Global Investigation into the Population Genetic Structure of *Cryptosporidium hominis* Based on a Whole Genome Multi-locus SNP-typing Scheme; Inferences about the Existence of Biogeographical Partitions

by

Jill Marie Williamson
B.Sc., University of British Columbia, 2000

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

(Pathology and Laboratory Medicine)

THE UNIVERSITY OF BRITISH COLUMBIA
(Vancouver)

April 2009

ABSTRACT

Previously considered a disease of importance strictly to veterinary medicine *Cryptosporidium* has emerged as a highly successful opportunistic parasitic protozoan posing a significant threat to public health. Intricate transmission dynamics, a complex epidemiology, and parasite robustness and persistence have all hampered efforts for the prevention and control of *Cryptosporidium*. Genetic diversity is a prerequisite to better understand the role of parasite variation in disease etiology and pathobiology. The extent of genetic structure among *C. hominis* and *C. parvum*, the two most prevalent species of *Cryptosporidium*, is insufficiently understood with the population structure still largely suspect

We report on the distribution of genetic diversity and possible existence of geographic partitions among *C. hominis* subpopulations from Australia, Kenya, Peru and Scotland. We studied *C. hominis* population genetic structure using a multi-locus SNP-type (MlSt) established from 45 single nucleotide polymorphic loci covering 13 bio-functionally relevant proteins. A total of 77 isolates from 4 intercontinental subpopulations were genetically typed. Twenty-four unique MlSt's were identified, 25% of which were found to be located within one or more subpopulations. Diversity statistical tests to discern the degree of intra-population and inter-population diversity, genetic distance, and genetic identity variation were used to examine the population genetic structure. Within-population differences among subpopulations account for 69.6% of genetic variation; differentiation among subpopulations constitute 30.4%. Genetic distances among subpopulations averaged 0.048 and varied from 0.034 between the Australian and Scotland subpopulations to 0.061 between Scotland and Kenya. More broadly, our results argue that too wide of a geographic boundary can impede rather than advance genetic population studies and that the practice of sampling more regional subpopulations be adopted.

A fifth subpopulation, a combination of *C. hominis* and *C. parvum* isolates, was drawn upon to determine whether or not a pre-defined allelic profile of single nucleotide polymorphisms (SNPs) was an efficient and reliable means for species specific identification. Results showed the SNP-typing approach's ability to distinguish between different species as well as being capable of uncovering potential novel SNPs within an individual isolate.

We propose that the patterns of genetic variation are influenced by geography and that the identification of host adapted geographically conserved sub-genotypes within a defined geographic cohort versus widespread dissemination of genetically stable isolates could ultimately provide a valuable basis for the predictive epidemiology of *Cryptosporidium* infection. Our findings provide an alternative method for species detection, a crucial element to epidemiological investigations.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

A	Australia
aa	amino acid
AIDS	Acquired Immunodeficiency Disease
APR	Apoptosis Related Protein
BC	British Columbia
bp	basepair
BT	β-tubulin
COWP	*Cryptosporidium* oocyst wall protein
Cp23	*Cryptosporidium* protein
DHFR	Di-hydrofolate Reductase
D(st)	average among populations diversity
EID	emerging infectious disease
EMAAg	erythrocyte membrane associated antigen
Fst	Wright's F statistics
Gst	Nei's analog for Wright's (Fst) statistics
H(s)	average within population diversity
H(t)	total genetic diversity
HIV	human immunodeficiency virus
HSP	heat shock protein
GKO	genetic knockout
Gp60	glycoprotein 60
K	Kenya
LDH	lactate dehydrogenase
MDH	malate dehydrogenase
MLG	multi-locus genotype
MlSt	multi-locus SNP-type
NIAID	National Institute Allergy and Infectious Disease
NJM	neighbour joining method
NS-SNP	non- synonymous SNP
nt	nucleotide
ORF	open reading frame
P	Peru

PAGE	poly-acrylamide gel electrophoresis
PCR	polymerase chain reaction
S	Scotland
SAAP	single amino acid polymorphism
SBE	single base extension
SNP	single nucleotide polymorphism
S-SNP	synonymous SNP
θst	Weir & Cockerham analog for Wright's statistics
UPGMA	unweighted pair group method w/ arithmetic mean
UPRTase	Uracil phosphoribotase
WHO	World Health Organization

GLOSSARY

Allele - alternative form of a gene.

Allopatric - an organism whose range is entirely separate. In regards to speciation it refers to biological populations that are physically isolated by an extrinsic barrier and evolve genetic reproductive isolation.

Antigen - any substance that causes your immune system to produce antibodies against; it may be a foreign substance from the environment such as chemicals, bacteria, viruses, parasites or pollen.

B cell - lymphocytes that play role in the humoral immune response whose principal functions are to make antibodies against antigens.

Bioinformatics - collection of methods utilized for the analysis of molecular biology data through a computer.

Brush border - microvilli-covered surface of epithelium cells found in intestinal tract of the body.

ClustalW - multiple sequence alignment program.

Codon - a group of three adjacent nucleotides that encode an amino acid.

Comparative genomics - analysis based on the comparisons of whole genomes.

Differentiation (genetic) – differences in allele frequencies among populations.

Distance Matrix - a pairwise 'distance' between taxa, for molecular data, it could be the observed number of nucleotide differences between the pairs of taxa.

Enterocyte - intestinal absorptive cells, simple columnar epithelial cells found in the small intestine and colon.

Eukaryote - organisms with a complex cell structure and cell nucleus.

Exon - coding part of a gene/protein.

Extension - process in PCR by which nucleotides are extended from the initial binding site of a primer by the action of a polymerase.

Gene flow - also known as migration, any movement of genes from one population to another, result of which decrease inter-population variation and increase intra-population variation.

Genetic drift - random fluctuations in allele frequency which occur by chance, particularly in small subpopulations, as a result of sampling error.

Genome - full set of chromosomes carried by a given organism.

Genotype - genetic characteristics of a cell or organism according to its entire genome or a specific set of genetic loci (allele)

Gnotobiotic - animals born in aseptic conditions, removed by Caesarean section, are exposed only to the microorganisms researchers wish to be present in the animal.

Haploid - possessing only one copy of each chromosome in a genome, in contrast to diploid where two copies of each chromosome are present.

Homolog - gene or morphological character that shares a common ancestry with a different gene or morphological character.

Humoral - the part of immunity or the immune response that involves antibodies secreted by B cells and circulating in bodily fluids.

Intra-population - within a population.

Inter-population - among or between different populations.

Intron - non-coding part of a gene/protein.

Isolate - single sample of a species from a given population

Locus (pl. loci) - specific location on a chromosome.

Macrodiversity – genotype diversity of an organism revealed by a strain-typing method designed to detect changes throughout the genome.

Meiosis - cell division in sexually reproducing organisms that reduces amount of genetic information by half.

Metapopulation - group of populations connected by some level of gene flow, also referred to as population.

Microdiversity - genotype diversity of an organism revealed by a strain-typing method designed to detect nucleotide changes in a restricted part of the genome.

Microvilli - microscopic cellular membrane protrusions that increase the surface area of cell.

Migration - see gene flow.

Mitosis - simple cell division without a reduction in chromosome number.

Monophyletic group - group of organisms with the same taxonomic title that are shown phylogenetically to share a common ancestor that is exclusive to these organisms.

Multi-locus SNP type (MlSt) – genetic typing of an isolate based on the allelic profile of all molecular markers examined, most studies discuss similar results in terms of a multi-locus genotype (MLG), we chose the acronym MlSt based on the SNP foundation of the study.

Mutation - change in the nucleotide sequence of the genetic material of an organism.

Neighbour-Joining (NJ) - algorithm for inferring a branching tree diagram from a distance matrix, by successively clustering pairs of taxa together.

Non-synonymous substitutions (NS) - substitutions that occur in protein-coding genes that result in a change at the amino acid level.

Oocyst - thick-walled spore phase of *Cryptosporidium,* is the infectious form of the parasite.

Opportunistic microorganism - pathogenic organism that exploits an immunocompromised immune system to establish infection.

ORF - open reading frame, DNA sequence without stop codons thus allowing for the translation of a protein sequence.

Ortholog - homologous genes found in two different taxa that are performing the same function in each taxon.

Outbreak – acute appearance of a cluster of illnesses caused by a single pathogen that occurs in numbers of excess of what is expected for that time and place.

Pathogenesis - development of a disease; the origin of a disease and the chain of events leading to that disease.

Phenotype – observable characteristics expressed by an organism, including drug resistance, virulence, and morphology.

Potable - water considered to be of sufficiently high quality so it can be consumed or utilized without risk of immediate or long term harm.

Prepatent - period between the time of exposure to a parasite and the time when the parasite can be detected in the blood or stool.

Primer – short piece of nucleic acid that binds to a complementary target or template DNA strand, serving as the starting point for the addition or extension of complementary nucleotides along the rest of the template strand.

Purine - Adenine (A) or Guanine (G) nucleotide.

Pyrimidine – Cytosine (C) or Thymidine (T) nucleotide.

Subpopulation - subgroup of a metapopulation or total population, commonly annotated as subpopulation, local populations, or demes.

Symbiotic - a relationship that is mutualistic, parasitic, or commensal in nature.

Synonymous substitutions (S) - substitutions that do not change the identity of the encoded amino acid.

T cell - cells belonging to a group of white blood cells known as lymphocytes that play a central role in cell-mediated immunity and are produced in the thymus.

Taxonomy - scientific discipline of naming organisms.

Transition - substitution of a purine for a purine or a pyrimidine for a pyrimidine (like-for-like).

Transversion - substitution of a purine for a pyrimidine or vice versa.

ACKNOWLEDGMENTS

I offer my enduring gratitude to the faculty, staff and fellow students at UBC who have inspired me to pursue my work in this field throughout both my undergraduate and graduate careers.

As I moved into post-graduate studies I was beyond blessed with the exemplary mentoring, guidance and enduring support, both emotionally and academically of Dr. David Walker, without whom I am not sure I would be where I am today. My sincerest gratitude for all his hard work within the Department of Pathology as a whole and his more personal dedication to me, it is more appreciated than I could ever express.

I further give great credit for my success to one of the most efficient, knowledgeable and patient people I know, Penny Woo. Without her the Department of Pathology and Laboratory Medicine would be at a tremendous loss.

I was very fortunate to be situated within University of British Columbia's research division of the British Columbia Center for Disease Control. This state of the art facility provided me access to superior technologies, world-class scientists and some of the best trained technicians and resource staff there is. In particular I must extend much thanks for all their guidance, technical expertise and resource support to; Kathy Adie, Ruben Chen, Alan McNabb, Diane Eisler, Glenna Geddes, Theresa Low, Yvonne Santa-Cruz, and Quantine Wong.

This study would not have been possible without the incredible privilege I was afforded through the collaborations with five of the top scientists in the field of *Cryptosporidium* research. I had the incredible opportunity to work alongside them, have them available as a resource, receive their contributions as co-authors and use them as generous sources of advice and guidance. This is something most people would hope to achieve in their entire careers, let alone their graduate work. With great thanks to Dr. Huw Smith (Scottish Parasite Diagnostic Laboratory, Stobhill Hospital, Glasgow, Scotland), Dr. Una Ryan (Murdoch University of Western Australia, Perth, Australia), Dr. Lihua Xiao and Dr. Vitaliano Cama (Center for Disease Control, Atlanta, Georgia, USA) and Dr. Wangeci Gatei (Kenya Medical Research Institute; Nairobi, Kenya).

I was again beyond privileged with the make-up of my supervisory committee which included some of the most accredited and renowned researchers in their respective fields; Dr Judy Isaac-Renton, Dr. Martin Adamson, Dr. Craig Stephens, and Dr.David Walker. As a whole their different perspectives and suggestions regarding my studies provided a real robustness to my academic career and really

allowed me to look outside the narrowness of one research topic and think of how I might apply my knowledge to other aspects of disease research or to different concepts all together. Their questions and ideas not only challenged me but allowed me the opportunity to mature my own thinking to a level where I became more and more confident and proud of the progress I had made. Each in their own way, their contributions to my work will have a profound effect on all that I have yet to accomplish in my career and I have such gratitude for that.

Last, but certainly not least, I have to give the greatest of thanks and accolades to my supervisor, Dr. Corinne Ong. She taught me to think for myself, work for myself and expand myself to a new level of academic standards yet at all times I never doubted that she was right their alongside me to guide me, challenge me and encourage me. I cannot thank her enough for her patience with my work, her respect for my work methods and her enduring support through thick and thin. She is a true scientist, with expertise not only in her specific field but also in how good, reputable research should be conducted. She was not just a mentor for my work but also for what I strive to achieve as an all around medical professional in regards to work ethic, professionalism, reputation and contributions to and collaborations in science. After all she has done for me, how hard she has fought for me and stood up for me and my work there are not enough words to acknowledge what all her dedication to me has meant. What she has taught me will be the foundation of what I hope to achieve in my career. Anyone would be so blessed to have the chance to work with her.

DEDICATION

From elementary school, to secondary school to undergraduate academics to graduate academics there have been countless people alongside me that whom without with I am sure I would not have been able to achieve what I have thus far in my career. I give profound thanks for all their support and inspiration and I continue to be in awe that I have been so blessed with such incredible people in both my personal and academic lives.

The encouragement and support from my two amazing older sisters is one of the greatest gifts they could have ever been given to me. From as early as I can remember I have always looked up to them, been inspired by them and challenged myself to be more like them. Their courageous spirits continue to awe me and I am so humbled to have such incredible people in my life for me to look up to. I have learnt so much from them and I am so proud of the people they have become and all they have accomplished. Furthermore I am indebted to my Auntie Trish, who was always there with her unconditional love and endless supply of funny mail when all I had to read were "boring" papers and journals.

The greatest thanks of all must go to my mother. No words will ever be powerful enough to impress upon her how much I love her and how much I appreciate all she has done for me and sacrificed for me. She cheered me on when I was already up and lifted me up in times I was down. I have no doubt that I would not have gone into the field I have if it weren't for her. My mom is a highly accomplished and respected nurse. As a small child I would wait on bated breath for her to get home from work. Though exhausted from a long 12 hour shift and likely just wanting 10 minutes of peace in a hot bath before resuming her life as wife and mother of three active daughters I would pounce on her the minute she was in the door. I would pepper her with questions about her patients, their illnesses, the causes, their prognoses, what she did to treat them, how she helped them and so on and so on and so on. Most days I would draw the bath for her ahead of time so I would not have to wait any longer than the minutes it would take her to get in the door draw it herself. While she tried to unwind I chatted non-stop and yet her patience never wavered, not once, she never turned me away, she never asked me to be quiet or to leave. It is those stories, that came at 730 pm post-day shift or 730 am post-morning shift that created and nurtured my interest in the medical field. It is the compassion with which she told those stories that inspired me, drove me, and created the conviction within me to be involved in this field that strives to create a greater good for people of all walks of life; to help all of mankind no matter who you are, where you live or what you have. I am eternally grateful to have had the life and upbringing that enabled me to hear these stories. It is these stories and their storyteller to whom I dedicate my past and future career to. It is these stories and their storyteller to whom I dedicate this dissertation to.

CHAPTER 1

INTRODUCTION TO *Cryptosporidium*

- Biological Concepts of *Cryptosporidium* -

Chapter Summary – Parasitism is a symbiotic relationship between organisms of different species where the parasite benefits at the expense of the host. Parasitic diseases account for a large proportion of both human and animal morbidity and mortality. Long thought to be a disease of importance strictly to veterinary medicine *Cryptosporidium* has recently emerged as an important group of parasitic protozoa posing a serious threat to public health. Discussed is a comprehensive review of the biology, pathogenesis, and epidemiology of *Cryptosporidium.*

1.1 Parasitology

1.1.1 Parasitism

According to the latest WHO[240] estimates worldwide infectious disease accounts for nearly 30% of 56 million annual deaths. Even in the face of amazing advancements in medicine, science ,and technology in the last twenty years modern medicine has seen a spectacular resurgence or novel emergence of several pathogens thought to have been be eradicated or contained, parasites included. This is likely due to a combination of factors: human demographics and behaviour, technology, economic development and land use, international travel and commerce, microbial adaptation, and changes to and the breakdown of public health measures. One of the greatest examples is the increased resistance to drugs and insecticides which has proven to be a major cause in the resurgence of malarial diseases[46, 185]. The changing global climate and ecology could create new environments more favourable to pathogenic organisms in addition to hosts and vectors of pathogenic organisms. Furthermore the rising global population is forcing the expansion of human habitats into the niches of potentially virulent organisms.

Parasitism is defined by the relationship between two organisms: a parasite, most often the smaller of the two, and a host upon which the parasite is physiologically dependent[46, 154]. The host-parasite interaction is a dynamic balance between the two organisms; the virulence of the pathogen and the resistance of the host are constantly changing[46]. Equilibrium between host and parasite is necessary to ensure the survival of both partners and thus sustaining the endurance of the relationship. Hosts and parasites therefore co-evolve. Pathogens constitute selective pressures in the evolution of the host just as hosts are to pathogens.

Globally, eukaryotic parasitic infections account for a significantly higher incidence of morbidity and mortality than disease produced by any other group of organisms[46]. In the developing world this is further exacerbated by additional factors such as other diseases rampant within a given population, poor diet, socioeconomic status, age, war, and other similar stressors. Opportunistic parasites minimally pathogenic to immunocompetent organisms, can exploit decreased or compromised cellular immunity to induce serious host damage. While perhaps the most well known example is the immunosuppression brought on by the HIV/AIDS epidemic centralizing around Africa and Asia, compromised immunity is also common in those individuals undergoing chemotherapy, organ transplantation, with auto-immune diseases or with poor health due to nutrition or starvation. Examples of such opportunistic parasites include *Plasmodium* spp., *Toxoplasma* spp., *Leishmania* spp., *Giardia* spp., *Pneumocystis* spp. and *Cryptosporidium* spp., the focus of this study[6, 20, 22, 48, 60, 89, 92, 115, 137, 154, 183, 222].

1.1.2 The Emergence of *Cryptosporidium*

Primarily because of the growth rate of the global human population, the expanding appetite for resources of all types has led to the dissolution of many ecological barriers important to the natural control of disease. More specifically, for parasitic zoonoses acquired from wildlife habitats and vice versa, a shift in the interface between wildlife and people from often sporadic and fragile to more permanent and substantial provides significant opportunities for parasite transmission[55, 87, 123, 131, 150, 167, 176, 224]. Dasak et al. (2000) reviewed an array of emerging infectious diseases (EID) affecting people and discussed some of the underlying factors mediating their emergence. They went on to address the consequences such EIDs have on humans, domestic animals and wildlife health in addition to the consequences EIDS pose for biodiversity as a whole. Of the 18 EIDs identified only cryptosporidiosis is named as a zoonotic pathogen of major importance to public health.

Long thought to be a disease restricted to veterinary medicine, the *Cryptosporidium* genus with its various species has now been established as a significant emerging opportunistic global enteropathogen to humans. Tyzzer[226] (1907) first published evidence on the one-celled organism known as *Cryptosporidium muris* found in the gastric glands of mice[111, 226]. Five years later, in 1912, he reported a second species, *Cryptosporidium parvum*, from laboratory mice which differed from the type species in both the localization of infection and the developmental morphology of the organism[63]. It was not until the late 1970's early 1980's, a time period coinciding with the forefront of the AIDS epidemic that the host switching by *Cryptosporidium* to humans was recognized and its pathogenic potential appreciated. In the advent of new detection methods human cases of *Cryptosporidium* infection became more apparent. In 1976 *Cryptosporidium* was diagnosed in a previously healthy three year old child[63]. Two months later a second case arose in an immunosuppressed individual undergoing drug therapy[229]. From 1976 to 1982 the disease was rarely reported and primarily occurred in immunocompromised persons. In 1982, the number of reported cases began to increase dramatically, though still relatively limited to immunocompromised persons. With the aid of newly developed laboratory diagnostic techniques, outbreaks in immunocompetent persons began to be recognized[56]. In the 1990's, the application of molecular techniques to the identification of isolates brought about both clarification and complexity to our understanding of *Cryptosporidium* spp. and host specificity. The WHO[240] now defines *Cryptosporidium* as the most notable example of an emerging disease caused by a protozoan parasite in the last 25 years. It has been put forth by some that the catalyst in the emergence of human *Cryptosporidium* infections was the altered host environment to be found in a new population of HIV (+) individuals[174]. As the demographic and temporal conditions became suitable for the propagation of an organism such as *Cryptosporidium* natural selection would likely have facilitated its evolution.

Cryptosporidium has caused many recent large-scale outbreaks in the developed and developing world[63, 90, 112, 130, 151, 200]. The pathogenic success of *Cryptosporidium* within the host is in part attributed to its low infectious dose, as few as one oocyst, the infectious form of *Cryptosporidium*, is capable of establishing infection[171]. The intracellular but extra-cytoplasmic location and monoxenous life cycle also contribute to the parasite's success. Oocysts are omnipresent, highly stable, and occur in diverse ecological situations, having shown an environmental persistence of up to six months. Coupled with the fact that *Cryptosporidium* is now classified by the NIAID as a class B bioterrorism agent a re-examination of the public health threat from *Cryptosporidium* spp. is warranted, thus generating a major shift in research focus towards the organism[37].

1.1.3 *Cryptosporidium* Biology: Life Cycle & Propagation

Cryptosporidium's life cycle involves two asexual stages (merogony) and a sexual stage (gametogony), all three of which occur intracellularly[64, 230]. This has impaired efforts to clarify the exact mechanisms of the developmental stages of *Cryptosporidium* maturation[238]. Throughout the majority of their life cycles *Cryptosporidium* exists as haploid organisms, multiplying asexually. A transient sexual stage results in a diploid state followed by meiotic division. Genetic recombination between *C. parvum* and *C. hominis* has been shown but the exact contribution to genetic diversity within and between each species remains unclear[3, 216, 217].

The oocyst contains four sporozoites, unique considering most Apicomplexans contain 8, that are released upon host cell colonization[64, 230]. The sporozoite differentiates into a spherical trophozoite and asexual development begins, forming two types of meronts. Typically type one meronts contain 6-8 nuclei which eventually incorporate into 6-8 first generation merozoites as the meront matures. Each merozoite is capable of infecting a new host cell and subsequently developing into either a new type one meront or into a type two meront. Upon maturation type two meronts contain four second generation merozoites which also invade new host cells[64, 230]. It's these second generation merozoites that initiate sexual reproduction by differentiating into either male (microgametocyte) or female (macrogamont) stages. Microgametes mature and form a microgametocyte which upon excysting from its host cell goes on to invade another host cell containing a female macrogamont. Fertilization ensues and a zygote is developed and then eventually an oocyst which undergoes meiosis (sporogeny) within the host. Once completed each oocyst comprises of four potentially infectious sporozoites.

Oocysts either sporulate ***in situ*** and release sporozoites for autoinfection or are expelled from the body in the feces[64, 230]. From this life cycle two types of oocysts are formed, thick walled and thin walled oocysts. Thick walled oocysts make up about 80% of those produced and with the tough, hearty outer shell are those that are released into the environment via the intestinal tract. These oocysts are therefore responsible for the long-term survival of *Cryptosporidium* in the environment. In the environment each oocyst is in its infectious form, haploid, and contains eight chromosomes. The remaining 20% of oocysts produced are considered thin walled which excyst endogenously, infect new host cells thus auto infecting the host, and perpetuating the life cycle. It is this autoinfection that can cause *Cryptosporidium* to develop into a chronic disease in immunosuppressed hosts[64, 230].

It takes approximately 12-14 hours for a generation of parasites to develop and mature[64, 230]. The rapid life cycle exacerbated by its multiple modes of autoinfection and monoxenous nature creates a heavy burden upon the host. This parasite burden can lead to the development of secondary infection sites within the intestinal tract thus creating chronic, relentless infections which are often seen in the immunocompromised, elderly and young. Such a high propagation rate further contributes to *Cryptosporidium's* pathogenic success by flooding the natural world with oocysts. As many as 20 billion oocysts have been collected over a 24hr time period from experimentally infected cattle making the affliction these oocysts can have on the environment apparent[10, 11, 64, 230].

Figure 1.1 *Cryptosporidium* life cycle.

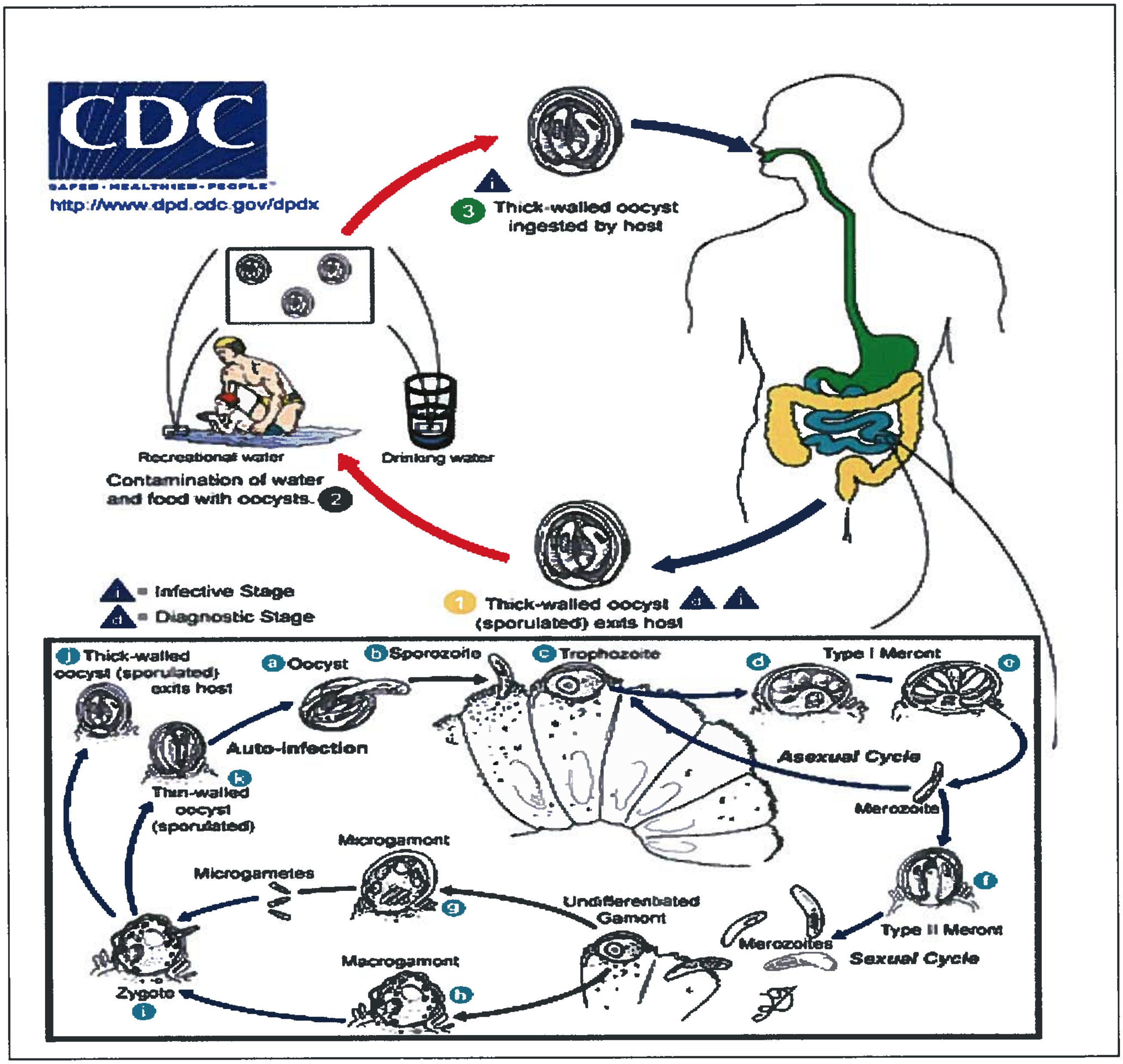


Figure 1.1. Depiction of the basic life cycle of *Cryptosporidium* within and outside of the human host[37].

1.2 Clinical Pathogenesis

1.2.1 Host Colonization

Oocysts are encountered through a fecal-oral transmission dynamic. Once ingested oocysts excyst or break open releasing four individual parasites known as sporozoites. Sporozoites have a predilection for the gastro-intestinal tract, principally to the ileum or jejunum of the lower small intestine where they enter and parasitize the luminal space of the epithelial cells of the brush border of the microvillus (Figure 1.2) [16, 56, 105, 119, 133, 138]. *Cryptosporidium,* like all Apicomplexans, is considered an intracellular but extra-cytoplasmic organism[1]. This positioning could enable it to evade the host's standard immune surveillance giving it time to colonize and establish disease. Studies have shown that these parasites can colonize other tissues such as respiratory tissues, different regions of the digestive tract or the conjunctiva[2, 3, 31, 35]. Some research indicates that the clinical presentation of cryptosporidiosis can vary in severity depending on where precisely it has localized to[171]. Periera et al. (2002) used a gnotobiotic piglet model of *C. hominis* and *C. parvum* infection and found that the parasite consistently invaded the ileum and colon when challenged with *C. hominis* versus the jejunum, duodenum, and ileum when challenged with *C. parvum*[171]. The clinical outcome of each infection was significantly different leading to the hypothesis that the colonization site within the host could be a determinant of clinical manifestation by that host.

Figure 1.2 Brush border of human intestinal tract.

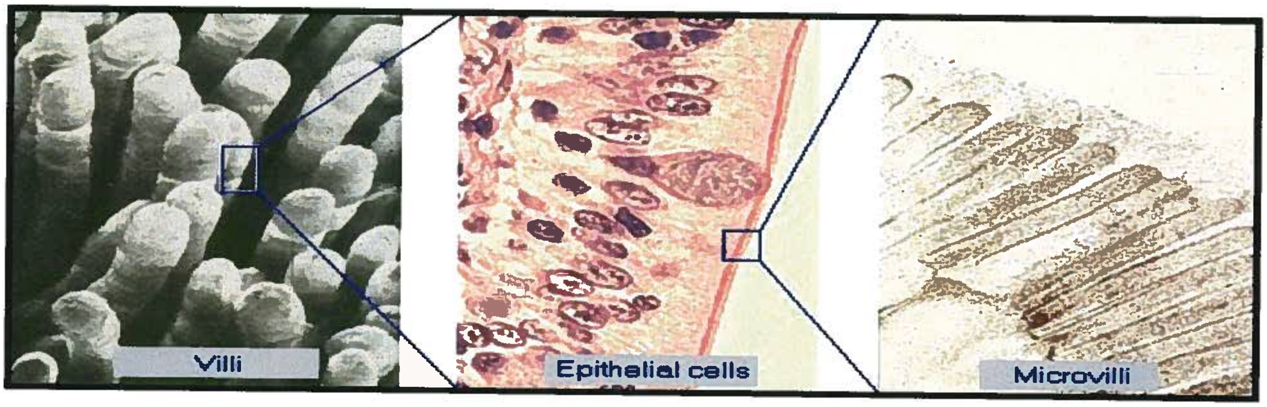


Figure 1.2. Scans of the microvilli lining the gastrointestinal tract, illustrating the predominant site of human host colonization of *Cryptosporidium* once ingested. Image at: capra.iespana.es/.../intestino/intestinoing.htm 627 x 300

1.2.2 Clinical Manifestation

Asymptomatic infections of *Cryptosporidium* have been reported[57, 62, 63]. Clinical cryptosporidiosis presents as an acute gastroenteritis with classic symptoms such as nausea, cramping, fever, weight loss, fatigue, and most notably profuse secretory diarrheal episodes[56, 63, 98, 137, 153, 156,191]. The volume of diarrhea can be extreme, with 3L/day being common and with reports of up to 17L/day[42, 63, 151]. When infected, humans can excrete up to 10^{10} oocysts per gram of feces[63]. The prepatent period, time from ingestion of oocysts to the excretion of oocysts following completion of the life cycle, can be anywhere from 3-5 days up to two weeks. Duration of infection is largely dependent on the immune status of the patient though typically symptoms remit in 30 days or so. While self-limiting in the immunocompetent host, infection can result in chronic disease with an intractable diarrhea in the immunocompromised host, making it an opportunist pathogen[62, 63, 99, 120, 137]. The greatest impact of *Cryptosporidium* is seen on the HIV(+) community where infection has become an AIDS defining diagnoses with a more than 2-fold hazard of death than other AIDS defining diagnosis's[71, 99, 156, 170]. Livestock and domesticated or companion animals exhibit the same prominent clinical signs of infection as humans, voluminous watery diarrhea but more severe cases more often result in mortality[62, 63].

The molecular basis for pathogenicity is not well understood and no specific virulence factors have been unequivocally shown to cause direct or indirect damage to host tissues. Cell death comes about as a direct result of the parasites invasion, multiplication, and extrusion[157, 158]. Pathologically, the increased turnover of mature intestinal epithelial cells with immature cells results in a loss of absorptive capacity of the epithelium causing a reduced ability of the host to digest and absorb fats and fat soluble vitamins[78]. The subsequent release of inflammatory cell mediators stimulates electrolyte secretion and diarrhea. Morphologically cell damage is a result of villous atrophy, lengthening of the crypt, mitochondrial changes and an upsurge of lysozomal activity in infected cells due to T-cell mediated inflammation[227].

1.2.3 Immune System Response

The human host immune response to *Cryptosporidium* has not been extensively studied and is still for the most part poorly understood. The parasite appears to make little effort in evading the immune system of the host. Many of *Cryptosporidium's* surface proteins, glycoproteins, and phospholipids are strongly immunogenic and antigenically cross-reactive[172, 175, 230]. It seems plausible that this high immunologic profile may represent a survival strategy of the organism.

The primary mechanism of host defence appears to be cellular immunity though to some degree humoral immunity is also known to be involved[21, 57, 178, 179, 199]. Mouse models of infection have suggested that IL-12, an important interferon-γ inducer, could play a critical role in determining resistance to *C. parvum* infection[21]. Hunter et al (2002) confirmed Tcell, specifically CD4 T-Lymphocyte, involvement in the mechanisms of immunity and demonstrated that Tcell deficient or impaired mice consistently presented with increased susceptibility and a more severe course of disease. In contrast studies on human cases of *Cryptosporidium* have shown that primary exposure is not sufficient to protect against future bouts of disease[57, 158]. Combined these two studies would suggest a predominant T-lymphocyte involvement and minimal involvement of B-lymphocytes or memory introduced into the immune system.

1.2.4 Diagnosis

When *C. parvum* was first diagnosed as a human pathogen diagnosis was made by a biopsy of intestinal tissue[111]. Methodologies for the routine monitoring of *Cryptosporidium* are only semi-quantitative as they do not provide information on the viability or human infectivity[33, 34]. A variety of tests such as ELISA (Enzyme-Linked Immuno Assay) and IFA (immunoflourescence assay) can detect anti-Cryptosporidial IgM, IgG and IgA antibodies but they are unable to distinguish between different pathogenic species or distinct genotypes of a given species. In recent years the advancement of PCR and related techniques has proven reliable for accurate species identification and distinction[4, 15, 62].

Medical laboratory diagnosis of cryptosporidiosis is relatively simple[7, 34]. Stool samples are collected and sugar flotation or a comparable technique is used to concentrate the organism[7]. Acid-fast staining methods, with or without stool concentration, are most frequently used in clinical laboratories. For greatest sensitivity and specificity, immunoflourescence microscopy is the method of choice, followed closely by enzyme immunoassays[33, 34]. Niehlsen acid fast staining or IFA staining of oocysts in fecal smears are proficient for indicating the presence of parasites (Figure 1.3). Histological sections from a biopsy of intestinal epithelium indicating any stages of the organism can also provide a positive identification of the parasite. However a majority of cases are not confirmed and/or reported to health officials as not all patients seek treatment or health professionals fail to submit specimens for *Cryptosporidium* specific diagnosis. The result of this, especially in an outbreak situation, leads to a skewed representation of actual cases as few are definitively confirmed through a diagnostic laboratory and many are simply resolved with the speculation of *Cryptosporidium* infection.

Figure 1.3 *Cryptosporidium* detection; acid-fast staining method.

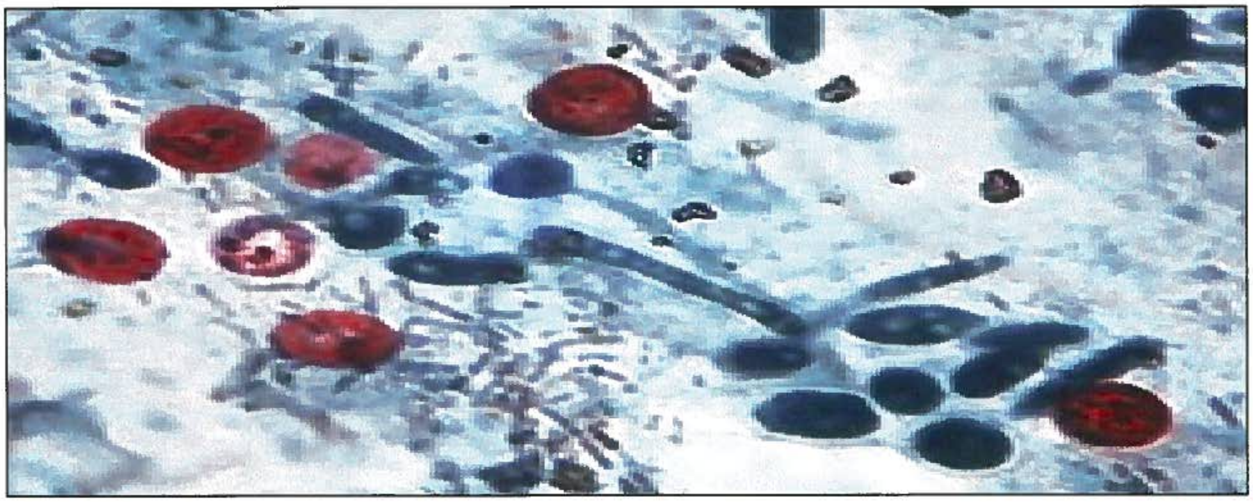


Figure 1.3. Micrograph of a direct fecal smear stained to detect *Cryptosporidium* using modified cold Kinyoun acid-fast staining technique. *Cryptosporidium* oocysts are stained red. Source: CDC/Dr. Pearl Ma.

1.2.5 Chemotherapy & Treatment

There is no definitive cure for cryptosporidiosis and efficacious chemotherapies or vaccines have yet to be described in literature[61, 107, 108, 110, 118]. Despite more than 120 drugs tested against *Cryptosporidium* effective anti-microbial treatments for the disease are still lacking. Some have proven toxic to the patient at the doses required to reduce parasite multiplication while others have shown efficacy only in animal models and most have shown no efficacy at all. Though limited in scope some studies have indicated moderate success with the use of macrolide antibiotics (Figure 1.4). Currently the treatment of choice is symptomatic based. Oral rehydration, anti-diarrhea medications and electrolyte replacement are all recommended[63, 65, 240].

Figure 1.4. Macrolide antibiotics recently explored as *Cryptosporidium* therapeutics[63, 65].

Nitazoxanide - treatment demonstrates some effectiveness when administered to immunocompetent patients with a dramatically more severe clinical manifestation of disease. Nitazanoxide has recently been approved by the United States Food and Drug Administration for treatment of *Cryptosporidiosis* in 1-11 year old children (USFDA, 2002).

Azithromycin - another drug that has demonstrated some improvement in diarrhea symptoms when given to immunosuppressed children.

Octreotide - though it has no effect against the organism itself, Octreotide also appears to help control diarrhea symptoms as alternative uses have shown it effective against watery diarrhea and in the reduction of flushing.

Parmomycin – studies have proven a mild effectiveness against the actual parasite when used in HIV (+) individuals. A substantial decline in the number of infectious oocysts, a decrease in intensity of the disease and improvements in intestinal function and morphology were reported[78].

Figure 1.4 Major macrolide antibiotics used as chemotherapeutic agents against *Cryptosporidium* infection with varying degrees of success.

1.3 Epidemiology

1.3.1 Exposure

Person-to-person contact, agriculture, livestock, wildlife and drinking or recreational water sources are all principle points of exposure to *Cryptosporidium*[17, 90, 125, 165, 181]. *Cryptosporidium* is considered to be a highly contagious and communicable disease[17, 191]. By some accounts *Cryptosporidium* is one of the most important parasitic causes of diarrheal disease[154]. The success of *Cryptosporidium* is attributed to multiple factors. *Cryptosporidium* oocysts, the infectious form of *Cryptosporidium,* have a very resistant nature. Oocysts are omnipresent, highly stable and have been isolated from a diverse array of ecological situations: lakes, rivers, streams, ponds, marshlands and saline rich coastlines. These tiny spore-like bodies are surrounded by a tough protective wall and can remain in their infectious state outside of the host whereas many parasites are only in their infectious state once inside their host. The thick outer wall of an oocyst measures just 4-5um in diameter, about half that of a normal red blood cell[7, 15]. Oocysts are shed in the feces and have been shown to have an environmental persistence of up to 6 months[62, 63, 64, 65]. Also contributing to the parasite's success is the fact that only a few oocysts are required to establish infection[62, 64]. As many as 100 different mammals can serve as a reservoir host for infectious *Cryptosporidium* species making it even more universal [148, 177, 181, 210, 227]. Humans and livestock, particularly livestock neonates, are considered to be the most significant source of oocysts[42, 165, 181].

1.3.2 Transmission Dynamics

With the host range of the organism being so broad the transmission routes of *Cryptosporidium* oocysts becomes multi-faceted and very complex. This is further complicated by the ability of different transmission routes to interlace with one another (Figure 1.5). This makes tracking infection sources and subsequent transmission routes arduous. *Cryptosporidium* is transferred through zoonotic transmission (animal to human)[37, 62, 63], anthroponotic transmission[111] (human to human), through contact with fecally contaminated surfaces[35, 63], and the ingestion of fecally contaminated food[111, 143, 180] (food borne) and water (waterborne)[86, 90, 99, 234, 240] and though rare *via* aerosol[94] transmission. Regardless of the mode of transmission the transfer of parasitic oocysts from host-to-host is mediated by the fecal-oral transmission route.

Zoonotic Transmission

C. parvum is capable of infecting most species of mammals[62, 63, 230]. In many zoonotic diseases a vector agent is required however *Cryptosporidium* can pass from animal to human through direct fecal contact with infected animals. Most of these situations arise on farms, petting zoos or direct contact with wildlife[10, 12, 134].

Bovines are the primary reservoirs for *C. parvum* and they play a central role in maintaining and disseminating oocysts because of their high susceptibility to disease and extensive diarrheal episodes[10, 11, 208]. The disease is most prevalent in neonates but the definitive source of infection and the direction of transmission between calf and adult or adult and calf are still unclear. Calves can excrete10^{10} oocysts per day[93, 197]. Outbreaks of cryptosporidiosis have been associated with both beef and dairy cattle. The evidence for *C. parvum* transmission from calves to humans is unequivocal, particularly during the calving season. Besides direct contact with livestock human cases may also arise as agriculture or farm personnel and equipment become vehicles for transmission. Furthermore the exceedingly high prevalence of infected calves on dairy farms raises additional questions about the prudence of handling and drinking unpasteurized milk. While the human threat from *Cryptosporidium* within the public sector is of greatest concern the indirect correlation to the prevalence of *Cryptosporidium* on farm and agricultural lands cannot be ignored.

Many infectious agents, mostly parasites, are carried by wild animals[176, 196]. Direct fecal-oral transmission between people and wildlife on farms or petting zoos can facilitate zoonotic transfer of infectious oocysts[173, 176, 197]. The precise significance of wildlife as a reservoir for farm animal or human cases still needs to be elucidated. More recently, *Cryptosporidium* has been reported in mice, feral pigs, wild rabbits, foxes, squirrels, chipmunks, and muskrats[12, 40, 173, 184, 189, 209, 213]. Some of these animals can often be found in urban areas therefore increasing the opportunity for transmission to humans and domestic animals[6, 176]. Wild animals often share their habitats with farm animals and agricultural lands, providing an additional source for environmental contamination and livestock contamination which could ultimately carry on to man[181]. Humans have close interaction with companion animals. Sharing living spaces also means sharing microorganisms that can cause disease. Common pets include dogs, cats and birds and the more exotic fish, snakes, lizards and ferrets. Though suspected, reports regarding cases of transfer between household pets such as cats and dogs to humans are limited in the literature. The prospect of acquiring *Cryptosporidium* from a household pet is typically more serious for children, the elderly and the immunocompromised[45, 76, 111, 131].

