACN- 187
                                    *.*  : *.**    *.* : :*  *  :. *.

AAK94941              ECPAGTLVTDGVTPTYTVSLSQCVNCKAGFYQNSNFEAGKSQCNKCAVSK 350
SerLC_AAG38118        ----------DTNSSLKADNSACV-CKANFYGTPNAVAGGATG--CTACP 224
                                ....: ... * ** ***.** ..*  ** :    *:..

AAK94941              TGSASVPGNSATSATQCQNDCPAGTVVDDGTSTNFVALASECTKCQANFY 400
SerLC_AAG38118        TGTTSTAGTTVIGSCACP-----------DTNAALNSATPPVCQCKANFY 263
                      **::*..*.:. .:  *            .*.: : : :.   :*:****

AAK94941              ASKTSGFAAGTDTCTECSKKLTSGATAKVYAEATQKAQCASSTFAKFLSM 450
SerLC_AAG38118        GTPT---ASGASGCTACPSGQTAPAGS-----ATNVCKAASTSSTYILPI 305
                      .: *   *:*:. ** *..  *: * :     **: .:.**:: : :*.:

AAK94941              SLIFISFYLL 460
SerLC_AAG38118        ---------- 
```

B.

```
XP_001450224          CSAYVERLHCFKSRTNYCAWTDYSYCSQVNQPSECSFVTGLINLDHTKCQ 1000
SerH1_AAA91970        -------------------------------------------------


XP_001450224          LYHSSMALDVRSIRQLVLVMSATNNCATSGQGKCYFDGTDCMRFSNCASI 1050
SerH1_AAA91970        MQNKTLIICLIISQLLVSVFSVT----AGGAAQCPGTGANCNVAAACPVP 46
                      : :.:: : :   : ** *:*.*    :.* .:*   *::*   : *.

XP_001450224          TGTGLTTTICATYDAGCIANVDGTACQEKLATCDLYLTQKSCSTSAAAAT 1100
SerH1_AAA91970        TIQGTGTAACT---------------------------------WAAGTD 63
                      *  *  *: *:                                  **.:

XP_001450224          ADKCAWSGTACLAIVTANIATHCAYVTGTGLTDTTCAAYNVDCTANRAGT 1150
SerH1_AAA91970        LTQCTVTDCDCLTTGTVTGKTGLTDAFCT-----SCKGATQNLYANNAGT 108
                        :*: :.  **:  *..  *  : .  *     :* . . :  **.***

XP_001450224          ACQEQKATCDVYTTEATCSTSKATATADTCAWSGTDCRAVRTAYIATDCG 1200
SerH1_AAA91970        SCVAASKTCASGSRGTTAAN----------AWTAADC------------- 135
                      :*   . **   :  :*.:.          **:.:**

XP_001450224          YVIGTGLTDAICASYNPDCTANRAGTACQEQMSDCELYVTWLSCSTSKAA 1250
SerH1_AAA91970        ----------------------------------------LACTPATPI 144
                                                            *:*:.:..
```

```
XP_001450224          AIADKCAWSGTACLAIVTANIATHCAYVTGTGLTDTICAAYNADCTANRA 1300
SerH1_AAA91970        FVAAVSPATTTSCAACSTV----------TTGLTDSLCNACGTNASP--A 182
                       :*  .. : *:* *  *.           *****::* * .::.:.  *

XP_001450224          GTACQEQKATCDLYFTEATCSTSKAAATADKCVWSGTACLRVT-VVDIHC 1349
SerH1_AAA91970        NNKIFANAAGSACVASSLTCASGSRGTTAGN-AWTAADCLACTPATPIFV 231
                      ..    : * .    :. **::.. .:**.: .*:.: **  * .. *.

XP_001450224          AYVTETRLSDAICAQYNVECTVNKARTACQKKISICNEYADYSTCSQSFS 1399
SerH1_AAA91970        AAVSPS------------------------------------------- 237
                      * *: :

XP_001450224          SLCVWNGTSCISVTNATTDCEFITEIRLTDQICSMYNSGCISLKDGTACQ 1449
SerH1_AAA91970        -----ANTSCAACSTVTTG--------LTDSLCNACGTNASPANN----- 269
                            .*** : :..**.        ***.:*.  .:.. . ::

XP_001450224          EAKYKCQDYTTFNKCTIQTNDSRSCIWIDNSCYPISGVNCGAITGSGLDH 1499
SerH1_AAA91970        ---------------KIFANAAG------SACVASSATCAGGSRGTTLAN 298
                                    .* :* :       .:* . *.. .*.  *: * :

XP_001450224          AQCQAYSLSCTSIADGTRCQDFKQTCEQYPGTLLECTKSFTQKCYLVGST 1549
SerH1_AAA91970        AWTAADCLACTPATP---------------------------------- 313
                      *   * .*:**. :

XP_001450224          CITISDATTHCSKIFGAAGAITYEICQSYNPGCSVNRSRNACVLQQAQCY 1599
SerH1_AAA91970        -------------------------------------------------


XP_001450224          GYTELNSCYKSAAGLCITKNGGATCVAASSAATCEAIELGSGNYSSANCN 1649
SerH1_AAA91970        -----------------AVQLGASPATTSSCVACNTITSG---WTDANCN 343
                                      : : **: .::**..:*::*  *   ::.****

XP_001450224          EIKAGCTNNGTTGCVAKTCANAVDITFNHANCNNWINICTVNSGNTGCLQ 1699
SerH1_AAA91970        SCAMAASPQTKT-------------------------IVAKADGSACVA 367
                      .   ..: : .*                          ..::..:.*:

XP_001450224          MAFRCADQTQSSCLYSVEGECEIVGSSCVRRTCDTAPADINYDTDTECSS 1749
SerH1_AAA91970        AVYSCTQSARGS---------------------------NKWTDADCAA 389
                       .: *::.::.*                           *  **::*::

XP_001450224          YLQSCTVARSGGCQVRKACASYKSQIQCKLSSSGAKCFWNPTYKTCVDLT 1799
SerH1_AAA91970        CNGTAANANQYASADGSSCQATK--------ASGSSTFSGQIFVSILLVL 431
                         :.: *.. ..   .:* : *        :**:. * .  : : : :

XP_001450224          CSNIEATSLFDTHNKCVAVDSSLACTVKQSNGVAVPGCVERSRCSLYIVE 1849
SerH1_AAA91970        SALLI------------------------------------------- 436
                      .: :