Anthroponotic Transmission

Human to human contact involving an infected individual also facilitates spread of disease[72, 107]. Typically these cases migrate out to the immediate family or other household members who are likely to be exposed to the organism. *Cryptosporidium* transmission occurs with very high frequency in children's facilities such as daycares and schools[111]. Infants or young children are clustered in classrooms, share toilets, and common play areas or necessitate frequent diaper changing. Nosocomial infection or hospital acquired infection is another major opportunity for anthroponotic transmission making both staff and patients vulnerable[35]. The housing of multiple patients in close quarters increases the chance of cross-infection as do staff circulating about the hospital from patient to patient and ward to ward.

Foodborne transmission

Occasionally food sources, such as raw meat, unpasteurized products and fruits or vegetables, may serve as vehicles for transmission. This is presumably because of contamination through fecal matter in untreated water used to wash, irrigate or spray crops[133]. Furthermore foodborne transmission can result from improper food preparation and/or safety measures in the food itself. A *Cryptosporidium* outbreak in Maine was traced to children who drank fresh-pressed apple cider contaminated by animal feces at a county fair[143]. This is thought to be the first documented outbreak using this transmission mode. The handling of food in unsanitary conditions or by unsanitary workers is of great concern as they may unwittingly transfer oocysts to foods not cooked after handling thus creating the potential for large-scale outbreaks[180, 198].

Waterborne Transmission

In the last two decades enteric protozoa have become the leading cause of waterborne disease outbreaks for which an etiologic agent can be determined. Waterborne transmission of *Cryptosporidium* is the most significant route of exposure for sporadic and outbreak situations. The most common sources are contaminated drinking and recreational water sources.

Contaminated drinking water has shown to be responsible for many outbreaks[90, 99, 100, 101, 121, 122, 130, 165, 239]. General causes include inadequate treatment practices, contamination at treatment plants, and direct sewage contamination through pipe leakage, breakage, backsiphoning and cross-connections. Most municipalities throughout North America acquire drinking water from surface and groundwater resources[59]. Recent studies indicate that *Cryptosporidium* oocysts are present in 65-97% of surface waters (i.e. rivers, lakes, streams) tested throughout North America[59]. Concentrations of oocysts as high as 5800 per litre of surface water have been found. Groundwater is also impacted with estimates of 9.5%-22% of the United States ground water samples testing positive for *Cryptosporidium*[59]. The largest waterborne disease outbreak for any pathogen ever recorded resulted in approximately 403,000 cases of cryptosporidiosis in Milwaukee, Wisconsin in the early 1990's[86, 130]. Runoff from nearby dairy farms, drainage from an abattoir and other sources were all suspected, thus creating an in-direct component of zoonotic transmission.

An association to agriculture or wildlife in close proximity to water resource facilities is thought to be a major contributor to this transmission mode[95]. Water intake facilities located near agricultural or pasture land provides opportunity for oocysts to enter public water systems through run-off waters. During a confirmed waterborne outbreak in the British Columbia interior, oocysts were detected in 70% of the cattle fecal specimens collected in the watershed close to the reservoir intake[163].

Non-domesticated wildlife and livestock can also initiate drinking water contamination by gaining access to and releasing infectious oocysts in regions designated as protected water resources for human consumption. Graczyk et al (1997) found that feces from migratory Canada geese collected in 7 of the nine sites in Chesapeake Bay contained *Cryptosporidium* oocysts[80]. Oocysts from three of the sites were infectious to mice and identified as *C. parvum*. Based on this it would appear waterfowl can pick up infectious *Cryptosporidium* spp. from their habitat and deposit it into the environment, including drinking water supplies where it becomes accessible to humans.

Recreational water amenities are a second route for encounter, particularly those visited by small children. Unintentional fecal release from infected babies or toddlers could contaminate a pool, wading pool or hot tub enough that upon ingestion of water others would be exposed. This combined with oocysts resistance to chlorine, low infectious dose, and a high bather density creates optimal conditions for outbreak situations. In both the United States and Canada numerous outbreaks associated with swimming pools, waterslides and water parks have been documented[18, 104, 200].

According to the Natural Resources Defence Council of the United States, at least 33% of rivers and over 50% of lakes in North America are unfit for swimming, fishing and other activities. Contamination from uncontrollable animal and human sources contribute to the fecal burden making safeguards in these environments difficult or impossible to implement.

Figure 1.5. Interlacing of *Cryptosporidium* transmission dynamics.

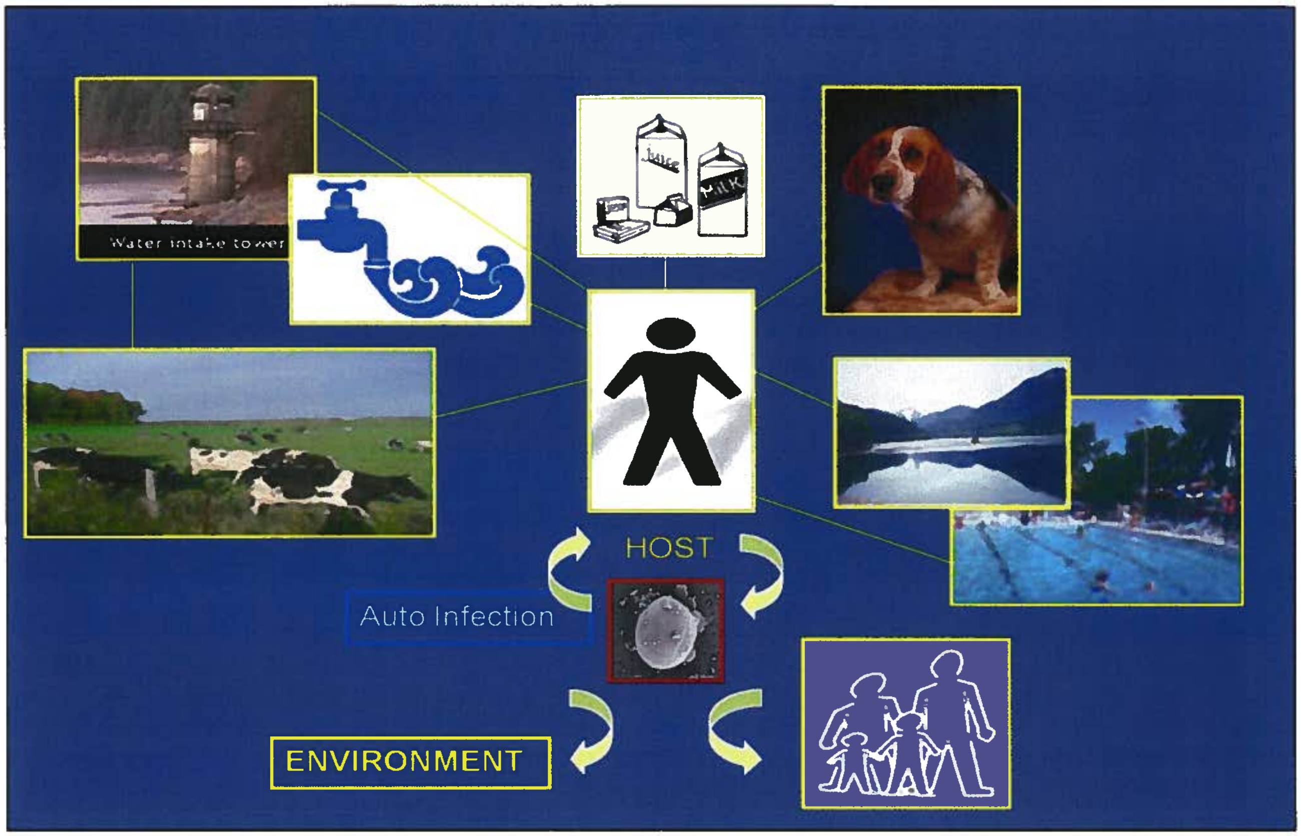


Figure 1.5. Simplified schematic *Cryptosporidium* transmission dynamics, illustrating sources of exposure and pathways of spread. Produced by author; JMW.

1.3.3 Prevalence

Infectious diarrheal diseases are the second leading cause of morbidity and mortality in the world. An estimated 200 to 375 million episodes occur each year in the U.S. alone, resulting in 73 million physician consultations, 1.8 million hospitalizations and 3,100 deaths[240]. Worldwide, there are 3.1 million deaths associated with diarrhea each year, more than 8,400 per day, mostly among children in developing countries[240]. Some estimates claim that 3-7% of all reported diarrheal diseases in the third world can be traced to *Cryptosporidium* species[19]. In industrialized nations it is estimated that somewhere around 0.4% of the population appears to be passing oocysts in the feces at any one time. Unfortunately most countries in the world have no testing protocols for *Cryptosporidium*, either on a routine basis or as a cause of death when diarrhea is implicated. The result is that many cases go unrecognized or are settled with speculation therefore skewing the actual rates of incidence for disease. Most experts collectively agree the actual number of cases is much higher than those reported on or documented.

At least nine molecularly different types of *Cryptosporidium* have been found to infect humans, whether immunocompetent or not[62, 63, 65]. The vast majority of cases are caused by *C. hominis* and *C. parvum* making them the biggest concern from a public health standpoint. In the United States and Australia *C. hominis* is responsible for greater than 75% and 85-92%, respectively, of all human infections[170, 212]. In contrast, it is reported that in the United Kingdom human cases are mainly the result of *C. parvum* infection, 61.5%, while *C. hominis* accounts for approximately 37.8% of all human cryptosporidiosis cases[139]. The difference is likely a result of the obvious separation between urban and rural populations in the US and Australia when compared to those of the UK, where agriculture plays a more significant role. In developing countries, diarrheal disease is much harder to contain and is worsened by the increase in migration and movement of populations in the last two decades enabling national boundaries to disappear as far as the transmission of disease is concerned[98, 240]. While the prevalence in these regions is likely extremely high the lack of infrastructure in place to document suspected cases prevents precise estimates. According to WHO the three most common causes of protozoan diarrhea are *Cryptosporidium parvum*, *Giardia intestinalis,* and *Entamoeba histolytica.*

Temporal circumstances also play an elemental role in *Cryptosporidium* epidemiology. *Cryptosporidium* has been shown to have a seasonal distribution based on geography, and temporal trends or patterns leading to an increase in parasite burden in the environment[10]. Many outbreaks have been dated to post-rainy seasons as increased rainfall and run off events are major factors affecting the total microorganism load in water sources. In the Northern hemisphere *Cryptosporidium* generally becomes a problem from March to June[59, 62, 63]. Typically during the spring season rains increase the run-off and the population of neonate animals is higher. In both Great Britain and the West Coast of Canada this season tends to be a little longer, beginning in February and lasting until mid-May[59, 62, 63]. Earth is experiencing increasingly more severe weather systems, in turn leading to more frequent and more extensive flooding. To emphasize the effects of climate change and weather patterns on engineered water systems we only have to look at the recent major flooding events of the Asian tsunami and Hurricane Katrina. During massive flooding events water systems become inundated resulting in their collapse and contamination with human and agricultural wastes which ultimately leads to a lack of potable water for human consumption. Flooding can lead to significant population displacements which compromise normal hygiene standards creating the perfect dynamics for large-scale outbreaks.

1.3.4 Demographics & Sociological Influences

Cryptosporidium is cosmopolitan in its distribution. All humans are presumed susceptible to infection regardless of age, race and gender[19, 111]. *Cryptosporidium* infection has been reported in persons from three days of age to ninety-five years of age[63]. The severity of *Cryptosporidium* typically varies according to age, immune status, and socioeconomic circumstances. There are extraneous factors that can increase the chances for an encounter with the organism or that will dictate the course of disease thus creating pockets of target populations for *Cryptosporidium* incidence (Table1.1).

While age is not a defining factor in *Cryptosporidium* epidemiology some of the most critical cases tend to appear in the age extremes of the elderly and young[20]. The immune system is either deteriorating with age or is just beginning to mature both of which create a vulnerability to opportunistic pathogens. Age can also impact the degree of personal hygiene among these populations as the elderly may not be physically able to tend to themselves properly or young children are unaware of the importance of proper sanitation.

Cryptosporidium is a serious illness in patients with suppressed immune systems making them a high risk target population for infection[32, 45]. Primary illnesses such as HIV (+) status, immune system deficiencies, and autoimmune disorders create a favourable niche for the parasite to colonize. Infection rates for AIDS patients are reported to be 4% and 2.5% in the United States and Canada respectively[240]. Patients undergoing immunosuppressive chemotherapies for cancers and organ transplant situations are also in a weakened immune state and have a greater susceptibility to the disease.

Since transmission is dominated by the fecal-oral route areas of inadequate sanitation and poor hygiene standards create another target population[8, 20, 169]. This renders cryptosporidiosis a disease of great socioeconomic status as these conditions are more likely to be found in poverty stricken regions or the developing world. According to the WHO worldwide approximately 1.1 billion people lack access to improved water sources and 2.4 billion have no basic sanitation. The scope of the problem is enormous as each year close to 4 billion cases of diarrhea occurs globally. In Southeast Asia and Africa diarrhea is responsible for as much as 8.5% of all deaths. The differences in diarrheal disease incidence in developed countries versus underdeveloped ones can be attributed to sanitation, access to potable water and personal and domestic hygiene[8]. A review of the geographic distribution and prevalence of *Cryptosporidium* based on oocyst detection and seroprevalance studies in humans from forty countries was compiled by Ungar et al (1990). Based on detection of oocysts in fecal specimens the prevalence of human infection in African countries (2.6-21.3%), Central and South American countries (3.2-31.5%), Asian countries (1.3-13.1) and others in the Pacific and Caribbean areas is considerably greater than that of Europe (0.1-14.1%) or North America (0.3-4.3%)[4, 62, 144].

Table 1.1

Target populations for *Cryptosporidium* exposure

Source	Likely Target Population	Rural/Urban
Daycares/schools	Infants, young kids, employees	either
Unfiltered/untreated drinking water	Small communities, farms, those using well based water	rural
Lambing, calving, muck spreading	Farmers, ranchers, ranch hands	rural
Sexual practices	Young to older adults, gay males	either
Nosocomial/health facilities	Elderly, patients, staff, visitors	either
Farm & zoo animals	Veterinarians, children, employees	either
Regions without water treatment standards	travelers	rural
Household pets (rare)	Household members	either

Table 1.1. Populations considered being at a slightly elevated risk of *Cryptosporidium* exposure.

1.3.5 Prevention & Control

Knowledge regarding clinical and ecological aspects of a pathogen are important if public health measures to control it are to be effective. Pathogens that have more complex, interlacing transmission dynamics demand a comprehensive approach to control and prevention strategies. *Cryptosporidium* is acquired through the ingestion or inhalation of infectious oocysts therefore control efforts are aimed at limiting host contact with the organism. Management of outbreaks calls for a multi-tiered approach encompassing scientific, medical, economic, political, and educational solutions. For *Cryptosporidium* this is three dimensional with efforts focused on public water safety, agricultural practices, and hygiene standards.

Cryptosporidium & Water Quality

It stands to reason that the role of water in the transmission of waterborne pathogens may increase substantially in importance and complexity as human and animal populations grow and the demand for potable water escalates. Contaminated water is commonly considered to be the most potential source of *Cryptosporidium* exposure making water purification the most important single measure available for ensuring public health[133]. In North America drinking water contaminated with oocysts has been blamed in a number of gastroenteritis outbreaks bringing about some apprehension about the safety of public water supplies[86, 90, 130, 163]. The 1993 Milwaukee outbreak, where over 400,000 people were affected, brought the issue of drinking water safety and standards to the forefront. The sheer magnitude of the outbreak and its association to water obtained from a municipal water plant that was operating within existing state and federal guidelines initiated questions as to the validity of current regulations. Not only did this outbreak emphasize the need for improved surveillance by public health agencies it also stimulated efforts to develop regulatory standards specific to *Cryptosporidium*. Here in Canada authorities have been forced to re-evaluate water treatment and monitoring practices following the tragic consequences of the waterborne outbreaks of *E. coli* 0157:H7 in Walkerton, Ontario and *C. parvum* in North Battleford, Saskatchewan[59, 90]. Both have awakened health officials here in Canada to our own vulnerability to such diseases despite our more "urbanized" utilities in place. Ignorance is no longer an option.

Water contamination can occur at any of the three major steps in water systems; source water, water treatment and water distribution[28]. The first link in the chain of providing access to clean safe water is ensuring the source water quality. Many municipalities in North America extract water from surface waters such as rivers, lakes and streams or groundwater resources. Many of the "supposedly" protected sources are susceptible to contamination from wildlife, accidents or contaminated groundwater flows[100]. Older water distribution systems are rapidly deteriorating. In any given city there are thousands of miles of piping and not only are the replacement costs extreme but the process is very slow, likely taking years. Although technologies are available to treat even the most contaminated water source it is an ongoing challenge for many communities to incur these costs and implement the more current or advanced methods.

Treatment of municipal drinking water is commonly done in two ways: through chemical treatments and through filtration. Chemically, chlorination is the most frequently used for disinfection of water by killing most viruses, bacteria and protozoa such as *Giardia*. Research shows that *Cryptosporidium* is 240,000 times more resistant to chlorination than *Giardia* and the actual amount needed to effectively kill *Cryptosporidium* oocysts would render water to toxic for consumption[101].

Filtration, using ultra-fine membranes, is a better bet for removing *Cryptosporidium* oocysts though the expense of such a system is a major issue[29]. Listed below are the contaminant removal parameters for biological agents in municipal water systems[28, 29].

- 5-100 microns, conventional filtration: removes human hair, the smallest particles visible to the naked eye and red blood cells.
- 0.1-5 microns, micro filtration: removes the smallest yeast cells, tobacco smoke and the smallest bacteria.
- 0.01-0.1 microns, ultra filtration: removes carbon black
- 0.001-0.01 microns, reverse osmosis: removes ionic particles such as polio virus, aqueous salts and metal ions.

With a size range of 4-6 microns for *Cryptosporidium* oocysts a minimum of micro filtration is required but preferable treatments would entail the use of ultra filtration or reverse osmosis[28]. Most current water systems are aging and in desperate need of upgrades to newer more technical systems such as filtration.

The use of ozone and ultra-violet (UV) lights has also been shown to disinfect water sources successfully with respect to *Cryptosporidium*[59, 63, 90]. Although ozonation of water demonstrates the ability to kill *Cryptosporidium* oocysts, the appropriate amounts of ozone needed to disinfect water at various temperatures and pH levels have not been clearly defined. In general, the amount of ozone needed to kill *Cryptosporidium* species is hundreds of times greater than that needed to kill bacterial contaminants[107].

For immunocompromised individuals avoiding contaminated water is particularly important[107]. The risks involved with tap water are still not clearly defined but are considered to be high enough that it is advised these people use properly filtered bottled water or boil water intended for drinking for a minimum of one minute. In-home purification and filtration systems can reduce the risk exposure providing it can remove particles 0.1-1 micron in size, filters via reverse osmosis or has an absolute 1 micrometer filter[107].

In an effort to get a handle on how best to kill *Cryptosporidium* oocysts a number of different techniques and chemicals have been tested (Table 1.2)[63].

Table 1.2

Deactivation Of *Cryptosporidium* Oocysts

Very Effective	**Somewhat effective**	**Not effective**
Boiling, $>73^{0}$, >1min	Ammonia	Phenol
Freezing, $<(-)2^{0}$C, >24hrs	Chlorine	Formaldehyde
Methyl bromide	UV light	Ethanol
Ethylene oxide	Iodine	Isopropyl alcohol
	Hydrogen peroxide	Lysol

Table 1.2. Efficacy of various methods for deactivation of infectious *Cryptosporidium* oocysts[56].

Despite all the advancements in understanding *Cryptosporidium* and how best to approach it in public water resources the relentless nature of the organism still creates problems. Most waterborne outbreaks of *Cryptosporidium* in North America have occurred in communities whose water facilities were compliant with governmental and health regulations. Although utility companies may adhere to guidelines, the guidelines themselves may not be sufficiently stringent for public protection. Recent surveys in the United States for the presence of *Cryptosporidium* oocysts in fully treated (disinfected and filtered) municipal water showed a small number of oocysts breached the barriers and could be isolated from tap water in nearly half of the communities evaluated[121, 122]. This gave birth to many questions. Whether or not a small number of oocysts in drinking water constitutes a large enough infectious dose to cause illness, are immunosuppressed persons more susceptible to lower doses, are there strains of *Cryptosporidium* that vary in infectious dose and infectivity?

Agriculture Approaches

Farming and agriculture lands may also play a role in introducing *Cryptosporidium* into the water systems, most often because of their locations to waterways and intake facilities. Because rainfall or snowmelts can transport contaminated fecal material from grazing fields cattle farms or feedlots should be located away from surface water sources such as rivers, lakes and streams. Stream bank fencing is recommended for landowners that pasture their livestock along these waters as it not only improves the overall water quality it protects the wildlife, fish and vegetation. Other practices encouraged are the reduction of stock density, separation of neonates from adult populations, minimal contact between personnel and calves and maintaining a relatively short calving period[11, 93].

Hygiene

With enteropathogens prevention centralizes around hygiene measures in any setting in attempt to interrupt fecal-oral transmission. Sanitation and personal hygiene standards are critical in the home and public places[45]. Epidemiological evidence suggests that hygiene and sanitation is at least as effective in preventing disease as is improved water supplies[133]. Regular hand washing is the number one recommended practice to prevent exposure to fecally transmitted microorganisms. The simple act of washing hands with soap and water can decrease diarrheal disease transmission by one third[240]. Vigilance is especially required after visits to hospitals, nursing homes and daycares or zoos. Avoidance of public pools and aquatic centers frequented by diapered or young children is encouraged. The amount of chlorine and types of filters used in public swimming pools do not prevent transmission from swimmers shedding infectious oocysts. The safe disposal of children's feces is critical as children are not only more likely to acquire diarrheal disease but they are most likely the source of infection also.

Implementation of the strategies for prevention and control of cryptosporidiosis are a more daunting challenge for the developing world[8, 19, 20]. The World Health Organization has listed *Cryptosporidium* as a "reference pathogen" for the monitoring of global water quality. A lack of fresh, clean water and poor sanitation conditions is a catalyst for the spread of disease making *Cryptosporidium* and similar pathogens endemic to these regions. The water supply and sanitation sectors, or lack of, will face enormous challenges over the coming decades as the urban populations of Africa, Asia, and Latin America are all expected to dramatically increase[20]. This will put great strain on an already failing system. In rural Africa, Asia and Latin America alone, just fewer than 2 billion, one third the total global population, are without access to improved sanitation. Approximately 1.1 billion are without improved water supply. Some believe these countries would be more appropriately referred to as "Thirst World" countries rather than third world. Innovative and cost effective source and treatment options, public initiatives, and monitoring programs directed towards the needs of these countries are essential if the global incidence of waterborne disease is to be crippled.

CHAPTER 2

GENOMICS OF *Cryptosporidium*

- Phylum, Genus, Species -

Chapter Summary - The *Cryptosporidium* genus, phylum Apicomplexa, has been undergoing constant revision and is the subject of great debate within the field. The accurate identification of a species or genotype is fundamental to the diagnosis, treatment, and prevention or control strategies of cryptosporidiosis in both humans and animals. The burden of disease attributable to a specific species is still elusive therefore hampering efforts to clarify the transmission dynamics and epidemiology of cryptosporidiosis. Herein the genomics of *Cryptosporidium* spp. as they are currently known are described.

2.1 Phylum Apicomplexa

Cryptosporidium has been classified as a genus within the Apicomplexa[17, 191]. Phylum Apicomplexa, previously known as Sporozoa, is large, complex, and consists of protists characterized by the presence of an apical complex. They are unicellular, spore-forming, and most often parasites of animals. All members are parasitic, have multifaceted life cycles involving both asexual and sexual reproduction and since most are intracellular lack any visible means of locomotion.

Apicomplexan parasites are eukaryotes and therefore share many metabolic pathways with their hosts. Because of this therapeutic target development becomes extremely difficult. An efficacious drug that harms an Apicomplexan parasite is also likely to cause harm or damage in its human or animal host. Biomedical research on these parasites is challenging because it is difficult, if not impossible, to maintain live parasite cultures in the laboratory and to genetically manipulate these organisms. This has impaired efforts to secure purified samples at different developmental stages. Research is forced to focus on intensive molecular studies to clarify the pathobiology. The most medically important and notorious Apicomplexa genus is *Plasmodium*, which includes the causative agents of malarial disease. Listed below are four of the classes within the Apicomplexan phylum and some genera they contain.

- Coccidia: *Cryptosporidium, Eimeria, Sarcocystis, Toxoplasma*
- Gregarinia: *Gregarina, Monocystis, Pseudomonocystis*
- Haemosporidians: *Hepatocystis, Plasmodium*
- Piroplasmids: *Babesia, Theileria*

In spite of its medical and veterinary importance *Cryptosporidium* has not been studied to the extent that other Apicomplexa, like *Plasmodium* spp., *Giardia* spp., and *Toxoplasma* spp., have. *Cryptosporidium* has the traditional hallmark features of an Apicomplexan organism however the differences between *Cryptosporidium* and other Apicomplexa have prevented the application of parallel scientific and therapeutic tactics.

2.2 Genus *Cryptosporidium*

Cryptosporidium is classified as a genus of protozoan parasites with multiple species capable of infecting mammals, reptiles, birds, fish and amphibians[65, 124, 136, 145, 146, 147, 227]. The taxonomic status of this genus is rapidly changing as new molecular data is published. The most widely recognized species, *C. parvum,* was once thought to be a single species with a broad host range whereas now several species have been identified. Presently there are 20 pathogenic species/genotypes of *Cryptosporidium* recognized (Table 2.1). With the exception of *C. parvum*, capable of infecting over 150 mammals, each species appears to have a more restricted host specificity or range. Initially studies indicated that species were limited to a single host. Molecular investigations have since challenged this concept. *Cryptosporidium muris*, *C. meleagridis, C. baileyi, C. canis,* and *C. felis*, considered to be contained to mice, turkeys, chickens, dogs, and cats respectively, have now all been documented in human infections[63, 230]. In the case of *C. parvum*, morphologically identical genotypes could eventually be accepted as separate species as biological and molecular data amasses, as was the case with *C. hominis.* The recent acceptance of *C. hominis* as a distinct human specific species has already been challenged. Though still incapable of establishing infection in mice *C. hominis* has now been shown to infect lambs, gnotobiotic pigs and higher primates under the appropriate laboratory conditions[2, 148, 2003]. Host-parasite co-evolution is also common in *Cryptosporidium,* as closely related hosts usually had related *Cryptosporidium* parasites[167]. The issue of host specificity is clearly multifarious and therefore is likely a fallible means for determining a species and its host range. Understanding the evolution of *Cryptosporidium* species is important not only for clarification of the taxonomy of the parasites but also for the assessment of the public health significance of *Cryptosporidium* parasites from animals. Until the full extent of intra-specific allelic variation in *Cryptosporidium* taxonomy is fully resolved numerous genotypes will likely continue to be described. The vagueness of the ancestral relationship of the *Cryptosporidium* genus as a whole has fuelled the need for extensive molecular and biological research into the organism.

Table 2.1

Recognized *Cryptosporidium* spp.		
Species/Genotype	Predominant Host	Reference
C. andersoni	Cattle; bovine (*bos Taurus*)	Lindsay et al., 2000
C. baileyi	Birds; chicken (*gallus gallus*)	Current et al., 1986
C. bovis	Cattle; bovine (*bos Taurus)*	Xiao et al., 2001
C. canis	Dog (*canis familiaris*)	Fayer et al., 2001
C. felis	Cat (*felis catis*)	Iseki, 1979; Pedraza-Diaz et al., 2000; Morgan et al., 2000
C. galli	Birds; (*Spermistidae fringillidae, G. Gallus*)	Pavlasek, 1999
C. hominis	Humans, higher primates (*homo sapiens*)	Morgan-Ryan et al., 2002; Xiao et al., 2001; McLaughlin et al., 1999, 2000
C. meleagridis	Birds; turkey (*meleagris galloparo*)	Slavin, 1955; Morgan et al., 2000; Xiao et al., 2001; Pedraza et al., 2001, Streter et al., 2000
C. molnari	Fish (*sparus aurata*)	Alvarez-Pellitero et al., 2002
C. muris	Rodents; mouse (*mus musculus*)	Katsumata et al., 2000; Gatei et al., 2003; Palmer et al., 2000; Tyzzer, 1907
C. nasorum	Fish (*naso literatus*)	Hoover et al., 1981
C. parvum	> 100 mammals; humans, livestock, wildlife	Xiao et al., 2001 & 2002; Tyzzer, 1912
C. saurophilum	Reptiles; lizards	Koudela et al., 1998
C. serpentis	Reptiles; snakes (Elaphe guttata)	Levine, 1980
C. suis	Pigs	Xiao et al., 2002
C. varanii	Reptiles; emerald monitor lizard	Pavlasek, 1995
C. wrairi	Guinea Pig (Cavia porcellus)	Vetterling et al., 1971
Cervine genotype	deer	Ong et al., 2002 & 2006
Rabbit genotype	Rabbits, hares	Xiao et al., 2002
Marsupial genotype	Marsupials	Morgan et al., 1999

Table 2.1. Accepted *Cryptosporidium* species and/or genotypes. The genus is undergoing constant review and revision and many of the species listed here were only recently identified, classified and accepted as a true species.

Abrahamsen et al. (2004) successfully sequenced and published the *C. parvum* genome in its entirety, a first in the *Cryptosporidium* research field[1]. This uncovered new genes/proteins and components of *Cryptosporidium's* biology, all of which could potentially be targeted or exploited in downstream studies. The genome is only 9.1 Mbp with 8 chromosomes all highly compacted and extremely gene dense[1]. This makes the *Cryptosporidium* genome 2.5x's smaller than that of *Plasmodium falciparum* but with 1.8x's greater gene density (Table 2.2, Appendix 1). Genome reduction has occurred predominantly through the shortening of intergenic regions, the loss and/or shortening of introns, and a reduction in the mean length of genes themselves[1]. The *C. parvum* genome was shown to have highly streamlined metabolic pathways rendering the organism dependent on the host for nutrient supply[1]. Of particular note was that in contrast to other Apicomplexans *Cryptosporidium* lacks a plastid and mitochondrial genome. In addition *Cryptosporidium* species demonstrate some peculiarities when compared to other Apicomplexa including an endogenous phase of development in microvilli of epithelial surfaces, two morpho-functional types of oocysts (thin and thick walled oocysts), and the smallest number of sporozoites per oocyst[1].

Table 2.2

***Cryptosporidium hominis* genome summary**

The genome	***C. hominis***	***C. parvum***	***P.falciparum***
Size (Mb)	9.16	9.11	22.85
Coding Regions$^+$			
G + C content (%)	32.3	31.9	23.7
Coding size (Mb)	6.29	6.80	12.03
Percentage coding	69	74	53
No. of genes	3,994	3,952	5,268
Mean gene length (bp)	1,576	1,720	2,283
Genes w/ introns (%) $^\pm$	5-20%	5%	54%
Intergenic Regions			
Non-coding size (Mb)	2.87	2.32	10.83
% non-coding	31	25	47
Mean length (bp)	716	585	1,694
Gene Ontology			
Biological processes	1,239	n.d.	1,613
Cellular component	1,265	n.d.	1,586
Molecular function	1,235	n.d.	1,625

Table 2.2. *C. hominis* genome summary in comparison to *C. parvum* and *P. falciparum* genomes, adapted from Xu et al., 2004[245]. Full table available in Appendix 1, Figure A.1. $^+$ Excluding introns. $^\pm$ Estimated intron content from expressed sequence tags.

2.3 *Cryptosporidium hominis* versus *Cryptosporidium parvum*

Though C. parvum has long thought to be the principal species, molecular investigations revealed that a genotype of *C. parvum* delineated into a separate molecular, host specificity, and transmission pattern[14, 170, 201, 202, 207, 232]. This genotype is now known as *C. hominis* which appears to be restricted to human hosts. Publication of the C. *hominis* genome following that of the *C. parvum* genome shows a highly similar gene complement with approximately 97% genetic identity between the two[1, 245]. The phenotypic variation of *Cryptosporidium* protein repertoires is likely due to these subtle genotypic differences. Parasite transmission, infectivity and host resistance could all be subject to key genetically determined variations between two closely related species or within the species itself.

Molecular evidence for the existence of two separate species has rapidly accumulated[2, 3, 26, 27, 30, 52, 170, 228, 241, 243]. Both are able to infect humans and have occasionally been simultaneously identified from the same host but a clear absence of recombinants is suggestive of reproductive incompatibility, a requirement of species distinction[2, 3]. In areas where both species are endemic genetic dimorphism is conserved[145]. Considering the sexual stage of its life cycle the lack of genetic recombinants is of interest. The development of the gnotobiotic pig model able to sustain *C. hominis* infection is highly significant as it is the first of its kind and has opened the door for comparative studies on the differences of biological behaviour between the two species[3, 171]. Using this model Pereira et al. showed that the prepatent period, disease severity and site of colonization all differed giving further evidence to the idea of two distinct species[171]. Akiyoshi (2003) used a gnotobiotic pig model system to demonstrate the inability of *C. hominis* to compete with *C. parvum*. When both were administered concurrently *C. parvum* consistently dominated or displaced C. *hominis*[3], implying type specific factors associated with virulence, transmission, and disease severity. Medically these differences could have considerable importance as indicators of specific risk factors for disease and patterns of epidemiology. Genetic variations correlated to clinically important parameters, specifically within attachment and invasion proteins, could explain why *C. hominis* preferentially infects humans.

2.4 Biogeographical Genetic Diversity of *Cryptosporidium*

C. hominis and *C. parvum* are responsible for greater than 90% of cryptosporidiosis cases in humans in most areas [244]. Geographic differences have been shown to exist among the species and the burden of disease attributable to them. In the United Kingdom, other parts of Europe and New Zealand, *C. parvum* is responsible for slightly more infections than *C. hominis*[40, 41, 85, 139]. In contrast, *C. hominis* is responsible for far more infections in North America, Australia, Japan, and developing countries where genotyping studies have been conducted [71, 120, 146, 163, 169, 170, 212]. The geographic prevalence of one species over another is likely a factor of transmission dynamics.

With the now common acceptance that *C. hominis* and *C. parvum* are in fact distinct species research has shifted to dissecting the intra-species pathogenomics. Evaluating intra-species variation will help elucidate the population structure which in turn will facilitate tracing outbreak sources and transmission routes. Such investigations can identify sub-species or sub-genotypes inhabiting a geographic subdivision of the range of a species[127]. In contrast to the original idea of a limited intra-genotypic variation within *C. hominis* and *C. parvum* sizeable intra-genotypic diversity has been proven and sub-genotypes identified[25, 40, 169, 186]. Sub-genotypes have revealed a wide-spread geographical distribution whereas others appear to be limited to a restricted geography.

A study of three outbreaks in Northern Ireland all attributed to drinking water contamination demonstrated the value of genotyping analysis, especially within a timely manner[75]. The research isolated and identified a sub-genotype of *C. hominis* in 2 of the 3 outbreaks. The same sub-genotype had been isolated in the United States, Canada, the United Kingdom, Portugal, and Peru suggesting there may be no correlation between strain and point of geographic origin. A second study focusing on the DHFR gene sequence in both *C. hominis* and *C. parvum* isolates from the United Kingdom, the United States, Canada and Guinea Bissau showed DNA sequence conservation indicating geography has no effect on intra-genotypic variation[9, 232]. A third study supporting the concept of independence from geography was conducted on the 18S rRNA locus using *C. hominis* and *C. parvum* isolates from patients with and without HIV, and living in Kenya, Malawi, Brazil, the United Kingdom or Vietnam[71]. Supported by phylogenetic analysis the results revealed a lack of specific variation in correlation to geography.

While geographically independent *C. hominis* and *C. parvum* sub-genotypes have been documented, other studies give evidence of genetic variation conserved to a defined geography. In Italy Caccio et al. (2001) used 2 microsatellite loci in zoonotic isolates of *C. parvum* from all over the country to examine sequence polymorphisms[27]. At the ML1 locus three alleles were discovered, two of which proved to be widespread and a third that was restricted to Southern Italy. Interpretation of this could argue for the non-random distribution which could be indicative of clonal populations adapted to a particular climate or environment. A similar study on *C. parvum*, done in Scotland using RFLP analysis of three different loci, revealed little to no evidence for geographical sub-structuring[82, 135]. It should be considered whether or not the conclusions reached from these studies are applicable to other geographic locales or are particular to the country studied. The analysis of samples from more diverse geographies could determine if a lack of geographical sub-structuring is due to a relatively small geographic area and the frequent movement of hosts between areas[135]. In an attempt to better define the geographic components of species distribution Gp60 sub-typing tools have been used. Results have revealed the complexity of *Cryptosporidium* transmission and could potentially have great significance because the Gp60 locus encodes two glycoproteins thought to be involved with attachment to and invasion of host cells. Numerous Gp60 sub-types of *C. parvum* and *C. hominis* have been seen in endemic areas of disease. In developing regions, the complexity of transmission is often reflected by the strong presence of many sub-types within *C. hominis*. An example of one study focusing on the highly polymorphic Gp60 locus recognized a geographically limited allele, Ie. This allele was one sub-type among many found predominantly in isolates from South Africa even though isolates from the United States, Brazil, Peru, Guatemala, and Zaire were genotyped[120]. High heterogeneity is likely an indicator of stable cryptosporidiosis transmission in the area while conversely in developed or industrialized nations the degree of heterogeneity is typically smaller with fewer sub-types predominating, suggestive of a more unstable transmission of cryptosporidiosis.

2.5 Theory of Clonal Population in *Cryptosporidium*

For many pathogenic organisms that utilize mainly asexual reproduction methods it is often unclear whether or not epidemics are the result of the emergence of pathogenic clones or environmentally determined increases in the population size of the organism[67, 88]. Descriptions of the genetic structures of epidemic populations are able to help distinguish between these competing ideas. Examinations of intra- and inter-genotype diversity have put forth the question of whether or not *Cryptosporidium* is in fact a clonal population. The occurrence of widespread genetically identical isolates, apparent parity between unlinked loci and infrequent genetic recombination suggests that it is. Whether or not these same conditions occur in nature and can occur between *C. hominis* and *C. parvum* has not yet been proven. A clonal population implies that meiotic recombination is rare but does not preclude the existence of complete meiosis or the occurrence of sexual reproduction[223]. Meiotic recombination bears great consequence for the generation and maintenance of variation within a species and can have long term evolutionary impact. Genetic variation can also be generated by intragenic recombination. While *Cryptosporidium* population structure is almost certainly highly clonal and dominated by the *C. parvum* and *C. hominis* lineages interlineage recombination can occur naturally producing mixed genotype progeny that are viable and infectious[217, 223]. There are no stable genetic differences among the lineages and this could hamper efforts to characterize molecularly diverse individual natural genotypes and sub-genotypes. The clones, not the species as wholes are distinctive evolving units therefore the hypothesis of clonality has great genetic and medical implications. Advances in our understanding of the population structure and pathogenic properties of *Cryptosporidium* will come from examining those polymorphisms capable of differentiating among isolates belonging to the same genotypic group.

Examining the genetic variation among isolates of the related *Giardia lamblia* one study showed a range of a virtual lack of genetic variation to extensive genetic variation[221]. The level at which variation does exist was left unclear, meaning the heterozygosity within an individual versus the polymorphism within a population. Some authors have argued for close relatedness of isolates throughout the world while others emphasize clonal lineages that are evolutionary independent. Similarly molecular evidence and arguments for both a genetically sub-structured and clonal population structure of *Cryptosporidium* has accumulated[13, 77,155, 223]. Frequent transmission from environmental sources could increase the probability of coinfections with genetically heterogeneous parasites thus favouring recombination. In countries where the sanitary conditions are better and diseases like HIV less prevalent, coinfections with heterogeneous parasites originating from environmental sources may be less frequent, and clonal propagation may prevail.

CHAPTER 3

STUDY DESIGN

- Research Concepts, Goals, Hypothesis & Significance -

Chapter Summary – Studies on population genetics provide insight into the genetic relationship of "difficult" complexes and taxonomic representations of a genus. The distribution of genetic variation in intercontinentally disjunct subpopulations may provide important information about the transmission dynamics, epidemiological behaviours and evolutionary patterns of *C. hominis.* The aim and design of the study enabled us to address research questions regarding the distribution of genetic variation in global *C. hominis* subpopulations, whether or not such subpopulations are genetically differentiated, and the efficacy of such a typing system for species specific identification.

3. 1 Aim

3.1.1 Central Research Questions

The specific aims for this study are threefold.

(1) What is the relationship of genetic diversity and genetic differentiation among international *C. hominis* subpopulations when partitioned into intra-population and inter-population genetic structures?

(2) Is the distribution pattern of genetic variation at the SNP level within a particular gene and/or genotype geographically conserved versus geographically widespread?

(3) Are single nucleotide polymorphisms (SNPs) mapped to the *C. hominis* genome a sound approach for molecular typing applications in epidemiological investigations; considering both established and/or novel genotypes in addition to sporadic and outbreak situations?

3.1.2 Hypothesis

We hypothesize that on the basis of the allelic profile of a panel of single nucleotide polymorphisms distributed throughout the *Cryptosporidium* genome, the degree and partitioning of genetic diversity within and among *C. hominis* subpopulations will be influenced geographically.

Secondly we hypothesize that through the use of a pre-defined pattern of single nucleotide polymorphisms mapped to multiple loci throughout the *Cryptosporidium* genome we can establish an efficient and reliable molecular tool for species distinction of *Cryptosporidium* isolates.

3.2 Experimental Objectives

To address our research questions three sequential experimental objectives had to be achieved. The initial objective was to identify, map and characterize SNPs from biologically relevant genes/proteins throughout the *Cryptosporidium* genome, allowing for the establishment of a data set of molecular markers. This provided the baseline foundation of the study setting the tone for all downstream processes. A panel of mutations needed to be assembled, creating a multi-locus SNP-type (MlSt) or haplotype, differing from the more commonly used and reported on approach of multi-locus genotyping for molecular epidemiology studies. To address different kinds of variation a combination of silent and expressed molecular markers, potential antigenic markers and those more likely to be under positive or diversifying selection pressures were used.

Our second objective was the design and development of a SNP-based molecular typing assay. The technological design had to be usable with the crude fecal specimens from which parasite DNA is isolated. Competition from other naturally or invasively occurring organisms is likely so it is essential the research design was capable of detecting and isolating *Cryptosporidium* DNA amongst that of other microorganisms. An efficient, standardized methodology was crucial as isolates that were kindly donated originated from facilities around the world and have all gone through various processing applications, depending on the lab's internal procedures and protocols.

To assess the full impact of geography on mutation profiles, the third objective was to establish as defined as possible geographic boundaries. Infectious agents do not obey national boundaries and given the opportunity, an organism will always spread. In light of this one application of genetic typing techniques is to track pathogens geographically. To allow for more rigorous epidemiological analyses to be made geographic boundaries with minimal overlap need to be established. This is particularly important when considering the globalization of the modern world. The more restricted or isolated the geography the lesser the likelihood that similar genotypes arise from human travel or urbanization.

3.3 Experimental Rationale

The purpose of this study was to evaluate the biogeographical distribution of genetic variation of *C. hominis* subpopulations. While we know *Cryptosporidium* populations have shown genetic substructuring we were interested in asking whether or not global *C. hominis* subpopulations have the same genetic structure. Examination of many isolates from different geographic origins, from both outbreak and sporadic cases using unlinked informative genetic markers is crucial to commanding a better understanding of the transmission route of disease and probable outbreak sources. Natural geographic barriers can directly impact the dispersion of the organism thus affecting the exchange of genetic material between species. The relative importance of zoonotic transmission patterns has already been shown to vary according to geography. The cohesive characteristics that discriminate between species of *Cryptosporidium* could be good predictors of range and potential for transmission. The extent or degree of genetic differentiation could be reflective of transmission intensities within restricted geographic boundaries, having a direct impact on the design and development of chemotherapies.

The Use of SNPs

Phylogenetic analyses are best done with neutral data but identifying genes that might characterize a transmission mode might well involve genes that are under selection. Allelic variation arises from many ways including random nucleotide mutations, diversifying selection, horizontal gene transfer and intragenic recombination events[54, 68, 81]. Single nucleotide polymorphisms may be synonymous and non-synonymous. Non-synonymous (NS) SNPs result in amino acid replacements and hence are targets for evolutionary selection. NS SNPs are also excellent markers when evaluating pathogenic species and the differences between them[68, 185]. In contrast, synonymous SNPs (S SNPs) do not alter the chemistry or structure of a protein and are therefore functionally and evolutionary neutral, or nearly so[23]. Synonymous SNPs occur with higher frequency and are therefore more accessible targets for genetic population studies to examine evolutionary relationships between species or sub-species[5, 185]. SNPs provide an efficient tool for making associations between whole genome comparisons and epidemiology[192]. The use of both antigenic relevant SNPs and more neutral molecular markers allows for deeper insight into the molecular epidemiology of *C. hominis.*

The worldwide threat from *Cryptosporidium* to human health emphasizes the need to develop rapid methods for the identification of genetic relationships among infectious strains, especially those responsible for mass outbreaks or more clinically severe disease. Restricted allelic variation can limit the utility of multi-locus sequence analysis for estimating phylogenetic relationships among strains or genotypes[192]. Hi-throughput SNP genotyping is an efficient way for assigning closely related strains to specific lineages, either identical or related by descent[83]. This removes a critical barrier to population and geographically based studies on the relationships between *Cryptosporidium* genotypes, where genetic variation is limited.

Compared with other molecular markers, SNPs exhibit extremely low mutation rates, making them rare in recently emerged pathogens and therefore extremely valuable from a phylogenetic standpoint [166, 185]. It is likely that the subtle genetic differences between *Cryptosporidium* species accounts for the variances observed in pathogenicity, host range, clinical presentation and response to therapies. If a clearer understanding of the biology and epidemiology of the emerging pathogen is to develop complete characterization of the few genetic polymorphisms that do exist is crucial.

With the many advantages of using SNPs for estimating and examining the taxonomy of *Cryptosporidium* it could become a tool routinely used to categorize species and strains of the genus. Hi-throughput SNP-typing is an attractive method for analysis of this type as hundreds to thousands of SNPs can be catalogued in relatively short time periods. It would be reasonable to assume that in an outbreak situation this magnitude of SNP genotyping would be a commodity.

To date studies focused on biogeographical SNP-typing of *Cryptosporidium* have been limited thus requiring further study into their use in *Cryptosporidium* genomics. While some have addressed the issue from a single locus, whole genome SNP analyses with a global perspective are so far grossly under-reported in the literature. A multi-locus, whole genome approach can better define the *Cryptosporidium* population structure.

CHAPTER 4

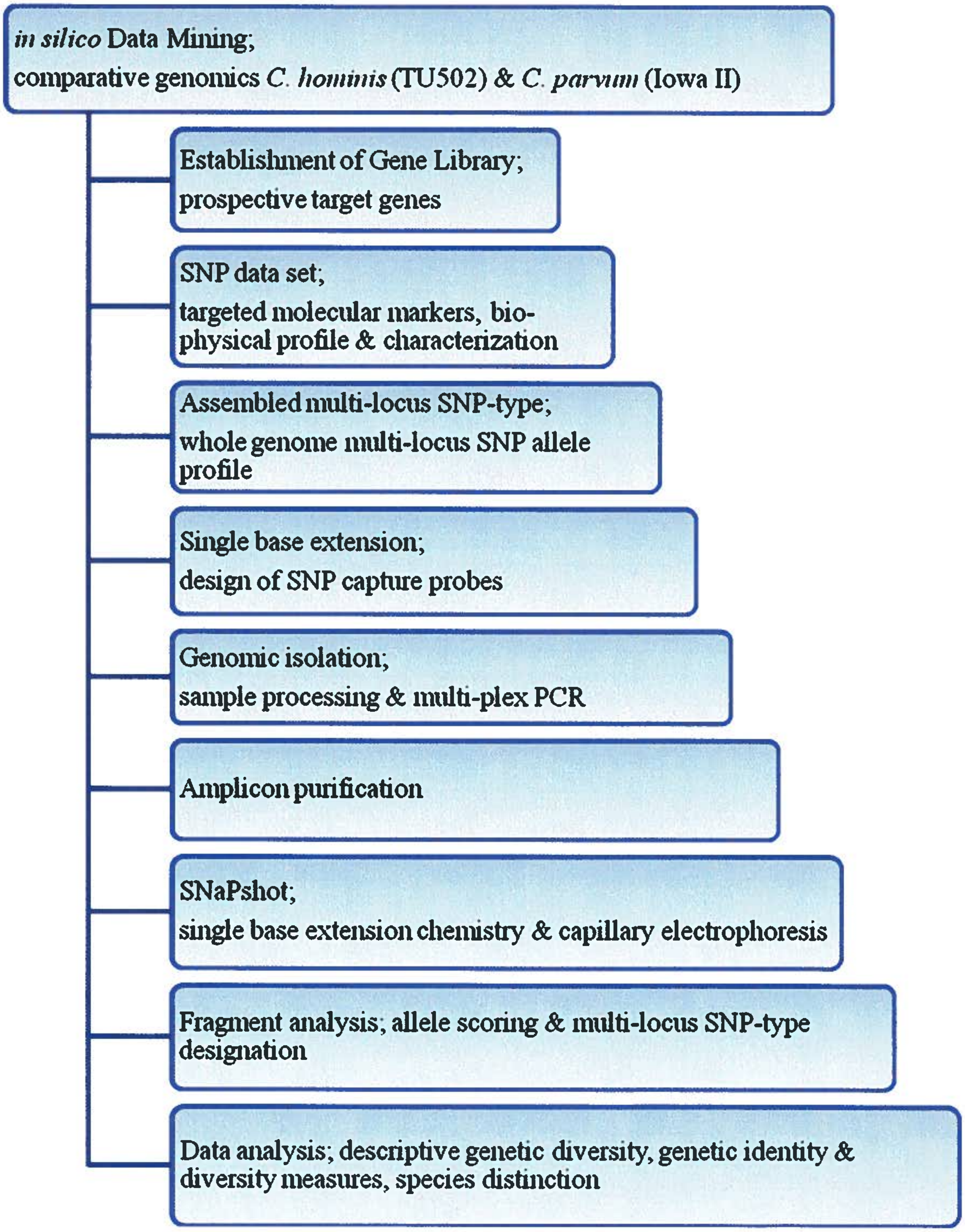
EXPERIMENTAL PLATFORM

- Methodology, Techniques, Instrumentation -

Chapter Summary – The experimental platform of this study involved a multiple method approach, uniting qualitative data with quantitative data, conducted in a sequential manner. A combination of genetic data mining, comparative genomics between the *C. hominis* and *C. parvum* genomes, and bioinformatics analyses of target loci led to the establishment of a 394 SNPs data set. A panel of 45 SNPs distributed throughout 13 loci was assembled establishing a multi-locus SNP-type (MlSt). A total of 77 international *C. hominis isolates* were processed and scored using the SNaPshot single base extension assay coupled to fragment analysis. Inferences about population structure were made using genetic data analysis software, GDA and Fstat. The ability of the SNP-typing scheme to discriminate between species was also assessed.

4.1 Experimental Workflow

Figure 4.1 Experimental Platform; sequential workflow.


in silico Data Mining;
comparative genomics *C. hominis* (TU502) & *C. parvum* (Iowa II)
Establishment of Gene Library;
prospective target genes
SNP data set;
targeted molecular markers, bio-physical profile & characterization
Assembled multi-locus SNP-type;
whole genome multi-locus SNP allele profile
Single base extension;
design of SNP capture probes
Genomic isolation;
sample processing & multi-plex PCR
Amplicon purification
SNaPshot;
single base extension chemistry & capillary electrophoresis
Fragment analysis; allele scoring & multi-locus SNP-type designation
Data analysis; descriptive genetic diversity, genetic identity & diversity measures, species distinction

4.2 Materials & Methods

4.2.1 Data Mining

From the published *C. hominis* and *C. parvum* genomes, based on reference strains TU502 and Iowa Type "II" respectively, exhaustive data mining was done to target genes for SNP-typing. Criteria were bio-functionality, genome location and indicators of selection pressures. A library of 161 candidate loci, covering all 8 chromosomes was compiled. Each gene was catalogued and compared to its ortholog in regards to chemical and physical properties, and genetic identity. For each locus all available sequences were used in alignment comparisons.

Nucleotide sequences were aligned followed by amino acid translation. Open reading frames (ORF) were determined by the presence of an in frame methionine residue. Highly hyper-variable genes and those that are completely conserved were discounted due to primer constraints and/or lack of genetic diversity. Gene sequences were aligned for primary sequence analysis and screened for SNPs using the SeqMan2™ module of Lasergene V5. Clustal W slow/accurate alignment at both the nucleotide and amino acid level for each gene was performed using the MegAlign™ module of Lasergene V5. Alignment reports generated were used to highlight single nucleotide polymorphisms conferring a single amino acid polymorphism (SAAP) within the peptide arrangement. On the basis of the inferred protein sequences SNPs were assigned as either synonymous (S) or non-synonymous (NS). Each gene or gene fragment was submitted as a BLAST query to ensure no significant similarity to other microorganisms that may be naturally or invasively occurring in fecal samples used in this study.

4.2.2 Bioinformatics Analyses: ORF Analysis & Biophysical Properties of Target Genes

ORF analysis of candidate genes/proteins consisted of profiling chemical, structural, and positional characteristics, with particular attention paid to those regions containing target SNPs. The Protean module of Lasergene V5 was used to generate graphical and numerical representations of hydropathic character, surface probability, antigenic indices, and predicted secondary structure (Appendix 4) [43, 70, 96, 103, 116].

The Kyte-Dolittle hydrophobicity plot is a graphical representation of the hydropathic score of a sequence of amino acid side chains in a protein[116]. Based on their ability to repel and attract water and to what degree each possible amino acid is assigned a number or score. Scores are biologically significant in that they are indicative of possible transmembrane domains. Proteins passing though the phospholipid bilayer of a cell interact with a region inside or outside of the cell, where they will find water, and will therefore have a hydrophobic region correlating to the hydrophobic region of the bilayer. Non-globular proteins, those without transmembrane domains, will be strictly hydrophilic in nature. With a scale set at (-) 4.5 – (+) 4.5 values greater than zero are suggestive of hydrophobic character while a value of two or more indicates strong hydrophobic region.

Emini Surface Probability indicates the probability of finding an amino acid residue on the surface of the protein molecule. With a threshold scale of 1-6 a value of 1 or greater increasingly predicts probability of protein surface location.

The computer software program, developed by Jameson-Wolf, is a multi-disciplinary index integrating hydrophilicity, surface probability, backbone flexibility, and secondary structure parameters to predict possible antigenic sites[103]. On a scale of (-) 1.7 – (+) 1.7 values approaching (+) 1.7 are reminiscent of antigenicity.

Complete Protean^{TM} profile examines the flexibility of the peptide backbone as predicted by the method of Karplus and Schulz (1985) is indicated. The propensity of the peptide chain to form various secondary structure conformations such as α-helix, β-sheet, and β -turn are calculated by both the Chou-Fasman and Garnier-Osguthorpe-Robson (GOR) algorithms[43, 70].

Quantitative evidence of diversifying selection for each locus was evaluated using the straightforward mathematical model that examines the ratio of non-synonymous to synonymous divergence[113, 114]. This ratio analysis is the most widely used criterion for detecting natural selection. A disproportionate number of NS: S substitutions would be indicative of regions under positive diversifying selection[128].

4.2.3 Target Loci: Multi-locus SNP-typing (MISt)

From 161 initially considered target genes 25 genes or families of genes were isolated for genetic typing (Table 4.1). These were organized in reaction sets (RS) comprised of two or three loci each to facilitate high-throughput and multiplexed PCR amplification. The remaining two loci used a single-plex PCR platform. From the 25 protein loci targeted 13 were established as initial targets. A pre-defined panel of 45 SNPs from these 13 was assembled for allele discrimination. The remaining 12 loci, RS's 6-9, were successfully amplified and prepared for multi-locus SNP-typing at a later date.

Table 4.1

MIS-typing Reaction Sets				
Reaction Set	**Gene Annotation**	**Gene Abbreviation**	**PCR Platform**	**Chromosomal Position**
1	Malate Dehydrogenase	MDH	Multi-plex	7
1	Lactate Dehydrogenase	LDH	Multi-plex	7
1	Uracil Phosphoribosyl Transferase	UPRTase	Multi-plex	1
2	Erythrocyte Membrane Associated Ag	EMAAg	Multi-plex	1
2	Apoptosis Related Protein	APR	Multi-plex	4
3	*Cryptosporidium* Oocyst Wall protein	COWP	Multi-plex	4
3	Beta-tubulin	B-tubulin	Multi-plex	6
4	Acetyl coA synthetase	ACoA	Multi-plex	8
4	Mucin-1	Mucin-1	Multi-plex	6
5	Cp23	Cp23	Multi-plex	4
5	18S rRNA	18s rRNA	Multi-plex	multi-copy
6	Cellcycle Regulator	CCR	Multi-plex	1
6	CTCL Tumor Ag	CTCL	Multi-plex	2
6	Aldahyde-Alcohol Dehyd'ase	AAD	Multi-plex	8
7	CLL Associated Ag-KW-2	CLL	Multi-plex	2
7	Sexual Stage Specific Kinase	SSSk	Multi-plex	3
7	FLJ31812/DHHC palmitoyl transferase	FLJ	Multi-plex	7
8	Transmembrane amino acid Transporter	TMaaT	Multi-plex	3
8	ABC multi-drug or ion efflux	ABC	Multi-plex	4
8	Thiolproteinase	Thiol	Multi-plex	7
9	Extracellular protein w/ 8 kazal repeats	Epro	Multi-plex	4
9	Seroreactive Ag BMN-19B related protein	SeroAg	Multi-plex	7
9	RIK protein w/? WD40 repeats	RIK	Multi-plex	8
10	Heat Shock Protein 70	HSP70	Single-plex	2
11	Gp60	Gp60	Single-plex	6

Table 4.1. The 25 genes targeted at the outset of the study, their abbreviated identifiers, and chromosomal position.

4.2.4 Primer Design

Gene sequences of regions spanning target SNPs were drawn from Genbank. Primers for multiplex PCR reactions were designed using Primer Select module of Lasergene V5 with criteria stipulated as: melting temperature of >52°C, length of 18-24 nucleotides, GC content within range of 40-60% (Table 4.2). Primers were synthesized by Sigma-Genosys at a scale of 0.5ul and were RP1 purified (Cambridge, UK). Primers were validated by single-plex amplification of using standard HotStar Taq (Qiagen, Mississauga, ON) PCR conditions: 1X PCR master mix, 800uM dNTPs, 1.5mM MgCl2 and 0.5uM each primer. Reactions consisted of an initial denaturation at 94 °C for 15m followed by 40 cycles of denaturation for 1m at 94°C, annealing for 1m at 55 °C, and extension for 1m at 72°C, with a final extension at 72 °C for 10m.

Single base extension (SBE), primer design was done manually using parameters set out in the corresponding SNaPshot^{TM} protocol (Applied Biosystems, Foster City, CA). An individual capture primer for each viable SNP target was developed. Primers containing neighbouring SNPs within 15 nucleotides upstream (5`) of the target SNP were excluded to ensure primer conservation. Primers all met the stipulations of: minimum melting temperature of 50 °C, GC content of less than 70-80%, as devoid as possible of secondary structure interactions. All SBE primers were PAGE purified. A non-annealing GACT nucleotide repeat tail was added to the 5` end of the oligomer in different size ranges to ensure separation in fragment analysis. A minimum primer length, including the GACT tail, of 36 nucleotides was designed for. All original primer sequences were modified with an initial GACT repeat sequence to total 36 nucleotides followed by varying increments of GACT repeats added thus ensuring a fragment separation of nucleotides between target SNPs and allowing for mobility shift due to different dye weights. SBE primers used in this study are listed in Table 4.3.

4.2.5 Multi-plex PCR Amplification

Genomic DNA was extracted from fecal *C. hominis* specimens. Target gene regions were amplified using Qiagen Multiplex PCR Kit (catalogue #206143). Primers and expected amplicon sizes are listed in Table 4.2. To ensure optimal PCR conditions the protocol stipulating the use of Q-solution was followed. Q-solution changes the melting temperature of DNA and improves sub-optimal PCR caused by templates that have a high degree of secondary structure or are GC% rich[141]. Reaction mixes were made of 2x Qiagen Multiplex PCR Master Mix (final concentration of 3mM MgCl2) at a 25ul volume, 10x primer mix (final concentration of 2uM per primer) at a 5ul volume, 5x Q-solution at a 5ul volume and RNase-free water at a 5ul volume. Ten microliters of template DNA (< 1ug DNA/50ul) was added to give a final reaction volume of 50ul. Using Applied Biosytem 2720 thermacycler an initial activation step of 95°C for 15m was followed by 40 cycles of: 94°C, 30s; 60°C, 90s; 72°C, 60s. After a final extension of 72°C for 10m samples were consumed immediately or stored at 4°C for up to 24hrs or at -20°C for longer than 24hrs. Samples were verified on a 1.8% agarose gel electrophoresis at 110volts for 1hr 15m alongside 100bp molecular weight marker supplied by Invitrogen (Figure 4.2).

Figure 4.2 Multi-plex PCR; gel electrophoresis of Cp23, COWP & β-tubulin loci.

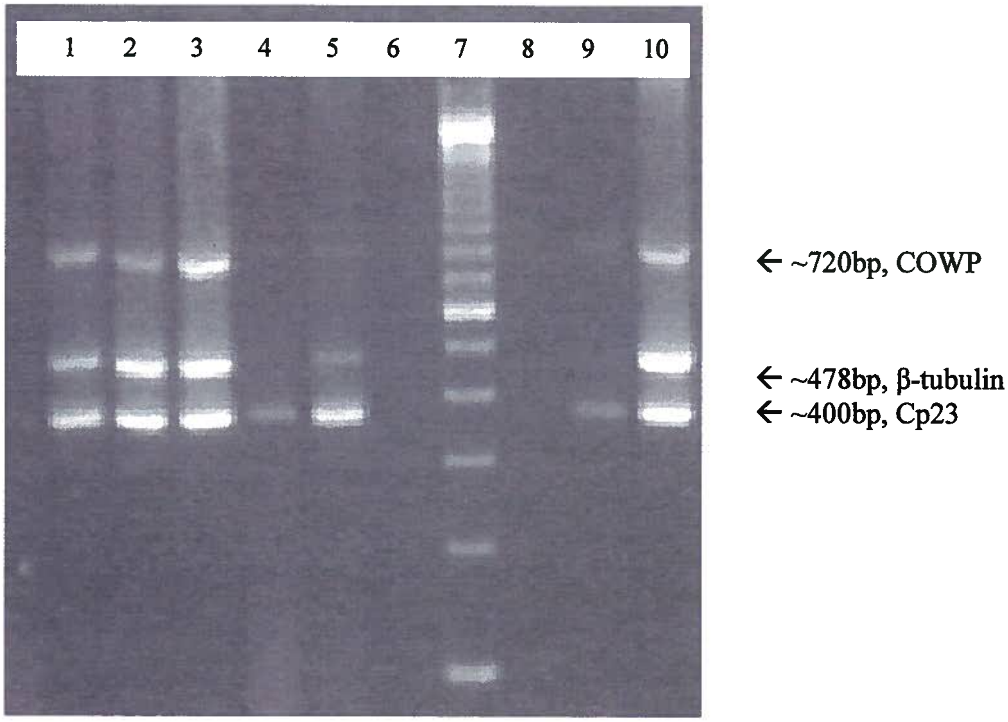


Figure 4.2. Gel electrophoresis (1.8%) of Qiagen multiplex PCR for Cp23, COWP and β-tubulin loci. Lanes are designated along the top. Isolates are all Canadian at: lane 1, BC1; lane2, BC2; lane 3, BC3; lane 4, BC4; lane 5, BC12; lane 6, empty; lane 7, 100bp molecular marker ladder; lane 8, empty; lane 9, BC13; lane 10, BC14.

Table 4.2

PCR Primer Pairs

Rxn Set	Gene	F Primer	F primer (5`-3`)	R Primer	R Primer (5`-3`)	bp^{+}	original/ published$^{\pm}$
1	MDH	MDH-f	TTCCAATGTTTGTTTCT	MDH-r	GTTGATAAATCTTGTAACTG	796	original
1	LDH	LDH-f	TTTCGAGAACAAAAA	LDH-r	CACAAAAATCTAACCATTA	510	original
1	UPRT	UPRT-f	CAACTTCTATTTATGCTTGCGATTG	UPRT-r	TGCTTTTGTTATTGTAATTGTCCAAA	460	original
2	EMAAg	EMA-f	TGGTTTTCAGTTGCAT	EMA-r	CAAGGGATAAATCCGCAGT	988	original
2	APR	APR-f	AAATCTCAAAGCAAGAGA	APR-r	CAACTGTGGAACATACCCAACT	297	original
3	COWP	COWPF	GACTCAATTATTGATCG	COWPR	CAGAGTACCAGCTTTTGT	720	original
3	B-tubulin	B-tub-5	AGGAACCCATGTGAATTTAAGAGA	B-tub-4	TGGCTCTGCAACAAGCTG	478	published
4	ACoA	AcoA-F1	GGACCTATTGAATTTGTCAAGG	AcoA-R1	GAGTAATTCTGTGTCTCTCCAC	298	published
4	Mucin-1	Muc1-F2	TTGATGATTCAGAATCATCTGACT	Muc1-R2	GTGAGTTCTTCTTCATCTGTATAG	650-900	published
5	Cp23	8170	AGGAACCCATGTGAATTTAAGAGA	8167	GAGTAATTCTGTGTCTCTCCAC	400	published
5	18S rRNA	CDC-F	GGACCTATTGAATTTGTCAAGG	CDC-R	GTGAGTTCTTCTTCATCTGTATAG	435	published
6	CCR	CCreg-F	GATGATATTCTGCTAGACCATTCAA	CCreg-R	CTTTCTTCTGTTGTTTTTGGTTG	953	original
6	CTCL	CTCL AgF	AGGCTATTCAGGTGGATGCT	CTCL Ag-R	CAAAAATGTTAAAGAGCGCAAT	678	original
6	AAD	AA D'aseF	TTCCAATGTTTGTGGCTTCT	AA D'aseR	GTTGATAAATCCTCCTTTGTAACTG	817	original
7	CLL	CLLaAg-F	TTTCGAGAATGAATGCAAAAA	CLLaAgR	CACAAAAATCTAAAATATCGCCATTA	983	original
7	SSSk	SSSkin-F	CAACTTCTATTTATGCTTGCGATTG	SSSkin-R	TGCTTTTGTTATTGTAATTGTCCAAA	507	original
7	FLJ	FLJ-F	TTTGCGAAGTGCATGGATAG	FLJ-R	GAAAAACAAGTTCTGATGGTATTCAA	849	original
8	TMaaT	TMaaT-F	TGGCTGGTTTTCAGTTGCAT	TMaaT-R	CAAGGGATAAATCGACGCAGT	850	original
8	ABC	ABCmd-F	AAACCTTTTCTCAAAGCAAGAGA	ABCmd-R	CAACTGTGGAACATACCCAACT	762	original
8	Thiol	Thiol-F	GACTCAATTATCGCTTGATCG	Thiol-R	CAGGCAGTACCAGCTTTTGT	398	original
9	Epro	Hypo-F	AGGAACCCATGTGAATTTAAGAGA	Hypo-R	TGGCTCTGCAACAAGCTGTA	949	original
9	SeroAg	SeroAg-F	GCAATTAAGAACATCGGGTTT	SeroAg-R	TTACAATCACAGGGGCAAAT	777	original
9	RIK	RIK-F	GCAAATACTTCATCGAACACCA	RIK-R	TCCATGTGGGACTTCATCAGA	573	original
solo	HSP70	HSP70-4F	AATTCTCAAAGCAAGAGA	HSP70-4R	CAGGCAACCAGCTTTTGT	745	published
solo	Gp60	Gp60-F1**	GACTCAATTATCTGATCG	Gp60-R1**	CAGGCAGTACCAGCTTGT	934	published
solo	Gp60	8175/IntF	AGGAACCCATTAAGAGA	8174/IntR	TGGGCAACAAGCTGTA	800-850	published

Table 4.2. Multi-plex PCR primer pairs used for gene specific amplification from fecal specimens and the expected size fragment of the amplified product. F primer; forward primer. R primer; reverse primer. +Bp; approximate expected fragment size from PCR. ±Original/Published; original primers were designed within the lab by the author.

4.2.6 Amplicon Purification

PCR amplicons were purified using Invitrogen's ChargeSwitch PCR Clean-Up Kit (Catalogue# CS12000). Using the supplied protocol, amplicons were bound using 10ul of magnetic beads, 50ul of supplied purification buffer, 50ul of PCR product and the MagnaRack. Each reaction was washed twice with the supplied buffer and subsequently eluted using the supplied elution buffer E5 (10mM Tris-HCl, pH 8.5). Purified PCR product was quantified by a Pharmacia Biotech GeneQuant 2 spectrophotometer.

4.2.7 SNaPshot: Single Base Extension (SBE) Chemistry

The SNaPshot primer extension method (Applied Biosystems, Foster City, CA) was used to analyze the 45 SNP-marker set with capture probes listed in Table 4.3. The SNaPshot technique is based on the addition of a single fluorescently labelled ddNTP to the 3` end of an unlabeled specific oligonucleotide primer that hybridizes to its target DNA located contiguous to the SNP of interest (Figure 4.3). A total volume of 0.5ul of purified PCR amplicon was used for dideoxy single base extension of unlabeled oligonucleotide primer following conditions recommended by manufacturer.

Figure 4.3 SNaPshot fragment analysis sequential procedure.

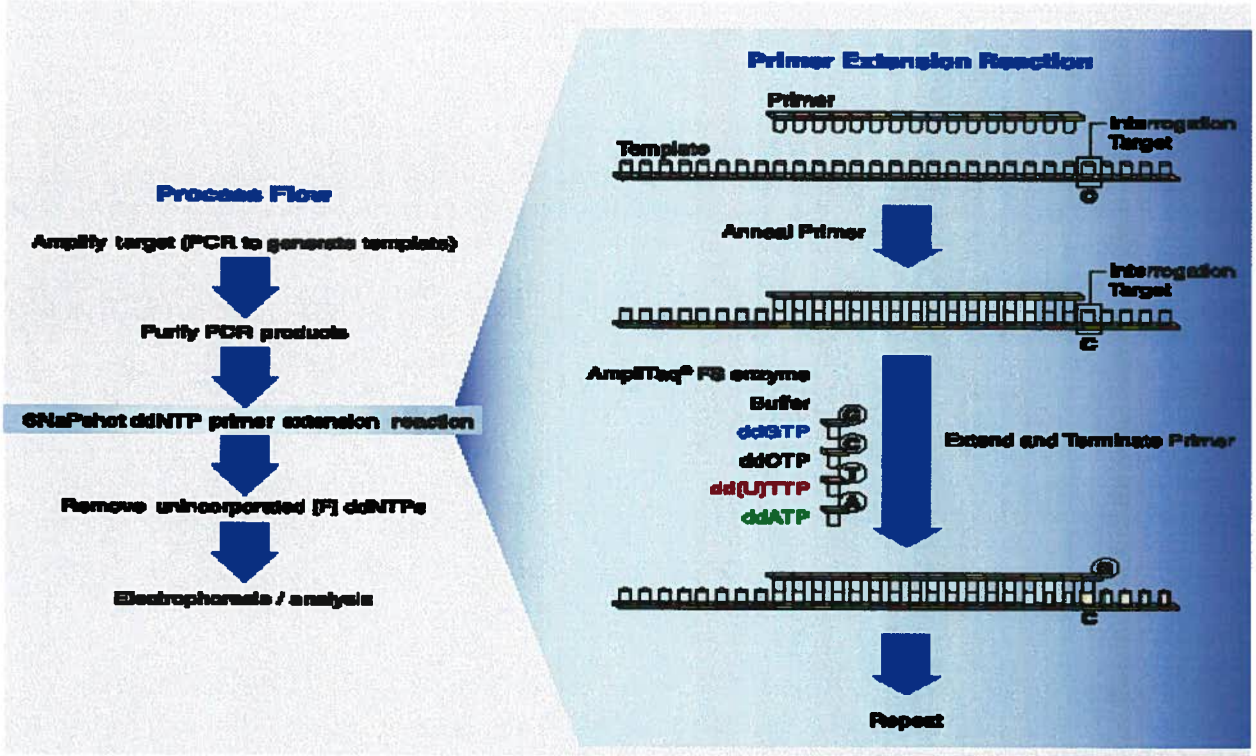

Process Flow
Amplify target (PCR to generate template)
Purify PCR products
SNaPshot ddNTP primer extension reaction
Remove unincorporated [F] ddNTPs
Electrophoresis / analysis
Primer Extension Reaction
Primer
Template
Interrogation Target
G
Anneal Primer
Interrogation Target
C
AmpliTaq® FS enzyme
Buffer
ddGTP
ddCTP
dd(U)TTP
ddATP
G
C
T
A
Extend and Terminate Primer
G
C
Repeat

Figure 4.3. Sequential methodology for single base extension chemistry for targeting specific SNP markers with a capture probe designed uniquely for it. Probes combined into the same reaction differed in sizes due to the presence of a non-annealing poly-GACT tail, not shown here. Image from Applied Biosystems, Foster City, California. https://products.appliedbiosystems.com

Mini-sequencing primers were designed so that the 3` end was situated one base upstream from the relevant polymorphism. Primers were tailed with a non-annealing poly-GACT tail to produce fragments with differing sizes. Extension was performed for 25 cycles at 96°C for 10s, 50°C for 5s and 60°C for 30s in a GeneAmp 2720 Thermacycler (Applied Biosystems). Inactivation of dNTPs and removal of primers from extension products were performed by incubation with 1unit Calf Intestinal Phosphatase (CIP) for 1hr at 37°C followed by enzyme inactivation for 15m at 75°C.

Table 4.3

Single Base Extension Probes for Fragment Analysis

Gene	SNP Marker+	SBE Capture Probe	Length(nt)	5`GACT tail(nt) ±	Total Fragment(nt) ∞
B-tubulin	1	TGAACTGAATAATTGACT	20	47	67
	4	AAGATACGTTCCAAGAGC	18	48	66
	3	TTTCTACAATGAAGCTTC	18	49	67
	5	CTGATAACTTCATTTTTGG	19	69	88
	7	GAGGGAGCTGAACTCTT	17	83	100
	8	GCAGAATCTTGTGATTGC	18	94	112
COWP	5	TCAATATCTCCCTGCAAA	18	23	41
	6	CCCACCAGGATATACAGA	20	20	42
	1	ATGTCCCCCAGGATTCGT	18	22	40
	3	TCCAGAATGTCCTCCAGG	18	34	52
	7	CATTTTACAAGGCCTCCA	18	46	64
Cp23	4	TCCTCCAGCTGCTGATGC	18	22	40
	3	AAAGAATCCAGCTCCAAT	21	31	52
	1	AAAGAATCCAGCTCCAAT	16	41	57
	5	CCAGCAGCCCAAGCTCCT	16	52	68
	6	ACAGGATAAGCCAGCTGA	18	62	80
18S rRNA	1	TATATAATATTAACATAATTCATATTACTAT	16	76	92
	3	GAATAATATTAAAGATTTTTATCTTT	26	78	104
HSP70	14	TGCTAATGGTATCTTGAATGT	21	19	40
	17	AGGGTGAGGATGAGCA	16	36	52
	19	GAGAACTACCTGTATAACATGAG	23	41	64
	20	TGAACATCAACAAAAGGA	18	58	76
	22	GAATGCCAGGTGG	13	75	88
ACoA	1	TATATACCTCAGGTAGTACTGG	22	42	64
	4	AGGATACTTACTTTATGCTGC	21	19	40
	6	TCAACATTCATCCTGG	16	48	64
	7	GTGCTGGAGATATTGG	16	60	76
Mucin-1	13	GACCTGATAACTTCATTTTTGG	22	66	88
	16	CAAAACCTGAAAAGGAG	17	83	100
Gp 60	80	CTGGTGAAGTTACATCTGTA	20	20	40
	108	GATTTGTTTGCCTTTAC	17	35	52
	126	CGGCGCAAACAG	12	52	64
	79	TCGTCTATGCACCTATAAAAGA	22	68	90
	98	ATCAAGATCAAGAAGATCACTC	22	78	100
	115	AGAATTGAAGTGGCTGT	17	95	112
LDH	10	TTGTGTGCCATACCAGAA	19	32	51
	3	AACATTCATTGCACAACA	16	40	56
MDH	8	ATTACTCATTCACAAATC	20	21	41
	7	TTCAGTTGCAGAAAATGT	17	29	46
UPRTase	2	ACCTTCTTCCTTATGATT	20	41	61
	3	TAAAACCCAAATGGAAT	16	50	66
EMAAg	29	TATAACAAACTCCCATAC	22	39	61
	27	TTCTATTGATGAGCTTGC	18	48	66
APR	2	ACGTTGGCCCGATTGAAA	15	36	51
	3	TCAAAAGAGAATAATTGA	20	36	56

Table 4.3. Markers labelled according to internal lab database. + Numbered according to internal lab data. ± Non-annealing poly-GACT tail to create size discrepancies for fragment separation and sizing. ∞ Expected fragment size including original capture probe, non-annealing poly-GACT tail, and single base extension.

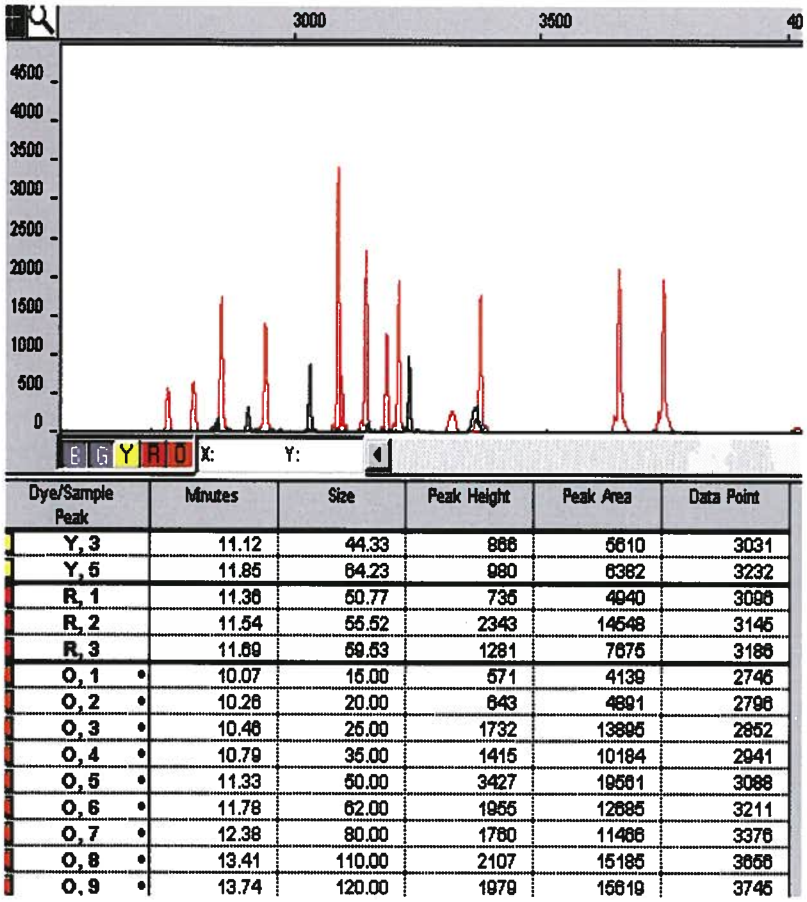
4.2.8 Capillary Electrophoresis & Electropherogram Analysis

Extension products, 0.5ul, were mixed with 0.5ul of internal size standard Liz120 (Applied Biosystems) and 9ul of Hi-Di formamide (Applied Biosystems), heated for 5m at 95^{0}C, and immediately quenched in an ice bath. SNP detection was carried out on an automated 3130xl 5-color sequencer (Applied Biosystems) with a capillary length of 36cm (Appendix 6). Performance optimized polymer-6 (POP-6) was injected into the capillary to serve as a dynamic coating for the capillary walls and to provide sieving medium for fragment analysis. Parameters for electrophoresis were performed for 15m with an injection time of 5s, voltage of 15kv and run temperature of 60^{0}C. Total run time for each sample was 24 minutes. Fluorescently labelled fragments were exported to GeneMapper analysis software v4.0 (Applied Biosystems). Peaks were scored and analyzed based on size and color using the local Southern method.

4.2.9 Allele Discrimination & Scoring

It can be difficult to predict mobility of an oligonucleotide since final mobility is determined not only by length, but also by molecular weight of the labelling dye for each base. Fragment sizes can be skewed to a slightly larger size than that expected based off the unique primer sequence alone. Primer dimers and secondary structures can skew the expected fragment size. These factors were all considered during fragment analysis. Resulting electropherograms consisted of peaks corresponding to the predetermined fragment size and allele discrimination was assigned according to DS-03 dye set (Figure 4.4). Fluorescent specific nucleotides were identified; green for adenine (A), black for cysteine (C), blue for guanine (G), red for thymine (T). An internal Liz 120 size standard (Applied Biosystems, Foster city, CA) with size markers at 15, 20, 25, 35, 60, 80, 100, and 120 nucleotides was used to analyze and size electropherogram data (Appendix 3). The Liz120 standard fluoresces orange at each of the above mentioned pre-sized fragments. To verify the efficiency and accuracy of the SBE protocol and subsequent allele scoring a collection of random SNPs from random samples were typed in a single-plex reaction before being multi-plexed. Allele scoring was done via GeneMapper V4.0 software (Applied Biosystems, Foster City, CA). Alleles were scored based on fluorescent labels and size ranges for each SNP marker individually.

Figure 4.4 Fragment analyses; example electropherogram representation of allele discrimination & scoring.



Dye/Sample Peak	Minutes	Size	Peak Height	Peak Area	Data Point
Y, 3	11.12	44.33	866	5610	3031
Y, 5	11.85	64.23	980	6362	3232
R, 1	11.36	50.77	735	4940	3096
R, 2	11.54	55.52	2343	14648	3146
R, 3	11.69	59.53	1281	7675	3186
O, 1 •	10.07	15.00	571	4139	2746
O, 2 •	10.26	20.00	643	4891	2796
O, 3 •	10.46	25.00	1732	13895	2852
O, 4 •	10.79	35.00	1415	10184	2941
O, 5 •	11.33	50.00	3427	19561	3088
O, 6 •	11.78	62.00	1955	12685	3211
O, 7 •	12.38	80.00	1760	11466	3376
O, 8 •	13.41	110.00	2107	15185	3656
O, 9 •	13.74	120.00	1979	15619	3745

Figure 4.4. Electropherogram representation of Canadian *C. hominis* isolate, BC12, scored at SNP markers COWP 5, COWP 6, Cp23 3, Cp23 1, and BT 1 with an expected *C. hominis* allelic profile of C-T-T-T-C respectively. Alleles discriminate as black for C, red for T, green for A, and blue for G. Sizes for each fragment are listed below the electropherogram and are compared against Liz120 dye standard, fluorescing orange with a pre-inscribed set of fragments. Sample peaks correspond as: Y, 3 as COWP 5 (C, black); R, 1 as COWP 6 (T, red); R, 2 as Cp23 3 (T, red); R, 3 Cp23 1 (T, red); Y, 5 as BT1 (C, black).

In the event of two alleles typed to a single SNP locus additional peaks present at >20% of the height of the main peak were scored as mixed or double allele calls. The main or predominant peak was used for MlS-typing under the assumption that the predominant peak at each SNP locus represents the actual genotype. The cut-off was applied to prevent the inclusion of possible false-positive peaks resulting from artifactual stutter peaks or adjacent pull-up peaks. The combination of alleles from all 45 SNP markers isolated from the 13 targeted loci was used to determine a multi-locus SNP-type (MlSt) for each isolate from Australia, Kenya, Peru, and Scotland subpopulations. SNP patterns from Western Canada were more varied across the 45 SNP markers and therefore reserved to decide about viability of the assay as a species distinction tool.

4.2.10 Diversity Statistical Analysis

MlSts determined for each isolate were used to make species distinctions and identify variant or novel alleles at any given SNP position. Allele frequencies were tabulated and used for analysis of genetic diversity and structure. Based on MlSt the number of unique SNP profiles was determined as was their geographic distribution. Qualitative SNP data was converted to quantitative data by creating appropriate mathematical input files and processed through a combination of software programs designed for genetic population studies. Using the genetic data analysis (GDA) program five standard genetic diversity parameters were estimated: percent polymorphic loci (P), mean number of alleles per polymorphic locus (AP), mean number of alleles per locus (A), effective number of alleles per locus (A_e), and expected heterozygosity or gene diversity (H_e)[126]. Intercontinental differentiation was compared among subpopulations and estimated by calculating Nei's genetic identity and distance. Dendogram representation based on genetic distances using the Neighbour Joining Group was done using TreeViewX version 0.5.0[187]. The Fstat^{TM} Statistics software package was used partition genetic diversity into within and between subpopulations based Fixation indexes[79]. In addition to descriptive statistics, total genetic diversity (Ht), genetic diversity within subpopulations (Hs), average among subpopulations genetic diversity (Dst), and the proportion of genetic diversity found among subpopulations (Gst) were calculated for subpopulations following Nei's (1973,1977) gene diversity formulae.

4.2.11 Geographic Boundaries & Study Subpopulations

In any phylogeographic study populations from opposite sides of obvious or suspected barriers are sampled. Parasite isolates were obtained from leading facilities around the world. With great appreciation isolates were donated by Drs. L. Xiao, W. Gatei, H. Smith, and U. Ryan from Peru, Kenya, Scotland and Australia respectively. Including isolates secured here in Canada our study currently covers 5 distinct global localities, climates and ecologies from 5 continents (Figure 4.5).

Figure 4.5 Geographical representation of intercontinental subpopulations used in study.

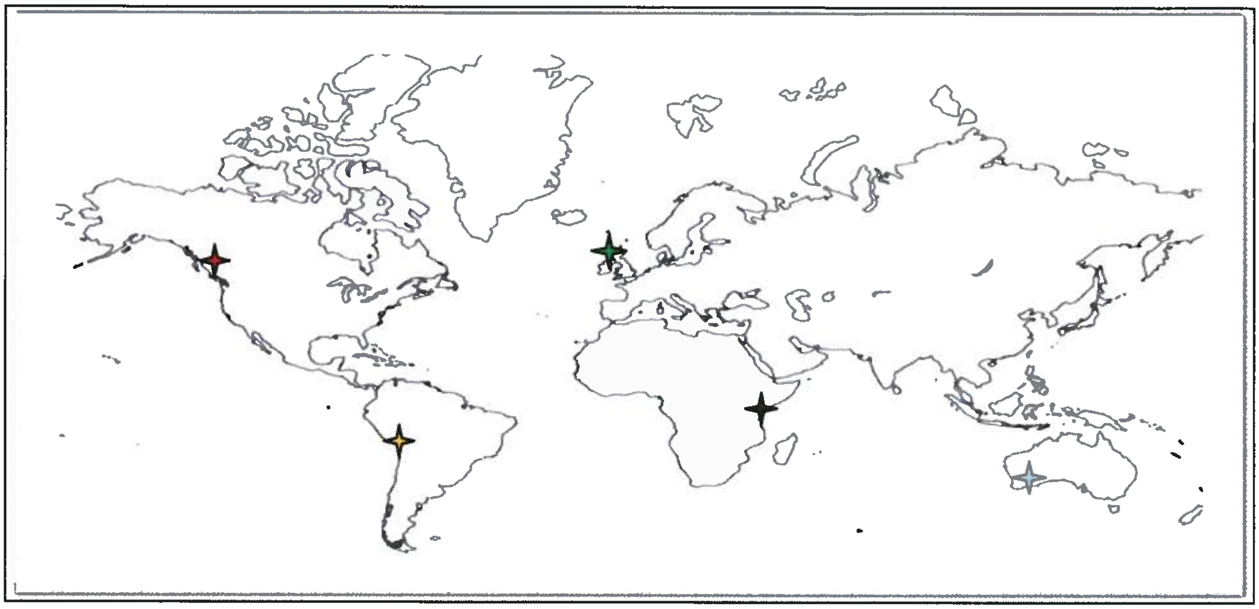


Figure 4.5. Pictorial illustration of the five subpopulations used in this study; Australia (blue), Canada (red), Kenya (black), Peru (yellow), Scotland (green).

Isolates were genetically identified as *C. hominis* in their home labs based on host, oocyst morphology and sequence data. All samples donated were isolated from human hosts of *Cryptosporidium*. Given their varied histories isolates from each geography were treated as a subpopulation. Isolates are referred by their country code throughout the remainder of this document. A total of 108 samples from the five subpopulations were used in this study: Australia, 15; Kenya, 20; Peru, 22; Scotland, 22; W. Canada, 31. Subpopulations for all five international regions are listed below, tables 4.4, 4.5, 4.6, 4.7, and 4.8, with a simplified schematic of subpopulations versus metapopulation in Appendix 2.