XP_001450224          EQCKTDANGGACFWNTDATVFFPVCQDMSCSHAPTDLSTHNECYAYYNTP 1899
SerH1_AAA91970        -------------------------------------------------
```

Figure 3.10 Pairwise alignment of *T. thermophila Ser*-encoded i-ag with weak blastp hit from other ciliates. All alignments were performed by ClustalW. (A) AAK94941, a protein from *I. multifiliis*, with *T. thermophila* SerLC. (B) XP_001450224, a surface antigen from *P. tetraurelia*, with *T. thermophila* SerH1. Flanking regions at N- and C-termini are omitted.

CHAPTER IV
DISCUSSION

T. thermophila i-ag was originally identified as variant surface antigen. Their high degree of polymorphism has prevented systematic identification of its *Ser* family based on sequence homology alone despite available genomic data. Previous studies expected more *Ser* in *T. thermophila* genomes in addition to thirteen molecularly characterized *Ser*. Early study of *Ser* genes based on restriction fragment length polymorphism (RFLP) found that the polymorphism was greater at the 5'-terminus half of the genes, suggesting that there are unidentified *Ser* genes in *T. thermophila* genome (Gerber et al. 2002). Subtype-J which only has one gene characterized was reported to have paralogs (Doerder 2000). In this study, an alternative strategy was adopted for identifying this gene family by selecting two common features of known *Ser* genes, namely, the cysteine-rich motif and GPI-anchor site. Their phylogenetic distributions were analyzed in order to characterize *Ser* candidates that are related to the known subtypes. The approach identified *Ser* candidates that can be grouped into the known subtypes. Subtype-H, subtype-L, and subtype-J candidates were identified in addition to the original set of thirteen characterized *Ser* genes. The known (accession AAF44706) subtype-J, which was characterized from a natural isolate of *T. thermophila* (ANF814-1), was not found from MAC genome of *T. thermophila* strain SB310. This suggests that each strain of *T. thermophila* may carry a distinct subtype species, and analysis on different *T. thermophila* isolates will uncover more *Ser* candidates.

One of two criteria for Ser candidate selection is the presence of cysteine pattern. Based on amino acid sequences of thirteen known Ser, the pattern was set to '$CX_{(\geq 6)}CX_{(\geq 1)}CX_{(\geq 1)}CX_{(\geq 1)}CX_{(\geq 1)}C$' (in regular expression as '"C[^C]{6,}C[^C]+C[^C] +C[^C]+C[^C]+C"') (C = cysteine; X = any amino acid except C). However, the pattern chosen in this study did not completely represent the alternating short and long span of residues between cysteine pair. Search pattern used in this study would match

any proteins that have one pair of cysteine between a long span of amino acid residues that is followed by four pairs of cysteines with any length of amino acid residues span. The downside of this search pattern is the lesser degree of specificity and probably the increase number of false positive output from the pattern search script. It is possible to write a regular expression representing 'six cysteines with alternating short and long amino acid residues spans'. However, it is also arguable whether the cysteine pattern observed from data of thirteen identified Ser that is limited to only a portion of subtypes in Ser candidate search. Despite the observation that the shortest span of amino acid residues of subtype-H, subtype-L, and subtype-J Ser do not exceed six residues, it is not known whether Ser of other unidentified subtypes would deviate from this trend. In addition, the downside was compensated by the use of GPI-anchored protein prediction as the second Ser selection criterion, which successfully reduced the number of candidates from 4,925 to a workable number of 222 proteins.

The pattern search script was developed to find any protein containing one or more matching pattern within a limited range, but it did not take the characteristic repeat block into consideration. The reasons that the feature of repeat as not implemented into the pattern search script were 1) Ser with only one repeat block was found (e.g. SerLE1 and SerLE2); 2) the complexity in developing an algorithm to detect the repeat block; and 3) the requirement of prior information of unique sequence feature of repeat block for each subtype. Although the unique sequence feature of repeat block for classified subtypes is described in this study, such information is unknown for unclassified subtype. Moreover, the detection of repeat block would require the script to compare all the matching patterns from one input data (i.e. one *T. thermophila* protein sequence) whether they are the repeat block, and eventually consumes more time to process the data.

GPI anchor signal prediction is one critical step for *Ser* prediction algorithm. FragAnchor was selected as GPI anchor prediction tool in this study due to high performance and ability to handle large data that is required for genome analysis. Sequences given highest probability score ('highly probable' class) from FragAnchor were chosen as Ser candidates in order to minimize the false positive effect. The obtained result shows that among all sequences that passed into GPI anchor signal probability score assignment by HMM, about half of them (222 out of 591 sequences,

37.6%) were assigned into 'highly probable' class. Another half (263 out of 591 sequences, 44.5%) was assigned with the lowest probability score ('potential false positive' class). Fewer were assigned into 'probable' class (71 out of 591 sequences, 17%) and into 'weakly probable' class (35 out of 591 sequences, 5.9%). With majority of sequences assigned into the so-called 'best' and 'worst' class, the distinction between high and low probability score for GPI anchor signal prediction is clear. Therefore, considering *Ser* candidates from 'highly probable' class alone should be ample for *Ser* candidate identification from *T. thermophila* genome. However, analysis of one gene from 'probable' class (TTHERM_00489460) showed that its sequence contains cysteine pair and sequence features unique to subtype-H. It is well aligned with subtype-H *Ser* candidates and is located in adjacent to one *Ser* tandem (on scaffold 60). Its C-terminus where GPI anchor signal is expected is not truncated and contains small hydrophobic residues similar to GPI anchor signal. The reason that TTHERM_00489460 was given lower probability score of GPI anchor signal could be the presence of lysine and glutamine at the C-terminus, which is uncommon in GPI anchor signal of *T. thermophila* i-ag. This observation emphasizes the importance of reliable GPI anchor signal prediction for i-ag prediction from genome data.

Ser candidates were selected by their sequence properties that passed both cysteine pattern criterion and GPI-anchored signal criterion. The order of which selection criterion is applied first should not alter the final result. In this study the determination of cysteine pattern was performed first because handling output from cysteine pattern search was simpler than the output from GPI-anchored signal determination. The cysteine pattern search script returns result as a report file and a fasta-formatted file of matched candidates that can be immediately subjected to further step. FragAnchor returns result as a report that requires further data processing step to extract the sequence of matched candidates. Moreover, candidates in lower GPI-anchor signal probability score class can be investigated immediately without any additional steps if the GPI-anchor signal prediction is performed after cysteine pattern determination.

55% of the genes could not be classified into any known subtype based on phylogenetic analysis and might belong to new subtypes. At present, the genes encoding several types of surface antigens such as SerT, SerS, SerM and SerI have not

been identified, indicating the existence of more Ser repertoires and subtypes. In addition, certain unclassified gene candidates exhibit properties similar to Ser, such as tendency to have their genes located in tandem and similar gene expression profile. It is possible that the missing *Ser* subtypes might belong to one of the unclassified gene families. Experiments with specific antibodies to *Ser* candidates will be required in order to prove that they are indeed i-ag proteins. Sequence comparison and synteny analysis of the highly polymorphic *Ser* genes from more *T. thermophila* isolates will confirm whether these genes are under positive selection which is a strong evolutionary driving force for mating proteins, molecular sensors and evasive decoys (Miao et al. 2009; Swanson 2003).

Our study took advantage of the available *T. thermophila* MAC genome and DNA microarray expression data to analyze the *Ser* gene family. *Ser* genes are organized in tandem arrays on several MAC scaffolds. These tandem arrays often belong to the same subtype, suggesting that they arose by gene duplication or genetic recombination. *T. thermophila* MAC genome analysis found extensive numbers of duplicated gene tandem suggesting duplication event, with majority of gene tandem containing less than five genes (Chookajorn et al. 2007). The number of genes in each *Ser* tandem has similar trend. In some *Ser* tandem arrays, pseudogene with sequence similar to *Ser* genes was found adjacent to the tandem array. The number of *Ser* genes found in each tandem array may correlate with the number of *Ser* genes in each subtype. Large gene tandem often consists of subtype-H and subtype-L *Ser* genes that have relatively large number of subtype member (Figure 3.5). Subtype with smaller number of subtype member, such as subtype-J (10 candidates) and J* (7 candidates), does not appear to form large gene tandem. Subtype-J *Ser* genes form two gene tandems (Figure 3.5, scaffold 84 and scaffold 47), each tandem with two genes. Two J* *Ser* genes form one gene tandem, and one gene form another tandem with unclassified *Ser* genes (Figure 3.5, scaffold 60 and scaffold 3705). This observation suggests gene duplication as a possible mechanism of *Ser* gene family expansion. Subtype-H and subtype-L *Ser* genes may be duplicated to the great degree via several duplication events, resulting in formation of gene tandem, and larger subtype group; while subtype-J *Ser* genes were subjected to lesser duplication event that resulted in a tandem of two genes. A remarkable case of *Ser* gene tandem is a group of thirteen

subtype-L Ser candidates on scaffold 38 (Figure 3.5). These genes have high sequence similarity especially among adjacent genes. However, such degree of *Ser* tandem expansion is not observed frequently.

Expression pattern cluster analysis does not explicitly indicate the role of the *Ser* gene family in any developmental stage in particular. *Ser* genes up-regulation is observed for all conditions (growth, starvation, conjugation). Available microarray data did not include every known culture conditions for inducing the expression of *Ser* genes such as temperature (SerH, SerL, SerT) and salt concentration (SerS). Therefore, the association between *Ser* subtypes and their expression condition cannot be concluded here, and the candidates assigned to known subtype with known expression condition (SerH and SerL) cannot be confirmed using this expression data. In addition, there are subtypes that no expression condition was described (e.g. subtype J and subtype K), and confirming candidates for these subtype is not possible via gene expression data. Despite no clear correlation between Ser subtype and expression, stage-specific expression patterns of several *Ser* transcripts at the same time point were observed. Therefore, mutual exclusion mechanism is not likely the only strategy underlying expression control of every *Ser* gene.

Because the *Ser* candidate identification is based mainly on the genomic data of *T. thermophila*, the quality of whole genome sequencing have impact on the quality of the result. During inspection of *Ser* gene tandem (on scaffold 1), one protein (TTHERM_00006490) that was not picked up as *Ser* candidate was found to contain several stop codons in the middle of its sequence. This protein is aligned well with adjacent *Ser* genes and located in the same orientation with the adjacent *Ser* genes. It has high sequence similarity to J* *Ser* genes, suggesting that it could be another *Ser* candidates. Its gene sequence, however, was not well sequenced and contains a string of undefined nucleotide in the middle. It should be noted that because of the low sequencing quality issue, it remains uncertain whether this protein is potential Ser candidate. The issue is expected to be improved as the quality of *T. thermophila* genome data is improved. However, this observation suggests that investigation of genes adjacent to identified Ser candidates might be one approach to extend Ser candidate search further.

Using experimentally characterized SerH3 amino acid sequence as query, homology search performed by blastp yielded 71 proteins that were all picked up by pattern search implemented in *Ser* prediction algorithm. Subtype-H Ser candidates found by homology search were also found by *Ser* prediction algorithm. Except TTHERM_00491130 (classified as J*), *Ser* candidates from other subtypes were not found by this approach. This result points out the limitation of Ser candidate search by sequence homology approach—it cannot uncover Ser subtypes that amino acid sequence is unknown. Because various Ser subtypes have no characterized amino acid sequences, homology search would be limited to a few characterized subtypes. *Ser* prediction algorithm presented in this study does not rely on sequence homology and can perform more extensive *Ser* candidate search.

Using the *Ser* candidates found by prediction algorithm as blastp query on NCBI protein database yielded sixteen *T. thermophila* proteins not previously included in the *Ser* candidates list (Table 3.3), but they either lack the GPI anchor motif or matching cysteine pattern. The phylogenetic analysis showed that these proteins are not grouped with any subtype-classified *Ser* candidate. *Ser* candidate identification based on sequence features yields more thorough result than homology search because the former automatically detect the key feature of Ser i-ag sequence without relying on sequence homology. However, both approaches may be applied in combination to achieve the best search result.