Table 4.4

Australia Subpopulation, n=15			
Country Code	**Isolate Code**	**Species**	**Sample Type**
A1	H131	*C. hominis*	fecal, whole DNA
A2	H139	*C. hominis*	fecal, whole DNA
A3	H140	*C. hominis*	fecal, whole DNA
A4	H141	*C. hominis*	fecal, whole DNA
A5	H142	*C. hominis*	fecal, whole DNA
A6	H158	*C. hominis*	fecal, whole DNA
A7	H160	*C. hominis*	fecal, whole DNA
A8	H161	*C. hominis*	fecal, whole DNA
A9	H162	*C. hominis*	fecal, whole DNA
A10	H163	*C. hominis*	fecal, whole DNA
A11	H164	*C. hominis*	fecal, whole DNA
A12	H165	*C. hominis*	fecal, whole DNA
A13	H166	*C. hominis*	fecal, whole DNA
A14*	H167	*C. parvum*	fecal, whole DNA
A15*	H168	*C. parvum*	fecal, whole DNA

* Confirmed *C. parvum* isolates in originating lab, U. Ryan.

Table 4.5

Kenya Subpopulation, n=20			
Country Code	**Isolate Code**	**Species**	**Sample Type**
K1	10962	*C. hominis*	fecal, whole DNA
K2	10963	*C. hominis*	fecal, whole DNA
K3	10965	*C. hominis*	fecal, whole DNA
K4	10966	*C. hominis*	fecal, whole DNA
K5	10967	*C. hominis*	fecal, whole DNA
K6	10972	*C. hominis*	fecal, whole DNA
K7	10973	*C. hominis*	fecal, whole DNA
K8	10974	*C. hominis*	fecal, whole DNA
K9	10975	*C. hominis*	fecal, whole DNA
K10	10976	*C. hominis*	fecal, whole DNA
K11	11144	*C. hominis*	fecal, whole DNA
K12	11145	*C. hominis*	fecal, whole DNA
K13	11146	*C. hominis*	fecal, whole DNA
K14	11148	*C. hominis*	fecal, whole DNA
K15	11149	*C. hominis*	fecal, whole DNA
K16	11150	*C. hominis*	fecal, whole DNA
K17	11151	*C. hominis*	fecal, whole DNA
K18	11152	*C. hominis*	fecal, whole DNA
K19	11153	*C. hominis*	fecal, whole DNA
K20	11154	*C. hominis*	fecal, whole DNA

Table 4.6

Peru Subpopulation, n=22			
Country Code	**Isolate Code**	**Species**	**Sample Type**
P1	11122	*C. hominis*	fecal, whole DNA
P2	11123	*C. hominis*	fecal, whole DNA
P3	11124	*C. hominis*	fecal, whole DNA
P4	11125	*C. hominis*	fecal, whole DNA
P5	11126	*C. hominis*	fecal, whole DNA
P6	11127	*C. hominis*	fecal, whole DNA
P7	11128	*C. hominis*	fecal, whole DNA
P8	11129	*C. hominis*	fecal, whole DNA
P9	11130	*C. hominis*	fecal, whole DNA
P10	11131	*C. hominis*	fecal, whole DNA
P11	11132	*C. hominis*	fecal, whole DNA
P12	11133	*C. hominis*	fecal, whole DNA
P13	11134	*C. hominis*	fecal, whole DNA
P14	11135	*C. hominis*	fecal, whole DNA
P15	11136	*C. hominis*	fecal, whole DNA
P16	11137	*C. hominis*	fecal, whole DNA
P17	11138	*C. hominis*	fecal, whole DNA
P18	11139	*C. hominis*	fecal, whole DNA
P19	11140	*C. hominis*	fecal, whole DNA
P20	11141	*C. hominis*	fecal, whole DNA
P21	11142	*C. hominis*	fecal, whole DNA
P22	11143	*C. hominis*	fecal, whole DNA

Table 4.7

Scotland Subpopulation, n=20			
Country Code	**Isolate Code**	**Species**	**Sample Type**
S1	2023	*C. hominis*	fecal, whole DNA
S2	2026	*C. hominis*	fecal, whole DNA
S3	2068	*C. hominis*	fecal, whole DNA
S4	2091	*C. hominis*	fecal, whole DNA
S5	2096	*C. hominis*	fecal, whole DNA
S6	2106	*C. hominis*	fecal, whole DNA
S7	2114	*C. hominis*	fecal, whole DNA
S8	2118	*C. hominis*	fecal, whole DNA
S9	2122	*C. hominis*	fecal, whole DNA
S10	2132	*C. hominis*	fecal, whole DNA
S11	2133	*C. hominis*	fecal, whole DNA
S12	2140	*C. hominis*	fecal, whole DNA
S13	2177	*C. hominis*	fecal, whole DNA
S14	2181	*C. hominis*	fecal, whole DNA
S15	2203	*C. hominis*	fecal, whole DNA
S16	2224	*C. hominis*	fecal, whole DNA
S17	2234	*C. hominis*	fecal, whole DNA
S18	2242	*C. hominis*	fecal, whole DNA
S19	2249	*C. hominis*	fecal, whole DNA
S20	2267	*C. hominis*	fecal, whole DNA

Table 4.8

Canada Subpopulation, n=31			
Country Code	**Isolate Code***	**Species**	**Sample Type**
BC1	nr	*C. parvum*	fecal, whole DNA
BC2	nr	*C. parvum*	fecal, whole DNA
BC3	nr	*C. parvum*	fecal, whole DNA
BC4	nr	*C. parvum*	fecal, whole DNA
BC5	nr	*C. parvum*	fecal, whole DNA
BC6	nr	*C. parvum*	fecal, whole DNA
BC7	nr	*C. parvum*	fecal, whole DNA
BC8	nr	*C. parvum*	fecal, whole DNA
BC9	nr	*C. parvum*	fecal, whole DNA
BC10	nr	*C. parvum*	fecal, whole DNA
BC11	nr	*C. hominis*	fecal, whole DNA
BC12	nr	*C. hominis*	fecal, whole DNA
BC13	nr	*C. hominis*	fecal, whole DNA
BC14	nr	*C. hominis*	fecal, whole DNA
BC15	nr	*C. hominis*	fecal, whole DNA
BC16	nr	*C. hominis*	fecal, whole DNA
BC17	nr	*C. hominis*	fecal, whole DNA
BC18	nr	*C. hominis*	fecal, whole DNA
BC19	nr	*C. hominis*	fecal, whole DNA
BC20	nr	*C. hominis*	fecal, whole DNA
BC21	nr	*C. hominis*	fecal, whole DNA
BC22	nr	*C. hominis*	fecal, whole DNA
BC23	nr	*C. hominis*	fecal, whole DNA
BC24	nr	*C. hominis*	fecal, whole DNA
BC25	nr	*C. parvum*	fecal, whole DNA
BC26	nr	*C. hominis*	fecal, whole DNA
BC27	nr	*C. hominis*	fecal, whole DNA
BC28	nr	*C. hominis*	fecal, whole DNA
BC29	nr	*C. hominis*	fecal, whole DNA
BC30	nr	*C. hominis*	fecal, whole DNA
BC31	nr	*C. hominis*	fecal, whole DNA

* Not reportable by law.

CHAPTER 5

RESULTS

- Descriptive Genomics, Multi-locus SNP-typing, Genetic Diversity Measures & Partition -

Summary – We report on genetic variation both within and between *C. hominis* subpopulations from Australia, Kenya, Peru, and Scotland. We examined ~18 500 bp and assembled a data set of 394 SNPs. Following comparative genomics and bio-physical profiling an expected haplotype, representing a set of 45 single nucleotide polymorphisms at individual loci was established. Molecular typing of 77 international isolates based on this haplotype or multi-locus SNP-type was done, twenty-four unique MlSt's were identified. Inferences about genetic relationships were made using genetic data analysis software programs to quantify and partition the genetic diversity into intra- and inter-population diversity and to discern genetic distances among subpopulations. Within population differences among subpopulations account for 69.6% of genetic variation; differentiation among subpopulations constitute 30.4%. Genetic distances among subpopulations averaged 0.048 and varied from 0.034 between the Australian and Scotland subpopulations to 0.061 between Scotland and Kenya. The potential use of a DNA-typing scheme based on SNPs to resolve *Cryptosporidium* epidemiology was examined. Established *C. hominis* and *C. parvum* subpopulations from Western Canada, in collaboration with international SNP-typing results, were successfully used to assess the efficacy of a mutation based platform for genetic typing and species distinction.

5.1 Comparative Genomics: Global Patterns of Single Nucleotide Polymorphisms (SNPs)

Of 3956 total genes annotated in the *C. hominis* genome[1, 245] 25 were targeted for bioinformatics analysis in the hopes of identifying ideal SNPs that would have the greatest research potential (Table 5.1). Genes examined consisted of four bio-functionalities; enzymatic or bio-synthesis related genes, structural or cellular related genes, virulence or antigenic determinant genes, and intron-housing genes. The use of various genes allows for more robust conclusions to be made from the compiled data sets.

Table 5.1

Target Gene Library

	Gene Annotation	**Abbreviation**	**Bio-functionality**	**Chromosome**
1	Malate Dehydrogenase	MDH	enzymatic	7
2	Lactate Dehydrogenase	LDH	enzymatic	7
3	Uracil Phosphoribosyl Transferase	UPRT	enzymatic	1
4	Erythrocyte Membrane Associated Ag	EMAAg	antigenic	1
5	Apoptosis Related Protein	APR	bio-synthesis	4
6	*Cryptosporidium* Oocyst Wall protein	COWP	structurally-related	4
7	Beta-tubulin	BT	intron-containing	6
8	Acetyl coA synthetase	AcoA	enzymatic	8
9	Mucin-1	Muc-1	structurally-related	6
10	Cp23	Cp23	antigenic	4
11	18S rRNA	18S rRNA	ribosomal (structural)	multi-copy
12	Cellcycle Regulator	CCR	enzymatic	1
13	CTCL Tumor Ag	CTCL	antigenic	2
14	Aldahyde-Alcohol Dehyd'ase	AAD	enzymatic	8
15	CLL Associated Ag-KW-2	CLLAg	antigenic	2
16	Sexual Stage Specific Kinase	SSK	bio-synthesis	3
17	FLJ31812/DHHC palmitoyl transferase	FLJ	enzymatic	7
18	Transmembrane amino acid Transporter	TMaaT	bio-synthesis	3
19	ABC multi-drug or ion efflux	ABC	bio-synthesis	4
20	Thiolproteinase	Thiol	enzymatic	7
21	Extracellular protein w/ 8 kazal repeats	ExP	unknown	4
22	Seroreactive Ag BMN-19B related protein	SeroAg	antigenic	7
23	RIK protein w/ ? WD40 repeats	RIK	unknown	8
24	Heat Shock Protein 70	HSP70	bio-synthesis	2
25	Glycoprotein60	Gp60	antigenic	6

Table 5.1. Genes targeted for genetic typing; abbreviated identifiers and putative molecular function and chromosomal position are all listed.

Thirteen of these (Table 5.2) were targeted for further characterization and profiled based on numerous criteria as discussed in the materials and methods section. This initial baseline analysis was crucial to pinpointing those SNPs that would best represent the gene for downstream SNP-typing of isolates. Each of the 13 genes used in the study are identified by their *C. hominis* annotation as described in the CryptoDB database (www.cryptodb.org). Clustal alignments generated from the Mercator and MAVID programs at the nucleotide level between the compiled *C. hominis* and *C. parvum* databases within the CryptoDB database were done to elucidate single base differences between the *C. hominis* reference strain, TU502, and the *C. parvum* reference strain, IowaII. These two strains are the parent strains of all molecular studies to date on *Cryptosporidium* and the only two strains to have had their genome sequenced in its entirety.

Table 5.2

Gene Panel for Multi-locus SNP-typing

Gene	CryptoDB ID (Ch.#####)	Contig # (GenBank)	Nucleotide Position	MW (Da)	Coding Seq. (bp)	Protein Seq. (aa)	*C. parvum* Ortholog ID
AcoA	10418	AAEL01000346	468-2555	77945	2088	695	cgd1_3710
APR	40253	AAEL01000018	38706-39098	15042	393	130	cgd4_2220
β-tub	50050	AAEL01000450	4053-4892	31939	840	279	cgd5_3220
COWP	40378	AAEL01000002	6200-8425	84131	2226	741	cgd4_3340
Cp23	40414	AAEL01000216	5348-5683	11262	336	111	cgd4_3620
EMAAg	30018	AAEL01000207	2119-4338	84923	2220	739	cgd3_90
Gp60	60138	AAEL01000025	26940-27974	35768	1035	344	cgd6_1080
HSP70	20010	AAEL01000158	12080-14113	73706	2034	677	cgd2_20
LDH	70063	AAEL01000175	8305-9270	33866	966	321	cgd7_480
MDH	70062	AAEL01000175	9686-10636	33589	951	316	cgd7_470
Muc-1	60622	AAEL01000244	4857-7199	85466	2343	780	cgd6_5400
18S rRNA	rrn013	AAEL01000005	6112-7866	66497	1689	563	cgd2_1375
UPRT	80328	AAEL01000030	7647-9020	51885	1374	457	cgd8_2810

Table 5.2. Thirteen genes used herein for SNP-typing of international subpopulations and their NCBI databank identification codes. Molecular weight (MW) in Daltons, coding sequence and peptide length listed.

To identify SNPs, we examined 18 495 bp representing the 13 target loci, which included 13 coding regions and one intron. A total of 394 SNPs were mapped creating a comprehensive data set; the density of SNPs was approximately 1 in 47 bp. Of 394 total SNPs, 392 were mapped to coding regions, while the remaining 2 were found in the non-coding or intron region of the β-tubulin gene. In total our SNP data set is comprised of 274 synonymous or non-expressed single base mutations compared to 120 non-synonymous or expressed mutations (Table 5.3). Overall there were almost half as many non-synonymous (NS) changes as synonymous (S) changes, producing a ratio of NS SNPs/S SNPs of 0.44.

Table 5.3

Genetic Organization of Gene Panel

Gene	# SNPs	Transitions^{+}	Transversions^{+}	# SAAPs$^{\pm}$	ω Ratio$^{\infty}$
AcoA	35	27	8	6	6:39
APR	9	7	2	3	3:6
β-tubulin	9	8	1	0	0:9
COWP	38	30	8	11	11:27
Cp23	7	5	2	1	1:6
EMAAg	40	36	4	19	19:21
Gp60	153	88	65	72	72:81
HSP70	32	26	6	1	1:31
LDH	20	16	4	1	1:19
MDH	15	8	7	2	2:13
Muc-1	2	0	2	0	0:2
18S rRNA	5	1	4	0	0:5
UPRT	29	23	6	4	4:25

Table 5.3. Descriptive makeup of SNPs mapped to target genes. $^{+}$Indicates number of transitions (purine to purine or pyrimidine to pyrimidine) or transversions (purine < > pyrimidine) present. $^{\pm}$ SAAPs are single amino acid polymorphisms introduced by a mutation conferring a change to the original peptide make up of the protein. $^{\infty}$ ω ratio represents the ratio of non-synonymous (expressed) to synonymous (silent) mutations and can be an indicator of selection pressure.

To ensure genomic comparisons were made to orthologous genes a blastp search was conducted against over 50 eukaryotic genomes from multiple phyla, important considering the prospect of other competing microorganisms within gastrointestinal space (Table 5.4). A tblastn, to infer any changes at the amino acid level that may occur within the each target protein and its closest ortholog, followed. A tblastn search is a modified blastn search that compares a given protein sequence against a nucleotide sequence database dynamically translated in all 6 reading frames for both strands. *C. hominis* protein sequences were searched against both the complete *C. parvum* and *C. muris* databases. In all cases the closest ortholog was from the *C. parvum* IowaII genomic database (Table 5.4). Results of the blast searches were used to decipher if particular SNPs mediated a single amino acid change to the peptide configuration of a given protein. Positions of SAAPs were determined based on the in-frame start or methionine codon according to the blast alignment.

Table 5.4

blastp Protein Queries					
Gene	**Length**	**Score (bit)**	**E value**	**Frame**	**Closest Ortholog**$^{+}$
AcoA	8117	1310.7	0.0000	(+) 3	*C. parvum* IowaII
APR	39459	233.9	1.3e-63	(+) 3	*C. parvum* IowaII
β-tub	844015	509.5	1.4e-145	(+) 3	*C. parvum* IowaII
COWP	579568	1278.0	0.0000	(+) 2	*C. parvum* IowaII
Cp23	579568	90.6	1.9e-20	(-) 3	*C. parvum* IowaII
EMAAg	84326	1219.2	0.0000	(+) 2	*C. parvum* IowaII
Gp60	1213436	289.1	1.9e-85	(-) 2	*C. parvum* IowaII
HSP70	97996	1056.2	0.0000	(+) 1	*C. parvum* IowaII
LDH	1278458	520.8	1.4e-148	(-) 3	*C. parvum* IowaII
MDH	1278458	533.1	1.2e-153	(-) 1	*C. parvum* IowaII
Muc-1	10947	312.4	1.5e-200	(+) 3	*C. parvum* IowaII
18S rRNA	887873	1150.2	0.0000	(+) 3	*C. parvum* IowaII
UPRT	1156729	795.7	1.1e-229	(-) 1	*C. parvum* IowaII

Table 5.4. Blast queries of thirteen target genes for sequence similarity. E value is the expectation value and calculates how many times you could have expected the result by chance alone, the lower the better. $^{+}$Based on comparative search encompassing over 50 genomes from 9 different Eukaryotic phyla.

5.2 Target Genes: Qualitative Characterization

Malate & Lactate Dehydrogenase

Malate dehydrogenase (MDH) and lactate dehydrogenase (LDH) are NAD dependent enzymes located adjacently to one another on chromosome 7. Both belong to the 2-Ketoacid: NAD (P)-dependent dehydrogenase superfamily responsible for the catalytic conversion of 2-Hydroxyacids to the corresponding 2-Ketoacid. MDH is found across all three domains (Eukarya, Bacteria and Archaea) of life while LDH is exclusive to the Eukarya and Bacteria. Both genes are reported to be syntenic single copy genes within the *Cryptosporidium* genome. In contrast genetic evidence for the closely related Apicomplexans *Plasmodium falciparum* and *Toxoplasma gondii* points to the presence of two LDH loci[132]. Genetic investigations indicate that independent of other Apicomplexans, *Cryptosporidium*'s LDH gene evolved from a LDH-like MDH gene[132]. This supports the concept that *Cryptosporidium* has a more distinct molecular and evolutionary divergence from its fellow Apicomplexans. Zhu et al. (2001) showed that both MDH and LDH are extremely substrate specific for oxoalacetate and pyruvate respectively[246]. Mainly responsible for this specificity are the residues at amino acid position 102. For both the LDH and MDH loci this is a conserved region between the two species of *C. hominis* and *C. parvum*. The divergence of *Cryptosporidium*'s LDH and MDH loci from their counterparts in other Apicomplexans, animals and humans supports the idea that these enzymes should be further investigated as rational *Cryptosporidium* specific prophylactic targets.

Figure 5.1 Phylogenetic relationship Apicomplexan parasites based on the MDH and LDH enzymatic genes.

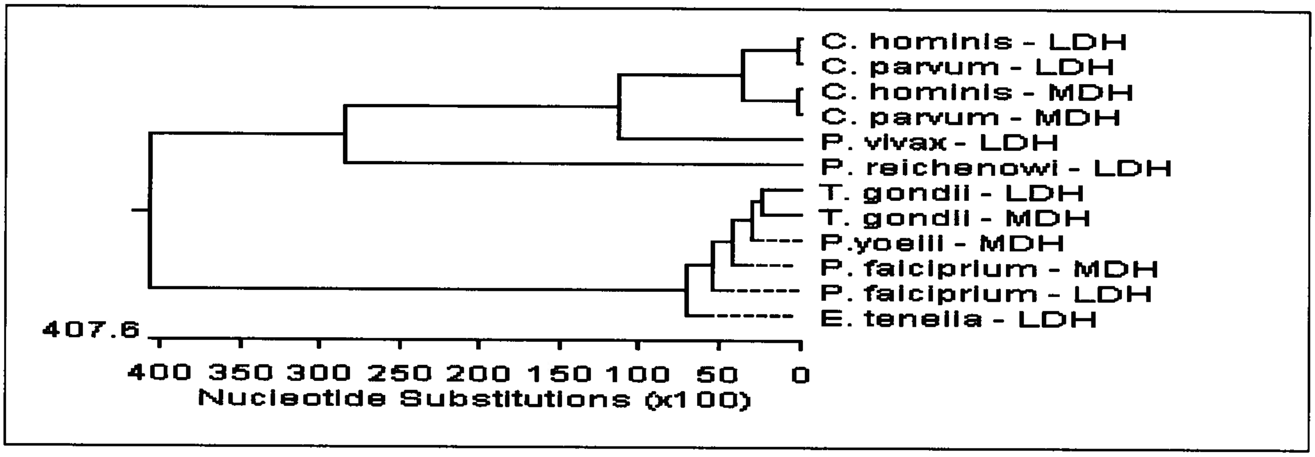


Figure 5.1. The lack of strong genetic similarity is clearly evident as *Cryptosporidium* clades distinctly within its own genus. Phylogeny based on GenBank sequences accession; *C. hominis*-LDH, AAEL01000175; *C. parvum*-LDH, AF274310; *C. hominis*-MDH, AAEL01000175; *C. parvum*-MDH, AY334274; *P. vivax*-LDH, AY486060; *P. reichenowi*-LDH, AB122147; *T. gondii*-LDH,U23207; *T. gondii*-MDH, AY650028; *P. yoelli*-MDH, AABL01000969; *P. falciaprum*-MDH, AY324107; *P. falciprium*-LDH, AF323520; *E. tenella*-LDH, AY143389. JMWilliamson, unpublished.

Lactate Dehydrogenase – Two biological processes are attributed to the LDH loci specifically. First it is thought to play a major role in glycolysis in addition to being involved in the tricarboxylic acid cycle. In terms of molecular function this gene is considered to have both L-lactate dehydrogenase activity as well as oxidoreductase activity. Sequence analysis of alignment of a 965bp fragment of the LDH gene showed 20 SNPs, only one of which appears to confer an amino acid polymorphism. The crucial 102nd amino acid position remained conserved with an aspartic acid residue suggesting no effect on substrate specificity. All 20 silent SNPs were located in the third or wobble position of the codon.

Malate Dehydrogenase – Gene ontology for the MDH loci also includes glycolysis, malate metabolic processes, and tricarboxylic acid cycle involvement in addition to L-malate dehydrogenase and oxidoreductase activity for biological and processes and molecular functions respectively. A 957 bp ClustalW alignment of *C. hominis* MDH and *C. parvum* MDH gene indicated 15 single nucleotide polymorphisms. Of the 15 total SNPs, 11 occur in the third codon position with the remaining 4 occurring in the first codon position. Seven of the 15 SNPs procured an amino acid change, none of which occur at the critical 102nd valine residue amino acid position previously mentioned.

Uracil PhosphoRibosyl Transferase (UPRT)

Located on chromosome 1, UPRTase, is an operative enzyme in *Cryptosporidium* amino acid metabolism. It is the key component of the pyrimidine salvage pathway. Its molecular functions include uridine kinase activity, ATP binding, and kinase activity. In theory hampering the expression of UPRTase could decrease the amount of functional transcripts produced hence diminishing the ability to assimilate. Alignment of a smaller 711bp fragment of the *C. hominis* UPRTase gene and its *C. parvum* ortholog was done. Within this targeted gene fragment there were 2 transitions and 7 transversions and only 1 SNP out of a total of 10 mediated an amino change in the primary peptide sequence. Of the 9 silent mutations 8 occur in the wobble position and the ninth occurs in the first codon position. The lone non-synonymous SNP is situated in the third codon position.

Apoptosis Related Protein (APR)

Apoptosis is characterized by a controlled cellular self-digestive process that functions to eliminate pathogen invaded host cells during the development of infectious organisms. Apoptosis, or cell death, is crucial with respect to host-parasite interactions. Host cell apoptosis on infected cells has a parasiticidal effect but can also kill neighbouring un-infected host cells. *Cryptosporidium* colonizes the gastro-intestinal tract where other histological consequences such as the loss of absorptive or goblet cells and crypt hyperplasia can result. Some studies suggest that the differences exerted on host cell apoptosis, whether parasite infected or not, are mediated by the developmental stages of the parasite. By halting apoptosis in a host cell the parasite is allowed to complete its development to fruition. Once fully formed and functional the parasites can excyst from the host cell, re-establishing apoptosis. The transcription factor NF-kappa B regulates many host derived genes that encode apoptosis inhibitor proteins. One theory is that by upregulating the activation pathway for NF-kappa B *Cryptosporidium* can prolong the life of its host cells giving it sufficient time to fully develop. The exact balance of apoptotic benefits or exploitations between host and parasite are still largely unclear.

For *C. hominis* on chromosome 4 Tzipori et al. 2004 have annotated a 393bp ORF as an apoptosis related protein. Little is known about this protein and its exact functions. To date there is extremely limited genetic documentation or literature on this specific protein thus making it a point of interest within this study. It is considered to be a smaller more conserved protein hypothesized to have a double stranded DNA-binding region. Upon comparison to the *C. parvum* ortholog 9 potential target SNPs were visualized at the basic genetic level. Further analysis at the amino acid sequence for the gene indicated the presence of 3 SAAPs.

Acetyl Co-enzyme A (AcoA)

Acetyl-CoA is an important molecule in metabolism, used in many biochemical reactions, in species of all types. Its main use is to convey the carbon atoms within the acetyl group to the Krebs cycle to be oxidized for energy production. In chemical structure, acetyl-CoA is the thioester between coenzyme A (a thiol) and acetic acid (an acyl group carrier).

In *Cryptosporidium* ACoA is a single-copy gene which has been reported to have genetic preference for thymidine and adenosine amino acid residues in the third codon position. As is the case with most other *Cryptosporidium* genes the gene is extremely dense, containing no intron regions. Gene ontology at the biological level is metabolic while at the molecular level it is thought to have catalytic activity functions as well as AMP binding functions. Previously Upton et al. reported on the degree of low usage codon within this protein. Their results found there to be relatively high number of them within the ACoA open reading frame (ORF). In all species, proteins containing a high percentage of low-usage codons could be characterized as cases where an excess of the protein could be detrimental, thus making it a suitable target for further molecular characterization. For the purpose of this study a smaller 344bp fragment was aligned between the *C. hominis* and *C. parvum* parent strains. The result of this led to the identification of 7 potential single base mutation targets.

Heat Shock Protein 70 (HSP70)

The HSP70 protein is classified as cytoplasmic protein that helps mediate ATP-binding processes. Previous literature has suggested the usefulness of this locus for phylogenetic analysis of the genus *Cryptosporidium.* The heat shock protein belongs to a multi-gene family that is highly conserved across prokaryotes and eukaryotes. These proteins appear to function as molecular chaperones for facilitating the folding of proteins in secretion and transport. Previous studies have shown their up-regulation under environmental stresses and their involvement in cellular protection.

The high incidence of synonymous or silent mutations suggests that this locus is more permissive of nucleotide changes and may be under selection pressure. The fact that these nucleotide mutations are spread over the entire sequence versus those proteins that cluster mutations close to one another indicates that this gene is a more robust target for molecular marking as well as genotyping and phylogenetic studies. Thirty-two single point mutations were mapped within the HSP70 domain, 26 transitions and 6 transversions. In all, only one SNP introduced a single amino acid polymorphism (SAAP).

Cryptosporidium Oocyst Wall Protein (COWP)

The durability of the outer oocyst membrane is largely attributed to a family of *Cryptosporidium* oocyst wall proteins (COWP). Their expression is significantly upregulated during the later stages of its life cycle, a time point coinciding with oocyst development stage. A great deal of the parasites resistance to chemical or environmental stresses comes from this outermost double layered membrane. The 2 layers are tightly connected through a protein-lipid-carbohydrate matrix. COWP-1 was the initial loci to be discovered and analyzed. With the completion of both the *C. hominis* and *C. parvum* genomes 8 additional COWP genes have been uncovered. Supporting the discovery of these other COWP loci was post-sequence analysis on the COWP-1 loci showing a pattern of cysteine residues every 10-12 amino acids. The COWP genes are scattered throughout the genome on multiple chromosomes and are therefore unlinked. Of particular future interest is the COWP-7 gene. Molecular analysis has shown that it houses intronic regions within its open reading frame, a rarity in *Cryptosporidium.*

Using a smaller 552nt fragment of the more well known and researched COWP-1 gene positioned on chromosome 4 alignments and analysis highlighted 8 potential target single nucleotide polymorphisms. Translation into an 184aa sequence indicated that only 1 of these introduced a change to its amino acid makeup. All remaining 7 SNPs involved a redundant amino acid change that occurred in the 3rd or wobble position.

18S rRNA

A central componenet to the ribosome in eukaryotic organisms is the small subunit, the 18S rRNA subunit. Ribosomal RNA's (rRNA's) are targets for many clinically relevant antibiotics, including parmomycin which has been used in varying success with *Cryptosporidium* infection. Typically the ribosomal subunits are among the most conserved genes in cells making them among the most studied and most used in clarifying the taxonomic status of an organism.

Due to its genetic stability 18S rRNA is one of the most reliable and early genes used for the identification and diagnosis of *Cryptosporidium.* The locus has the most molecular data accrued within all the major genetic databases giving an increased confidence in the true presence of what mutations may exist within it. Molecular markers isolated from the multi-copy *Cryptosporidium* 18S rRNA gene are expected to be useful for both species confirmation and as information for population genetic studies. Alignment of multiple partial *C. hominis* and *C. parvum* fragments ranging from 540bp to 745bp revealed only 3 single base mutations or SNPs all of which were silent.

Cp23

The Cp23 gene encodes an immunodominant surface protein. The immunodominant Cp23 locus has mainly been used as an investigative tool in immunoreactivity and serology studies. Characteristic serum immunoglobulin G (IgG) antibody responses to this 27kDa antigen have been shown to develop post-infection[21]. It has also been previously studied at the molecular level for the characterization of gene fragments encoding epitopes to which sporozoite neutralizing antibodies were directed. Between *C. hominis* and *C. parvum* the Cp23 loci is relatively stable genetically. This coupled with functional properties make it another excellent genotypic and clinical marker for investigating *Cryptosporidium* epidemiology.

Alignment of the *C. hominis* and *C. parvum* Cp23 locus revealed 7 nucleotide differences including 5 transitions and 2 transversions. Only 1 of these mutations conferred a single amino acid polymorphism upon translation into the peptide sequence. Of the 6 silent mutations 3 were to be found in the wobble spot with 2 in the first position and the remaining 1 in the middle codon position.

Erythrocyte Membrane Associated Antigen (EMAAg)

Scattered throughout the *Cryptosporidium* genome is a gene family encoding for a protein termed the erythrocyte membrane associated antigen. These loci range in size and degrees of genetic variation between species. The least known protein of the three antigenic proteins under investigation here, the erythrocyte membrane associated antigen was only recently annotated upon the publication of the *C. hominis* genome by Abrahamsen et al. (2004) hence literature on molecular investigations into its genetic structure and organization are almost nil.

A 2037nt fragment of the erythrocyte membrane associated antigen from chromosome one was aligned. In the *C. parvum* homologue there was a 4nt deletion present at positions 520-524 and a 14nt deletion at positions 525 through 539. In all 40 single nucleotide changes were observed, 19 of which were the precursors to a single amino acid change.

Glycoprotein60 (*Gp60)*

The *Cryptosporidium* glycoprotein 60, Gp60, also commonly known as Cpgp 40/15, is a single copy surface protein that in previous molecular studies has shown to be highly polymorphic in intra-species and inter-species population studies. Previous research focused on the molecular clarification of the high degree of genetic differentiation within this protein has led to the identification and classification of five allelic subgroups termed 1a, 1b, 1c, 1d and more recently 1e. Specifically the Gp60 locus encodes for a precursor protein that upon proteolytic cleavage yields two mature cell surface glycoproteins, gp40 and gp15. These two mature proteins are implicated in zoite attachment to and invasion of enterocytes thus in part mediating host cell invasion. Such an observation could imply that polymorphisms in this gene could contribute to differences in the host specificities and infection patterns of *C. hominis* versus *C. parvum*. The high degree of genetic polymorphisms within this protein support the notion that its gene products, gp40 and gp15, are in fact surface associated proteins that are hypothesized to be virulence determinants likely under host immune selection. It stands to reason that further investigation into the unprecedented degree of genetic polymorphism within this protein, particularly from a geographic standpoint, could prove to be very helpful in understanding the molecular epidemiology of cryptosporidiosis.

Comparative genomics for the Gp60 protein revealed extensive polymorphisms, parallel to that of other studies. Sequences from both the *C. hominis* and *C. parvum* genomes were isolated and aligned for visual comparison and SNP position identification. Mapped to Gp60 were 153 SNPs, 72 of which conferred amino acid polymorphisms to the primary protein sequence. This locus is the most variable of all the target proteins involved in the study and considered to be under the most selective pressure.

β-tubulin

A single copy structural protein, the β-tubulin gene is of particular interest as it is one of the few open reading frames in the *Cryptosporidium* genome containing an intron region. A single intron is sandwiched between two open reading frames, exons 1 and 2. This intron is 84nt long with a 2bp deletion in the *C. hominis* gene. A 457nt alignment between *C. hominis* and *C. parvum* showed 2 SNPs within the intron and 9 in the 5` region of the adjacent exon2. Translation into the 153aa sequence revealed all 9 of these to be silent. Both intron SNPs were identical involving a cysteine residue versus a thymidine residue for the *C. hominis* and *C. parvum* respectively. Genetically the gene is variable but stable. SNPs within this gene are hypothesized to be invaluable markers for differentiating unique or novel isolates.

5.3 Multi-locus SNP-type (MISt) Assembly

Compilation of the comparative genomics and bioinformatics criteria as outlined resulted in a pre-defined SNP-type and allelic profile of the *Cryptosporidium* genome based on target SNPs. Collectively 45 SNP loci from the 13 target genes were used to type isolates (Table 5.5).

Table 5.5

		Assembled SNP Marker Panel		
Gene	**SNP loci[+]**	***C. hominis* allele[±]**	***C. parvum* allele[±]**	**expressed/silent**
AcoA	1	C	A	silent
	4	C	T	silent
	6	C	G	silent
	7	A	T	silent
APR	2	T	C	expressed
	3	T	G	expressed
B-tubulin	1	C	T	intronic, silent
	4	A	G	silent
	3	T	G	silent
	5	C	T	silent
	7	G	A	silent
	8	C	T	silent
COWP	5	C	T	silent
	6	T	C	silent
	1	C	T	silent
	3	T	C	silent
	7	G	A	silent
Cp23	4	G	A	silent
	3	T	C	expressed
	1	T	C	silent
	5	C	G	expressed
	6	G	T	expressed
EMAAg	29	G	A	expressed
	27	G	C	silent
Gp60	80	T	A	silent
	108	T	C	expressed
	126	T	C	expressed
	79	T	C	expressed
	98	G	T	expressed
	115	G	A	expressed
Hsp70	14	A	G	silent
	17	A	G	silent
	19	A	G	silent
	20	A	G	silent
	22	A	T	silent
LDH	10	C	T	expressed
	3	C	T	silent
MDH	8	G	A	expressed
	7	G	C	silent
Mucin-1	13	T	G	expressed
	16	A	T	expressed
UPRTase	2	T	A	expressed
	3	T	C	silent
18S rRNA	1	T	A	silent
	3	T	C	silent

Table 5.5. [+] SNP loci identified by internal lab molecular data set. [±]Expected allele as annotated in *C. hominis* TU502 and *C. parvum* Iowa II reference genomes. The allelic profiles for the *C. hominis* and *C. parvum* genomes from 45 molecular markers and 13 gene loci for DNA subtyping of international subpopulations.

5.4 Bioinformatics Computations: Protein Topology & Biophysical Predictions in Relation to Polymorphism.

The distribution of a pathogen in a community of hosts may be determined by the pathogens own genetic content that can in turn be influenced by the genetic make-up of that host as well as to the environment in which the pathogen resides. Ascertaining those genes that may undergo selective pressures from these niches is one approach to the identification of genes putatively involved in a pathogen's unique epidemiological behaviour. In genetics the ratio of non-synonymous or expressed substitutions to synonymous or silent substitutions, can be used to infer the degree or likelihood of positive selection on any given protein-coding gene. A disproportionate number of NS:S substitutions would be indicative of regions under positive diversifying selection. Of the 13 protein encoding genes examined herein only 2 loci demonstrate a slight propensity towards positive selection, both of which are antigenic proteins (Table 5.3). First, is the Gp60 gene and secondly is the EMAAg gene, both considered to be as antigenic determinants in *Cryptosporidium* infection.

The behaviour of amino acids can be dramatically altered depending on their immediate surroundings. Multiple criteria were applied to each of the target genes in the hopes of better elucidating the SNPs within them having the potential to be of greater research value (Table 5.6). ORF analysis consisted of profiling their chemical, structural and positional characteristics, with particular attention paid to those regions containing target SNPs. When collectively considered this multi-faceted approach enabled better decisions about which SNPs were best as molecular targets.

Hydrophobicity is a popular analysis tool because it's a good biological indicator of transmembrane segments or core regions within a protein (Figure 5.2, Appendix 5). The Kyte-Dolittle hydropathic score was used to assess an individual amino acids ability to repel and attract water and to what degree. Finding one transmembrane segment at the N-terminus of a sequence may imply the protein is secreted whereas many transmembrane domains in one protein may indicate a channel. SNPs positioned within such domains could be theorized to be involved in such biological processes such as molecule transport. Protein motifs and any possible mutations within them that are positioned on either side of the cellular membrane could potentially be contributing to biological processes such as cell signalling, phosphorylation or invasion processes. Just as hydrophobic regions can be membrane spanning segments in proteins that anchor themselves into a membrane, hydrophilic stretches could be looping out at the surface of the protein. When coupled to the predictability of the Emini algorithm, probability of finding an amino acid residue on the surface of the protein molecule, we were better able to hypothesize the location of a particular SNP within a protein.

Figure 5.2 Example hydrophobicity plot for the Gp60 ORF returned using the Kyte-Doolittle algorithm.

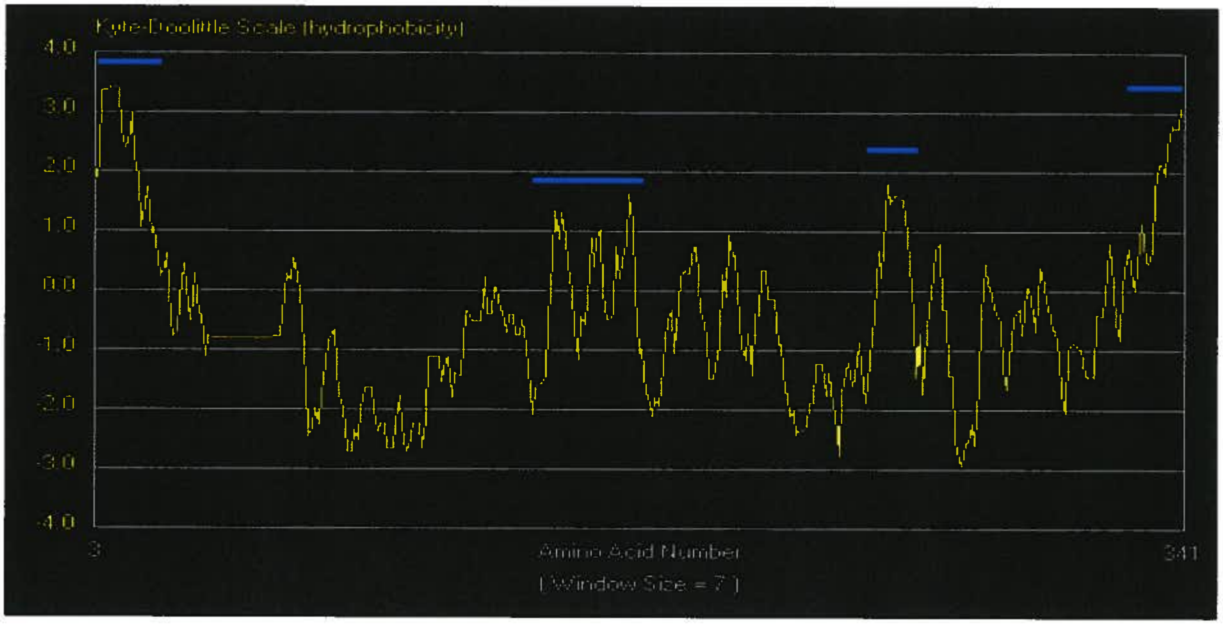


Figure 5.2. Hydrophobicity profile returned by Protean by using the Kyte-Doolittle scale for Gp60 protein locus (orientated 5` → 3`); values greater than zero are reminiscent of hydrophobic, membrane segments, with values of 2 indicating strong hydrophobic character. Regions of hydrophobic nature are highlighted beneath blue bars, with a clear indication that either ends or termini of the protein appear to be membrane anchored.

Further influencing molecular marker or target SNP selection were scores computed by the Jameson-Wolf antigenic index. SNPs evaluated as reminiscent of antigenicity are considered to be of greater research potential when considering the pathogenesis model of *Cryptosporidium*. On a threshold scale of (-) 1.7 to (+) 1.7 values approaching (+) 1.7 are evocative of antigenicity (Table 5.6, Figure A.5).

When undergoing SNP selection the secondary structure characteristics of a given protein were taken into consideration. This is the intermediary structure between the primary structure, the protein sequence, and the tertiary structure, the 3-D folded conformation that is the final shape of the protein. The propensity of the peptide chain to form various secondary structure conformations is basic protein chemistry; α-helices where protein residues seem to follow the shape of a spring, β-sheet or strands where residues are in line and successive residues turn their back to each other, and random coils or turns when the amino acid chain is neither helical nor extended. A mutation located within the inside fold of a turn versus one that may be on the outside of a turn may be under different positive pressures from the external or internal biological environments. `

Table 5.6

BioPhysical Properties of Targeted SNP Loci

Gene	Species	SNP	NS/S*	Marker	KD Score	JW Score	Emini	Chau-Fas	GOR
AcoA (1)	*C. hominis*	C	S	enzymatic	0.52	1.20	1.45	T	C
	C. parvum	A			0.52	1.09	1.45	-	-
AcoA (4)	*C. hominis*	C	S	enzymatic	-0.10	0.00	1.35	-	-
	C. parvum	T			0.01	0.00	1.35	-	-
AcoA (6)	*C. hominis*	C	S	enzymatic	-0.28	-0.10	0.12	-	-
	C. parvum	G			-0.22	-0.10	0.90	-	-
AcoA (7)	*C. hominis*	A	S	enzymatic	-0.14	-0.60	0.44	-	-
	C. parvum	T			-0.14	-0.44	0.84	-	-
APR (2)	*C. hominis*	T	NS	bio-synthesis	0.47	0.45	0.94	-	-
	C. parvum	C			0.47	0.45	0.94	-	-
APR (3)	*C. hominis*	T	NS	bio-synthesis	2.03	1.58	1.09	H/E	E
	C. parvum	G			2.03	0.90	1.13	H/E	H
β-tubulin (1)	*C. hominis*	C	S	introns (molecular)	0.32	0.30	0.51	H/E	E
	C. parvum	T			0.32	0.30	0.51	H/E	E
β-tubulin (4)	*C. hominis*	A	S	structural (molecular)	0.36	-0.60	0.33	-	-
	C. parvum	G			0.36	-0.60	0.33	-	-
β-tubulin (3)	*C. hominis*	T	S	structural (molecular)	0.26	0.45	0.43		
	C. parvum	G			0.22	0.45	0.43	T	C
β-tubulin (5)	*C. hominis*	C	S	structural (molecular)	1.00	0.85	0.58	T	C
	C. parvum	T			0.84	0.85	0.55	T	-
β-tubulin (7)	*C. hominis*	G	S	structural (molecular)	0.43	0.10	0.85	T	-
	C. parvum	A			0.43	0.01	0.85	T	-
β-tubulin (8)	*C. hominis*	C	S	structural (molecular)	-0.24	-0.60	0.60	-	-
	C. parvum	T			-0.23	-0.60	0.72	-	-
Cp23 (4)	*C. hominis*	T	S	antigenic	0.13	0.30	0.62	-	-
	C. parvum	C			0.13	0.30	0.73	-	-
Cp23 (3)	*C. hominis*	T	NS	antigenic	1.68	1.43	1.92	-	T
	C. parvum	C			1.77	1.39	2.21	-	T
Cp23 (1)	*C. hominis*	T	S	antigenic	0.71	1.28	0.94	H/E	T
	C. parvum	C			0.71	0.65	0.94	H/E	T
Cp23 (5)	*C. hominis*	T	NS	antigenic	0.13	-0.30	0.57	-	-
	C. parvum	C			0.13	0.01	0.57	-	-
Cp23 (6)	*C. hominis*	T	NS	antigenic	0.77	0.75	0.94	-	-
	C. parvum	C			0.77	0.75	1.02	-	-
COWP (5)	*C. hominis*	C	S	structural (molecular)	-0.08	0.50	0.43	-	-
	C. parvum	T			-0.08	0.50	0.43	-	-
COWP (6)	*C. hominis*	T	S	structural (molecular)	1.74	1.70	1.92	-	-
	C. parvum	C			1.74	1.70	1.92	-	-
COWP (1)	*C. hominis*	T	S	structural (molecular)	0.22	-0.60	0.29	-	-
	C. parvum	C			0.22	-0.58	0.34	-	-
COWP (3)	*C. hominis*	C	S	structural (molecular)	0..06	-0.05	0.45	T	-
	C. parvum	G			0.00	0.06	0.45	T	-
COWP (7)	*C. hominis*	G	S	structural (molecular)	-0.37	0.25	0.39	T	-
	C. parvum	T			-0.37	0.78	0.39	T	
EMAAg (29)	*C. hominis*	G	NS	antigenic	1.09	1.02	1.40	E	T
	C. parvum	A			1.09	0.59	1.35	-	T
EMAAg (27)	*C. hominis*	G	S	antigenic	-0.41	0.45	0.91	H/E	H
	C. parvum	C			-0.44	0.30	0.38	H/E	H
Gp60 (80)	*C. hominis*	T	S	antigenic	0.77	0.90	1.53	-	-
	C. parvum	A			0.78	0.90	1.53	-	-
Gp60 (108)	*C. hominis*	T	NS	antigenic	1.09	0.75	0.68	-	-
	C. parvum	C			1.09	1.01	0.60	-	-

*NS, non-synonymous; S, synonymous.