The analysis outside *T. thermophila* revealed two weak hits with the proteins from the other ciliates—*Ichthyophthirius multifiliis* AAK94941 (E-value: 0.008) with 21% identity to *Ser*L and *Paramecium tetraurelia* XP_001450224 (E-value: 2.1) with 22% identity to *Ser*H. *I. multifiliis* AAK94941, designated IAG52B, was characterized from *I. multifiliis* isolate D5 expressing serotype D (Eisen et al. 2006). I-ag of *I. multifiliis*, a virulent fish parasite, is the target of immune response (Lin et al. 2002). Infected fish may respond to the *I. multifiliis* by forcing the parasite to leave the host via antigen-antibody interaction, displaying host evasion role of surface protein of the parasite (Lin et al. 1996). There are only three i-ag genes characterized from multiple isolates of *I. multifiliis* (Clark et al. 1996).

Despite the weak sequence resemblance, i-ags of *I. multifiliis* and *T. thermophila* have comparable sequence features. Their sequences contain repetitive

blocks of 70 – 80 amino acid residues with variation in the quantity of block, periodic cysteine pairs alternating between short and long interval of non-cysteine residues ($CX_2CX_{20}CX_3CX_{20}CX_2C$), and ER-translocation signal sequence at N-terminus and hydrophobic signal sequence at C-terminus (Lin et al. 2002). I-ag of *I. multifiliis* was proved to be GPI-anchored protein by phospholipase digestion (Lin et al. 2002). Previous study that characterized *I. multifiliis* i-ag also reported the similarity of one *I. multifiliis* i-ag to *T. thermophila* subtype-L i-ag based on sequence homology search (Clark et al. 2001). MAC genome of *I. multifiliis* was sequenced and undergone comparative study with *T. thermophila* (Lin et al. 2002). Phylogenetic analysis of 17 i-ag candidates of *I. multifiliis* showed that they were not clustered with the three characterized i-ag (Coyne et al. 2011). They were also shown to form gene tandem array (Coyne et al. 2011), similarly to *T. thermophila Ser* genes.

Another weak blastp hit from *P. tetraurelia* (XP_001450224) with 22% identity to *Ser*H contains several regions called paramecium surface antigen repeat. It should be noted that i-ag of other Paramecium species, such as *P. primaurelia* (Coyne et al. 2011), was described but was not found from homology search using *T. thermophila* Ser as query. There are eleven serotypes described from *P. tetraurelia* (Prat 1990). A few genes encoding i-ag of *P. tetraurelia* were characterized (Simon and Schmidt 2007; Prat et al. 1986; Preer et al. 1985). The i-ag of Paramecium was reported to have similar features of *T. thermophila* i-ag, namely the periodic cysteine repeats (Prat et al. 1986) and GPI anchor signal (Prat et al. 1986). Study of *P. tetraurelia* surface antigen described that the protein contains 37 repeat blocks of about 75 amino acid residues with 8 cysteine residues in each block (Capdeville 2000). Amino acid sequences of XP_0014500224 aligned well with a set of characterized *P. tetraurelia* surface antigen (surface antigen A, A51, G, 51B) with approximately 60% – 70% sequence identity, suggesting that this protein might be *P. tetraurelia* surface antigen. Despite the result that *P. tetraurelia* i-ag is GPI-anchored protein based on radiolabelling and PI-PLC treatment (Prat et al. 1986), FragAnchor assigned low probability score to XP_0014500224 and this selected set of characterized *P. tetraaurelia* surface antigens. Manual inspection of XP_001450224 amino acid sequence revealed that its putative GPI anchor signal contains arginine, aspartate, and histidine, which are unlikely for GPI anchor signal that should be

composed of small, hydrophobic residues in order to fit into the active site of GPI transamidase which attaches GPI anchor to the target protein. Nevertheless, this observation suggested that GPI anchor signal prediction algorithm has limitation that may not be applicable for every organism. In this study, the algorithm successfully detected all previously characterized Ser and gave confidence in its power of candidate prediction for *T. thermophila* Ser. Verification that prediction algorithm can successfully detect known set of i-ag (positive control) in the genome of organism of interest is suggested for any applications of surface antigen prediction algorithm.

i-ag proteins were discovered based on their variation in response to antibodies directed at *T. thermophila* surface antigens. This is a hallmark for many antigenic variation phenomena among parasitic and free-living protozoa (Capdeville 2000). In parasitic protozoa such as *Plasmodium falciparum*, a family of proteins on infected red blood cells is needed for a parasitic adherence mechanism to human cells and tissue which is crucial for malaria pathogenesis (Templeton 2007). Antigenic variation in *P. falciparum* switches the variant pathogenic proteins in order to avoid immune detection (Miller et al. 2013). Other parasitic protozoa also exploit a similar system (Chookajorn et al. 2008; Taylor and Rudenko 2006). When antigenic variation is compromised, the parasite becomes vulnerable to the immune system (Lujan 2011).

Unlike parasitic protozoa, free-living protozoa are not subject to host immunity pressure, and the purpose for having surface antigenic variation remains unclear. Nevertheless, understanding the mechanism of antigenic variation in free-living organisms could provide new insights into the evolution and regulation control of antigenic variation in parasitic organisms. Thus, identifying the whole repertoire of the *Ser* gene family is the first step towards the exploration of antigenic variation in *T. thermophila*, an important model organism for many seminal discoveries in molecular biology.#

CHAPTER V
CONCLUSION

Ser prediction algorithm takes advantage of cysteine pattern search and GPI-anchored protein prediction to identify 216 *Ser* candidates. Subtype-H, subtype-L, and subtype-J *Ser* candidates were assigned in addition to the original dataset of known thirteen *Ser*. *Ser* candidates of potentially new subtypes were associated with known subtypes, and putative new subtypes were also uncovered. *Ser* candidates of all subtypes were found to be located as gene tandem array. *Ser* genes in the same tandem array tend to have similar gene expression profile. No clear correlation was found between subtype and gene expression. The work provides the basis for systematically studying *Ser* genes and their functions.

Performance of prediction algorithm was confirmed by its ability to pick up experimentally identified *Ser*. The advantage of this algorithm is its ability to perform extensive search regardless of availability of known *Ser* sequence, allowing it to pick up *Ser* candidates of unknown subtype or candidates with low sequence homology to known *Ser* genes. In comparison, conventional homology search requires known *Ser* sequence to be used as search query, and the result is limited to *Ser* candidates that are closely in the same subtype as the query.

Ser candidates reported here will be the starting point for testing *T. thermophila Ser* repertoires, which will be crucial in the understanding of antigenic variation in free-living *T. thermophila*. The study of antigenic variation in *T. thermophila* might provide an insight into antigenic variation in parasitic single-celled eukaryotes such as *Plasmodium*. Prediction algorithm might also be applied in other ciliates that exhibit variant surface i-ag with similar features such as i-ag of *I. multifiliis*.

REFERENCES

1. Abascal F, Zardoya R, Posada D (2005) ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21 (9):2104-2105. doi:10.1093/bioinformatics/bti263
2. Allen SL (1967) Genomic exclusion: a rapid means for inducing homozygous diploid lines in Tetrahymena pyriformis, syngen 1. Science 155 (3762): 575-577
3. Allen SL, Altschuler MI, Bruns PJ, Cohen J, Doerder FP, Gaertig J, Gorovsky M, Orias E, Turkewitz AP (1998) Proposed genetic nomenclature rules for Tetrahymena thermophila, Paramecium primaurelia and Paramecium tetraurelia. The Seventh International Meeting on Ciliate Molecular Biology Genetics Nomenclature. Genetics 149 (1):459-462
4. Altschuler MI, Yao MC (1985) Macronuclear DNA of Tetrahymena thermophila exists as defined subchromosomal-sized molecules. Nucleic acids research 13 (16):5817-5831
5. Avraham I, Schreier J, Dzikowski R (2012) Insulator-like pairing elements regulate silencing and mutually exclusive expression in the malaria parasite Plasmodium falciparum. Proceedings of the National Academy of Sciences of the United States of America 109 (52):E3678-3686. doi:10.1073/pnas.1214572109
6. Bangs JD, Hereld D, Krakow JL, Hart GW, Englund PT (1985) Rapid processing of the carboxyl terminus of a trypanosome variant surface glycoprotein. Proceedings of the National Academy of Sciences of the United States of America 82 (10):3207-3211
7. Baruch DI, Pasloske BL, Singh HB, Bi X, Ma XC, Feldman M, Taraschi TF, Howard RJ (1995) Cloning the P. falciparum gene encoding PfEMP1, a malarial variant antigen and adherence receptor on the surface of parasitized human erythrocytes. Cell 82 (1):77-87

8. Basmussen L, Orias E (1975) Tetrahymena: growth without phagocytosis. Science 190 (4213):464-465
9. Biggs BA, Gooze L, Wycherley K, Wollish W, Southwell B, Leech JH, Brown GV (1991) Antigenic variation in Plasmodium falciparum. Proceedings of the National Academy of Sciences of the United States of America 88 (20):9171-9174
10. Brown IN, Brown KN, Hills LA (1968) Immunity to malaria: the antibody response to antigenic variation by Plasmodium knowlesi. Immunology 14 (1):127-138
11. Bruns PJ, Brussard TB, Kavka AB (1976) Isolation of homozygous mutants after induced self-fertilization in Tetrahymena. Proceedings of the National Academy of Sciences of the United States of America 73 (9):3243-3247
12. Bruns PJ, Katzen AL, Martin L, Blackburn EH (1985) A drug-resistant mutation in the ribosomal DNA of Tetrahymena. Proceedings of the National Academy of Sciences of the United States of America 82 (9):2844-2846
13. Byrne BC, Brussard TB, Bruns PJ (1978) Induced resistance to 6-methylpurine and cycloheximide in tetrahymena. I. Germ line mutants of T. thermophila. Genetics 89 (4):695-702
14. Calderwood MS, Gannoun-Zaki L, Wellems TE, Deitsch KW (2003) Plasmodium falciparum var genes are regulated by two regions with separate promo -ters, one upstream of the coding region and a second within the intron. The Journal of biological chemistry 278 (36):34125-34132. doi:10.1074/jbc.M213065200
15. Capdeville Y (2000) Paramecium GPI proteins: variability of expression and localization. Protist 151 (2):161-169. doi:10.1078/1434-4610-00016
16. Cassidy-Hanley DM (2012) Tetrahymena in the laboratory: strain resources, methods for culture, maintenance, and storage. Methods in cell biology 109:237-276. doi:10.1016/B978-0-12-385967-9.00008-6
17. Chang J-Y (2011) Diverse Pathways of Oxidative Folding of Disulfide Proteins: Underlying Causes and Folding Models. Biochemistry 50 (17):3414-3431. doi:10.1021/bi200131j

18. Chantangsi C, Lynn DH (2008) Phylogenetic relationships within the genus Tetrahymena inferred from the cytochrome c oxidase subunit 1 and the small subunit ribosomal RNA genes. Molecular phylogenetics and evolution 49 (3):979-987. doi:10.1016/j.ympev.2008.09.017
19. Chaudhary K, Darling JA, Fohl LM, Sullivan WJ, Jr., Donald RG, Pfefferkorn ER, Ullman B, Roos DS (2004) Purine salvage pathways in the apicomplexan parasite Toxoplasma gondii. The Journal of biological chemistry 279 (30):31221-31227. doi:10.1074/jbc.M404232200
20. Chen Q, Fernandez V, Sundstrom A, Schlichtherle M, Datta S, Hagblom P, Wahlgren M (1998) Developmental selection of var gene expression in Plasmodium falciparum. Nature 394 (6691):392-395. doi:10.1038/28660
21. Chookajorn T, Costanzo MS, Hartl DL, Deitsch KW (2007) Malaria: a peek at the var variorum. Trends in parasitology 23 (12):563-565. doi:10.1016/j.pt.2007.08.022
22. Chookajorn T, Kachroo A, Ripoll DR, Clark AG, Nasrallah JB (2004) Specificity determinants and diversification of the Brassica self-incompatibility pollen ligand. Proceedings of the National Academy of Sciences of the United States of America 101 (4):911-917. doi:10.1073/pnas.2637116100
23. Chookajorn T, Ponsuwanna P, Cui L (2008) Mutually exclusive var gene expression in the malaria parasite: multiple layers of regulation. Trends in parasitology 24 (10):455-461. doi:10.1016/j.pt.2008.07.005
24. Clark TG, Gao Y, Gaertig J, Wang X, Cheng G (2001) The I-antigens of Ichthyophthirius multifiliis are GPI-anchored proteins. The Journal of eukaryotic microbiology 48 (3):332-337
25. Clark TG, Lin TL, Dickerson HW (1996) Surface antigen cross-linking triggers forced exit of a protozoan parasite from its host. Proceedings of the National Academy of Sciences of the United States of America 93 (13):6825-6829
26. Cole ES, Bruns PJ (1992) Uniparental cytogamy: a novel method for bringing micronuclear mutations of Tetrahymena into homozygous macronuclear expression with precocious sexual maturity. Genetics 132 (4):1017-1031

27. Collins K, Gorovsky MA (2005) Tetrahymena thermophila. Current biology : CB 15 (9):R317-318. doi:10.1016/j.cub.2005.04.039
28. Conover RK, Brunk CF (1986) Macronuclear DNA molecules of Tetrahymena thermophila. Molecular and cellular biology 6 (3):900-905
29. Coyne RS, Hannick L, Shanmugam D, Hostetler JB, Brami D, Joardar VS, Johnson J, Radune D, Singh I, Badger JH, Kumar U, Saier M, Wang Y, Cai H, Gu J, Mather MW, Vaidya AB, Wilkes DE, Rajagopalan V, Asai DJ, Pearson CG, Findly RC, Dickerson HW, Wu M, Martens C, Van de Peer Y, Roos DS, Cassidy-Hanley DM, Clark TG (2011) Comparative genomics of the pathogenic ciliate Ichthyophthirius multifiliis, its free-living relatives and a host species provide insights into adoption of a parasitic lifestyle and prospects for disease control. Genome biology 12 (10):R100. doi:10.1186/gb-2011-12-10-r100
30. Coyne RS, Thiagarajan M, Jones KM, Wortman JR, Tallon LJ, Haas BJ, Cassidy-Hanley DM, Wiley EA, Smith JJ, Collins K, Lee SR, Couvillion MT, Liu Y, Garg J, Pearlman RE, Hamilton EP, Orias E, Eisen JA, Methe BA (2008) Refined annotation and assembly of the Tetrahymena thermophila genome sequence through EST analysis, comparative genomic hybridization, and targeted gap closure. BMC genomics 9:562. doi:10.1186/1471-2164-9-562
31. Deak JC, Doerder FP (1995) Sequence, codon usage and cysteine periodicity of the SerH1 gene and in the encoded surface protein of Tetrahymena thermophila. Gene 164 (1):163-166
32. Deitsch KW, Calderwood MS, Wellems TE (2001) Malaria. Cooperative silencing elements in var genes. Nature 412 (6850):875-876. doi:10.1038/35091146