Table 5.6 continued

BioPhysical Properties of Targeted SNP Loci									
Gene	**Species**	**SNP**	**N/S***	**Marker**	**KD Score**	**JW Score**	**Emini**	**Chau-Fas**	**GOR**
Gp60 (126)	C. *hominis*	T	NS	antigenic	1.29	1.58	1.27	-	-
	C. *parvum*	C			1.54	1.42	1.30	-	-
Gp60 (79)	C. *hominis*	T	NS	antigenic	2.33	1.50	2.31	T	-
	C. *parvum*	C			2.10	1.66	2.45	T	-
Gp60 (98)	C. *hominis*	G	NS	antigenic	0.32	1.00	1.19	T	C
	C. *parvum*	T			0.30	1.06	1.19	T	C
Gp60 (115)	C. *hominis*	G	NS	antigenic	2.90	1.30	3.60	T	-
	C. *parvum*	A			2.85	1.30	3.61	T	-
HSP70 (14)	C. *hominis*	A	S	bio-synthesis	0.33	0.45	1.03	-	-
	C. *parvum*	G			0.27	0.45	1.03	-	-
HSP70 (17)	C. *hominis*	A	S	bio-synthesis	1.96	0.90	1.45	-	-
	C. *parvum*	G			1.78	0.85	1.44	-	-
HSP70 (19)	C. *hominis*	A	S	bio-synthesis	1.24	1.00	1.45	T	-
	C. *parvum*	G			1.24	0.62	1.45	-	-
HSP70 (20)	C. *hominis*	A	S	bio-synthesis	0.42	-0.73	0.59	-	-
	C. *parvum*	G			0.44	-0.71	0.59	-	-
HSP70 (22)	C. *hominis*	A	S	bio-synthesis	0.29	0.45	0.34	T	C
	C. *parvum*	T			0.29	0.45	0.56	T	C
LDH (10)	C. *hominis*	C	NS	enzymatic	1.46	0.60	1.99	H/E	H
	C. *parvum*	T			0.88	-0.30	0.97	H/E	H
LDH (3)	C. *hominis*	C	S	enzymatic	-0.66	-0.30	0.93	H/E	H
	C. *parvum*	T			-0.66	-0.30	0.93	H/E	H
MDH (8)	C. *hominis*	G	NS	enzymatic	0.54	0.45	0.40	H/E	H
	C. *parvum*	A			0.54	-0.30	0.38	H/E	H
MDH (7)	C. *hominis*	G	S	enzymatic	-0.01	0.30	1.47	H/E	T
	C. *parvum*	C			0.01	0.33	1.34	H/E	T
Mucin-1 (13)	C. *hominis*	T	NS	structural (molecular)	1.68	0.80	1.92	-	T
	C. *parvum*	G			1.40	0.62	2.21	-	T
Mucin-1 (16)	C. *hominis*	A	NS	structural (molecular)	0.71	1.28	0.94	H/E	T
	C. *parvum*	T			0.42	0.98	0.94	H/E	T
UPRTase (2)	C. *hominis*	T	NS	enzymatic	1.19	0.45	2.28	H/E	H
	C. *parvum*	A			0.73	0.45	1.26	H/E	E
UPRTase (3)	C. *hominis*	T	S	enzymatic	0.59	0.95	0.66	H/E	H
	C. *parvum*	C			0.61	0.95	0.66	H/E	H
18S rRNA (1)	C. *hominis*	T	S	ribosomal (structural)	0.44	1.09	0.59	-	-
	C. *parvum*	A			0.44	1.09	0.59	-	-
18S rRNA (3)	C. *hominis*	T	S	ribosomal (structural)	0.29	0.45	0.34	-	-
	C. *parvum*	C			0.33	0.45	1.03	-	-

Table 5.6. Kyte-Dolittle hydropathy score on a scale of regions > 0.0 indicative of hydrophobicity while regions < 0.0 are indicative of hydrophilicity. Jameson-Wolf algorithm for antigenicity, domains at or reaching towards 1.7 are typically considered antigenic. The Emini surface probability indicates the probability of finding an amino acid residue on the surface of the protein molecule; with a threshold scale of 1-6 a value of 1 or greater increasingly predicts probability of protein surface location. Chau-Fasman secondary structures are predicted based on helices (H), sheets (-) or turns (T). GOR predictions for secondary structure are based on helices (H), sheets (E), turns (T) or coils (C). NS, non-synonymous; S, synonymous.

5.5 Multi-locus SNP-typing

A panel of 45 SNP markers was assembled for the purpose of MlS-typing (Table 5.5). In the cases of Australia, Kenya, Peru, and Scotland 43 of these were consistently used to type samples from each subpopulation. Though well amplified through routine and multi-plex PCR methods the SNPs within the APR protein were not scored because of poor activity and/or resolution of SNP markers/loci upon downstream typing. This reduced a panel of 45 SNP markers to 43 SNP markers. The SNaPshot method relies on single base extension (SBE) reactions using dye-labelled, mobility modified detection probes to discriminate alleles. MlSt's were scored by peak size, peak color and peak height ratio. When scoring SNP alleles multiple features were taken into account such as mobility shift of labelling dyes, the presence of artifactual stutter peaks, fragment shift due to hairpins or secondary structures within SNP detection probes and the presence of mixed or double allele calls. In the case of double allele calls a predominant peak of significant and consistent height for the same marker was typically present. Mixed SNP-types were scored based on those as having a secondary allele call of at least a minimum of 20% that of the predominant allele call. For the case of data analysis in a haploid organism such as *Cryptosporidium* the predominant allele call was used for the designation of MlSt based genotyping.

5.5.1 Australia: Multi-locus SNP-typing

Thirteen *C. hominis* isolates and two *C. parvum* reference isolates (A_{14}, A_{15}) were collected from South Western Australia. Of the possible 559 (13*43) SNP loci 422, or 75.4%, were successfully typed (Table 5.7). Eighty-eight allele variants were uncovered from all 12 genes. There were 31 novel allele variants, that is those of neither the *C. hominis* or *C. parvum* expected MlSt. These were uncovered at three SNP positions; position 3 of 18S rRNA, identified in 13 isolates, position 98 and position 115 of Gp60, identified in 11 and 7 isolates respectively. For SNP loci 3 of the 18S rRNA gene isolates A_1-A_{13} were novel. At SNP position 98 of the Gp60 locus samples A $_{1\text{-}6,\ 8,\ 9,\ 10,\ 11,\ 13}$ were novel variants from either of the species specific alleles. Also at SNP marker 115 of the Gp60 locus samples $A_{1,\ 2,\ 3,5,6,7,\ 10}$ revealed unique allele types. Results are not surprising since the Gp60 antigenic protein is one of the most hyper-variable studied. The EMAAg protein was the only gene to show complete genetic stability in all SNPs successfully typed, no *C. parvum* or novel allele variants were seen. The most obvious observation in the Australian subpopulation is the variability of A_6, showing almost a complete *C. parvum* MlSt (Figure 5.3, C). This brings forth the concept of a possible mixed infection within this isolate. Also to be considered is the possibility that it is in fact a *C. parvum* sample displaying *C. hominis* allele variances, keeping in mind samples A_1-A_{13} were designated as confirmed *C. hominis* isolates from the donor country. With the exception of sample A_6 the Australian subpopulation had four genetically stable proteins including COWP, Cp23, HSP70 and EMAAg.

Table 5.7

Australia Subpopulation MlSt & Allele Variants

	B-tubulin						COWP					Cp23					18S		HSP70					AcoA				Muc-1		Gp60						LDH		MDH		EMA		UPRT	
SNP	1	4	3	5	7	8	5	6	1	3	7	4	3	1	5	6	1	3	14	17	19	20	22	1	4	6	7	13	16	80	108	126	79	98	115	10	3	8	7	29	27	2	3
Ch	C	A	T	C	G	C	C	T	C	T	G	G	T	T	C	G	T	T	A	A	A	A	A	C	C	C	A	T	A	T	T	T	T	G	G	C	C	G	G	G	G	T	T
Cp	T	G	G	T	A	T	T	C	T	C	A	A	C	C	G	T	A	C	G	G	G	G	T	A	T	G	T	G	T	A	C	C	C	T	A	T	T	A	C	A	C	A	C
A1	.	.	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	-	-	-	-	-	-	A	C	.	.	A	C	.	-	-	-	.	.	-	.
A2	.	.	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	-	-	-	-	-	-	A	C	.	-	A	C	.	.	.	-	.	.	.	.
A3	.	.	G.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	-	-	-	-	-	-	A	C	.	-	A	C	.	.	.	.	.	.	.	.
A4	.	.	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	-	-	-	-	-	-	A	C	.	.	A	-	-	-	-	-	.	.	-	.
A5	.	.	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	-	-	-	-	-	-	A	C	.	-	A	C	-	-	-	-	.	.	-	.
A6	T	.	G	-	A	T	T	C	T	C	.	A	C	C	G	T	.	A	G	G	G	G	T	-	T	G	T	-	T	A	C	.	-	A	C	T	-	A	C	.	-	A	-
A7	.	.	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	-	-	-	-	-	-	A	C	.	-	.	C	T	T	A	C	.	.	A	C
A8	.	.	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	-	.	.	-	-	-	A	C	.	-	A	.	-	-	-	-	.	.	-	-
A9	.	.	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	-	-	-	-	-	-	A	C	.	-	A	.	.	.	.	.	.	.	.	.
A10	.	-	.	-	.	.	.	-	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	-	-	-	-	-	-	A	C	.	-	A	C	.	.	.	.	.	.	.	.
A11	.	.	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	-	-	-	-	-	-	A	C	.	-	A	.	-	-	-	-	.	.	-	-
A12	.	.	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	-	-	-	.	-	-	.	.	.	-	.	.	-	-	-	-	.	.	-	-
A13	.	.	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	-	-	-	.	-	-	A	C	.	-	A	.	.	.	.	.	.	.	.	.

Table 5.7. Ch; *C. hominis*. Cp; *C. parvum*. (.) denotes alleles scored in agreement with the expected *C. hominis* SNP subtype (top). (-) denotes markers unsuccessfully typed or scored. Allele designations, A, C, T, G, represent allele variants that deviate from the expected *C. hominis* MlSt.

In the Australian subpopulation, 2 known *C. parvum* isolates were received as reference samples from the donating colleague (Table 5.8). These proved to be beneficial in providing a comparative differential from the *C. hominis* data. They allowed for the testing of species distinction via fragment analysis and point mutations on international isolates. Both *C. parvum* isolates showed the presence of a novel allele at SNP position 3 of the 18S rRNA locus. This molecular marker therefore gave the same novel allele call in all Australian samples, whether *C. hominis* or *C. parvum*. Also of note, though not completely unexpected given the hyper-variable nature of the locus, differential allele calls were seen in 4 of the Gp60 SNP markers; Gp60 126, and 79, 98,126 and 115 for isolates A_{14} and A_{15} respectively.

Table 5.8

Australia MlSt of *C. parvum* Isolates

	B-tubulin						COWP					Cp23					18S		HSP70					AcoA				Muc-1		Gp60						LDH		MDH		EMA		UPRT	
SNP	1	4	3	5	7	8	5	6	1	3	7	4	3	1	5	6	1	3	14	17	19	20	22	1	4	6	7	13	16	80	108	126	79	98	115	10	3	8	7	29	27	2	3
Ch	C	A	T	C	G	C	C	T	C	T	G	G	T	T	C	G	T	T	A	A	A	A	A	C	C	C	A	T	A	T	T	T	T	G	G	C	C	G	G	G	G	T	T
Cp	T	G	G	T	A	T	T	C	T	C	A	A	C	C	G	T	A	C	G	G	G	G	T	A	T	G	T	G	T	A	C	C	C	T	A	T	T	A	C	A	C	A	C
A14	T	G	G	-	A	T	T	C	T	C	A	A	C	C	G	T	A	A	G	G	G	G	T	A	T	G	T	-	T	A	C	T	-	-	-	T	T	A	C	A	-	A	C
A15	T	G	G	-	A	T	T	C	T	C	A	A	C	C	G	T	A	A	G	G	G	G	T	-	-	-	-	-	-	A	C	T	T	A	C	T	-	A	C	A	-	A	C

Table 5.8. Ch; *C. hominis*. Cp; *C. parvum*. (.) denotes alleles scored in agreement with the expected *C. hominis* SNP subtype (top). (-) denotes markers unsuccessfully typed or scored. Allele designations, A, C, T, G, represent allele variants that deviate from the expected *C. hominis* MlSt.

Figure 5.3 Australia, electropherogram representations; HSP70 locus SNP markers 14, 17, 19, 20, 22 (left to right).

A. Isolate A₉

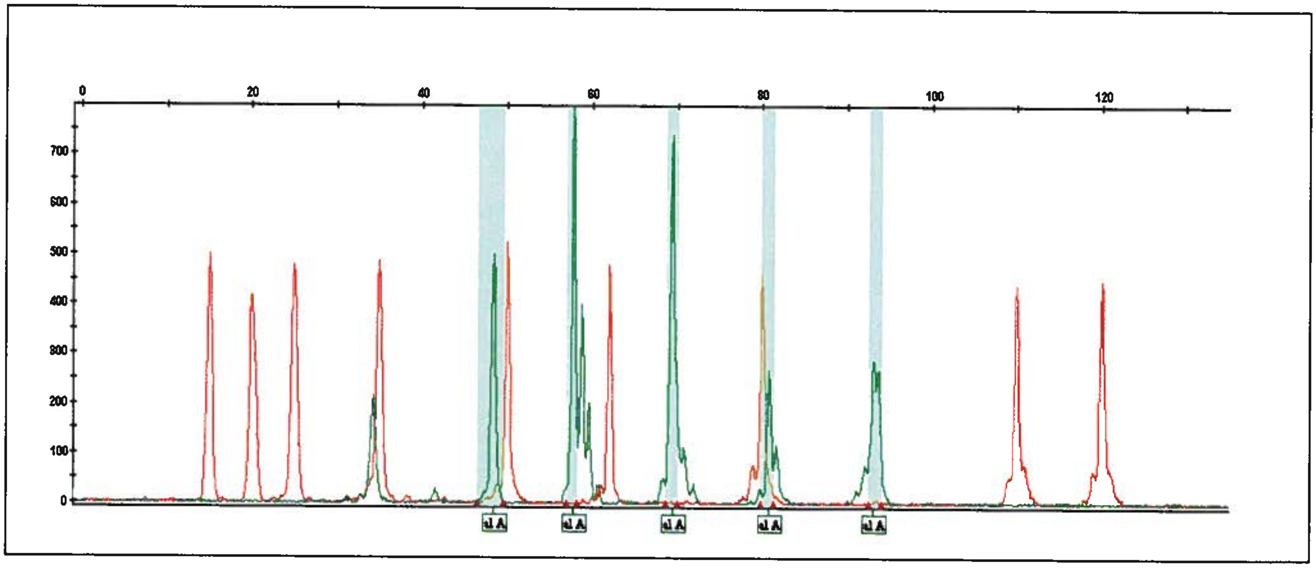


Figure 5.3, A. Expected *C. hominis* profile as seen by 5 adenine alleles scored to each marker (A, fluoresces green).

Figure 5.3 continued. Australia electropherogram representations; HSP70 locus SNP markers 14, 17, 19, 20, 22 (left to right).

B. Isolate A_3.

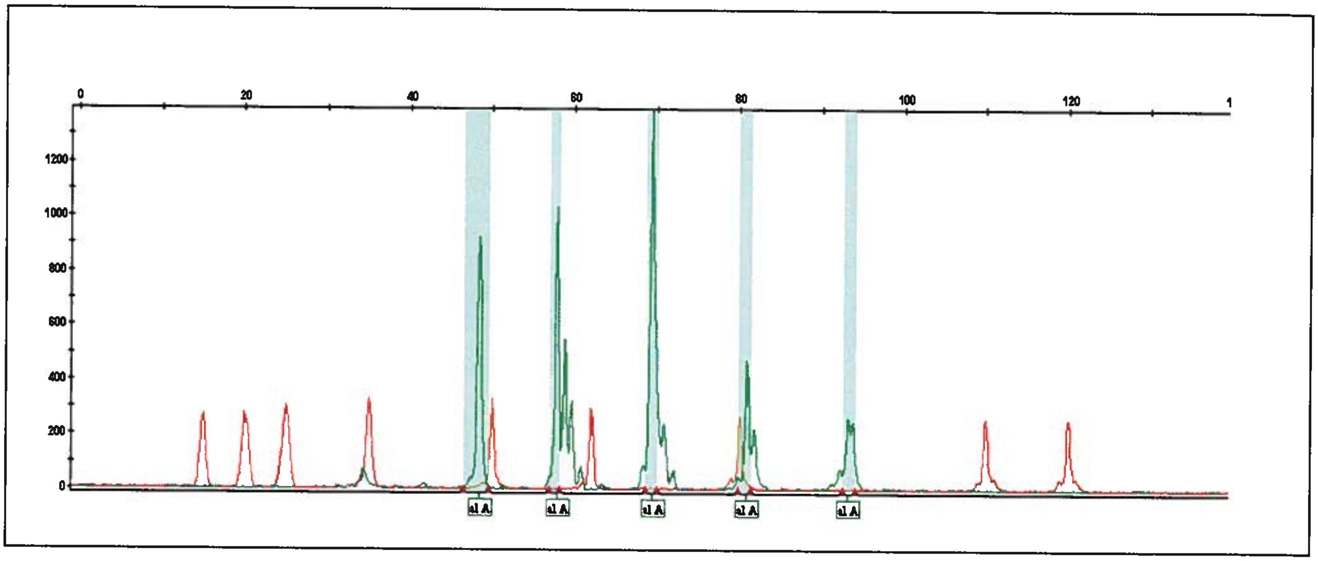


Figure 5.3, B. Expected *C. hominis* profile as seen by 5 adenine alleles scored to each marker (A, fluoresces green)

C. Isolate A_6.

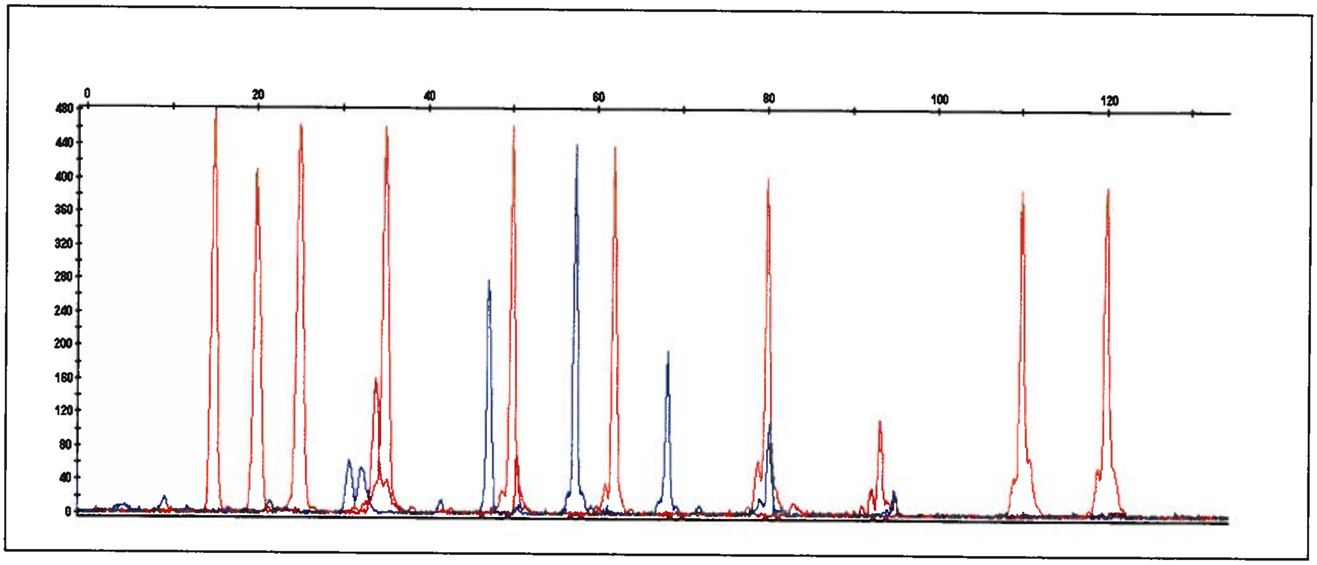


Figure 5.3, C. HSP70 profile for A_6, illustrating the *C. parvum* profile within *C. hominis* isolate, illustrated by the 4 guanine stretch followed by a single thymine allele (G, fluoresces blue; T, fluoresces red).

Figure 5.3 continued. Australia electropherogram representation; HSP70 locus SNP markers 14, 17, 19, 20, 22 (left to right).

D. Isolate A_{14}.

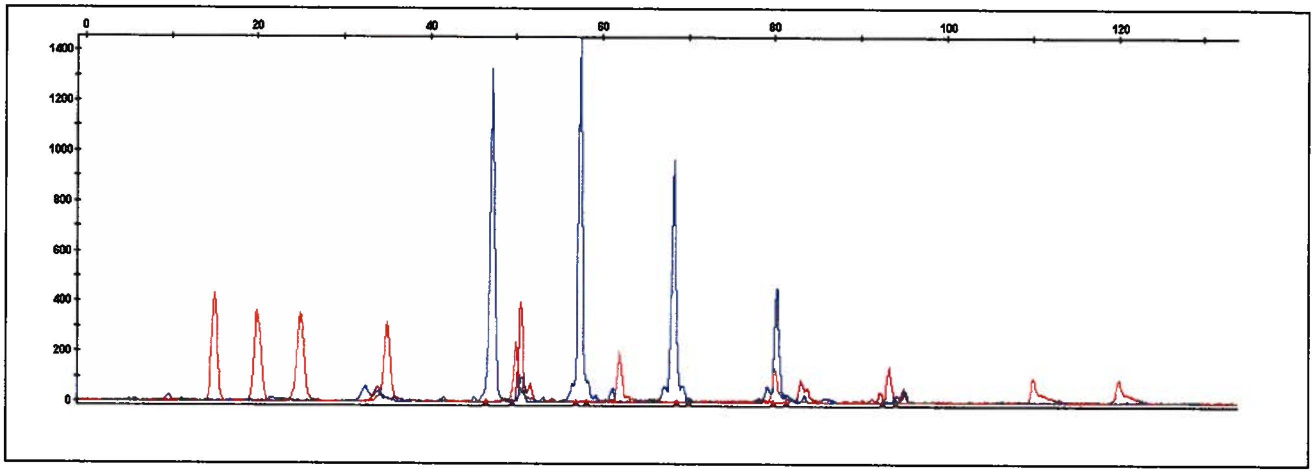


Figure 5.3, D. HSP70 profile for A_{14}, illustrating the *C. parvum* profile within *C. hominis* isolate, illustrated by the 4 guanine stretch followed by a single thymine allele (G, fluoresces blue; T, fluoresces red).

5.5.2 Kenya: Multi-locus SNP-typing

Of a possible 860 SNP positions to be typed (20*43) 583, or 67.8 %, were successfully scored for a specific allele; 74 were variant or not of the expected *C. hominis* SNP-type (Table 5.9). Six genes demonstrated complete genetic stability; COWP, Cp23, LDH, MDH, EMAAg, and UPRT. The EMAAg gene was also completely stable in the Australian subpopulation and with the exception of A_6. Minus K_{19} the HSP70 locus was genetically stable at all 5 SNP loci. When looking at all the differential alleles there is a slight predilection seen towards novel alleles versus those of the closely related *C. parvum* MlSt. There were 39 novel alleles in total for the Kenyan subpopulation. The remainder allele scores (35) were that of the *C. parvum* genotype. The same novel allele variant, adenine, was seen at SNP marker 3 of the 18S rRNA gene in isolates $K_{1,\,3\text{-}5,\,7\text{-}16,\,18\text{-}20}$. There was a novel allele, in a sole Kenyan sample, K_{11}, at SNP marker 7 of the β-tubulin locus. This was not seen in any of the other 3 intercontinental subpopulations. The remaining novel alleles were typed to the hyper-variable Gp60 locus in $K_{4,8,11}$ at position 79, in $K_{1,4,17,18}$ at position 98 and in $K_{1,\,4\text{-}7,\,10\text{-}13,\,15\text{-}17,\,19,\,20}$ at SNP position 115. Although variant, alleles scored at SNP markers 80 and 108 for Gp60 were consistently that of the *C. parvum* MlSt. A result of note for this particular protein, also mimicked in the *C. hominis* isolates of the Australian subpopulation, is the lack of genetic differentiation at SNP position 126. In such a variable gene, this could provide early indications of a specific biofunctionality, its expression or lack of, or even its position within the primary folding conformation of the Gp60 protein.

Table 5.9

Kenya Subpopulation MlSt & Allele Variants

	B-tubulin						COWP					Cp23					18S		HSP70					AcoA				Muc-1		Gp60						LDH		MDH		EMA		UPRT	
SNP	1	4	3	5	7	8	5	6	1	3	7	4	3	1	5	6	1	3	14	17	19	20	22	1	4	6	7	13	16	80	108	126	79	98	115	10	3	8	7	29	27	2	3
Ch	C	A	T	C	G	C	C	T	C	T	G	G	T	T	C	G	T	T	A	A	A	A	A	C	C	C	A	T	A	T	T	T	T	G	G	C	C	G	G	G	G	T	T
Cp	T	G	G	T	A	T	T	C	T	C	A	A	C	C	G	T	A	C	G	G	G	G	T	A	T	G	T	G	T	A	C	C	C	T	A	T	T	A	C	A	C	A	C
K1	.	.	-	-	.	-	-	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	-	-	-	-	-	-	A	C	.	-	A	C	.	-	.	.	.	.	.	.
K2	.	.	.	-	.	.	.	.	.	.	.	-	-	-	-	-	-	.	.	.	.	.	.	-	.	.	-	.	-	A	C	.	.	.	.	.	-	.	-	.	.	.	.
K3	-	-	-	-	-	-	-	-	-	-	-	.	.	.	.	.	.	A	.	.	.	.	.	-	.	.	-	-	-	A	-	.	-	-	-	-	-	.	-	.	.	.	.
K4	-	-	-	-	.	-	-	-	-	-	-	.	.	.	.	.	.	A	.	.	.	.	.	-	-	-	-	-	-	A	C	.	G	C	C	.	.	.	.	.	.	.	.
K5	.	.	-	-	-	-	.	.	.	.	.	-	-	-	-	-	.	A	.	.	.	.	.	-	.	.	-	-	-	-	-	.	-	-	C	.	-	.	.	.	.	.	.
K6	.	.	G	-	.	.	.	.	.	.	.	-	-	-	-	-	-	-	.	.	.	.	.	-	-	-	-	-	-	A	C	-	C	.	C	.	.	.	.	.	.	.	.
K7	.	.	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	-	-	-	-	-	-	A	C	.	-	.	C	.	.	.	.	.	.	.	.
K8	.	.	G	-	.	.	.	.	-	-	.	.	.	.	.	.	.	A	.	.	.	.	.	-	-	-	-	-	-	-	C	.	G	.	.	.	.	.	.	.	.	.	.
K9	.	.	-	-	-	-	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	-	.	.	-	-	-	.	.	.	-	.	.	.	.	.	.	.	.	.	.
K10	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	-	-	-	-	-	-	A	C	.	.	.	C	-	.	.	.	-	-	.	.
K11	-	-	-	.	C	-	-	-	-	-	-	-	-	-	-	-	.	A	.	.	.	.	.	A	-	-	-	-	-	A	-	.	G	.	C	.	-	-	-	-	-	-	-
K12	.	.	.	.	-	-	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	-	.	.	-	-	T	.	.	.	.	.	C	.	-	.	.	.	.	.	.
K13	-	.	-	.	.	-	-	.	.	-	.	.	.	.	.	.	.	A	.	.	.	.	.	-	-	-	-	-	-	A	C	.	.	.	C	-	-	-	-	-	-	-	-
K14	.	.	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	-	.	.	-	-	-	.	.	.	-	.	.	.	.	.	.	.	.	.	.
K15	.	.	.	-	.	.	.	-	.	-	.	.	.	.	.	.	.	A	.	.	.	.	.	-	.	.	-	-	-	A	C	.	-	.	C	-	-	.	.	-	-	.	.
K16	.	.	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	-	.	-	.	.	-	-	-	A	C	.	.	.	C	-	-	-	-	.	.	-	-
K17	.	.	.	-	.	.	.	.	.	.	.	-	-	-	-	-	-	.	-	-	-	-	-	-	-	-	-	-	-	A	C	.	-	A	C	-	-	.	.	.	.	.	.
K18	.	.	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	-	-	-	-	-	-	-	C	.	-	A	A	.	-	.	.	-	-	.	.
K19	.	.	.	-	.	.	-	.	.	.	.	.	.	.	.	.	.	A	.	G	.	.	.	-	-	-	-	-	-	A	C	.	-	-	C	-	-	-	-	-	-	-	-
K20	.	.	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	-	-	-	-	-	-	A	C	.	-	-	C	.	-	.	.	-	-	.	.

Table 5.9. Ch; *C. hominis*. Cp; *C. parvum*. (.) denotes alleles scored in agreement with the expected *C. hominis* SNP subtype (top). (-) denotes markers unsuccessfully typed or scored. Allele designations, A, C, T, G, represent allele variants that deviate from the expected *C. hominis* MlSt.

Figure 5.4 Kenya electropherogram representation; isolate K_9, COWP locus SNP markers 1, 3, 7.

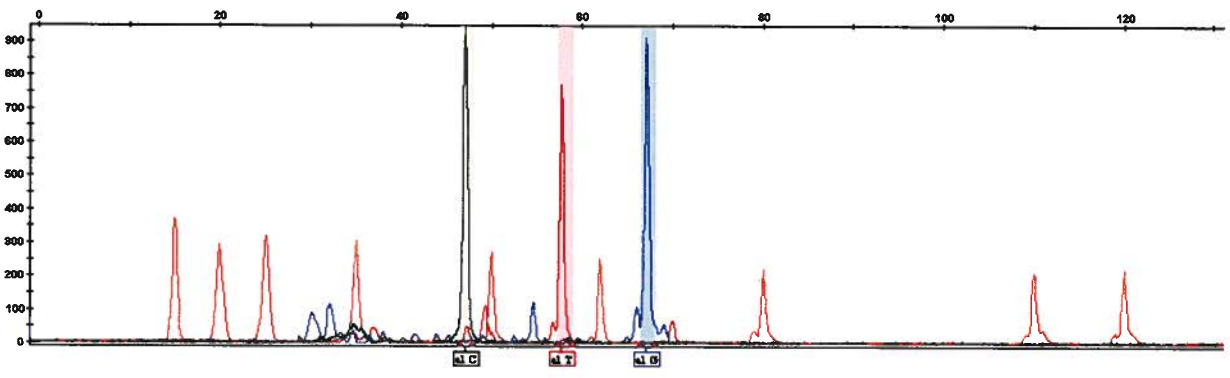


Figure 5.4. Electropherogram of Kenyan isolate, K_9, COWP SNP markers 1, 3, 7, alleles C, T and G respectively (left to right). Liz120 size standard seen in orange.

5.5.3 Peru: Multi-locus SNP-typing

With a subpopulation of 22 isolates there is the potential for 946 (22*43) SNP markers to be scored. At a success rate of 71.5 %, 676 were typed, of these only 10.8% or 73 in total were variant. Variant alleles were only seen in 4 loci; β-tubulin, COWP, 18S rRNA, and Gp60. Five of the remaining proteins, Cp23, HSP70, LDH, MDH, EMAAg and UPRT, showed complete genetic stability while AcoA and Muc-1 had too limited typing results to base any conclusive inferences. The Peruvian subpopulation shows the first report of a novel allele being detected in the β-tubulin locus at SNP position 3 (Table 5.10) as demonstrated in isolates $P_{5, 8,9,14, 21}$. Variants at the same SNP loci seen in Australia and Kenya were of the *C. parvum* MlSt. This is also the first report of 2 novel alleles revealed in the COWP locus which can be seen in samples $P_{5,8 \text{ and } 21}$ at SNP marker COWP6 and again in P_{21} at marker COWP3. As was the case in Australia and Kenya, the novel allele of A, was again seen 18S rRNA, position 3. If in fact a true novel allele it appears to be stable despite geography although Kenya had two incidences of an honest *C. hominis* allele type, K_2 and K_{17}. Multiple allele differences can be seen throughout the hyper-variable Gp60 gene. Variant alleles scored at SNP positions 80 and 108 for Gp60 were that of the *C. parvum* expected genotype. For SNP 126 marker all isolates are again stable except for P_2 showing a *C. parvum* allele call. As will be discussed this is also the case for the Scottish subpopulation. The stability of Gp60 marker 126 in all four subpopulations could suggest it would be a suitable SNP marker from such a variable protein for species distinction, phylogenetic studies or one to monitor for possible mutations evolving in the future. The other variant alleles for the Gp60 protein are spread throughout markers 98 and 115. Only 9 samples were successfully typed for Gp60 position 79 and there were no variances seen. As also seen in Kenya loci Cp23, LDH, MDH, EMAAg and UPRTase are all genetically stable for all markers typed. Markers mapped to the HSP70 locus gave a complete *C. hominis* MlSt profile with 95 of a possible 110 SNPs successfully typed. The only exceptions result from a failure to type samples $P_{4, 5, 15}$ for all five HSP70 SNP loci.

Table 5.10

Peru Subpopulation MlSt & Allele Variants

	B-tubulin						COWP					Cp23					18S		HSP70					AcoA				Muc-1		Gp60						LDH		MDH		EMA		UPRT	
SNP	1	4	3	5	7	8	5	6	1	3	7	4	3	1	5	6	1	3	14	17	19	20	22	1	4	6	7	13	16	80	108	126	79	98	115	10	3	8	7	29	27	2	3
Ch	C	A	T	C	G	C	C	T	C	T	G	G	T	T	C	G	T	T	A	A	A	A	A	C	C	C	A	T	A	T	T	T	T	G	G	C	C	G	G	G	G	T	T
Cp	T	G	G	T	A	T	T	C	T	C	A	A	C	C	G	T	A	C	G	G	G	G	T	A	T	G	T	G	T	A	C	C	C	T	A	T	T	A	C	A	C	A	C
P1	.	.	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	-	-	-	-	-	-	A	C	.	.	A	A	.	.	.	.	.	.	.	.
P2	.	.	.	-	.	.	.	.	.	.	.	-	-	-	-	-	.	A	.	.	.	.	.	-	.	.	-	-	-	.	C	C	.	A	A	.	.	.	.	.	.	.	.
P3	.	.	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	-	.	.	-	-	-	A	C	.	.	C	C	.	.	.	.	.	.	.	.
P4	.	.	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	A	-	-	-	-	-	-	-	-	-	-	-	A	C	.	-	A	.	.	-	.	-	.	.	.	-
P5	.	.	A	-	.	.	-	A	.	-	-	-	-	-	-	-	.	A	-	-	-	-	-	-	-	-	-	-	-	-	-	.	-	-	C	-	-	-	-	-	-	-	-
P6	.	-	.	-	.	-	.	-	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	-	-	-	-	-	-	A	C	.	-	A	C	.	-	.	.	.	.	.	.
P7	.	.	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	-	.	.	-	-	-	A	C	.	.	A	A	-	.	-	.	.	.	-	.
P8	.	.	A	-	.	-	-	A	-	-	.	-	-	-	-	-	.	A	.	.	.	.	.	-	-	-	-	-	-	-	-	.	-	-	C	-	-	.	-	.	.	.	.
P9	.	.	A	-	.	-	-	.	.	-	-	-	-	-	-	-	-	-	.	.	.	.	.	-	.	.	-	-	-	-	-	.	-	-	C	-	.	-	-	-	-	.	.
P10	.	.	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	-	.	.	-	-	-	.	.	.	-	.	.	.	.	.	.	.	.	.	.
P11	.	.	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	-	-	-	-	-	-	.	.	.	.	.	.	.	-	.	.	-	-	.	.
P12	.	.	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	-	-	-	-	-	-	.	.	.	.	A	.	.	.	.	.	.	.	.	.
P13	.	.	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	-	-	-	-	-	-	.	.	.	-	.	.	-	-	-	-	-	-	-	.
P14	.	.	A	-	.	-	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	-	-	-	-	-	-	-	C	.	-	A	C	-	-	-	-	-	-	-	-
P15	.	.	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	A	-	-	-	-	-	-	.	-	-	-	-	A	C	.	-	A	.	.	.	.	.	.	.	.	.
P16	.	-	.	-	.	.	.	-	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	-	-	-	-	-	-	A	C	.	.	C	C	-	-	.	.	.	.	.	.
P17	.	.	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	-	.	.	-	.	-	.	.	.	-	.	C	.	.	.	.	.	.	.	.
P18	.	.	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	-	.	.	-	-	-	.	.	.	.	.	.	.	-	.	.	.	.	.	.
P19	.	.	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	-	-	-	-	-	-	.	.	.	.	.	.	-	-	-	-	.	.	.	.
P20	.	.	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	-	-	-	-	-	-	.	.	.	-	.	.	-	.	-	-	.	.	.	.
P21	.	.	A	.	.	.	.	A	.	G	.	.	.	.	.	.	.	A	.	.	.	.	.	-	-	-	-	-	-	.	.	.	-	-	C	.	.	.	.	-	.	.	.
P22	.	.	.	-	.	.	-	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	-	-	-	-	-	-	A	C	.	-	A	C	.	.	.	.	.	.	.	.

Table 5.10. Ch; *C. hominis*. Cp; *C. parvum*. (.) denotes alleles scored in agreement with the expected *C. hominis* SNP subtype (top). (-) denotes markers unsuccessfully typed or scored. Allele designations, A, C, T, G, represent allele variants that deviate from the expected *C. hominis* MlSt.

Figure 5.5 Peru electropherogram representations; isolate P_6, of Cp23 and 18S rRNA loci SNP markers.

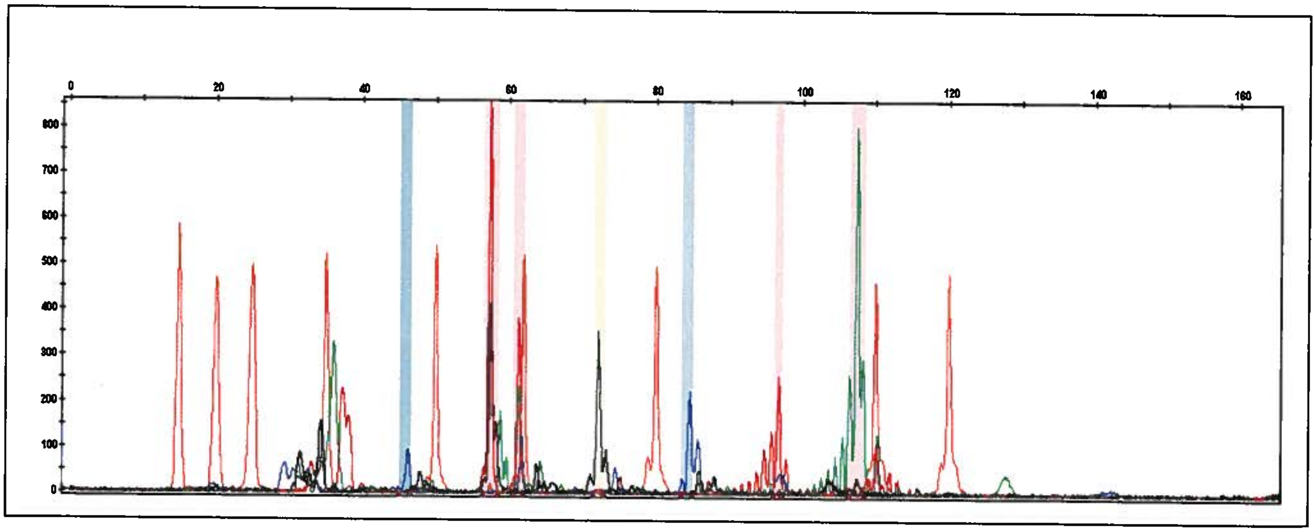


Figure 5.5. Peruvian isolate, P_6, reaction set 5 marker profile comprised of Cp23 markers 4,3,1,5,6 and 18S rRNA markers 1 and 3 (left to right). Allelic profile for expected *C. hominis* genotype would be G-T-T-C-G for COWP and T-T for 18S rRNA (Table 5.10). Shown is 18S rRNA's presence of a strong A allele at marker 3, far right. G (fluoresce blue), T (fluoresce red), C (fluoresce black), A (fluoresce green). Liz120 size standard seen in orange.

Figure 5.6 Peru electropherogram representations; isolate P_{21}, COWP locus SNP 3 novel allele variant, G.

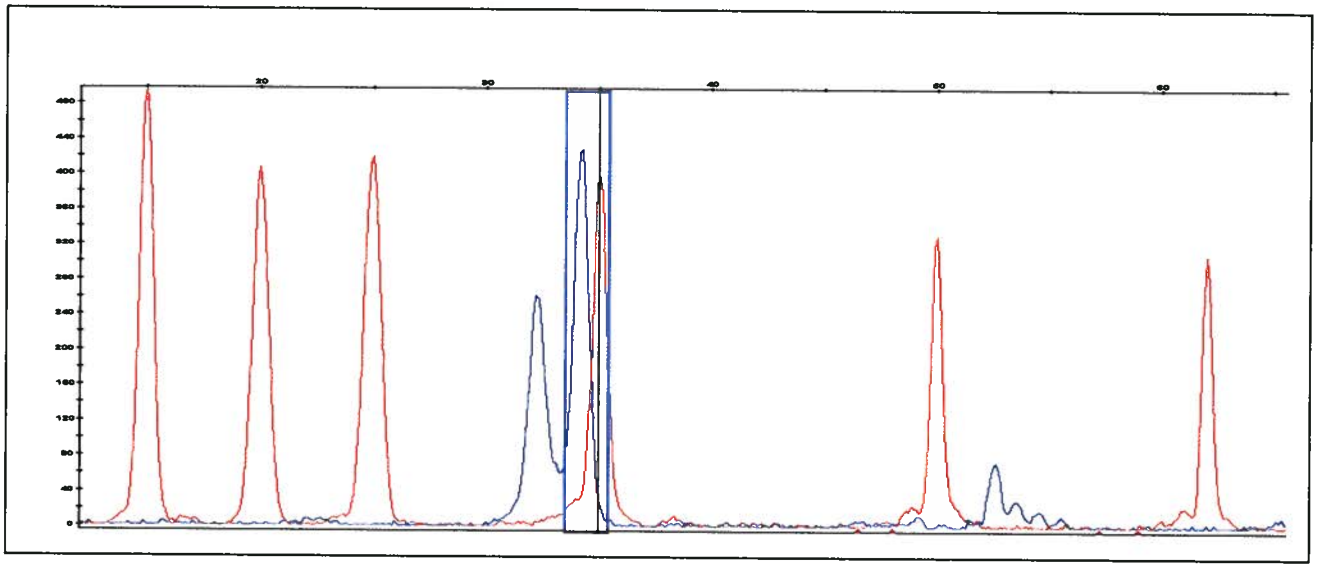


Figure 5.6.Electropherogram depicting G (fluoresce blue) variant allele at SNP marker COWP 3, expected *C. hominis* and *C. parvum* allele would be T or C respectively. Liz120 size standard seen in orange.

5.5.4 Scotland: Multi-locus SNP-typing

With a sample subpopulation of 20 there is the potential for 860 SNP markers to be typed. We were able to achieve 622 of these, a success rate of 71.1% (Table 5.11). Immediate observations reveal genetic stability for the COWP, Cp23, HSP70, LDH, MDH, EMAAg and UPRTase loci. This is akin to the results of SNP typing for both the Australian and Kenyan subpopulations and with the exception of the COWP gene the Peruvian subpopulation as well. Only a handful of markers were able to be reliably typed to the ACoA and Mucin-1 genes; those that were displayed genetic homology to the expected *C. hominis* MlSt.

Ninety-three different alleles were observed, 49 of which were novel. A novel allele variant seen in 5 Peruvian samples ($P_{5, 8, 9, 14, 21}$) at SNP locus 3 of the β-tubulin gene is also seen in the Scotland subpopulation in samples $S_{2 \text{ and } 10}$. The novel allele seen at position 3 in 18S rRNA in the previous three subpopulations is also clearly present in Scotland. All four subpopulations had the same variant allele in 71 of the 73 total samples successfully scored for this particular marker in the 18S rRNA locus. The SNP position and allelic profile was re-examined and confirmed for the expected *C. hominis* or *C. parvum* genotype. Given that A_6, an isolate that gave a very dominant *C. parvum* profile and both of the confirmed

C. parvum isolates, $A_{14, 15}$, also scored the same novel allele as the *C. hominis* isolates could imply that is a true novel allele.

The Gp60 gene marker 126 shows a genetically stable *C. hominis* MlSt. Markers 80 and 108 were genetically stable though of the *C. parvum* genotype. SNP marker 98 was stable as well albeit of a novel allele variant, A, which was also seen in 11 of the 13 Australian isolates. There were no mixed allele calls for this particular protein in any of the Scotland samples. In the Australian, Kenyan, and Peruvian subpopulations this protein had numerous double allele calls, of which the dominant one was used.

Table 5.11

Scotland Subpopulation MlSt & Allele Variants

	B-tubulin						COWP					Cp23					18S		HSP70					AcoA				Muc-1		Gp60						LDH		MDH		EMA		UPRT	
SNP	1	4	3	5	7	8	5	6	1	3	7	4	3	1	5	6	1	3	14	17	19	20	22	1	4	6	7	13	16	80	108	126	79	98	115	10	3	8	7	29	27	2	3
Ch	C	A	T	C	G	C	C	T	C	T	G	G	T	T	C	G	T	T	A	A	A	A	A	C	C	C	A	T	A	T	T	T	T	G	G	C	C	G	G	G	G	T	T
Cp	T	G	G	T	A	T	T	C	T	C	A	A	C	C	G	T	A	C	G	G	G	G	T	A	T	G	T	G	T	A	C	C	C	T	A	T	T	A	C	A	C	A	C
S1	.	.	.	.	.	.	.	-	.	.	.	.	.	.	.	.	.	A	.	.	.	.	-	-	.	.	-	-	-	A	C	.	.	A	A	.	.	-	-	.	.	.	.
S2	-	-	A	-	.	T	-	.	.	-	.	.	.	.	.	.	.	A	.	.	.	.	.	-	-	-	-	-	-	-	C	.	-	A	A	-	-	-	-	.	.	-	-
S3	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	-	.	.	-	.	-	A	C	.	-	A	.	-	-	-	-	-	-	-	-
S4	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	-	-	-	-	-	-	A	C	.	-	A	A	.	.	.	.	.	.	.	.
S5	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	-	-	-	-	-	-	A	C	.	-	A	A	-	.	-	-	-	-	-	.
S6	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	-	-	-	-	.	-	-	C	.	-	A	C	.	.	.	.	.	.	.	.
S7	.	.	-	-	-	-	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	-	-	-	-	-	-	-	C	.	-	A	A	-	.	-	.	.	.	-	.
S8	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	-	-	-	-	-	-	-	C	.	-	A	A	-	-	-	-	.	.	-	-
S9	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	-	-	-	-	-	-	A	C	.	-	A	C	.	.	.	.	.	.	.	.
S10	-	-	A	-	.	-	-	-	.	-	.	.	.	.	.	.	.	A	.	.	.	.	.	-	-	-	-	-	-	-	C	.	-	A	C	-	-	-	-	.	.	-	-
S11	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	-	-	-	-	-	-	A	C	.	-	A	A	-	.	-	.	.	.	-	.
S12	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	-	-	-	-	-	-	-	-	-	-	-	A	-	.	-	-	C	-	.	.	.	.	.	-	.
S13	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	-	.	.	-	.	-	-	C	.	-	A	A	-	.	-	.	.	.	-	.
S14	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	-	.	.	-	-	-	-	C	.	-	A	A	-	.	-	.	.	.	-	.
S15	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	-	-	-	-	-	-	-	C	.	-	A	C	-	.	-	.	-	-	-	.
S16	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	-	-	-	-	-	-	A	C	.	-	A	A	.	.	.	.	.	.	.	.
S17	.	.	.	.	.	.	.	.	.	.	.	-	-	-	-	-	-	A	.	.	.	.	.	-	.	.	-	-	-	A	C	.	-	A	A	-	-	-	-	.	.	-	-
S18	.	-	.	.	.	.	-	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	-	-	-	-	-	-	A	C	.	-	A	A	.	.	.	.	.	.	.	.
S19	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	-	-	-	-	-	-	A	C	.	-	A	C	-	-	-	-	.	.	-	-
S20	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	-	.	.	-	-	-	A	C	.	-	A	A	.	.	.	.	.	.	.	.

Table 5.11. Ch; *C. hominis*. Cp; *C. parvum*. (.) denotes alleles scored in agreement with the expected *C. hominis* SNP subtype (top). (-) denotes markers unsuccessfully typed or scored. Allele designations, A, C, T, G, represent allele variants that deviate from the expected *C. hominis* MlSt.

Figure 5.7 Scotland electropherogram representations; Cp23 and 18S rRNA loci SNP markers.

A. Isolate S_3.

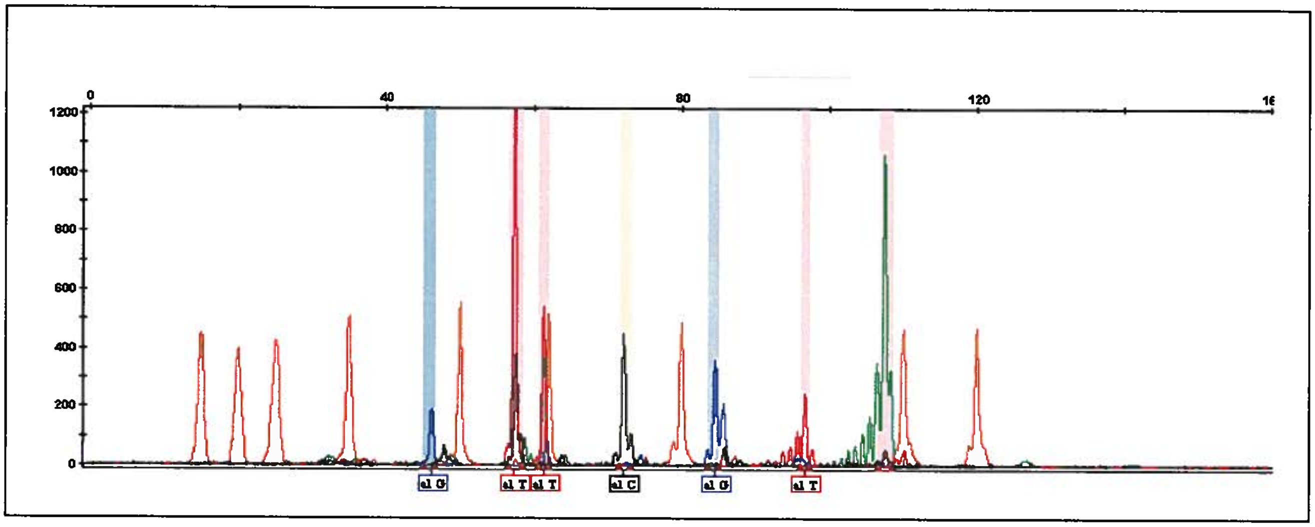


Figure 5.7, A. Scotland isolate, S_3, reaction set 5 marker profile comprised of Cp23 locus markers 4,3,1,5,6 and 18S rRNA locus markers 1 and 3 (left to right). Allelic profile for expected *C. hominis* genotype G-T-T- C-G for COWP and T-T for 18S rRNA (Table 5.11). Shown is 18S rRNA's presence of strong A allele at marker 3, far right.

B. Isolate S_6.

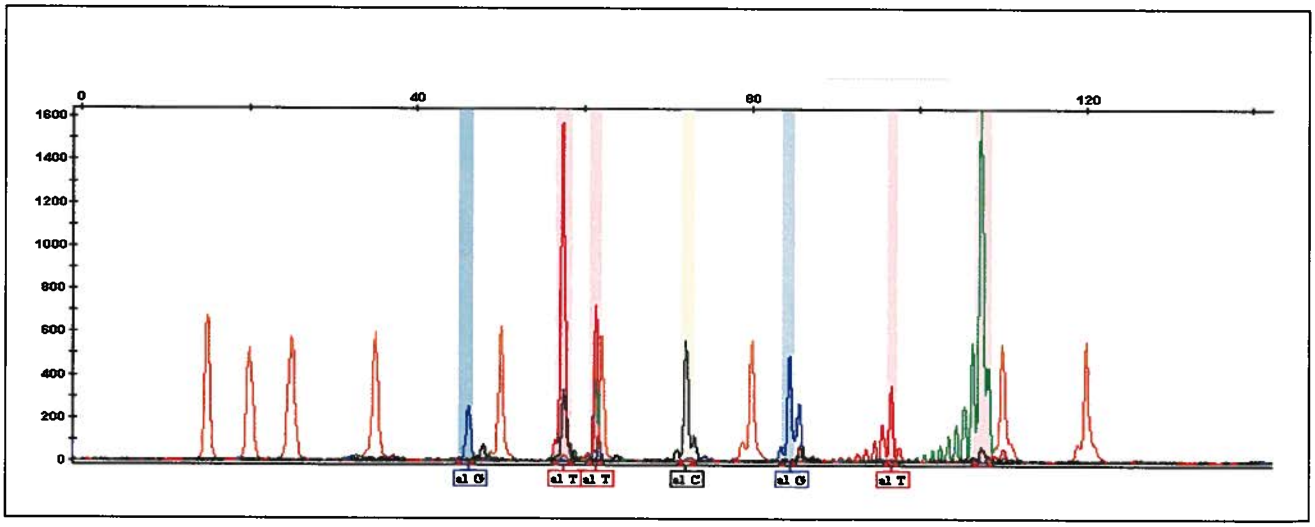


Figure 5.7, B. Scotland isolate, S_3 and S_6, reaction set 5 marker profile comprised of Cp23 locus markers 4,3,1,5,6 and 18S rRNA locus markers 1 and 3 (left to right). Allelic profile for expected *C. hominis* genotype would be G, T, T, C, G for COWP and T, T for 18S rRNA (Table 5.11). Shown is 18S rRNA's presence of a strong A allele at marker 3, far right. G (fluoresce blue), T (fluoresce red), C (fluoresce black), A (fluoresce green). Liz120 size standard (15nt-120, left to right) seen in orange.

5.6 Mixed Genotypes

In recent years the concept of mixed genotypes in *Cryptosporidium* has gained considerable momentum and is becoming increasingly more accepted. In light of this the presence of true mixed alleles was addressed. Multi-locus SNP-types were defined for each sample based on the allele calls (i.e., allele scored) within range of an expected fragment size. To conduct population genetic analysis it is imperative that each molecular marker can be assigned an allele to construct either a Multi-locus genotype (MLG) or an MlSt (Multi-locus SNP-type). The presence of mixed alleles is a common occurrence in any population genetics study on microparasites. In studies of a haploid organism it is essential that each sample can be assigned a single dominate allele in order to assemble a MLG or MlSt for individual isolates. To facilitate this, the assumption that the predominant peak (i.e. allele scored) at each SNP locus represents the actual genotype must be made. It is these alleles that are used to establish the MlSts described above.

The genotypic data suggests a high proportion of mixed SNP-types at numerous SNP positions within the 43 marker assembled SNP panel. Because *Cryptosporidium* is haploid lacking multicopy genes (except for the ribosomal genes); the presence of more than one peak fluorescing at a given fragment size could imply a mixed population which may reflect the transmission intensity within a given population. For the purpose of this study mixed alleles were scored based on peak height ratio. Those having a secondary peak within the expected fragment size range of at least 20% the height of the predominant allele call were scored as mixed. While peak height ratio changed between individual SNP markers, it remained constant for a given marker.

Figure 5.8 Electropherogram representation; mixed alleles scored isolates A_1 for Cp23 markers 3 and 1.

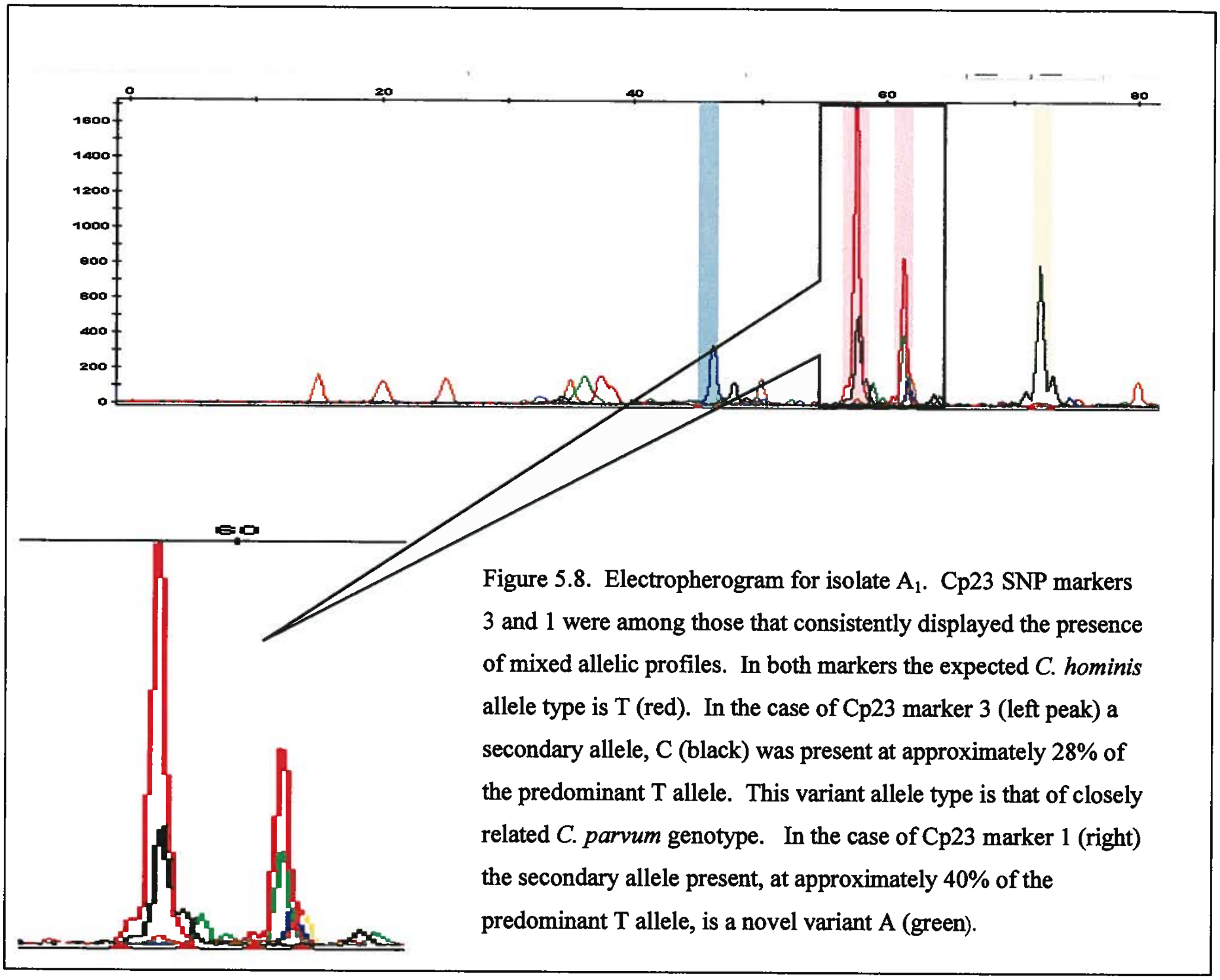


Figure 5.8. Electropherogram for isolate A_1. Cp23 SNP markers 3 and 1 were among those that consistently displayed the presence of mixed allelic profiles. In both markers the expected *C. hominis* allele type is T (red). In the case of Cp23 marker 3 (left peak) a secondary allele, C (black) was present at approximately 28% of the predominant T allele. This variant allele type is that of closely related *C. parvum* genotype. In the case of Cp23 marker 1 (right) the secondary allele present, at approximately 40% of the predominant T allele, is a novel variant A (green).

How these mixed alleles are distributed based on individual SNP marker among the Australia, Kenya, Peru and Scotland populations as a collective is shown in Figure 5.9. Of the total 43 SNP markers in this study, 13 displayed a propensity for mixed alleles. The SNP marker with the most mixed alleles scored is marker 3 in the COWP locus. Sixty-two, 82.7%, of a total of 75 *C. hominis* isolates (Australia, 13; Kenya, 20; Peru, 22; Scotland, 20), had mixed alleles scored at this particular SNP position. This was closely followed by 53, 70.6%, samples having mixed alleles at marker 22 of the HSP70 locus.

Figure 5.9 Mixed allele distribution based on individual SNP marker.

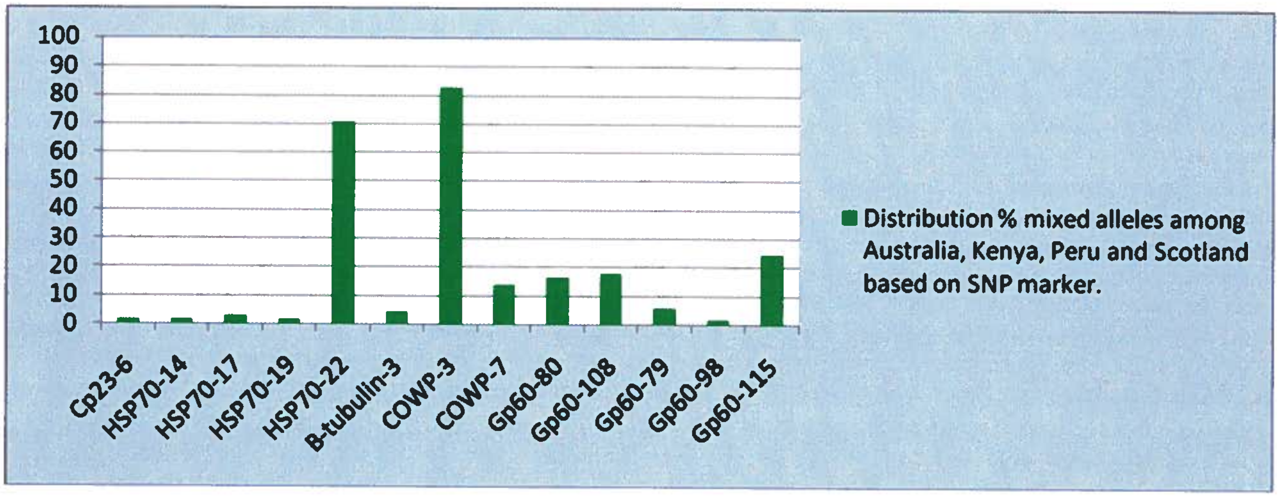


Figure 5.9. The *x* axis represents the 13 SNP markers that were observed as having mixed alleles scored at their position and the total percentage of their distribution throughout all four subpopulations combined.

Fragment analysis of the AcoA and Mucin-1 proteins resulted in a disproportionately high number of bi-allelic or mixed allele calls for almost all SNP markers and isolates tested. Those that could be scored a dominant allele were done so but for most resolving one allele over another was not possible; thus, were left un-scored for multi-locus SNP-typing. Data from these two genes was also omitted from downstream quantitative analysis to prevent a skew based on unreliable predominant allele calls. Despite a tri-peat of the same samples, using both exact as well as varying conditions, the same result ensued.

A higher number of genotypically mixed MlSt's seen in one subpopulation versus another could imply a less stable population structure. A prerequisite for genetic exchange is met by the concept of mixed infections. When considering the distribution of mixed alleles among the four subpopulations of Australia, Kenya, Peru, and Scotland there are no significant differences (Figure 5.10). Scotland comes out on top with 12.67% of all SNPs successfully typed as having mixed alleles though is closely followed by Australia at 11.35%, Peru at 10.18% and Kenya at 8.12%. In the context of our study these figures appear to refute the notion of a less stable genetic subpopulation in one country versus another. As sample sizes increase and more molecular markers are examined this could change.

Figure 5.10 Distribution mixed alleles according to geographic boundary.

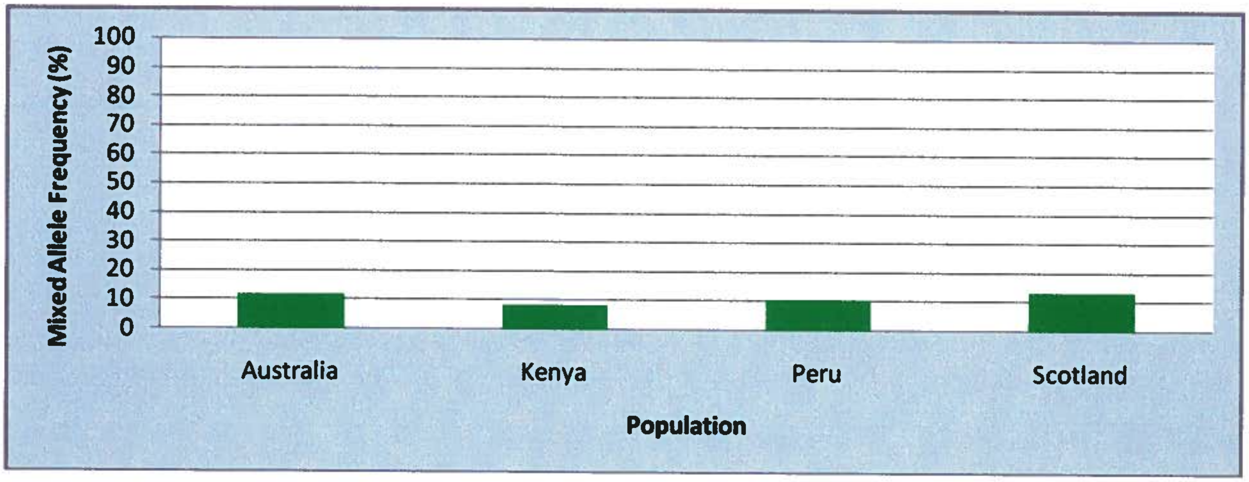


Figure 5.10. Distribution of mixed alleles when partitioned by geography, Scotland showing to have the most (slightly) with the fewest incidences of mixed alleles seen in Kenya.

5.7 Distribution of Multi-locus SNP-types

Multilocus SNP-type was determined based on the allelic profiles for each isolate from all four international geographies. Twenty-four unique MlSt profiles were encountered from 72 of a total 75 *C. hominis* isolates across the 4 international *C. hominis* subpopulations surveyed. MlSt's of the remaining three isolates (isolates $K_{3,5,17}$) were deemed too incomplete to designate with significance. Observed were a small number of highly abundant MlSts and a large number of singletons, consistent with previous data from Scotland[134]. Of 72 isolates scored for a quasi-complete MlSt sixteen (22.2%) belonged to the most abundant MlSt, MlSt1, which is closely followed by MlSt15 with thirteen (18.0%) isolates.

The single most frequent MlSt was found situated within the geographic boundary of Scotland, MlSt, with 9 isolates scored. This MlSt was also identified in one Kenyan and two Peruvian isolates (Figure 5.11) though not in the Australian subpopulation. Of the 24 MlSt's identified, 25% were found to be located within one or more geographies. The most widespread Mlst, Mlst1, occurred within all four geographies and contains novel SNPs.

Each biogeography contained at least one or more unique MlSt having private alleles, alleles detected only within one subpopulation. Kenya had the most followed closely by Peru at 7 and 6 respectively. There were 3 MlSt's unique to Australia though only 1 in Scotland. Each geographically diverse subpopulation displayed a wide variety of MlSt's, whether shared or unique. Within the boundary of Australia there were 6 different MlSt's identified with 11 in Kenya, 11 in Peru, and 6 in Scotland.

Pairwise comparison of all four geographically distinct subpopulations shows that Peru and Scotland have the most MlSts in common, sharing four of the 24 identified. Australia and Scotland, Kenya and Peru and Kenya and Scotland all share 3 common MlSts. Australia only shares 1 and 2 with Kenya and Peru respectively. Results suggest that the repertoires of MlSts circulating amongst all four subpopulations show only a slight overlap with one another. The broad range of MlSt's contained in varying degrees within all four geographies suggests that the intra-population genetic diversity plays a more significant role in population sub-structuring.

Figure 5.11 Geographic distribution of MlSt's identified.

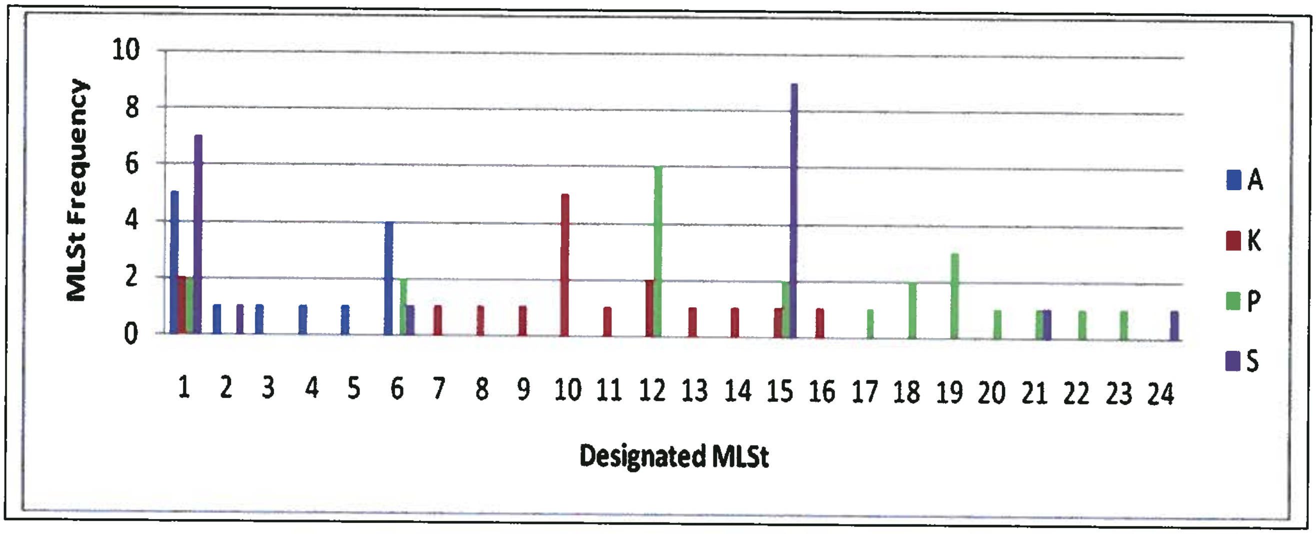


Figure 5.11. The 24 multi-locus SNP-types identified (numbered 1-24) and their distribution based on geography; Australia (A) seen in blue, Kenya (K) in red, Peru (P) in green, and Scotland (S) in purple.

5.8 Descriptive Statistics & Measures of Genetic Variability

For analysis of genetic diversity at a single point mutation level, a locus was considered polymorphic if two or more alleles were detected, regardless of their frequencies. Of a possible 37, Australia had 28 SNP loci that scored more than one allele type. This is the most out of all four geographies suggesting the most intra-population diversity exists in Australia when compared to the other three subpopulations, though we must keep in mind to the allelic profile of isolate A_6. In terms of allele frequency, Scotland was the most stable of the four geographies; only three SNP markers displayed more than one allele type: BT 3, BT 8, and GP60 115. All remaining SNP markers, whether novel, variant or of the expected genotype were genetically stable for this subpopulation. Nine separate marker loci were multi-allelic in the Peru subpopulation and 8 in the Kenyan subpopulation.

Standard genetic diversity parameters were estimated and the mean number of alleles per locus (A), frequency of polymorphic loci (P), observed heterozygosity (H_o) and expected heterozygosity (H_e) (Table 5.12) varied among the 4 subpopulations, with values of A ranging from 1.135 in Scotland to 1.778 in Australia; P from 8.10% in Scotland to 77.8%, in Australia; H_o from 0.097 in Scotland to 0.182 in Australia and H_e from 0.086 in Scotland to 0.165 in Australia. The four international subpopulations contained genetic diversity of varying degrees. With the extremely polymorphic profile of isolate A_6 when compared to the expected *C. hominis* expected allele type the threshold relationship between Australia and the other three international subpopulations may be skewed and if omitted would likely be reduced to levels comparable of the other three subpopulations. Scotland clearly showed the smallest level of variation.

Table 5.12

Diversity Indices for 4 International Subpopulations of *C. hominis*.

Subpopulation	N	# Mlsts	P[a]	PD	A	AP	H_e[b] (±)	H_o (±)
Australia	13	6	0.778	0.462	1.778	2.000	0.165 (0.079)	0.182 (0.049)
Kenya	20	11	0.243	0.550	1.297	2.222	0.071 (0.015)	0.068 (0.045)
Peru	22	11	0.243	0.550	1.297	2.222	0.085 (0.001)	0.109 (0.004)
Scotland	20	6	0.081	0.462	1.135	1.667	0.086 (0.000)	0.097 (0.016)
Mean		8.5	0.336	0.506	1.376	2.278	0.086	0.113

Table 5.12. [a]A locus is considered polymorphic if the frequency of the most common allele does not exceed 0.99. [b]Unbiased estimate (Nei 1978). N, sample size; P, frequency of polymorphic loci; A, mean number of alleles per locus; AP, mean number of alleles per polymorphic locus; H_e, expected heterozygosity, H_o, observed heterozygosity. PD is proportion of distinguishable MlSts. (±) Standard deviation.

Clonal diversity, measured as the proportion of distinguishable genotypes (PD), was relatively uniform across all four subpopulations with a value of only 0.088. Private alleles (Table 5.13) at low frequency were encountered in all four subpopulations though under different circumstances. With the high degree of variability isolated to A_6 in 28 SNP loci of 37 it can be considered to have a high proportion of private alleles, though all are of the *C. parvum* genotype. Excluding this isolate there are 6 private alleles seen in isolate A_7. Though each variant allele is of the *C. parvum* genotype they are contained to the LDH, MDH, and UPRT loci. There was genetic stability in the Kenya, Peru, and Scotland subpopulations for these three specific genes. Kenya had two discrete alleles reserved to two different SNP loci. First was a cysteine residue at marker BT 7 in K_{11}which was not only novel but no other subpopulation had any genetic diversity at this position (excluding A_6). Secondly while not variant isolate K_2 and K_{17} had what may be considered a private allele at loci 18S rRNA 3. This is the only subpopulation to show the expected *C. hominis* allele type at this particular marker, all other isolates and subpopulations had the novel Adenine allele variant. Allelic variances reserved to the Kenyan subpopulation were also seen at the Gp60 locus marker 79 but this particular SNP was not successfully typed in any of the other three subpopulations therefore this particular finding must be taken with some reserve. In three different isolates Peru had the same private allele at the COWP 6 locus, which was distinct from either the *C. hominis* or *C. parvum* genotype. Again in the Peruvian subpopulation there was the novel SNP type, guanine, at SNP marker COWP 3.

In contrast to variant alleles there are a number of SNP markers that demonstrate genetic consistency across all four international subpopulations; two from the β-tubulin locus (markers 4, 5), one from the COWP locus (COWP 7) and both EMAAg markers (EMAAg 29, 27). If the A_6 hyper-variable isolate is removed there are 12 additional genetically stable SNP loci and if A_7 is excluded there are a further 6 SNP loci with genetic constancy.

Based on the typing results gene diversity per locus and subpopulation were computed with zero indicating a complete lack of differentiation or a total presence of genetic stability. Table 5.14 describes the degree or level of gene diversity per locus and subpopulation. Increasing values toward a threshold of one indicate greater genetic variation at that particular SNP position within individual subpopulations. Scotland had the greatest number of genetically stable SNP-markers at 32 followed by Kenya and Peru at 58 and 27 respectively. Australia was only genetically stable at 8 of a possible 37 SNP markers. Four SNP markers were shown to have no genetic diversity across all four subpopulations; BT4, 18S rRNA1, EMA29 and EMA27. All other SNP markers were variable at least once in one or more subpopulations. SNP marker 18S rRNA 3 does show a lack of differentiation among the four subpopulations of Australia, Kenya, Peru, and Scotland in terms of the allele scored but as it is for a novel allele variant. If it is indeed a true novel allele it is one that is stable among globally diverse parasite subpopulations. SNP markers Gp60 98 and 115 showed the greatest amount of diversity in all four subpopulations, at least 2 or 3 different alleles were scored to this SNP position.

Table 5.13

Allele Frequencies at 31Polymorphic SNP Loci of International Populations of *C. hominis*

SNP marker	Australia	Kenya	Peru	Scotland
BT1-A_2	0.924			
BT1-A_3	0.076			
BT3-A_1			0.227	0.105
BT3-A_3	0.846	0.846	0.773	0.842
BT3-A_4	0.154	0.154		0.053
BT7-A_1	0.076			
BT7-A_2		0.067		
BT7-A_4	0.924	0.933		
BT8-A_2	0.924			0.944
BT8-A_3	0.076			0.056
COWP5-A_1	0.924			
COWP5-A_1	0.076			
COWP6-A_1			0.150	
COWP6-A_2	0.076			
COWP6-A_3	0.924		0.850	
COWP1-A_2	0.924			
COWP1-A_3	0.076			
COWP3-A_2	0.076		0.950	
COWP3-A_3	0.924			
COWP3-A_4			0.050	
Cp(23)4-A_1	0.076			
Cp(23)4-A_4	0.924			
Cp(23)3-A_2	0.076			
Cp(23)3-A_3	0.924			
Cp(23)1-A_2	0.076			
Cp(23)1-A_3	0.924			
Cp(23)5-A_2	0.924			
Cp(23)5-A_4	0.076			
Cp(23)6-A_3	0.076			
Cp(23)6-A_4	0.924			
18S rRNA3-A_1	1.000	0.895	1.000	1.000
18S rRNA3-A_3		0.105		
HSP(70) 14-A_1	0.924			
HSP(70) 14-A_4	0.076			
HSP(70) 17-A_1	0.924	0.947		
HSP(70) 17-A_4	0.076	0.053		

Table 5.13

Allele Frequencies at 31 Polymorphic SNP Loci of International Populations of *C. hominis*

SNP marker	Australia	Kenya	Peru	Scotland
HSP(70) 19-A_1	0.924			
HSP(70) 19-A_4	0.076			
HSP(70) 20-A_1	0.924			
HSP(70) 2-A_4	0.076			
HSP(70) 22-A_1	0.924			
HSP(70) 22-A_3	0.076			
Gp(60) 80-A_1	0.924	0.824	0.444	1.000
Gp(60) 80-A_3	0.076	0.176	0.556	
Gp(60) 108-A_2	0.924	0.883	0.500	1.000
Gp(60) 108-A_3	0.076	0.117	0.500	
Gp(60) 126-A_2			0.045	
Gp(60) 126-A_3			0.955	
Gp(60) 79-A_2		0.111		
Gp(60) 79-A_3		0.556		
Gp(60) 79-A_4		0.333		
Gp(60) 98-A_1	0.846	0.187	0.500	1.000
Gp(60) 98-A_2		0.063	0.111	
Gp(60) 98-A_4	0.154	0.750	0.389	
Gp(60) 115-A_1		0.053	0.136	0.650
Gp(60) 115-A_2	0.583	0.737	0.454	0.300
Gp(60) 115-A_4	0.417	0.210	0.410	0.050
LDH10-A_2	0.750			
LDH10-A_3	0.250			
LDH3-A_2	0.858			
LDH3-A_3	0.142			
MDH8-A_1	0.250			
MDH8-A_4	0.750			
MDH7-A_2	0.286			
MDH7-A_4	0.714			
UPRT2-A_1	0.250			
UPRT2-A_3	0.750			
UPRT3-A_2	0.100			
UPRT3-A_3	0.900			

Table 5.13. Allelic designation; A_1, Adenine; A_2, Cysteine; A_3, Thymine; A_4, Guanine.

5.9 Genetic Data Analysis

The ultimate goal in quantifying population genetic structure is to understand variation among species and to determine whether there are patterns within or among different populations of organisms and biogeography. Determining how genetic variation is distributed within versus among populations provides insight into genetic population structure, gene flow, historic population parameters and hints of speciation. Genetic differentiation is essentially defined as the level of differences in inter-population allele frequencies, the differences among populations. Intra-population, or within population variation is essentially a measure of heterozygosity for a given population. The two are not mutually exclusive as both can be influenced by shifts in one or another.

5.9.1 Distribution of Genetic Variation

Diversity measures were calculated by Nei's (1973) index and ranged from H=0 to H=0.642 (Table 5.14). Averaged over all markers Scotland was found to be the least diverse. Australia showed the highest level of diversity while Kenya and Peru revealed intermediate diversity values (Table 5.14). The low diversity seen in Scotland is largely a result of the fact that although six different SNP markers were polymorphic when compared to the expected *C. hominis* SNP-type three of these were genetically stable in that polymorphism; 18S rRNA marker 3, and Gp60 markers 80 and 108. The allelic profile of that polymorphism was consistent throughout all isolates typed. The other three SNP markers had bi-allelic profiles. Australia, the most diverse subpopulation, was polymorphic at almost all SNP loci except for BT4, COWP7, Gp60 126, and EMA 27 and 29. This is largely result of the A_6 and A_7 typing results and upon their omission the values would likely be more akin to that of Kenya and Peru. In respect of this both isolates were confirmed with the originating laboratory regarding species distinction, typing results were repeated in triplicate and standard gene sequencing methods within our lab were done to confirm their *C. hominis* status to be true.

Table 5.14

Gene diversity within Intercontinental *C. hominis* Populations estimated by Nei's (1973) Diversity Measure for 37 SNP Markers.

SNP locus	Australia	Kenya	Peru	Scotland
BT1	0.154	0	0	0
BT4	0	0	0	0
BT3	0.282	0.282	0.368	0.199
BT5	NA	0	0.00	0
BT7	0.154	0.133	0.000	0.000
BT8	0.154	0.000	0.000	0.111
COWP5	0.154	0.000	0.000	0.000
COWP6	0.167	0.000	0.209	0.000
COWP1	0.154	0.000	0.000	0.000
COWP3	0.154	0.000	0.105	0.000
COWP7	0.000	0.000	0.100	0.000
23Cp4	0.154	0.000	0.000	0.000
23Cp3	0.154	0.000	0.000	0.000
23Cp1	0.154	0.000	0.000	0.000
23Cp5	0.154	0.000	0.000	0.000
23Cp6	0.154	0.000	0.000	0.000
18S rRNA1	0.000	0.000	0.000	0.000
18S rRNA3	0.000	0.282	0.000	0.000
HSP14	0.154	0.000	0.000	0.000
HSP17	0.154	0.105	0.105	0.000
HSP19	0.154	0.000	0.000	0.000
HSP20	0.154	0.000	0.000	0.000
HSP22	0.154	0.000	0.000	0.000
60Gp80	0.167	0.343	0.515	0.000
60Gp108	0.182	0.385	0.529	0.000
60Gp126	0.000	0.000	0.091	0.000
60Gp79	0.000	0.500	0.000	0.000
60Gp98	0.389	0.425	0.642	0.000
60Gp115	0.53	0.433	0.636	0.511
LDH10	0.429	0.000	0.000	0.000
LDH3	0.286	0.000	0.000	0.000
MDH8	0.429	0.000	0.000	0.000
MDH7	0.476	0.000	0.000	0.000
EMAg29	0.000	0.000	0.000	0.000
EMAg27	0.000	0.000	0.000	0.000
UPRTase	0.429	0.000	0.000	0.000
UPRTase	0.200	0.000	0.000	0.000
Means	**0.177**	**0.078**	**0.089**	**0.022**

Table 5.14. Gene diversity scaled from 0 – 1, with a zero value telling of a total lack of diversity and 1 representing complete diversity.

Genetic diversity is a measurement of differentiation among closely related taxa by using a mathematical measure to understand the degree of genetic separation between species at the molecular level. The long-standing method of deciphering the amount of genetic differentiation that exists has been the use of F-statistics, as described by Wright (1943, 1951, and 1965). Wright's F-statistics (Fst) provides an integrated view of genetic variation at three hierarchal levels of population structure: within subpopulations, among subpopulations and the total variation in the metapopulation. All of the measures made under the guidelines of F-statistics are based on losses of heterozygosity and work to partition heterozygote deficiency into a within and among population component. Fst is a measurement value of the amount of genetic variation in the total samples that is due to differences among populations comprising that sample. This proportion can range from zero indicating genetically identical populations to one, indicating completely isolated populations. There are some constraints on Wrights original fixation indices resulting in several analogs being introduced to help circumvent such limitations. Nei's analog (1972, 1973) averaged Fst over alleles and pairs of populations and enabled its application to any population without many key assumptions. It operates under a disregard to patterns of evolutionary forces, sexual or asexual reproduction and ploidy as long as allele frequencies can be estimated. Nei's functional equivalent, Gst, is essentially the ratio of inter-subpopulational gene diversity (Dst) to the total gene diversity (Ht).

Nei's algorithms using the statistics Hs, Ht, Dst, Gst were estimated for each locus and overall based on SNP-typing results; Hs represents the within sample gene diversity, Ht the overall gene diversity, and Dst is the amount of gene diversity among samples (I.e. The average of genetic diversity among populations). The quantity Dst has been refined and is independent of the number of samples and used. Gst is an estimator of the proportion of total gene diversity partitioned among populations and again is independent of the number of samples used. While it is often cited or argued that Gst cannot be negative later more refined versions now allow for this[235].

Table 5.15

Apportionment of Genetic Diversity into Within and Between Intercontinental *C. hominis* Populations

Marker	Hs	Ht	Dst	Gst
BT1	0.038	0.038	0.001	0.018
BT4	0.000	0.000	0.000	0.000
BT3	0.283	0.286	0.004	0.015
BT5	0.000	0.667	1.000	1.000
BT7	0.071	0.071	0.000	0.000
BT8	0.066	0.065	0.001	-0.023
COWP5	0.038	0.038	0.000	0.013
COWP6	0.094	0.095	0.002	0.020
COWP1	0.038	0.038	0.001	0.019
COWP3	0.065	0.064	-0.000	-0.007
COWP7	0.025	0.025	-0.000	-0.009
23Cp4	0.038	0.038	0.001	0.015
23Cp3	0.038	0.038	0.001	0.015
23Cp1	0.038	0.038	0.001	0.015
23Cp5	0.038	0.038	0.001	0.015
23Cp6	0.038	0.038	0.001	0.015
18S rRNA1	0.000	0.000	0.000	0.000
18S rRNA3	0.000	0.003	0.001	0.002
HSP14	0.038	0.038	0.001	0.019
HSP17	0.064	0.064	-0.001	-0.012
HSP19	0.038	0.038	0.001	0.019
HSP20	0.038	0.038	0.001	0.019
HSP22	0.038	0.038	0.001	0.019
60Gp80	0.343	0.42	0.116	0.253
60Gp10	0.274	0.331	0.076	0.218
60Gp126	0.023	0.023	-0.000	-0.011
60Gp79	0.195	0.174	-0.028	-0.171
60Gp98	0.363	0.519	0.209	0.365
60Gp115	0.500	0.621	0.124	0.191
LDH10	0.105	0.120	0.020	0.162
LDH3	0.069	0.071	0.003	0.037
MDH8	0.105	0.120	0.020	0.162
MDH7	0.112	0.135	0.031	0.214
EMAg29	0.000	0.000	0.000	0.000
EMAg27	0.000	0.000	0.000	0.000
UPRTase	0.104	0.120	0.021	0.170
UPRTase	0.048	0.050	0.002	0.032
Overall	**0.092**	**0.122**	**0.040**	**0.304**

Table 5.15. Gene diversities calculated; average diversity within population (Hs), total diversity (Ht), mean level of genetic differentiation (Gst), and average among population diversity (Dst).