33. Doerder FP (2000) Sequence and expression of the SerJ immobilization antigen gene of Tetrahymena thermophila regulated by dominant epistasis. Gene 257 (2):319-326
34. Doerder FP (2014) Abandoning sex: multiple origins of asexuality in the ciliate Tetrahymena. BMC evolutionary biology 14:112. doi:10.1186/1471-2148-14-112

35. Doerder FP, Berkowitz MS (1987) Nucleo-cytoplasmic interaction during macronuclear differentiation in ciliate protists: genetic basis for cytoplasmic control of SerH expression during macronuclear development in Tetrahymena thermophila. Genetics 117 (1):13-23
36. Doerder FP, Berkowitz MS, Skalican-Crowe J (1985) Isolation and genetic analysis of mutations at the SerH immobilization antigen locus of Tetrahymena thermophila. Genetics 111 (2):273-286
37. Doerder FP, Gates MA, Eberhardt FP, Arslanyolu M (1995) High frequency of sex and equal frequencies of mating types in natural populations of the ciliate Tetrahymena thermophila. Proceedings of the National Academy of Sciences of the United States of America 92 (19):8715-8718
38. Doerder FP, Gerber CA (2000) Molecular characterization of the SerL paralogs of Tetrahymena thermophila. Biochemical and biophysical research communications 278 (3):621-626. doi:10.1006/bbrc.2000.3857
39. Duffy MF, Brown GV, Basuki W, Krejany EO, Noviyanti R, Cowman AF, Reeder JC (2002) Transcription of multiple var genes by individual, trophozoite-stage Plasmodium falciparum cells expressing a chondroitin sulphate A binding phenotype. Molecular microbiology 43 (5):1285-1293
40. Dzikowski R, Li F, Amulic B, Eisberg A, Frank M, Patel S, Wellems TE, Deitsch KW (2007) Mechanisms underlying mutually exclusive expression of virulence genes by malaria parasites. EMBO reports 8 (10):959-965. doi:10.1038/sj.embor.7401063
41. Eisen JA, Coyne RS, Wu M, Wu D, Thiagarajan M, Wortman JR, Badger JH, Ren Q, Amedeo P, Jones KM, Tallon LJ, Delcher AL, Salzberg SL, Silva JC, Haas BJ, Majoros WH, Farzad M, Carlton JM, Smith RK, Jr., Garg J, Pearlman RE, Karrer KM, Sun L, Manning G, Elde NC, Turkewitz AP, Asai DJ, Wilkes DE, Wang Y, Cai H, Collins K, Stewart BA, Lee SR, Wilamowska K, Weinberg Z, Ruzzo WL, Wloga D, Gaertig J, Frankel J, Tsao CC, Gorovsky MA, Keeling PJ, Waller RF, Patron NJ, Cherry JM, Stover NA, Krieger CJ, del Toro C, Ryder HF, Williamson SC, Barbeau RA, Hamilton EP, Orias E (2006) Macronuclear genome sequence of the

ciliate Tetrahymena thermophila, a model eukaryote. PLoS biology 4 (9):e286. doi:10.1371/journal.pbio.0040286

42. Eisenhaber B, Bork P, Eisenhaber F (1998) Sequence properties of GPI-anchored proteins near the omega-site: constraints for the polypeptide binding site of the putative transamidase. Protein engineering 11 (12):1155-1161
43. Eisenhaber B, Bork P, Eisenhaber F (1999) Prediction of potential GPI-modification sites in proprotein sequences. Journal of molecular biology 292 (3):741-758. doi:10.1006/jmbi.1999.3069
44. Elliott AM (1959) Biology of Tetrahymena. Annual Review of Microbiology 13 (1):79-96. doi:10.1146/annurev.mi.13.100159.000455
45. Epp C, Li F, Howitt CA, Chookajorn T, Deitsch KW (2009) Chromatin associated sense and antisense noncoding RNAs are transcribed from the var gene family of virulence genes of the malaria parasite Plasmodium falciparum. Rna 15 (1):116-127. doi:10.1261/rna.1080109
46. Fankhauser N, Maser P (2005) Identification of GPI anchor attachment signals by a Kohonen self-organizing map. Bioinformatics 21 (9):1846-1852. doi:10.1093/bioinformatics/bti299
47. Ferguson MA, Homans SW, Dwek RA, Rademacher TW (1988) Glycosyl-phosphatidylinositol moiety that anchors Trypanosoma brucei variant surface glycoprotein to the membrane. Science 239 (4841 Pt 1):753-759
48. Ferguson MK, T. Hart, GW. (2009) Glycosylphosphatidylinositol Anchors. In: Varki AC, RD. Esko, JD. et al. (ed) Essentials of Glycobiology. 2nd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY),
49. Florin-Christensen J, Florin-Christensen M, Tiedtke A, Rasmussen L (1990) The role of secreted acid hydrolases in the utilization of complex nutrients byTetrahymena. Microbial ecology 19 (3):311-316. doi:10.1007/BF02017175
50. Forney JD, Epstein LM, Preer LB, Rudman BM, Widmayer DJ, Klein WH, Preer JR, Jr. (1983) Structure and expression of genes for surface proteins in Paramecium. Molecular and cellular biology 3 (3):466-474
51. Frankel J (2000) Cell biology of Tetrahymena thermophila. Methods in cell biology 62:27-125

52. Gall JG (1974) Free ribosomal RNA genes in the macronucleus of Tetrahymena. Proceedings of the National Academy of Sciences of the United States of America 71 (8):3078-3081
53. Gerber CA, Lopez AB, Shook SJ, Doerder FP (2002) Polymorphism and selection at the SerH immobilization antigen locus in natural populations of Tetrahymena thermophila. Genetics 160 (4):1469-1479
54. Gerber LD, Kodukula K, Udenfriend S (1992) Phosphatidylinositol glycan (PI-G) anchored membrane proteins. Amino acid requirements adjacent to the site of cleavage and PI-G attachment in the COOH-terminal signal peptide. The Journal of biological chemistry 267 (17):12168-12173
55. Gould SB, Kraft LG, van Dooren GG, Goodman CD, Ford KL, Cassin AM, Bacic A, McFadden GI, Waller RF (2011) Ciliate pellicular proteome identifies novel protein families with characteristic repeat motifs that are common to alveolates. Molecular biology and evolution 28 (3):1319-1331. doi:10.1093/molbev/msq321
56. Grass FS (1972) An immobilization antigen in Tetrahymena pyriformis expressed under conditions of high salt stress. J Protozool 19 (3):505-511
57. Gruchy DF (1955) The Breeding System and Distribution ofTetrahymena pyriformis.*. The Journal of Protozoology 2 (4):178-185. doi:10.1111/j.1550-7408.1955.tb02419.x
58. Hall MS, Katz LA (2011) On the nature of species: insights from Paramecium and other ciliates. Genetica 139 (5):677-684. doi:10.1007/s10709-011-9571-3
59. Hallberg RL, Kraus KW, Findly RC (1984) Starved Tetrahymena thermophila cells that are unable to mount an effective heat shock response selectively degrade their rRNA. Molecular and cellular biology 4 (10):2170-2179
60. Hamilton EP, Williamson S, Dunn S, Merriam V, Lin C, Vong L, Russell-Colantonio J, Orias E (2006) The highly conserved family of Tetrahymena thermophila chromosome breakage elements contains an invariant 10-base-pair core. Eukaryotic cell 5 (4):771-780. doi:10.1128/EC.5.4.771-780.2006

61. Hansma HG, Kung C (1975) Studies of the cell surface of Paramecium. Ciliary membrane proteins and immobilization antigens. The Biochemical journal 152 (3):523-528
62. Harrison GS, Karrer KM (1985) DNA synthesis, methylation and degradation during conjugation in Tetrahymena thermophila. Nucleic acids research 13 (1):73-87
63. Hoffman GL, Landolt M, Camper JE, Coats DW, Stookey JL, Burek JD (1975) A disease of freshwater fishes caused by Tetrahymena corlissi Thompson, 1955, and a key for identification of holotrich ciliates of freshwater fishes. The Journal of parasitology 61 (2):217-223
64. Hommel M, David PH, Oligino LD (1983) Surface alterations of erythrocytes in Plasmodium falciparum malaria. Antigenic variation, antigenic diversity, and the role of the spleen. The Journal of experimental medicine 157 (4):1137-1148