The average diversity within subpopulation (Hs) was 0.092 and the total diversity (Ht) amounted to 0.122 (Table 5.15). The mean level of genetic differentiation (Gst), diversity between subpopulations overall loci was 0.304. This indicates that almost a third or 30.4% proportion of the total genetic variation existed among subpopulations, compared to diversity within subpopulations at 69.6%.

According to Wright Fst values for most organisms is typically 0.15 or less, though values upward of 0.7 have been recorded. Using the established and accepted but arbitrary guidelines for Fst values of Wright's statistics, which also apply to subsequent analogs, values of Fst or Gst greater than 0.25 are considered significant[47, 235]. Each SNP marker/loci contributed differently to the observed degree of subpopulation differentiation, varying from a low of 0 for BT4, BT7, 18S rRNA, and both EMAAg markers to a high of 36.5% for Gp60 marker 98. The average within subpopulation diversity (Hs) is greater than the average among population genetic diversity (Dst) at 0.092 and 0.040 respectively. This is indicative of intra-population diversity being more imposing than inter-population diversity.

The Weir and Cockerham (1984) method (θst) is another estimator of Fst/Gst. The main difference between the two methods is that Nei's approach weights all samples equally regardless of sample size whereas Weir & Cockerham weight samples according to sample size. Having a range of sample sizes from 13 to 22 we calculated the heterozygote deficit using the Weir & Cockerham (1984) parameters for Fstatistics which weight allele frequencies according to sample size. The results, Gst versus θst, were in agreement 0.304 and 0.319 respectively.

5.9.2 Genetic Identity Measures

Using the genetic data analysis program genetic identity values and distances were calculated from the Nei's (1978) gene diversity index between each of the subpopulations. Genetic identity values measure the degree of closeness based on allele frequencies between pairs of populations and range from 0, indicating no shared alleles between populations, to 1, indicating that the two populations have the same alleles in identical frequencies. Nei's unbiased genetic identities were computed to alleviate any bias caused by small sample size, for example, fewer than 50 individuals. Genetic identity values (Table 5.16) ranged from 0.942 between Kenya and Scotland to 0.984 Australia and Scotland. From the minima to the maxima the difference amounts to 4.2%. The mean identity between all pairwise comparisons is 0.963; on average there is a genetic identity of 96.3% between all four global subpopulations. Genetic distances averaged 0.048 and varied from 0.034 between Scotland and Australia and 0.061 between

Scotland and Kenya. A dendrogram (Figure 5.12) constructed on the basis of Nei's genetic distance was done using the Neighbour Joining method (NJM).

Table 5.16

Matrix of Nei's Unbiased Genetic Identity/Distance Measures Based on 37 Loci among 4 Global Subpopulations[+]

	Australia	Kenya	Peru	Scotland
Australia	-	0.976	0.977	0.984
Kenya	0.044	-	0.955	0.942
Peru	0.045	0.047	-	0.945
Scotland	0.034	0.061	0.058	-

Table 5.16. [+]Nei (1978) identity above the diagonal, Nei (1972) distance below the diagonal.

The two subpopulations at the minima of genetic distance and clad to one another are two of the farthest apart based on physical geographic distance (Table 5.17). In contrast, the two subpopulations closest in terms of geography showed to have the greatest genetic distance relationship.

Table 5.17

Approximate Distances[+] Between Intercontinental Populations (km)

	$\mathrm{A}_{\mathrm{Perth}}$	$\mathrm{K}_{\mathrm{Nairobi}}$	$\mathrm{P}_{\mathrm{Lima}}$
$\mathrm{A}_{\mathrm{Perth}}$			
$\mathrm{K}_{\mathrm{Nairobi}}$	8 896.15		
$\mathrm{P}_{\mathrm{Lima}}$	14 934.84	12 565.73	
$\mathrm{S}_{\mathrm{Glasgow}}$	14 740.90	7 352.15	10 072.83

Table 5.17. [+]Distances are only estimates and calculated according to location of originating laboratory of samples, which may or may not be the exact location of an isolates origin (data unknown).

Figure 5.12 Neighbour-joining phylogenetic analyses of 4 intercontinental subpopulations of *C. hominis*; Australia, Kenya, Peru, and Scotland, based on genetic distance.

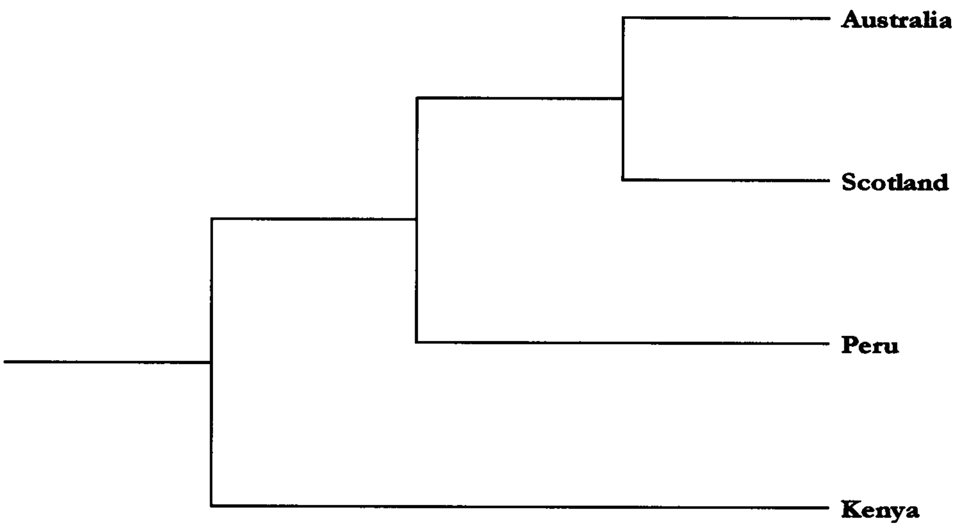


Figure 5.12. Dendrogram provides a visual account of how closely related one species is to another. The more alleles in common, the closer they are related. Dendrogram representation is on the basis that the shorter the distance the greater the number of shared alleles in contrast to the longer the distance representing the fewer number of shared allele.

5.10 Canada: Multi-locus SNP-typing

SNP-typing results for Canadian isolates are more scattered than those previously seen in Australia, Kenya, Peru and Scotland. Though the sample size is the largest (N=31) only 22 of a possible 45 SNP markers were typed with confidence. The reasons for this are three-fold. First, samples were more readily available and dispensable since it is the home location for the study therefore Canadian isolates were used to test the design of our experimental platform. Samples received from other countries were limited in supply and reserved until the methodology was reliably confirmed. Second, because of the supply of both *C. hominis* and *C. parvum* isolates in our lab we were able to put more effort into testing both species. Aside from the two *C. parvum* isolates received from Australia all isolates donated in kind were of the *C. hominis* genotype. Thirdly, both laboratory and financial resources compounded by time limitations prevented us from currently going back to further test Canadian isolates more in depth for the remaining 23 SNP markers. In light of these factors Canadian isolates are unlikely to make significant contributions to genetic population data analysis so were omitted from genetic diversity indices.

The importance of this subpopulation can be seen in other aspects of the study. First, as mentioned we had the freedom to test, manipulate and finesse our experimental approach before moving onto the more indispensable samples of our global populations. Next, the use of reliable but even more importantly confirmed and documented samples of the *C. hominis* and *C. parvum* genotype enabled us to address the question of whether MIS-typing would be a dependable and efficient tool for species distinction. Lastly, while limited there is molecular marker data available for the APR locus for the first time. This indicates to us that our primers were well designed, a welcomed indication considering the time and financial costs of SBE primer construction.

SNP-typing results for those samples and markers tested are shown below in Table 5.18. Isolates BC_1 through BC_{10}, and BC_{25} are confirmed *C. parvum* genotypes. BC_1 demonstrated a complete *C. parvum* MISt profile with the exception of two particular markers. First SNP marker 4 located in the β-tubulin locus reveals a *C. hominis* allele call, recall SNP marker 4 is located within the coding region of the β-tubulin locus. No variant alleles are seen at this particular SNP position in any of the other subpopulations studied. The second allele variant is located at SNP position 126 in the Gp60 locus, the same allele consistently seen in the other four subpopulations and again in multiple BC isolates (Table 5.18). Another confirmed *C. parvum* genotype, sample BC_5, showed two variant allele calls of the *C. hominis* genotype at markers 1 and 4 in the β-tubulin locus. With the exception of A_6, this is the only other example of a variant allele within the intron region of the β-tubulin locus. Samples $BC_{11-24, 26-31}$ are confirmed *C. hominis* isolates. A collection of allele variants, novel or of the *C. parvum* genotype, are seen at various markers throughout these isolates typed at the Gp60 gene, which are also present in the other four subpopulations. In contrast results for BC_{14} show the same allele variant (C) is as that seen in only one other population and isolate, A_6, at marker 6 in the COWP locus. In the APR locus 3 *C. parvum* alleles are seen in three confirmed *C. hominis* genotypes, $BC_{14, 20, 21}$. In the case of BC_{14} this is the second case of a *C. parvum* allele variant in its MISt, the first being COWP marker 6.

Table 5.18

Canada Subpopulation MlSt & Allele Variants

	B-tubulin		COWP		Cp23		Gp60						LDH		MDH		EMAAg		UPRT		APR	
SNP	1	4	5	6	3	1	80	108	126	79	98	115	10	3	8	7	29	27	2	3	2	3
Ch	C	A	C	T	T	T	T	T	T	T	G	G	C	C	G	G	G	G	T	T	T	T
Cp	T	G	T	C	C	C	A	C	C	C	T	A	T	T	A	C	A	C	A	C	C	C
BC1	T	A	T	C	C	C	A	C	T	-	T	-	T	-	A	C	-	-	A	C	C	G
BC2	T	G	T	C	C	-	A	C	T	-	-	C	-	-	-	-	-	-	-	-	-	-
BC3	T	G	T	C	C	C	A	C	T	-	-	C	-	-	-	-	-	-	-	-	-	-
BC4	-	-	-	-	-	-	A	C	T	-	-	C	-	-	-	-	-	-	-	-	-	-
BC5	C	A	T	C	C	C	-	C	T	-	-	C	-	-	-	-	-	-	-	-	-	-
BC6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
BC7	-	-	-	-	-	-	A	C	T	-	T	C	-	-	-	-	-	-	-	-	-	-
BC8	-	-	-	-	-	-	-	-	T	-	-	C	-	-	-	-	-	-	-	-	-	-
BC9	-	-	-	-	-	-	-	-	T	-	-	C	-	-	-	-	-	-	-	-	-	-
BC10	-	-	-	-	-	-	-	-	T	-	-	C	-	-	-	-	-	-	-	-	-	-
BC11	C	A	C	T	T	T	-	C	T	-	A	C	-	-	-	-	-	-	-	-	-	-
BC12	C	A	C	T	T	T	-	-	-	-	-	-	-	-	G	G	G	G	-	-	-	-
BC13	C	A	C	T	T	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
BC14	C	A	C	C	T	T	-	-	-	-	-	-	-	-	G	G	-	-	T	T	C	-
BC15	C	A	C	T	T	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
BC16	C	A	C	T	T	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
BC17	-		C	T	-	-	-	C	T	T	A	C	-	-	-	-	-	-	-	-	-	-
BC18	-		C	T	-	-	-	-	T	-	-	C	-	-	-	-	-	-	-	-	-	-
BC19	-	-	-	-	-	-	-	-	-	-	-	-	-	-	G		G	G	-	-	-	-
BC20	-	-	-	-	-	-	-	-	-	-	-	-	C	-	G	G	G	G	T	T	C	-
BC21	-	-	-	-	-	-	-	-	-	-	-	-	C	-	G	G	G	G	T	T	C	-
BC22	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T
BC23	-	-	-	-	-	-	-	C	T	-	A	G	-	-	-	-	-	-	-	-	-	-
BC24	-	-	-	-	-	-	-	C	T	-	A	C	-	-	-	-	-	-	-	-	-	-
BC25	-	-	-	-	-	-	A	C	C	G	T	A	-	-	-	-	-	-	-	-	-	-
BC26	-	-	-	-	-	-	-	C	T	-	A	G	-	-	-	-	-	-	-	-	-	-
BC27	-	-	-	-	-	-	-	-	T	G	C	C	-	-	-	-	-	-	-	-	-	-
BC28	-	-	-	-	-	-	-	-	T	-	-	C	-	-	-	-	-	-	-	-	-	-
BC29	-	-	-	-	-	-	-	C	T	-	A	C	-	-	-	-	-	-	-	-	-	-
BC30	-	-	-	-	-	-	-	C	T	-	C	C	-	-	-	-	-	-	-	-	-	-
BC31	-	-	-	-	-	-	-	C	T	-	-	C	-	-	-	-	-	-	-	-	-	-

Table 5.18. Ch; *C. hominis*. Cp; *C. parvum*. (-) denotes markers unsuccessfully typed or scored. Allele designations, A, C, T, G, represent alleles scored at that particular SNP loci.

Figure 5.13 Canada electropherogram representations; BC1, of COWP, Cp23 and β-tubulin loci molecular markers.

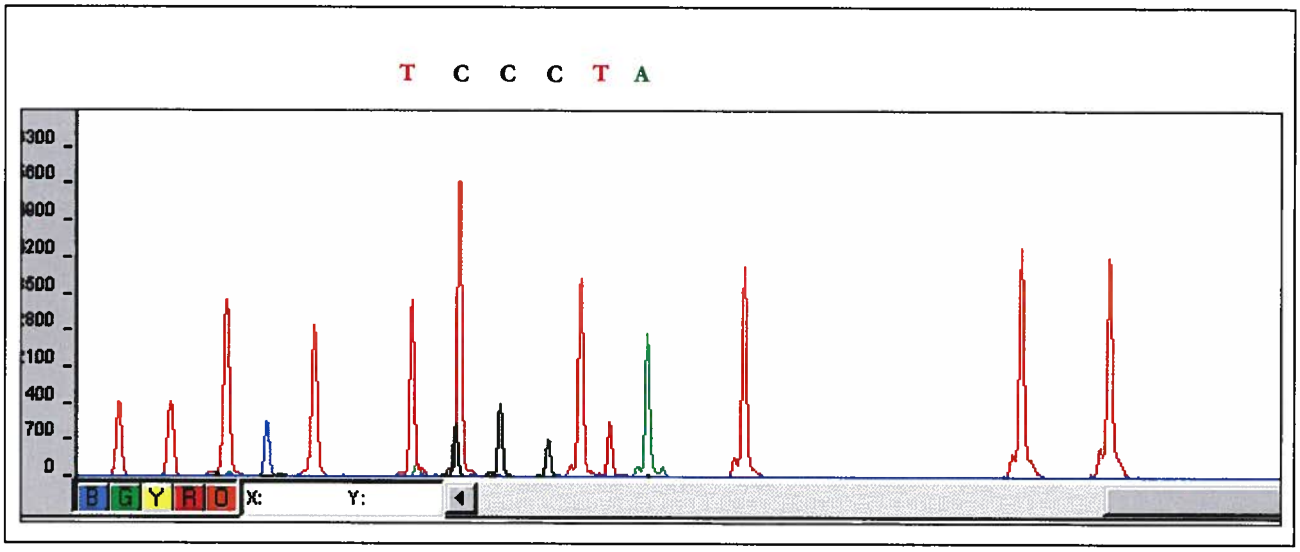


Figure 5.13. Electropherogram of Western Canada isolate, BC1, a confirmed *C. parvum* isolate, including from left to right molecular markers 5 (T, red) and 6 (C, black) for the COWP locus, molecular markers 3 (C, black) and 1 (C, black) for Cp23 and molecular marker 1 (T, red) for β-tubulin. Expected at SNP marker β-tubulin 4 for *C. parvum* is an allele type G, (blue) however this *C. parvum* isolate indicates the presence of a *C. hominis* allele (A) as the fluorescing green peak denotes (far right). Liz120 size standard (15nt-120nt, left to right) seen in orange.

5.11 Species Distinction

Identifying relationships between organisms involves grouping them according to a defined set of characteristics. In epidemiology studies just as crucial is the ability to exclude one organism from another for diagnostic purposes and the tracking of transmission routes and infection sources. Typing systems based on genomic material are designed to compare differences at the nucleotide level either in designated regions of the genome (microdiversity) or the entire genome itself (macrodiversity). The most optimal comparisons are those done by the typing and analysis of the entire genome sequence of every strain or isolate. Whole genome DNA sequencing is still not a widely accessible or affordable option for many countries and laboratories. In lieu of this high-throughput SNP-typing is an attractive alternative.

With the ability to multi-locus SNP-type Canadian isolates of both species using our methodology we were able to reliably examine whether or not MlS-typing can be used as a species distinction tool. The results of isolates tested, when compounded by the results from the international subpopulations support this finding. The high throughput, time and cost efficient protocol of our experimental platform makes our method a very attractive alternative to standard genotyping practices.

MIS-typing results of the Australian, Kenyan, Peruvian, and Scottish sample populations allow us to make further inferences about MIS-typing as an acceptable methodology for species identification. The genetic stability among four globally diverse populations of certain proteins and SNP markers implicates those that may have the best potential for species classification. The complete absence or almost complete absence of allele variants at SNP position 1 in 18S rRNA and all 5 SNP positions in HSP70 indicates they would be dependable markers for species differentiation. The stability of the enzymatic proteins LDH, MDH, and UPRT also suggests their usefulness as markers for species differentiation. As enzymatic proteins essential to *Cryptosporidium's* biosynthesis processes and therefore survival these proteins make excellent targets for phylogenetic studies as well as potential downstream applications focused on the development of chemotherapeutics. The immunodominant/antigenic yet genetically stable proteins Cp23 and EMAAG make for a sixth and seventh gene that should be explored for rapid genotyping via MIS-typing.

CHAPTER 6

DISCUSSION

Summary – In this study a battery of *C. hominis* isolates collected from 4 intercontinental regions was genetically typed based on a mutation or SNP profile, allowing for inferences on the global population structure of the parasite to be made. The aim of the study at large was to ascertain whether or not *C. hominis* populations are partitioned based on geography. The within-population component of genetic variation exceeds the average proportion of genetic differences among populations of *C. hominis*. Thus far our data appears to do little to indicate population substructuring. Previous studies have argued for more globally diverse biogeographic investigations into genetic variation of the *Cryptosporidium* genome. Our results argue that a too wide of a geographic boundary can impede rather than advance such population studies. Furthermore the high throughput, time and cost efficient protocol of our experimental platform makes our method a very attractive alternative to standard genotyping practices. Such an approach could ultimately help bridge the gap between *Cryptosporidium* detection versus specific genotype or strain identification.

6.1 Comparative & Computational Whole Genome Analysis

Optimally comparisons of similarity or diversity between organisms are best done using entire genomes, which to date are available for a limited number of organisms. The ability to sequence entire genomes of pathogens has engendered a new discipline termed comparative genomics. Though most often used in phylogenetic studies its potential for application in epidemiology studies is becoming more evident as new assumptions can be made about the nucleic acid sequences used to type and classify such pathogens. In *Cryptosporidium* research efforts have intensified in this respect but there are still insufficient data available from which to draw robust genomic comparisons.

In the post-genomic era attention has shifted to comparative genomics focused on the differences between genomes. Identifying genetic relationships between populations involves grouping organisms according to a defined set of characteristics. *Cryptosporidium* species are phenotypically very similar rendering it difficult to distinguish species based on morphology; molecular markers are therefore in demand. At the lowest genetic level are single point mutations or SNPs affecting individual nucleotides. Compared with other molecular markers, single-nucleotide polymorphisms exhibit extremely low mutation rates, making them rarer in recently emerged pathogens. A prerequisite for this study was mapping SNPs throughout the *Cryptosporidium* genome which involved comparing reference strains TU502 for *C. hominis* and Iowa II for *C. parvum*. The construction of the genomic and SNP library described yielded hundreds of initial targets for whole genome SNP-typing of *Cryptosporidium*. Thirteen genes were targeted from this group and used to generate a multi-locus SNP-type, representing a set of SNPs at 45 individual loci, which was subsequently used to partition the genetic relationships among and within globally distinct *C. hominis* subpopulations.

Multiple criteria were used to evaluate genes and the SNPs within them for the study. The initial focus was on 2 major demes of gene type. First were those hypothesized to be under positive or diversifying selection pressures, most often being antigenic determinant genes. Secondly we looked for putative genes thought to be bio-functionally relevant to the success of *Cryptosporidium*. Comparison of the ~9Mb *Cryptosporidium* genome sequences led to 13 targeted open reading frames. Alignment of target genes using the reference genomes for *C. hominis* and *C. parvum* was done to identify polymorphic sites at the nucleotide level. Results showed the presence of 394 single point mutations. On the basis of the protein sequences inferred each polymorphic site was differentiated as synonymous or silent versus non-synonymous, resulting in a single amino acid polymorphism.

While those conferring an amino acid change to the primary protein sequence are more likely to be clinically relevant genetic markers, synonymous mutations were used as molecular markers for phylogeography implications. Mutations that result in an amino acid substitution provide a substrate for evolutionary selection and have a greater chance of having a profound effect on the protein's function than silent ones do. They may be harmful with a greater chance of causing a deleterious effect on the function of the protein and as a result most species evolve to eliminate them from the population through selection processes. In contrast they may improve protein function and advantageous selection plays a major role. Because synonymous mutations have greater potential of being neutral a larger proportion of them will become fixed in the population making them excellent molecular targets for deciphering the genetic structure of parasite populations.

Mutations may affect organisms in multiple ways. A complete range of altered phenotypes from mild to phenotypically silent effects to minor advantageous traits to detrimental effects can be seen[36, 160]. If the global populations of *Cryptosporidium* do in fact share recent ancestry the account for the occurrence and ratio of silent versus nonsilent polymorphisms needs to be addressed. Most organisms, including many bacteria, viruses and eukaryotes carry it in abundance. Only rarely does replacement variation exceed synonymous variation. It can arise by strong diversifying selection, an event that might be highly anticipated in antigenic or virulent determinants of a pathogen. The most straightforward approach to determine this is to examine the ratio of non-synonymous versus synonymous mutations.

From our initial comparative genome analysis the overall amount of synonymous substitution was 70% versus 30% for non-synonymous from the total 394 SNPs mapped, proposing a higher level of conservation of protein sequences in *Cryptosporidium*. From this SNP data set there were almost half as many NS SNPs as there were S SNPs with a ratio of 0.44 found at the 13 target gene loci. This would result in a higher level of conservation of protein sequences compared to that found in the Apicomplexans *P. falciparum* (2.34 ratio of NS SNPs / S SNPs), *P. vivax* (1.75 ratio of NS SNPs / S SNPs), and even the human populations (0.89 ratio of NS SNPs / S SNPs)[221]. This implies a predilection for functional constraint, as was recently shown for *G. lamblia*[221]. The genetic stability of the *Cryptosporidium* genome is clear; thus, underlining the importance of examining the subtle genetic diversity that does exist to better understand a specific species' host range and transmission dynamics.

When looking at each gene alone the observations were diverse in regards to the concepts of natural selection pressure. Proteins involved in interactions with the host milieu are often rapidly evolving and can be identified by the comparison of silent versus expressed mutations. The EMAAg and Gp60 genes, two genes thought to be antigenically relevant were the only two that came close to having a predisposition to non-synonymous SNPs. While the hyper-variable nature of Gp60 was consistent with the allelic profiles seen from SNP-typing results for all four international subpopulations the EMAAg locus was highly conserved or genetically stable.

Alternatively the location of a protein and consequentially its exposure to the host environment can also influence selective pressure exerted upon it. The COWP gene, which due to its positioning is constantly exposed to external pressures from the host, showed only a slight propensity for expressed mutations versus silent mutations. This may be result of the COWP protein being an integral part of the hearty nature of the oocyst wall to ensure its environmental persistence. Alternatively it may be a reflection of the different niches occupied by the parasite within their respective hosts. Conversely genes representing biological processes such as cell growth, maintenance and metabolism have a much lower occurrence of expressed SNPs. The SNPs mapped to such genes for this study are in agreement with this.

In contrast to earlier research, our study investigates the sequence diversity between the two *Cryptosporidium* species at both the nucleic acid and amino acid level in addition to the biochemical and biophysical impact of such mutations. Scientific research has drastically evolved with the development and implementation of highly specific software programs. Computational methods are now widely used to make inferences about the coding regions of a gene or genome, polymorphic sites and the subsequent impact on genotype and phenotype and ultimately evolutionary relationships. The use of a bioinformatics approach with the application of computational methods for more comprehensive studies on the polymorphic nature of such proteins provides valuable groundwork for more extensive downstream applications and studies. It is under-reported on in the field of *Cryptosporidium* research and therefore warranted. The aim here specifically was to provide a comprehensive account of the molecular and biochemical properties of genes and SNPs targeted for molecular typing. Similarities or dissimilarities within and among such genes could be beneficial in the design of therapeutic approaches.

Bio-synthesis & Enzymatic Proteins

The ability to accurately determine the genetic relatedness of isolates is fundamental to molecular epidemiological and evolutionary studies. The use of nucleotide variation at multiple housekeeping loci is an excellent approach to strain characterization, as it has advantages for inferring levels of relatedness between strains and the reconstruction of evolutionary events. Housekeeping genes often include those crucial to biological processes of an organism, processes such as metabolic and cellular pathways. Examined herein were six proteins suspected to be part of major bio-synthetic pathways within *Cryptosporidium* machinery; APR, AcoA, HSP70, LDH, MDH, and UPRT. Unfortunately the resolution of SNP typing results for the APR and AcoA protein was insufficient to allow inferences to be made. However the remaining proteins involved in major biosynthesis processes of *Cryptosporidium*, acetyl coenzymeA (AcoA), lactate dehydrogenase (LDH), malate dehydrogenase (MDH), and uracil phosphoribosyl transferase (UPRT), were genetically stable amongst all four international populations. SNPs within these genes were typed with great reliability and they proved to be excellent identifying molecular markers for species determination.

Functionally crucial enzymes for parasite survival are unlikely to show significant interspecies variation. SNPs within these genes, as was the case here, could be utilized as suitable species discriminating or SNP-typing markers. Nucleotide biosynthetic pathways provide the precursors for DNA and RNA synthesis, essential processes to any pathogen, and are therefore are an excellent source of drug and/or vaccine targets. The metabolic machinery of *Cryptosporidium* is highly streamlined and is unique in that both mitochondrial and chloroplast DNA appear to be missing[9, 53, 109, 161, 246]. Whereas most parasitic protozoa salvage purines from their host and synthesize prymidines de novo, *Cryptosporidium* is dependent on amino acid salvage for both purines and prymidines, lacking the ability to synthesize them de novo[1, 205, 245]. Crucial to the import of nucleosides and amino acids are the transmembrane transporter proteins[117]. Located in the parasites plasma membrane they provide substrate specific permeation routes for preferred amino acids from the host. It is likely these transporters are required for parasite viability at most if not all of its life cycle stages. Transport proteins are not only vital for providing the parasite with the necessary nutrients but for the generation of electrochemical gradients, cell signalling pathways and the maintenance of ion homeostasis. Inhibition of transporter proteins would severely impair the metabolic and energy producing pathways of the organism leading to the parasite's starvation and eventual death. Enzymatic proteins such as kinases and those involved in self-induced apoptosis are also ideal candidates within this same respect.

Cryptosporidium's metabolic processes are unique from those of other Apicomplexan parasites making it a phylogenetic enigma within the phylum. The organism lacks a functional mitochondrion and a chloroplast. It is now apparent that *C. hominis* and *C. parvum* rely solely upon their host for the provision of nucleic acid precursors[205]. Besides being the primary units of nucleic acids, nucleotides contribute too many other crucial cellular processes such as cell signalling, replication and transcription. The biosynthetic pathways of *Cryptosporidium* are drastically simplified suggesting the metabolic function of each protein or enzyme involved is critical. Inhibition or deactivation of these proteins could severely hamper the biofunctionality of the organism.

Structural Proteins

The study of parasite and host cell-tissue interactions is focused on the identification of structural and/or surface proteins that contribute to infection and disease pathogenesis as well as parasite propagation. Such proteins are of fundamental importance to the success of an organism as they are responsible for attachment, invasion and interactions with the host's cellular niche[31, 73, 136]. *Cryptosporidium* invasion of the intestinal epithelium microvillus involves the apical complex of the organism and results in a parasitiphorous vacuole. *Cryptosporidium* localizes to the intracellular but extra-cytoplasmic region of epithelial cells making up the brush border of the microvillus lining the intestinal. Proteins involved in this process are ideal targets for immunoprophylaxis as their inhibition could greatly hamper or prevent host colonization. Polymorphisms in these genes may reflect host receptor specificities. As a collective genetic variation within attachment and invasion proteins may reveal the underlying factors of the differences in *C. hominis*' and *C. parvum*'s host range.

A major part of *Cryptosporidium*'s pathogenic success is due to its complex homoxenous (single host) life cycle. Research is hampered by the inability to isolate the specific stages of it in vitro. The life cycle consists of two asexual stages and a sexual stage resulting in mature infectious oocysts. In the case of *Cryptosporidium* there is no requirement of an external period for sporulation thus making direct fecal-oral transmission feasible. In other words oocysts are infective, and remain infective, immediately upon leaving the host. Ingestion of oocysts results in their breaking open and releasing four sporozoites which invade new host cells. Asexual maturation ensues and produces four second generation merozoites that excyst and invade new host cells. Sexual reproduction gives rise to zygotes that go through a second asexual stage to develop into mature oocysts, containing four haploid sporozoites, which are passed into the environment or auto-infect the host. Loci encoding proteins expressed at the surface of the merozoites or sporozoites would be expected to be more diverse than those expressed during the sexual stages or inside the parasite. Surface associated proteins shared by or unique to the sporozoites and merozoites make attractive targets for drugs based on the interception of attachment to host cells.

Three proteins considered to be involved in maintaining the structural integrity of *Cryptosporidium* were targeted for allelic profiling at mutation positions; COWP, Mucin-1, and 18S rRNA. The COWP (*Cryptosporidium* oocyst wall protein) locus is a crucial protein to the tough exterior shell of the infectious oocyst and hence helps it maintain its environmental persistence in a range of severe ecological conditions. Though genetically conserved within Australia, Kenya, and Scotland, two different molecular markers, COWP 6 and 1, revealed novel allele types of neither the *C. hominis* nor *C. parvum* standard in the Peruvian subpopulation. Whether a product of adaptive nature to varying temporal or environmental circumstances is not known and should be pursued. SNP-typing results from Mucin-1, across all four international parasite populations were left unresolved hence unavailable for interpretation. The 18S rRNA gene has long been the standard of genetic typing for *Cryptosporidium* due to its reliability and ease of amplification for genetic testing. We looked at two SNP loci within this gene and the results displayed elements of similarity and drastic difference. The first SNP position, 18S rRNA SNP 1, was stable across all isolates for all four subpopulations for the expected *C. hominis* allele type. Similarly the second SNP, 18S rRNA SNP 3, was genetically stable across all four international subpopulations however the discriminate allele was a novel allele, not *C. hominis* nor *C. parvum*. The majority of the variant alleles were most often not shared among isolates from either within or between subpopulations. Therefore before we could consider this a fixed variant we had to first consider the multi-copy nature of the gene. A fixed difference, in particular a common fixed difference, implies a nucleotide change that would be present in each sequence from isolates of each subpopulation. Having confirmed the SNP location and the SNP-specific probe design typing results consistently showed the same novel allele variant. Since genes/proteins of interest were amplified from specimens containing whole genome DNA we accept that in all likelihood the multi-copy nature of the protein is a factor though argue that further exploration into this particular SNP position be done.

Putative Antigenic Determinants

Basic immunology dictates that microorganisms exploit and manipulate their host to prevent recognition and attack from its pathogen defence mechanisms. Their ability to do so successfully likely emerges from the proteins involved having conserved amino acid sequences throughout the evolutionary history as selective pressures have eliminated genetic variations that have proven unsuccessful. It would be logical to expect that nucleotide changes introduced between *C. hominis* and *C. parvum* especially within pathogenically crucial proteins were a result of positive selection and could explain the phenotypic differences between the two species. Differences between host populations have a definite impact on this natural positive selection.

Using SNPs to investigate the genetic basis of *Cryptosporidium* virulence is of great value to the identification design of potential immunotherapy targets. Surface proteins with antigenic potential are ideal candidates as they are vital to establishing and maintaining infection within the hostile environment of the host. It is possible that SNPs or patterns of SNPs within genes encoding antigenic determinants could help better define the host-parasite interplay.

Intron-containing Proteins

A fourth type of protein considered are those that contain intronic regions. The *Cryptosporidium* genome is highly compact and gene dense[1, 245]. *C. hominis* and *C. parvum* have considerably fewer introns than Apicomplexa such as *Plasmodium falciparum*[245]. Its genome is only 9.1 Mbp with 8 chromosomes all highly compacted and extremely gene dense. In comparison the *Cryptosporidium* genome is 2.5xs smaller but has 1.8 x greater gene density. Genome reduction is thought to have occurred predominantly through the shortening of intergenic regions, loss and shortening of introns and a reduction in the mean length of genes[1, 245]. Intron regions have great potential for playing a role in gene transfer, exon shuffling and ultimately genetic drift[106]. SNPs or other mutations in these intragenic regions could serve as reliable genetic markers for genotyping of unknown or novel isolates.

The idea of introns and their origin, purpose or effect on gene organization has been heavily debated for years. To establish and retain intronic regions within a genome is difficult in the face of opposing evolutionary forces[68, 129, 231]. For an intron to persist and evolve with an emerging organism such as *Cryptosporidium* sufficient positive selection pressures that favour its presence must exist, especially if it is to avoid further mutational challenges.

Cells are programmed to splice out non-coding DNA (introns) through the recognition of identifiable nucleotide sequences dictating proper excision[97]. Mutation within this specific region or any other sites critical to intron processing could affect the splicing of an intron. Mutations within an intron can affect the neighbouring coding DNA exons by introducing a loss or gain of amino acids which alters the reading frame of the gene. This could render a change in function for the gene or affect the regulatory factors exerted upon it.

Examination of SNPs within intronic regions is also useful because SNPs within these introns could be crucial when examining the evolutionary relationships of *Cryptosporidium.* Silent SNPs are often completely or nearly so adaptively neutral and not subject to direct natural selection. Because of this they have great potential to give an honest reflection of the mutation rate and time elapsed since the organism diverged from its most common ancestor. Moreover polymorphisms located in untranslated regions could be suitable for genotype tagging allowing for faster and more efficient means of identifying isolates.

Bioinformatics Characterization

In studies investigating association between the epidemiology and a genome, it is inefficient and impractical to genotype every single nucleotide polymorphism[24, 225]. It is therefore necessary to target those mutations expected to be relevant in terms of pathology, taxonomy or epidemiology. Comparative genomics is greatly supported by the use of bioinformatics applications for molecular studies as an essential data mining tool. This is especially appreciated in light of limited sequence data or potential erroneous sequence data being available for isolates from different sources as is the case with *Cryptosporidium.* Formally SNPs can be defined as alleles in a population. Through the use of multiple bioinformatics algorithms we are able to elucidate those SNPs that are of greatest potential research value to allow for inferences about the genetic relationship between *C. hominis* populations to be made.

SNPs can exhibit a range of altered phenotypes from mild and phenotypically silent effects to minor advantageous traits to detrimental effects[68, 185]. The likelihood of a SNP having impact on a protein depends on where it occurs within the protein and the nature of the phenotype[36, 54, 68]. To respond to this we employed multiple predictive bioinformatics algorithms to enable a more explicit evaluation of prospective SNPs.

The profile of a protein's hydrophobic character can be useful in predicting membrane-spanning domains, potential antigenic sites and regions that are likely exposed on the protein's surface. Hydrophilic regions are more likely exposed on the surface of a protein and therefore are potentially antigenic. In contrast such analysis has the goal of predicting membrane-spanning segments which have a strong hydrophobic character. Proteins passing though the phospholipid bilayer of a cell interact with a region inside or outside of the cell, where they will find water, and will therefore have a hydrophobic region correlating to the hydrophobic region of the bilayer. Non-globular proteins and/or those without transmembrane domains will be strictly hydrophilic in nature. With a scale set at (-) 4.5 – (+) 4.5 a value greater than zero is suggestive of hydrophobic character while a value of two or more indicates a strong hydrophobic region.

The analysis of properties such as secondary structure can suggest disease-causing mutations are associated with extreme changes in the value of parameters relating to protein stability[66]. Secondary structure prediction methods attempt to use the statistical preference of amino acid residues for secondary structures with the sequence to predict the secondary structure of each residue. The Kyte-Dolittle method of using hydrophobicity plots to assess topology predictions or transmembrane domains is an earlier method that lacks the robustness of more recent and computationally more advanced methods such as TOPPRED or MEMSAT. For the purposes of this study, whose main focus was not one of predicting membrane protein topology, it was considered sufficient. Membrane proteins are those proteins that span a lipid bilayer. The exterior surface that is in contact with the lipid hydrocarbon tails is highly enriched in hydrophobic residues. Water-exposed surfaces on either side of the membrane are dominated by polar and charged residues while the residues in the membrane-water interface region often tend to be aromatic. When evaluating SNPs it is worth considering where these amino acids are positioned within the structure of the protein as this will dictate what external pressures they are exposed to. The primary structure of a protein consists of the amino acids that compose the protein. Different regions of this sequence then form local secondary structures, such as alpha helices and beta strands. The packing of these secondary structural elements into one or several compact globular units or domains defines a proteins tertiary structure. Depending on their location within this tertiary structure could again dictate what outside pressures a given amino acid may be exposed to which in turn will influence potential SNP mutations. The secondary structure level of a folded protein is the most important. All downstream folding conformations are based on this level. Amino acid residues prefer certain structures. For example some amino acid residues have a propensity to form helices while others do not[51, 152]. SNP mutations in genes that code for a particular globular domain that confer a more drastic amino acid change, i.e., from a highly hydrophobic amino residue to a strongly hydrophilic residue, could impact the folding conformation of a protein[152]. Ultimately this could influence the proteins bio-functionality and/or stability.

6.2 Population Geographical Substructuring

SNP-typing & Allelic Discrimination

Genetic variation within a population occurs when there is more than one allele present in a population at a given locus[47], in our case this locus is a SNP molecular marker. Just as there are variable alleles at a given loci there are fixed alleles. These can be useful for species designation or if a fixed allele is unique to a specific geography the identification of such a marker would be highly beneficial to any epidemiology study. If genetic variation occurs between populations or sub-populations it is considered genetic differentiation, which can be defined as the differences in allele frequencies among populations.

Many studies have employed multiple methods of comparative genomics to score allelic variation at various loci throughout a given organism's genome[58, 69, 102, 135, 159, 182, 204, 221]. In particular micro- and mini-satellites, allozyme analysis and oligonucleotide arrays have been used extensively in work with Apicomplexans providing a great increase in the understanding of the population genetic structure and epidemiology of these parasites[72, 135, 149, 164, 169, 195,211, 233]. The biological differences between *Cryptosporidium* and other Apicomplexans prevent comparisons from such studies to be made. There is little doubt that the use of similar genetic markers will be important tools for clarifying the population structure of *Cryptosporidium*. We have used a novel set of such molecular markers to analyse the population structure of *C. hominis* in four geographical subpopulations. The methodology utilized single base extension (SBE) chemistry for SNP-typing. SBE is one of the best biochemistries for genetic variation studies as it allows for large-scale association studies and population studies on the ever-growing SNP data bases being developed for organisms of all types[44].

From all the SNP-typing data accumulated thus far, using the baseline SNP panel established through comparative genomics and bioinformatics analysis, a MlSt defined by the combination of SNP alleles at the 43 SNP loci for each isolate from 4 international populations was generated. In total 24 distinct MlSts were found; 6, 11, 11, and 6 for Australia, Kenya, Peru, and Scotland respectively. Only one, MlSt-1, was found in all four subpopulations while 6 were shared by two or more subpopulations; 75% of the MlSts are unique to one subpopulation. This argues for a limited gene flow between such distant populations. What heterogeneity that does exist in parasite populations is likely occurring at a more regional level, implying that investigations into the levels of inter-specific genetic variation among more localised populations would better help elucidate the intensity of geographic barriers in *Cryptosporidium* population structure.

Environments where the frequency of transmission is high and/or conditions are more conducive to exposure competitive interactions may select for particular genotypes or SNP-types. Genetically isolated populations have an increased potential for local adaptation to specific environmental conditions as well as temporal or ecological circumstances. Differences such as access to potable water supplies, agriculture, hygiene standards or practices, diminished immune status of the population due to higher incidences of primary infections such as HIV, tuberculosis and malaria could all play a role in higher rates of transmission or more opportunities for exposure. One could argue this would certainly be the case in Kenya and Peru, where the nation's infrastructure for water supply in addition to socioeconomic circumstances likely differs from that of the more developed regions of Australia and Scotland. In a population of 20, Scotland had 9 occurrences of MlSt15 and 7 of MlSt1, the highest of any one MlSt among all four populations (Figure 5.10). In fact Scotland has had one of the greatest histories of cryptosporidiosis. However this may be accounted for by a more vigilant level of reported and documentation of such cases when compared to other nations. Following this, there were 6 cases of MlSt12 in Peru and 5 of MlSt10 in Kenya, both of which are found in greater numbers than in any other geographical location for these MlSts.

Variant alleles for the same SNP loci, 18S rRNA, BT 3, were found in more than one population and could suggest a global distribution. Where these alleles originated from would be of particular interest from an epidemiological standpoint but beyond the scope of this study. Perhaps most importantly is the identification of novel SNPs within the COWP protein in the Peruvian subpopulation. This could be suggestive of a unique genetic profile to Peru, or even South America, in a protein that is crucial to the attachment and invasion of host cells. A larger sample size would be needed to further investigate this as alternatively it could be a rare allele that occurs globally.

SNP-typing results showed the genetic stability of six of the 13 proteins investigated. First off Cp23 and EMAAg, two antigenic proteins, showed distinct genetic stability whether intra-population or inter-population. This could potentially be of great research value due to their suspected roles as immunodominant or antigenic proteins that is proteins that elicit an immune response from the host. Both proteins and their respective SNPs were genetically stable and therefore reliable for molecular studies, particularly in monitoring the species complex of *Cryptosporidium* for evolutionary events that may occur due to selective or immune pressures. In contrast, the lack of variation within these two genes may suggest they are under selection though does not conclusively prove it. If eventually proven to be truly genetically conserved despite geography one could hypothesize that a drug or vaccine developed to target either Cp23 or EMAAg could be universally applicable.

Four other proteins; MDH, LDH, UPRTase, and HSP70, showed significant genetic stability both within and between populations. This is not particular surprising in the case of the enzymatic proteins as they are housekeeping genes crucial to metabolic processes therefore a necessity to the pathobiology of *Cryptosporidium*. Of potential benefit from this is their candidacy to be neutral molecular targets for tracking evolutionary events or further species delineations that may occur within the phylum. It could also be inferred that the lack of variation in a global context would also render them excellent molecular targets for the disruption of essential *Cryptosporidium* biological processes.

Visual examination of SNP-typing and allele discrimination results resolves two major goals of this study. First is the usefulness of our experimental platform in the identification of allelic diversity from sample to sample. This proved to be the case whether the allele was that of the very closely related *C. parvum* genotype or a novel allele altogether. This can be further extended to the investigation of suspected species co-infections, a concept once ignored in the field but gaining momentum as perhaps being more common than originally thought. Secondly, we have brought to light the usefulness of certain proteins and their mutation profiles as a genetic typing tool for species specific distinction. Not only is the methodology more rapid than that of standard gene sequencing it is more cost effective with higher throughput than most molecular methods. Theoretically up to 12 SNPs in 96 different samples could be processed in less than 8 hours. As a collective all of these observations would be of great benefit to epidemiological studies into *Cryptosporidium* behaviour, particularly in a large-scale outbreak situation.