65. Howard EA, Blackburn EH (1985) Reproducible and variable genomic rearrangements occur in the developing somatic nucleus of the ciliate Tetrahymena thermophila. Molecular and cellular biology 5 (8):2039-2050
66. Huson DH, Richter DC, Rausch C, Dezulian T, Franz M, Rupp R (2007) Dendroscope: An interactive viewer for large phylogenetic trees. BMC bioinformatics 8:460. doi:10.1186/1471-2105-8-460
67. Jacobs ME, DeSouza LV, Samaranayake H, Pearlman RE, Siu KW, Klobutcher LA (2006) The Tetrahymena thermophila phagosome proteome. Eukaryotic cell 5 (12):1990-2000. doi:10.1128/EC.00195-06
68. Juergensmeyer EB (1969) Serotype expression and transformation in Tetrahymena pyriformis. J Protozool 16 (2):344-352
69. Karrer KM (2000) Tetrahymena genetics: two nuclei are better than one. Methods in cell biology 62:127-186
70. Kiy T, Tiedtke A (1992) Continuous high-cell-density fermentation of the ciliated protozoon Tetrahymena in a perfused bioreactor. Applied microbiology and biotechnology 38 (2):141-146

71. Ko YG, Thompson GA, Jr. (1992) Immobilization antigens from Tetrahymena thermophila are glycosyl-phosphatidylinositol-linked proteins. J Protozool 39 (6):719-723
72. Kodukula K, Gerber LD, Amthauer R, Brink L, Udenfriend S (1993) Biosynthesis of glycosylphosphatidylinositol (GPI)-anchored membrane proteins in intact cells: specific amino acid requirements adjacent to the site of cleavage and GPI attachment. The Journal of cell biology 120 (3):657-664
73. Kusch J, Schmidt HJ (2001) Genetically controlled expression of surface variant antigens in free-living protozoa. The Journal of membrane biology 180 (2):101-109
74. Kyes S, Horrocks P, Newbold C (2001) Antigenic variation at the infected red cell surface in malaria. Annu Rev Microbiol 55:673-707. doi:10.1146/ annurev.micro.55.1.673
75. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23 (21):2947-2948. doi:10.1093/bioinformatics/btm404
76. Lin TL, Clark TG, Dickerson H (1996) Passive immunization of channel catfish (Ictalurus punctatus) against the ciliated protozoan parasite Ichthyo -phthirius multifiliis by use of murine monoclonal antibodies. Infection and immunity 64 (10):4085-4090
77. Lin Y, Lin TL, Wang CC, Wang X, Stieger K, Klopfleisch R, Clarke TG (2002) Variation in primary sequence and tandem repeat copy number among i-antigens of Ichthyophthirius multifiliis. Molecular and biochemical parasitology 120 (1):93-106
78. Love HD, Jr., Allen-Nash A, Zhao QA, Bannon GA (1988) mRNA stability plays a major role in regulating the temperature-specific expression of a Tetrahymena thermophila surface protein. Molecular and cellular biology 8 (1):427-432
79. Low MG (1987) Biochemistry of the glycosyl-phosphatidylinositol membrane protein anchors. The Biochemical journal 244 (1):1-13

80. Lujan HD (2011) Mechanisms of adaptation in the intestinal parasite Giardia lamblia. Essays in biochemistry 51:177-191. doi:10.1042/bse0510177
81. Margolin P, Loefer JB, Owen RD (1959) Immobilizing Antigens ofTetrahymena pyriformis*. The Journal of Protozoology 6 (3):207-215. doi:10.1111/j.1550-7408.1959.tb04359.x
82. Mayo KA, Orias E (1981) Further evidence for lack of gene expression in the Tetrahymena micronucleus. Genetics 98 (4):747-762
83. McMillan PJ, Stanley JS, Bannon GA (1995) Evidence for the requirement of protein synthesis and protein kinase activity in the temperature regulated stability of a Tetrahymena surface protein mRNA. Nucleic acids research 23 (6):942-948
84. Miao W, Xiong J, Bowen J, Wang W, Liu Y, Braguinets O, Grigull J, Pearlman RE, Orias E, Gorovsky MA (2009) Microarray analyses of gene expression during the Tetrahymena thermophila life cycle. PloS one 4 (2):e4429. doi:10.1371/journal.pone.0004429
85. Micanovic R, Gerber LD, Berger J, Kodukula K, Udenfriend S (1990) Selectivity of the cleavage/attachment site of phosphatidylinositol-glycan-anchored membrane proteins determined by site-specific mutagenesis at Asp-484 of placental alkaline phosphatase. Proceedings of the National Academy of Sciences of the United States of America 87 (1):157-161
86. Miller LH, Ackerman HC, Su X-z, Wellems TE (2013) Malaria biology and disease pathogenesis: insights for new treatments. Nature Medicine 19 (2):156-167. doi:10.1038/nm.3073
87. Miller MP, W. Schwartz, T. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In: Proceedings of the Gateway Computing Environments Workshop (GCE), New Orleans, LA, 14 Nov 2010. pp 1 - 8
88. Moran P, Raab H, Kohr WJ, Caras IW (1991) Glycophospholipid membrane anchor attachment. Molecular analysis of the cleavage/attachment site. The Journal of biological chemistry 266 (2):1250-1257
89. Nanney DL, Dubert JM (1960) The Genetics of the H Serotype System in Variety 1 of Tetrahymena Pyriformis. Genetics 45 (10):1335-1349

90. Nanney DL, Park C, Preparata R, Simon EM (1998) Comparison of sequence differences in a variable 23S rRNA domain among sets of cryptic species of ciliated protozoa. The Journal of eukaryotic microbiology 45 (1):91-100
91. Nanney DL, Simon EM (2000) Laboratory and evolutionary history of Tetrahymena thermophila. Methods in cell biology 62:3-25
92. Noviyanti R, Brown GV, Wickham ME, Duffy MF, Cowman AF, Reeder JC (2001) Multiple var gene transcripts are expressed in Plasmodium falciparum infected erythrocytes selected for adhesion. Molecular and biochemical parasitology 114 (2):227-237
93. Orias E (2000) Toward sequencing the Tetrahymena genome: exploiting the gift of nuclear dimorphism. The Journal of eukaryotic microbiology 47 (4):328-333
94. Orias E, Cervantes MD, Hamilton EP (2011) Tetrahymena thermophila, a unicellular eukaryote with separate germline and somatic genomes. Research in microbiology 162 (6):578-586. doi:10.1016/j.resmic. 2011.05.001
95. Orias E, Hamilton EP (1979) Cytogamy: An Inducible, Alternate Pathway of Conjugation in TETRAHYMENA THERMOPHILA. Genetics 91 (4):657-671
96. Orias E, Hamilton EP, Orias JD (2000) Tetrahymena as a laboratory organism: useful strains, cell culture, and cell line maintenance. Methods in cell biology 62:189-211
97. Paulick MG, Bertozzi CR (2008) The glycosylphosphatidylinositol anchor: a complex membrane-anchoring structure for proteins. Biochemistry 47 (27):6991-7000. doi:10.1021/bi8006324
98. Phillips RB (1967) Inheritance of T serotypes in Tetrahymena. Genetics 56 (4):667-681
99. Poisson G, Chauve C, Chen X, Bergeron A (2007) FragAnchor: a large-scale predictor of glycosylphosphatidylinositol anchors in eukaryote protein sequences by qualitative scoring. Genomics, proteomics & bioinformatics 5 (2):121-130. doi:10.1016/S1672-0229(07)60022-9

100. Prat A (1990) Conserved sequences flank variable tandem repeats in two alleles of the G surface protein of Paramecium primaurelia. Journal of molecular biology 211 (3):521-535. doi:10.1016/0022-2836(90)90263-L

101. Prat A, Katinka M, Caron F, Meyer E (1986) Nucleotide sequence of the Paramecium primaurelia G surface protein. A huge protein with a highly periodic structure. Journal of molecular biology 189 (1):47-60

102. Preer JR, Jr., Preer LB, Rudman BM, Barnett AJ (1985) Deviation from the universal code shown by the gene for surface protein 51A in Paramecium. Nature 314 (6007):188-190

103. Rifkin MR, Landsberger FR (1990) Trypanosome variant surface glycoprotein transfer to target membranes: a model for the pathogenesis of trypanosomiasis. Proceedings of the National Academy of Sciences of the United States of America 87 (2):801-805

104. Roberts WL, Myher JJ, Kuksis A, Low MG, Rosenberry TL (1988) Lipid analysis of the glycoinositol phospholipid membrane anchor of human erythrocyte acetylcholinesterase. Palmitoylation of inositol results in resistance to phosphatidylinositol-specific phospholipase C. The Journal of biological chemistry 263 (35):18766-18775

105. Ron A, Williams NE, Doerder FP (1992) The immobilization antigens of Tetrahymena thermophila are glycoproteins. J Protozool 39 (4):508-510

106. Saad YM, Paul Doerder F (1995) Immobilization antigen variation in natural isolates of Tetrahymena thermophila. European Journal of Protistology 31 (1):45-53. doi:http://dx.doi.org/10.1016/S0932-4739(11)80355-X

107. Satir BH, Reichman M, Orias E (1986) Conjugation rescue of an exocytosis-competent membrane microdomain in Tetrahymena thermophila mutants. Proceedings of the National Academy of Sciences of the United States of America 83 (21):8221-8225

108. Scherf A, Hernandez-Rivas R, Buffet P, Bottius E, Benatar C, Pouvelle B, Gysin J, Lanzer M (1998) Antigenic variation in malaria: in situ switching, relaxed and mutually exclusive transcription of var genes during intra-erythrocytic development in Plasmodium falciparum. The EMBO journal 17 (18):5418-5426. doi:10.1093/emboj/17.18.5418

109. Simon EM, Nanney DL, Doerder FP (2007) The “Tetrahymena pyriformis” complex of cryptic species. Biodiversity and Conservation 17 (2):365-380. doi:10.1007/s10531-007-9255-6

110. Simon MC, Schmidt HJ (2007) Antigenic variation in ciliates: antigen structure, function, expression. The Journal of eukaryotic microbiology 54 (1):1-7. doi:10.1111/j.1550-7408.2006.00226.x

111. Smith DG, Gawryluk RM, Spencer DF, Pearlman RE, Siu KW, Gray MW (2007) Exploring the mitochondrial proteome of the ciliate protozoon Tetrahymena thermophila: direct analysis by tandem mass spectrometry. Journal of molecular biology 374 (3):837-863. doi:10.1016/j.jmb. 2007.09.051

112. Smith DL, Berkowitz MS, Potoczak D, Krause M, Raab C, Quinn F, Doerder FP (1992) Characterization of the T, L, I, S, M and P cell surface (im -mobilization) antigens of Tetrahymena thermophila: molecular weights, isoforms, and cross-reactivity of antisera. J Protozool 39 (3):420-428

113. Smith JD, Chitnis CE, Craig AG, Roberts DJ, Hudson-Taylor DE, Peterson DS, Pinches R, Newbold CI, Miller LH (1995) Switches in expression of Plasmodium falciparum var genes correlate with changes in antigenic and cytoadherent phenotypes of infected erythrocytes. Cell 82 (1):101-110

114. Sogin ML, Ingold A, Karlok M, Nielsen H, Engberg J (1986) Phylogenetic evidence for the acquisition of ribosomal RNA introns subsequent to the divergence of some of the major Tetrahymena groups. The EMBO journal 5 (13):3625-3630

115. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H, Lehvaslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD, Birney E (2002) The Bioperl toolkit: Perl modules for the life sciences. Genome research 12 (10):1611-1618. doi:10.1101/gr.361602

116. Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22 (21):2688-2690. doi:10.1093/bioinformatics/btl446

117. Stelly N, Halpern S, Nicolas G, Fragu P, Adoutte A (1995) Direct visualization of a vast cortical calcium compartment in Paramecium by secondary ion mass spectrometry (SIMS) microscopy: possible involvement in exo -cytosis. Journal of cell science 108 (Pt 5):1895-1909
118. Struder-Kypke MC, Wright AD, Jerome CA, Lynn DH (2001) Parallel evolution of histophagy in ciliates of the genus Tetrahymena. BMC evolutionary biology 1:5
119. Su XZ, Heatwole VM, Wertheimer SP, Guinet F, Herrfeldt JA, Peterson DS, Ravetch JA, Wellems TE (1995) The large diverse gene family var encodes proteins involved in cytoadherence and antigenic variation of Plasmodium falciparum-infected erythrocytes. Cell 82 (1):89-100
120. Swanson WJ (2003) Adaptive evolution of genes and gene families. Current opinion in genetics & development 13 (6):617-622
121. Taylor HM, Kyes SA, Harris D, Kriek N, Newbold CI (2000) A study of var gene transcription in vitro using universal var gene primers. Molecular and biochemical parasitology 105 (1):13-23
122. Taylor JE, Rudenko G (2006) Switching trypanosome coats: what's in the wardrobe? Trends in genetics : TIG 22 (11):614-620. doi:10.1016/j.tig. 2006.08.003
123. Templeton TJ (2007) Whole-genome natural histories of apicomplexan surface proteins. Trends in parasitology 23 (5):205-212. doi:10.1016/j.pt. 2007.03.001
124. Tian M, Chen X, Xiong Q, Xiong J, Xiao C, Ge F, Yang F, Miao W (2014) Phosphoproteomic analysis of protein phosphorylation networks in Tetrahymena thermophila, a model single-celled organism. Molecular & cellular proteomics : MCP 13 (2):503-519. doi:10.1074/mcp.M112.026575
125. Tondravi MM, Willis RL, Love HD, Jr., Bannon GA (1990) Molecular characterization of SerH3, a Tetrahymena thermophila gene encoding a temperature-regulated surface antigen. Molecular and cellular biology 10 (11):6091-6096

126. Tourancheau AB, Tsao N, Klobutcher LA, Pearlman RE, Adoutte A (1995) Genetic code deviations in the ciliates: evidence for multiple and independent events. The EMBO journal 14 (13):3262-3267
127. Turkewitz AP (2004) Out with a bang! Tetrahymena as a model system to study secretory granule biogenesis. Traffic 5 (2):63-68
128. Udenfriend S, Kodukula K (1995) How glycosylphosphatidylinositol-anchored membrane proteins are made. Annual review of biochemistry 64:563-591. doi:10.1146/annurev.bi.64.070195.003023
129. Williams NE (1983) Surface membrane regeneration in deciliated Tetrahymena. Journal of cell science 62:407-417
130. Williams NE, Doerder FP, Ron A (1985) Expression of a cell surface im -mobilization antigen during serotype transformation in Tetrahymena thermophila. Molecular and cellular biology 5 (8):1925-1932
131. Xiong J, Lu Y, Feng J, Yuan D, Tian M, Chang Y, Fu C, Wang G, Zeng H, Miao W (2013) Tetrahymena functional genomics database (TetraFGD): an integrated resource for Tetrahymena functional genomics. Database : the journal of biological databases and curation 2013:bat008. doi:10.1093/database/bat008
132. Xiong J, Yuan D, Fillingham JS, Garg J, Lu X, Chang Y, Liu Y, Fu C, Pearlman RE, Miao W (2011) Gene network landscape of the ciliate Tetrahymena thermophila. PloS one 6 (5):e20124. doi:10.1371/journal.pone.0020124
133. Yao MC, Gorovsky MA (1974) Comparison of the sequences of macro- and micronuclear DNA of Tetrahymena pyriformis. Chromosoma 48 (1):1-18

BIOGRAPHY

NAME Miss Patrath Ponsuwanna

DATE OF BIRTH 30 January 1984

PLACE OF BIRTH Bangkok, Thailand

INSTITUTIONS ATTENDED Mahidol University, 2002 – 2006
Bachelor of Science (Biology)
Mahidol University, 2007 – 2015
Doctor of Philosophy (Biochemistry)

HOME ADDRESS 180/7 Sukhumvit 81, Bangjak, Prakanong, BKK 10260
E-mail: patrath.pons@gmail.com

AWARD RECEIVED Academic Achievement award from department of Biochemistry, Faculty of Science, Mahidol University

PUBLICATIONS

1. Ponsuwanna P, Kumpornsin K, Chookajorn T. Genome-wide prediction of the polymorphic *Ser* gene family in *Tetrahymena thermophila* based on motif analysis. PloS one. 2014;9(8):e105201.
2. Chookajorn T, Ponsuwanna P, Cui L. Mutually exclusive *var* gene expression in the malaria parasite: multiple layers of regulation. Trends in parasitology. 2008 Oct;24(10):455-61.