In contrast to the use of stable allelic profiles sequence or MlSt comparison revealed that polymorphisms were in many cases not shared between individual isolates, let alone within or between subpopulations. If SNPs unique to one isolate over another is fixed within that isolate they could ultimately be used to define it. For an allele variant be strictly unique to one isolate it would be ever present in each sequence from that isolate but would be absent in sequences from other isolates. Ultimately this could lead to the identification of a novel allelic structure of an isolate within a group of isolates.

Mixed Infections

The presence of mixed alleles is a common occurrence in population genetics studies on microparasites. In the case of mixed allele calls at SNP markers the assumption was made that predominant peaks at each locus represent the actual genotype or SNP-type. This assumption is a necessity when dealing with haploid organisms and genetic data analysis. It is commonly cited in the literature for studies based on other related Apicomplexa[44, 136, 164]. Other than the isolation of individual oocysts and genotyping or SNP-typing these, the research community commonly accepts that there is no other way around the issue of mixed alleles.

Based on our experimental approach the decision to use 20% as a threshold value was made though can be left open for interpretation. The presence of such alleles or mixed genotypes is one that anyone dealing with microparasite infections runs into. To some extent this threshold can be considered arbitrary and most certainly is dependent on the reliability of the detection system used. Tait et al (2003) conducted a study on the geographical sub-structuring of *C. hominis* and *C. parvum* genotypes in the United Kingdom using a combination of micro and mini-satellite markers. Having encountered the problem of deciphering multi-locus genotypes in the presence of mixed alleles they reasoned that any secondary peak being at least 10% the height of the main peak could be scored as real. A second study by Anderson et al., (2005), using microsatellite regions for an investigation into the population structure of *Plasmodium falciparum* argued the use of a 33% cut off because of their experimental platform. Microsatellite loci often have stutter peaks one repeat length away from the main peak. If a cut off less than 33% was used, these artifactual stutter peaks could be confused with additional alleles. In contrast to using satellite markers, when dealing with SNPs, minor peaks can typically be detected at a lower threshold.

The presence of multiple peaks at certain SNP molecular markers could result from multiple infections within a given sample. If the case we would expect to see multiple peaks at multiple SNP positions within a single sample. Also to be considered is the sample source. Multiple infections are more likely to occur in geographic areas where cryptosporidiosis is highly endemic. Frequent transmission from environmental sources increases the probability of coinfections occurring with genetically heterogeneous parasites, favouring recombination. Though ubiquitous, on the global scale, certain localities within a given country may be more susceptible to cryptosporidiosis outbreaks. This could be from many different factors as discussed in chapter one, factors such as agricultural practices, potable water resources, hygiene standards and host immune status. For example in developed countries where sanitary practices are more stringent and HIV incidence rates are lower, coinfections with heterogeneous parasites originating from environmental sources may be less frequent thus allowing for clonal propagation to prevail.

Samples used in this study were crude fecal specimens known to contain genomic *Cryptosporidium* DNA. While typed or diagnosed as *C. hominis* the presence of other species cannot be ruled out unless individual oocysts are isolated and typed for a specific species. With the argument of clonality in *Cryptosporidium* genetic populations, a potential alternative could be a mutation within clonal infections. In this case we hypothesize that just one or perhaps two SNPs would have multiple peaks in any infection. Our data showed 19 SNP markers with bi-allelic typing results.

The results of the study presented here show there to be mixed alleles, at 13 (19 if the AcoA and Mucin1 loci are considered) different SNP loci within 5 (or 7) genes. These 5 genes are located on four different chromosomes of the *Cryptosporidium* genome. In the case of mixed genotypes, we can hypothesize that a higher number of such genotypically mixed MlSt's seen in one population versus another could be suggestive of a less stable population structure. This could be result of multiple factors. Such analysis would be hampered as most of these can only be answered by detailed retrospective studies within other public health districts,

In the case of Peru and Kenya these are likely very primitive if established at all. Since the majority of the secondary allele calls are that of the *C. parvum* genotype it is possible that humans co-habiting with or relatively near to bovine hosts could cause this. Also worth examining would be the exact etiological source of infection meaning questioning the infection source of water, versus food, versus swimming pool versus petting zoo and so forth. Because samples used were isolated from human hosts the human to human transmission of *C. parvum* may be more important than has been previously assumed. Alternatively as the temporal, ecological or host immune status conditions vary from country to country different conditions may favour the appearance of certain genotypes. It is here that data on the exact location of each isolate's isolation within the geographic territory it came from would be especially useful.

The preliminary results of our work show no predilection for mixed genotypes for one subpopulation over another. This may imply that Australia, Kenya, Peru, and Scotland have similar epidemiological circumstances, circumstances that would affect the prevalence of mixed genotypes. Further examination and more precise and stringent species specific molecular tools would be necessary to resolve this. The implication from the evidence of mixed infections occurring in the isolates tested is that rather than conforming to a strict paradigm of either clonal or panmictic species the data is suggestive of the cooccurence of both pathways which is consistent with other studies[134, 135, 218].

Genetic Diversity Indices

Understanding and quantifying the genetic structure of a natural population of any organism has been a long standing objective in evolutionary biology. Only by defining how genetic diversity is distributed within (intra) versus among (inter) populations can insight into genetic population structure, levels of gene flow, historic population parameters and even the early periods of speciation be provided. These studies are of particular importance when considering either a re-emerging infectious disease or a newly emerging infectious disease, as is the case with *Cryptosporidium*.

One of the ultimate goals in quantifying population genetic structure is to not only understand variation among species, or sub-populations of a given species, but to determine whether or not any patterns exist among different populations and life zones/geographies exist. At the genetic level, barriers to dispersal and subsequent genetic exchange between populations allow for their divergence because of local adaptation, gene flow or random genetic drift[47, 188, 236]. Spatial or geographic population structure is most readily estimated by evaluating the degree of genetic differentiation of genetic marker, neutral or expressed, among geographically separated populations.

Herein the genetic relationship between four international populations of *Cryptosporidium*, Australia, Kenya, Peru and Scotland, were investigated using such DNA markers. The molecular typing data indicate relatively higher levels of genetic variability within populations. Gene flow, or the movement of genes between populations, can be a potent force in the reduction of genetic differentiation among populations. Excessive gene migration among populations can convert genetic differentiation into an increase in genetic variation within a population. If gene migration is restricted, which can be the case in isolated geographic regions; genetic differentiation will increase as allelic frequencies within a population become more fixed or stable. In contrast genetic differentiation can be created through the process of random genetic drift which refers to the fluctuations in allele frequency occurring by chance. This is especially true within smaller sub-populations. As time goes by these allele frequencies become fixed or stable within a population, leading to an increase in population differentiation. While changes in allele frequencies within populations, most likely caused by natural selection, can lead to adaptation it is the genetic differentiation that can ultimately lead to major evolutionary events.

Multilocus sequence analysis of a panel of 43 molecular markers, SNPs, from *Cryptosporidium* isolates revealed that, on a global scale, genetic diversity is more considerable within a given population than it is among populations. The identification of such a biogeographical trend has implications when investigating the population structure of a microorganism. Previous studies into the biogeography of *Cryptosporidium* have uncovered evidence of both geographically restricted genetic sequences as well as more ubiquitous ones. However the geographic dispersal of genetic population structure within these studies was quite limited and questioned whether similar patterns would be seen in other locales and how large of an impact the movement of parasite hosts within the respective locale made. This study is one of the first to approach the pathogenomics of *Cryptosporidium* from such a global perspective. Considering all populations are situated on different continents, hence the obvious geographic boundaries of oceans and mountain ranges, it would stand to reason that the exchange of genetic information, or gene flow, would be more restricted. If this were a true factor one would expect to see a high degree of genetic differentiation between populations. Because mutation is the ultimate source of all genetic variation, it increases variation within subpopulations[47, 235]. Mutation also leads to an increase in differentiation (inter-population diversity), because the chance of the same mutations occurring within the same subpopulations is low. Gene flow works to convert inter-population genetic variation (differentiation) to intra-population variation, while genetic drift tends works oppositely, converting intra-population into inter-population variation[47]. With the physical geographic boundaries of Australia, Kenya, Peru, and Scotland limited gene flow would be expected resulting in a higher degree if intra-population diversity.

The Australian population was considerably more variable for all gene diversity measures when compared to the others, Hs=0.177 (Table 5.13), in particular Scotland which revealed to have the least genetic diversity averaged at Hs=0.022. Despite the pronounced difference the clonal diversity, PD (Table 5.12), was quite uniform across all four international populations. An estimation of Gst=0.304 indicates that 30.4% of the total genetic variation exists among the four international *C. hominis* populations. The inter-population differentiation estimated to be 30.4% is suggestive of limited but still present gene flow between populations. In other words gene flow was not sufficient to erase genetic divergence amongst these geographically separated subpopulations. Because the probability of gene flow between populations would be expected to be lowered when considering the relatively isolated habitats of the four intercontinental populations the practice of sampling more individuals within a given population should be adopted. If the degree of variation among populations was estimated to be significantly higher than variation within a population it would then be more prudent to sample more populations or geographies, with less focus on the number of individuals.

This result is in agreement with a more recent study that undertook the assessment of population structure on a global scale by Tanriverdi et al., 2008[218]. The study involved both *C. hominis* and *C. parvum* isolates and also showed insufficient gene flow to erase genetic divergence. Though considering the degree of travel occurring in the modern world today the author must point out that precisely defined and isolated geographic boundaries from an epidemiological standpoint are almost non-existent, thus creating opportunity for gene flow through travel and import/export of foods.

Taking into account results here we put forth that the degree of genetic variation partitioned among populations may be better examined by sampling more populations within the given territories of Australia, Kenya, Peru, and Scotland respectively. It must also be taken into account that while the origins of parasite populations donated are documented to be from with each country as confirmed *C. hominis* specimens' information on the exact location within their country of origin is not known. This fact could be key when considering the vastness of a country such as Australia. Through the use of very geographically separated populations we hoped to negate issues surrounding the travel of hosts among the four localities though we cannot ignore the intensive nature of human travel in the modern world. It is therefore of great interest that in the future we would hope to be able to access to data surrounding the exact circumstances of each isolate in terms of host and point of origin. In doing so we hypothesize we could limit a biased epidemiological structure by omitting isolates from patients reporting travel. This would perhaps result in a more regional assessment of inter-population differentiation being made which in return may help better determine the true impact of genetic drift at the subpopulation level. The severe isolation of a finite population will cause random genetic drift to become relatively more important than gene flow[47]. With the clonality of *Cryptosporidium* populations still being debated this would be worth exploring as in many clonal species the majority of genetic variation is often found among populations. Also supporting the idea that more "mini-subpopulations" within a locality should be evaluated is shown by the genetic distances measured among all four international populations. Despite being so geographically removed from one another the highest genetic distance relationship was only 0.061 between Kenya and Scotland.

Also a factor of consideration was the potential for selection for local adaptation of the four intercontinental sub-populations. The environments our four subpopulations inhabit, as is often the case, differ in terms of light, temperature, agriculture, population immunity, host density, and so on. As local adaptation occurs an increase in differentiation occurs[47, 185]. Conversely, selection events that do not differ between subpopulations, due to similar environments or fundamental features of a species, will lead to a decrease in differentiation[47]. Though overall results indicate that differentiation is low in comparison to intra-population variation thus implying minimal effect from local adaptation selection does not act on the genome as a whole. While genetic drift and gene flow affect all loci, selection can be more targeted. As mentioned the novel allele variants in the Peruvian subpopulation may be evidence of such a local adaptation upon further examination.

A common obstacle in genetic investigations into any new pathogen is the establishment of baseline genetics from which inferences, comparisons and differences can be ascertained. Some argue for close relatedness of *C. hominis* isolates throughout the world, other emphasize that clonal lineages within *C. hominis* are evolutionary independent. While there appears to be a monophyly of *C. hominis* there is extensive substructure so *Cryptosporidium* should be considered a species complex. Though data is starting to accumulate we don't know the baseline measures of genetic diversity or population differentiation for *Cryptosporidium*, especially at the bio-geographic level[218, 219].

Our data compliments those of other studies done previously on *Cryptosporidium* parasite populations using various methods, different loci, and in a less global manner[25, 27, 38, 82, 134, 135, 206, 219]. To date most studies have focused on a few loci, versus a multi-loci whole genome approach, which limit the ability to accurately capture the true genetic structure of a population. An earlier study conducted by Smith et al. (2003) showed through the analysis of a combination of micro- and mini-satellite markers there was no evidence to support geographic or temporal substructuring of *C. parvum* populations within Scotland. A lack of geographical sub-structuring was evident by both Wright's Fst values and Nei's genetic distance values. While the results of the Smith study are limited to the geographic boundaries of Scotland, a similar study done in Italy by Caccio et al. (2000) showed evidence for the non-random geographical distribution of specific alleles within a protein, the ML1 protein. Their study involved Italy and other Northern European samples and indicates that perhaps geographical sub-structuring is more evident when samples from a wider area are used. A third, more recent study aimed at geographic linkage and variation in *C. hominis* was done by Hunter et al. in the UK[40]. They assessed the geographic population structure from a standard genotyping approach using the Gp60 locus. This marker is especially appealing due to its functional relevance and extensive sequence polymorphism. There are some differences between their study and ours. One major discrepancy between the two studies is instead of using distinct geographic populations they were looking into the transmission dynamics from a movement of hosts out of and into the UK. Their conclusion of the relationship between travel outside of Europe and Gp60 subtypes was 37.08% with no other epidemiological associations present. That is differentiation or inter-population diversity constitutes 37.08% of genetic variation. Even though it was conducted through a different methodology and perspective of geography this was somewhat complementary to our finding of inter-population averaging at 30%.

Thus far it appears from our data that there is little to indicate population substructuring with part of the problem being that we don't have specifics on how the samples used relate to the geographies they represent. If they all came from the same location they would likely underestimate the diversity. Migration is an influential mover of genetic change. Whenever it is involved it is hard to maintain population sub-structuring. In contrast, selection and drift move much more slowly and can be easily swamped by minimal migration.

The minimal amount of global baseline data akin to this study in the literature leaves the researcher as having to take the results with confidence at face value until further studies are accomplished. In the interim it is likely that by increasing the number of SNP loci and samples, using well defined sub-populations and improving estimates of allele frequencies and divergence with more sophisticated data analysis methods can improve upon the study at hand. Ultimately patterns of modern *C. hominis* population structure discussed here could be used to guide construction of historical models of migration and admixture which would be useful in inferential studies of *Cryptosporidium* genetic history.

CHAPTER 7

FUTURE DIRECTIONS

- Current Work, Study Extensions –

7.1 Current Work

As eluded to we investigated a total of 25 target proteins for SNP-based genetic typing of globally distinct *C. hominis* subpopulations. While 13 of these were addressed in the study at hand, 12 are awaiting multi-locus SNP-typing. These 12 include: Cellcycle Regulator, CTCL Tumor Ag, Aldahyde-Alcohol Dehydrogenase, CLL Associated Ag-KW-2, Sexual Stage Specific Kinase, FLJ31812/DHHC palmitoyl transferase, Transmembrane amino acid Transporter, ABC multi-drug or ion efflux, Thiolproteinase, Extracellular protein w/ 8 kazal repeats, Seroreactive Ag BMN-19B related protein and RIK protein w/? WD40 repeats. These 12 proteins represent 6 of the total 8 chromosomes of the *Cryptosporidium* genome: 1 of which is situated on chromosome one, 2 from chromosomes 2, 3, 4 and 8 and 3 of which represent chromosome 7. These 12 proteins had bio-functionalities ranging from biosynthesis, enzymatic, metabolic and antigenic properties.

Of most interesting note is, with the exception of the genome sequencing project, there is either a limited or a complete lack of molecular data on any of the 12 above mentioned proteins, making them attractive novel target proteins for multi-locus SNP-typing. Using the restricted sequence details that were available through *in silico* data mining we designed original, un-published upon, PCR amplification primers. All 12 of these proved to be highly successful in generating large amounts of DNA amplicons from crude fecal specimens (Appendix 6). To date there is no published data on successful PCR amplification of any of these 12 proteins thus indicating our primers will be a valued addition to the molecular field of *Cryptosporidium* research. Using isolates from all five subpopulations MIS-typing of these target proteins is currently awaiting completion. It is believed that the addition of 12 new proteins and a minimum of 24 new SNP loci would add great robustness to the study at hand and allow for more conclusive correlations to be inferred.

7.2 Future Study Extensions

The success of efforts to design and develop efficacious vaccines or chemotherapies for *Cryptosporidium* is contingent on characterizing the extent and nature of genetic diversity within its genome. Just as important is the identification of the mechanisms by which such diversity is generated and able to persist in parasite populations. Our study is a preliminary investigation that could be extrapolated to address this. Keeping in mind the high degree of genetic similarity between the *C. hominis* and *C. parvum* genomes (~97%) it stands to reason that differences in their pathogenic behaviours, from host specificity to mode of transmission to disease severity, is most likely due to those subtle genetic differences that do exist. Further, more complete characterization and evaluation of the genetic make-up and organization between the two at the mutation level is necessary.

Molecular studies are vital for refining the host specificity, interlaced transmission dynamics and infection sources of *Cryptosporidium*. To put into an ecological context more studies need to be undertaken. Studies could provide important insights into the effects that anthropogenic activities like waste treatment, water supply treatments, farming and agricultural practices and public health or hygiene issues have on the overall epidemiology of *Cryptosporidium*.

7.2.1 Continuation of Current Multi-locus SNP Data: Examination of More Genes & Molecular Markers

Our data suggests through the evaluation of isolates located from more regional areas within the territories of Australia, Kenya, Peru, Scotland and Canada the exact degree of intra-population diversity could be better defined. It is important that a balance between geographically diverse populations and population structure conclusions be made. If populations examined are too close, results may be skewed by too narrow of a geographic boundary and the potential for increased movement of hosts within it.

In molecular research the more data accumulated the more robust the conclusions that can be made. The logical immediate extension of this study would be to continue to examine more proteins and SNPs. As a greater number SNPs and proteins are examined the stronger the associations to geography can be made. A comparable study was done on the human genome, estimated as having 20,000 to 25,000 genes, covering three distinct populations specifically designed to detect the number of SNP loci to infer population structure[225]. Results showed that just over 65 random SNP loci were required for identifying distinct geographically separated populations. The *Cryptosporidium* genome, estimated to contain approximately10, 000 genes, is just 0.17 that of the human genome and we are currently using 43 SNP loci. While our results are interesting it remains that the use of either more SNPs overall or particular SNPs not yet identified that are crucial to population studies could make a considerable impact.

Also to consider would be incorporating more intron regions and possible SNPs within them into the study. At the initial design of the study, 5 years ago, there were only 6 hypothesized genes containing intron regions. With the completion of the *C. hominis* and *C. parvum* genomes and the increasing accumulation of molecular data there is now an estimated 200-800 genes with putative intron regions (Figure A.1). Non-coding regions, like introns, are expected to have fewer functional constraints compared to coding regions. The levels of genetic variation within these regions could have significant implications. A low genetic variation could imply influences exerted upon them by which a gene's frequency changes due to selection operating upon a linked gene; proximity on a chromosome may allow genes to be dragged through the selection process due to an advantageous gene nearby. Alternatively low genetic variation could be suggestive of conserved functional roles usually involved with introns, such as splicing machinery. With respect to *Cryptosporidium* this is very likely as the splicing machinery of the genome has been shown to be drastically reduced or streamlined[1, 245]. The examination of allelic variation within these regions could be very informative.

7.2.2 Inferring Patterns of Evolutionary Descent

To obtain further ideas about the nature of population structure for *Cryptosporidium* authors would like to extend current multi-locus SNP data and future data generated to alternative analysis methods or approaches. The use of multi-locus molecular marker data for the accurate cataloguing of isolates of parasitic pathogens has a marked impact on both routine epidemiological surveillance and population biology. In both fields, a requirement for exploiting this resource is the ability to differentiate the relatedness and patterns of evolutionary descent among isolates with similar genotypes. Though valuable in their own right most clustering methods, such as dendrograms, tend to provide a poor representation of recent evolutionary events as they are inclined to rebuild relationships in the absence of a realistic model in which parasite populations emerge and diversify. Dendrograms typically represent multi-locus genetic data on the basis of a matrix of pairwise differences in the allelic profiles of the isolates studied. While a convenient means of identifying isolates that may be identical or closely related the topology of dendrogram representation can be arbitrary, providing little information on the patterns of evolutionary descent of the isolates.

In view of the cosmopolitan distribution of *Cryptosporidium* spp. which may easily travel between countries a more detailed account of the host from which each isolate was obtained would be attractive. This type of data would be crucial as to whether or not an epidemic structure or bias can be ruled out due to the presence of imported, thus reproductively isolated, MlSts in C. *hominis* isolates from the subpopulations studied. This would require a retrospective approach to the study as patient information regarding travel behaviours should be obtained in order to discern those isolates that may have come from hosts having travelled outside of the geographic boundaries. In theory the ecology or environment of cryptosporidiosis in different geographies may have selected for phenotypes best adapted to each environment. If true one would expect that imported parasites would be unlikely to spread if the environmental factors or transmission patterns are hostile.

In consideration of clonality for *Cryptosporidium*, of future interest for study may be the single locus variants, those allelic profiles or MlSt's that differ by just one molecular marker. In the simplest of terms the emergence of a clonal population is that an initial genotype increases in frequency in the population. This is likely a result of fitness advantage or random genetic drift thus enabling it to become a predominant genotype. As its frequency increases over time, this genotype will gradually diversify. Ultimately variants in the allelic profile of descendents of this genotype will arise, by point mutation or recombination. This may start with a single allele variant but in time can lead into multiple allele variants as further diversification occurs.

To address these concepts the authors would like to evaluate the allelic profiles of isolates from each of the subpopulations used in this study using a more recent program known as eBURST™ (http://eburst.mlst.net). The eBURST™ algorithm works to identify mutually exclusive groups of related genotypes in the population and attempts to identify the founding genotype or sequence type (ST) of each group. The algorithm then predicts the descent from the predicted founding genotype to the other genotypes in the group displaying the output as a radial diagram, centered on the predicted founding genotype. The primary founder of a group is defined as the sequence type (ST) that differs from the largest number of other STs at only a single locus. The eBURST diagrams display the patterns of descent within each group from the predicted founding ST (Appendix 9). The assignment of the founding ST does not take into account the number of isolates of each ST; this makes the procedure relatively robust to sampling bias.

7.2.3 The Potential Implications to the Range of Vaccines or Chemotherapies Targeted to Specific Mutations within the *Cryptosporidium* Genome.

Ultimately an attractive downstream approach to SNP diversification among *Cryptosporidium* populations would be to determine a more precise picture of the stability of SNPs known to be under intense immune or diversifying pressure, in particular surface antigens of the parasite. Needed would be epidemiological settings for *Cryptosporidium* that are suitable to test whether polymorphisms evolve rapidly because limited human movement among defined geographic regions and low transmission levels limit the diversity of parasite populations. While this was addressed in this study we cannot ignore the "global population" of the modern world created by transportation and travel practices. To get a true representation of SNP stability using such epidemiological conditions it is likely that the study would have to expand to a relatively remote geography that would support the epidemiological settings mentioned and be extended over a significant period of time. In theory SNPs that did not show any sequential or stepwise changes among alleles and/or alleles found in more than geography could imply SNPs originating outside of such a locale. Furthermore this would hint at novel SNPs having not evolved within that geography, thus indicating stable SNPs. In contrast if SNPs varied stepwise in terms of chronology inferences about the age of the observed SNPs could be made. In the big picture of the future of *Cryptosporidium* research, if an efficacious vaccine or drug regimen targeted to known antigenic molecular markers were to be developed the presence of these stable SNPs would suggest that they will be more effective where there is a limited gene pool, as in heavily isolated populations. Studies could also then be conducted radiating outwards to increasingly larger geographic boundaries to determine the range of usefulness of such therapies.

CHAPTER 8

EXECUTIVE SUMMARY

By some estimates water is responsible for approximately 80% of all infectious disease. One of the most prevalent causative agents is Apicomplexan organisms such as *Cryptosporidium*. Morbidity and mortality due to infectious Apicomplexa protozoa is of growing concern, especially in the era of AIDS. Mounting rates of infection and numerous large scale outbreaks coupled with ineffective therapies, their side effects and emerging resistance among organisms proves there is a tangible need for the development of novel therapeutics targeting these protozoan parasites.

Cryptosporidium is a globally ubiquitous enteropathogen of great importance to public health exacerbated by factors such as socioeconomic status, access to potable water, proximity to agricultural practices and wildlife, and personal immune health statuses. The environmental stability when coupled to the complex, interlacing transmission dynamics of this pathogenic parasite renders *Cryptosporidium* a highly successful microorganism for which there is currently no efficacious vaccine or prophylactic treatment.

A detailed analysis and characterization of the subtle differences between the emerging infectious species of *Cryptosporidium* is a crucial and important step towards the rational design of novel therapies and more effective intervention policies. It is foreseeable that the widespread occurrence of similar genomic regions considered potential vaccine targets but having high rates of mutability will impact the probability of success of protective vaccines. There is a need for longitudinal studies that link population based genetics with clinical end points. The potential implications of this are that prevention or treatment strategies may need to differ for different geographical areas where genetic variations are conserved in order to be more effective.

Extensive research into the genetics and etiopathogenesis of *Cryptosporidium* are being conducted in facilities all over the world. Even with all the progress the pathobiology of *Cryptosporidium* is still largely unclear. As analysis of the completed genomes proceeds the discovery of new genes and proteins will arise. Inevitably so will questions that address their degree of variability, how this variation is generated and maintained, and in what way can genetic diversification affect intervention efforts.

Cryptosporidium and the growing number of novel species being identified is an excellent example of how the parallels between wildlife or ecological circumstances and an emerging infectious disease in the human population can be associated with increased interactions with zoonotic pathogens coupled to the host-parasite paradigm. When investigating an organism, like *Cryptosporidium*, whose pathobiology is directly linked to the environment identifying any correlations between the emergence of disease and casual factors such as microbial adaptation and the degree of genetic diversity from a biogeographical perspective is crucial if a better understanding of the epidemiology of *Cryptosporidium* is to evolve.

We report on genetic variation both within and between *C. hominis* subpopulations from Australia, Kenya, Peru, and Scotland. We examined ~18 500 bp and assembled a data set of 394 SNPs. Employing comparative genomics and bio-physical profiling an expected haplotype, representing a set of 45 single nucleotide polymorphisms at individual loci was established. Molecular typing of 77 international isolates based on this haplotype or multi-locus SNP-type was done, twenty-four unique MlSt's were identified. Inferences about genetic relationships were made using genetic data analysis software programs to quantify and partition the genetic diversity into intra- and inter-population diversity and to discern genetic distances among subpopulations.

Our aim was to answer the question what level of genetic variation exists within geographically distinct populations. The possibility of exclusive "geo-types" would suggest *Cryptosporidium* parasites harbour substantially greater biodiversity and species richness than current estimates imply.

Our data suggests little to argue for population substructuring. Depending on the locus and isolate studied, the results ranged from a virtual lack of to more extensive genetic variation. Within population differences among subpopulations account for 69.6% of genetic variation; differentiation among subpopulations constitute 30.4%. Genetic distances among subpopulations averaged 0.048 and varied from 0.034 between the Australian and Scotland subpopulations to 0.061 between Scotland and Kenya.

The potential use of a DNA-typing scheme based on SNPs to resolve *Cryptosporidium* epidemiology was examined. Using the experimental methodology that we did enabled us to demonstrate the ability to genotype an isolate based on a particular mutation profile. Rapid and reliable species distinction is crucial to any epidemiological outbreak investigation. We identified four genetically stable SNP profiles within four different proteins that would be excellent candidates for study into this. Shown was the ability to clarify the presence of standard or novel allelic variation at a specific SNP locus. We identified private allele variants unique to one population, Peru, within a protein crucial to the invasion and attachment strategy of *Cryptosporidium*. The concept of mixed infections possibly being more common than once thought has garnered more attention. Implications about using SNPs as a molecular tool to reveal the presence of mixed infections could be made from the data generated.

This study is one of the first to report on international biogeographical diversity using a SNP profile as a DNA-typing scheme. Some of the proteins and SNPs are discussed for the first time within the field, offering some excellent baseline possibilities. To more precisely clarify the species complex as a whole, the evolutionary forces behind the emergence of new species and the subsequent consequences to human population health further molecular research is certainly warranted.

LITERATURE CITED

1. Abrahamsen M, Templeton T, Kapur V, et al. 2004. Complete genome sequence of the Apicomplexan, *Cryptosporidium parvum.* Science 304: 441-444.

2. Akiyoshi D, Feng X, Tzipori S, et al. 2002. Genetic Analysis of a *Cryptosporidium parvum* Human Genotype 1 Isolate Passaged through Different Host Species. Infection and Immunity 70(10): 5670-5675.

3. Akiyoshi D, Siobahn M, Tzipori S. 2003. Rapid Displacement of *Cryptosporidium parvum* Type 1 by Type 2 in Mixed Infections in Piglets. Infection and Immunity 71(10): 5765-5771.

4. Amar C, Dear P, McLaughlin J. 2003. Detection and identification by real time PCR/RFLP analyses of *Cryptosporidium* species from human feces. Society for App. Microbiology 38: 217-222.

5. Anderson T, Nair S, Nosten F, et al. 2005. Geographical Distribution of Selected and Putatively Neutral SNPs in Southeast Asian Malaria Parasites. Molecular Biology and Evolution 22(12):2362-2374.

6. Applebee A, Thompson A, Olson M. 2005. *Giardia* and *Cryptosporidium* in mammalian wildlife – current status and future needs. TRENDS in Parasitology 21(8):370-375.

7. Arrowood M. 1997. Diagnosis in *Cryptosporidium* and Cryptosporidiosis. Fayer R. CRC press: 43-64.

8. Ashbolt N. 2004. Microbial contamination of drinking water and disease outcomes in developing regions. Toxicology 198: 231-238.

9. Atreya C, Anderson K. 2004. Kinetic Characterization of Biofunctional Thymidylate synthase-dihydrofolate reductase from *C. hominis*. J. of Biol. Chemistry 279(18):18314-18322.

10. Atwill E, Johnson D, Frost W, et al. 1999. Age, geographic and temporal distribution of fecal shedding of *Cryptosporidium parvum* oocysts in cow-calf herds. American J. of Veterinary Research 60: 420-425.

11. Atwill E, Johnson D, Pereira M. 1999. Association of herd composition, stocking rate and duration of calving season with fecal shedding of *Cryptosporidium parvum* oocysts in beef herds. J. American Veterinary Med. Association 215:1833-1838.

12. Atwill E, Sweitzer R, Boyce W, et al. 1997. Prevalence of and associated risk factors for shedding *Cryptosporidium parvum* oocysts and *Giardia* cysts within feral pig populations in California. Appl. and Environmental Microbiology 63:3946-3949.

13. Awad-El-Kariem F. 1999. Does *Cryptosporidium parvum* have a Clonal Population Structure? Parasitology Today 15(12)502-504.

14. `Awad-El-Kariem F, Robinson H, Casemore D, et al. 1998. Differentiation between human and animal isolates of *Cryptosporidium parvum* using molecular and biological markers. Parasitology Research 84: 297-301.

15. Balabat A, Jordan G, Tang Y, Silva J. 1996. Detection of *Cryptosporidium* DNA in human feces by nested PCR. J. of Clinical Microbiology 34: 1769-1772.

16. Barnes D, Bonnin A, Huang J, et al. 1998. A novel multi-domain mucin like glycoprotein of *C. parvum* mediates invasion. Mol. & Bioch. Parasitology 96:93-110.

17. Beck J, Davies J. 1981. Medical Parasitology; 3rd Edition. C. V. Mosby Company.

18. Bell A, Meeds D, Farley J, et al. 1993. A swimming pool-associated outbreak of Cryptosporidiosis in British Columbia Canada. Canadian J. Public Health 84:334-337.

19. Black. 1996. Lecture in Infectious Disease Epidemiology. Johns Hopkins School of Public Health.

20. Bogitsh B, Carter C, Oeltmann T. 2005. Human Parasitology; 3rd Edition. Elsevier Academic Press.

21. Bonafonte M, Smith L, Mead J. 2000. A 23-kDa recombinant antigen of *Cryptosporidium parvum* induces a cellular immune response on in vitro stimulated spleen and mesenteric lymph node cells from infected mice. Exp. Parasitology 96(1):32-41.

22. Bourgon R, Delorenzi M, Sargeant T, et al. 2004. The serine repeat antigen gene family phylogeny in *Plasmodium*: the impact of GC content and reconciliation of gene and species trees. Molecular Biol. & Evolution 21(11):2161-2171.

23. Brookes A. 1999. The essence of SNPs. Gene 234 (2):177-186.

24. Butler J, Bishop T, Barrett J. 2005. Strategies for selecting subsets of single-nucleotide polymorphisms to genotype in association studies. BMC Genetics 6(Suppl 1):S72.

25. Caccio S, Homan W, Camilli R, Traldi G, Kortbeek T, Pozio E. 2000. A microsatellite marker reveals population heterogeneity within human and animal genotypes of *Cryptosporidium parvum*. Parasitology 120(Pt.3):237–244.

26. Caccio S, Homan W, van Dijk K, Pozio K. 1999. Genetic polymorphism at the b-tubulin locus among human and animal isolates of *C. parvum*. FEMS Microbiology Letters 170(1):173-179.

27. Caccio S, Spano F, Pozio E. 2001. Large sequence variation at two microsatellite loci among zoonotic (genotype C) isolates of *Cryptosporidium parvum*. International J. for Parasitology 31: 1082-1086.

28. Camp, Dresser and McKee. 1995. "Summary of the Mt. Vernon, Ohio, Membrane Softening Pilot Plant." December 14, 1995.

29. Campbell I, Tzipori S, Hutchinson G, Angus K. 1982. Effect of disinfectants on survival of *Cryptosporidium* oocysts. Veterinary Research 111: 414-415.

30. Carroway M, Tzipori S, Widmer G. 1996. Identification of genetic heterogeneity in the *Cryptosporidium parvum* ribosomal repeat. App. and Env. Microbiology 62(2): 712-716.

31. Casedevall A, Pirofski L. 2000. Host-pathogen Interactions: Basic concepts of microbial commensalisms, colonization, infection and disease. Infection & Immunity 68(12):6511-6518.

32. Casedevall A, Pirofski L. 1999. Host-pathogen Interactions: redefining the basics of concepts of virulence and pathogenicity. Infection & Immunity 67(8):3703-3713.

33. Casemore D. Molecular and Antigenic aspects of *Cryptosporidium* and Cryptosporidiosis, a brief review. Public Health Laboratory Service, *Cryptosporidium* Reference Unit, Wales, UK. Appendix 7:137-142.

34. Casemore D, Armstrong M, Sands R. 1985. Laboratory Diagnosis of *Cryptosporidium.* J. of Clinical Microbiology 38: 1337-1341.

35. Casemore D, Garder C, O'mahony C. 1994. Cryptosporidial infection, with special reference to Nosocomial transmission of *C parvum*: a review. Folia Parasitology 41(1):17-21.

36. Cavallo A, Martin A. 2004. Mapping SNPs to protein sequence and structure data. Bioinformatics 21(8):1443-1450.

37. Center for Disease Control; Atlanta, United States. www.cdc.com

38. Cevallos A, Bhat N, Verdon R, et al. 2000. Mediation of *Cryptosporidium parvum* infection in vitro by mucin-like glycoproteins defined by a neutralizing monoclonal antibody. Infection & Immunity 68(9): 5167-5175.

39. Cevallos A, Zhang X, Waldor M, et al. 2000. Molecular cloning and expression of a gene encoding *Cryptosporidium parvum* glycoproteins gp40 and gp15. Infection & Immunity 68(7): 4108-4116.

40. Chalmers R, Hadfield S, Jackson C, Elwin K, Xiao L, Hunter P. 2008. Geographic Linkage and Variation on *Cryptosporidium hominis*. Emerging Infectious Disease 14(3):496-498.

41. Chalmers R, Sturdee A, Bull S, Miller A, Wright E. 1997. The prevalence of *Cryptosporidium parvum* and *C. muris* in *Mus domesticus, Apodemus sylvaticus* and *Clethrionomys glareolus* in agricultural system. Parasitology Res. 83:478-482.

42. Chappell C, Okhuysen P, Sterling R, DuPont H. 1995. *Cryptosporidium parvum*: intensity of infection and oocyst excretion patterns in healthy volunteers. J. Infect Dis. 173(1):232-6.

43. Chou, P.Y. & Fasman, G.D. 1974. Prediction of protein conformation. Biochemistry 13:222–245.

44. Che Y, Chen X. 2003. A multiplexing single nucleotide polymorphism typing method based on restriction-enzyme-mediated single-base extension and capillary electrophoresis. Analytical Biochemistry 329(2):220-229.

45. Chin J. 2000. Control of Communicable Diseases Manual; 17th Edition. United Book Press.

46. Combes C. 2001. Parasitism; the Ecology and Evolutions of Intimate Interactions. University of Chicago Press.

47. Conner J, Hartl D. 2004. A Primer of Ecological Genetics. Sinauer Associates. Sunderland, Massachusetts, U.S.A.

48. Cortes A, Mellombo M, Mueller I, Benet A, Reeder J, Anders R. 2003. Geographical Structure of Diversity and Differences between Symptomatic and Asymptomatic Infections for *Plasmodium falciprium* Vaccine Candidate AMA1. Infection & Immunity 71(3):1416-1426.

49. Culley T, Wallace L, Gengler-Nowak K, Crawford D. 2001. A comparison of two methods of calculating Gst, a genetic measure of population differentiation. American J. of Botany 89:460-465.

50. Current W, Reese N, Weinstein W, et al. 1983. Human cryptosporidiosis in immunocompetent and immunodeficient persons. Studies of an outbreak and experimental transmission. New England J. of Medicine 308: 1252-1257.

51. Dawson J, Weinger J, Engelman D. 2002. Motifs of Serine and Threonine can Drive Association of Transmembrane Helices. J. of Molecular Biology 316:799-805.

52. Deng M, Templeton T, Abrahamsen J, et al. 2002. *C. parvum* genes containing thrombospondin type-1 domains. Infection & Immunity 70(12):6987-6995.

53. Denton H, Brown S, Coombs G, et al. 1996. Comparison of the phosphofructokinase and pyruvate kinase activities of *C. parvum, E. tenella* and *T. gondii*. Mol. and Biochem. Parasitology 76:23-29.

54. Deonier R, Tavare S, Waterman M. 2005. Computational Genome Analysis. Springer Science.

55. Dronamraju K. 2004. Infectious Disease and Host-Pathogen Evolution. Cambridge University Press. Cambridge, United Kingdom.

56. Dubey J, Speer C, Fayer R, et al. 1990. Cryptosporidiosis of man and animals. Boston: CRC Press, 1990:1-199.

57. Dupont H, Chappell C, Sterling C, Jakubowski W, et al. 1995. The infectivity of *Cryptosporidium parvum* in healthy volunteers. New England J. Medicine 332:855-926a.

58. El-Sayed N, Myler P, Blandin G, Hall N, et al. 2005. Comparative Genomics of Trypanosomatid Parasitic Protozoa. Science 309:408-409.

59. Environment Canada. www.ec.gc.ca/environment

60. Etkin N. 2003. The co-evolution of people, plants, and parasites: biological and cultural adaptations to malaria. Proceedings of the Nutrition Society 62:311-317.

61. Fayer R, Andrews C, Ungar B, Blagburn B. 1989. Efficacy of hyper immune bovine colostrums for prophylaxis of cryptosporidiosis in neonatal calves. J. of Parasitology 75: 393-397.

62. Fayer R, Morgan U, Upton S. 2000. Epidemiology of Crypto: transmission, detection and identification. International J. of Parasitology 30:1305-1322.

63. Fayer R, Speer C, Dubey J. 1997. The general biology of *Cryptosporidium*: 1-41. *In* R. Fayer (ed.), *Cryptosporidium* and cryptosporidiosis. CRC Press, Boca Raton, Fla.

64. Fayer R, Unger B. 1986. *Cryptosporidium spp* and Cryptosporidiosis. Microbiology reviews 50:458-483.

65. Feng X, Rich S, Tzipori S, Widmer G. 2002. Experimental evidence for genetic recombination in the opportunistic pathogen *Cryptosporidium parvum*. Mol. and Bio. Parasitology 119: 55-62.

66. Ferrer-Costa C, Orozco M, de le Cruz X. 2001. Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. J. of Molecular Biology 315(4):771-786.

67. Fisher M, Koenig G, White T, Taylor J. 2000. Pathogenic Clones versus Environmentally Driven Population Increase: Analysis of an Epidemic of the Human Fungal Pathogen *Coccodioides immitis*. J. of Clinical Microbiology 38(2): 807-813.

68. Forsdyke D. 2006. Evolutionary Bioinformatics. Springer Science. New York, New York, U.S.A.

69. Gao L, Ge, S, Hong D. 2000. Low Levels of Genetic Diversity within Populations and High Differentiation Among Populations of Wild Rice, *Oryza Granulata* Nees et Arn. Ex. Watt., From China. International J. of Plant Science 161(4):691-697.

70. Garnier, Osguthorpe and Robson. 1978. J. of Molecular Biology 120:97-120.

71. Gatei W, Greensill J, Hart A, et al. 2003. Molecular Analysis of the 18s rRNA Gene of *Cryptosporidium* Parasites from Patients with or without Human Immunodeficiency Virus Infections Living in Kenya, Malawi, Brazil, the United Kingdom, and Vietnam. J. of Clinical Microbiology 41(4): 1458-1462.

72. Gasser R, Abs EL-Osta Y, Chalmers R. 2003. Electrophoretic Analysis of Genetic Variability within *Cryptosporidium parvum* from Imported and Autochthonous Cases of Human Cryptosporidiosis in the United Kingdom. App. and Environmental Microbiology 69(5): 2719-2730.

73. Gaur D, Mayer G, Miller L. 2004. Parasite ligand-host receptor interactions during invasion of erythrocytes by *Plasmodium* merozoites. International J. for Parasitology 34(13, 14):1413-1429.

74. Gavrilescu C, Denkers E. 2003. Apoptosis and the balance of homeostatic and pathologic responses to protozoan infection. Infection & Immunity 71(11):6109-6115.

75. Glaberman S, Moore J, Xiao L, et al. 2002. Three Drinking Water-Associated Cryptosporidiosis Outbreaks, Northern Ireland. Emerging Infectious Diseases 8(6): 631-633.

76. Glaser C, Safrin S, Reingold A, Newman T. 1998. Association between *Cryptosporidium* infection and animal exposure in HIV-infected individuals. J. Acquired Immune Deficiency Syndrome. Hum. Retrovirology 17:79-82.

77. Gibbons C, Gazzard B, Awad-El-Kariem F, et al. 1998. Correlation between markers of strain variation in *Cryptosporidium parvum*: Evidence of clonality. Parasitology International 47: 139-147.

78. Goodgame R. 1996. Understanding intestinal spore forming protozoa: *cryptosporidia, microsporidia, Isospora* and *Cyclospora*. Ann intern. Medicine 124(4):429-441

79. Goudet J. FSTAT (Version 1.2): A Computer Program to Calculate F-Statistics. www.2.unil.ch/popgen/softwares/fstat.htm

80. Graczyk TK, Fayer R, Cranfield MR, Owens R. 1997. Infectivity of *Cryptosporidium parvum* oocysts is retained upon intestinal passage through a migratory water-fowl species (Canada goose, *Branta canadensis*). J. Parasitology 83(1):111-4.

81. Graur D, Li W. 2000. Fundamentals of Molecular Evolution; 2nd Edition. Sinauer Associates.

82. Grinberg A, Lopez-Villalobos N, Pomroy W, Widmer G, Smith H, Tait A. 2008. Host-shaped segregation of the *Cryptosporidium parvum* multilocus genotype repertoire. Epidemiology and Infection 136:273-278.

83. Gut I. 2001. Automation in genotyping of single nucleotide polymorphisms. Human Mutations 17:475-492.

84. Gutacker M, Smoot J, Musser J., et al. 2002. Genome-wide Analysis of Synonymous Single Nucleotide Polymorphisms in *Mycobacterium tuberculosis* Complex Organisms: Resolution of Genetic Relationship among Closely Related Strains. Genetics 162:1533-1543.

85. Guyot K, Follet-Dumoulin E, Dei-Cas E, et al. 2001. Molecular Characterization of *Cryptosporidium* Isolates Obtained from Humans in France. J. of Clinical Microbiology 39(10): 3472-3480.

86. Haas C, Rose J. 1994. Reconciliation of microbial risk assessment and epidemiology: the case of the Milwaukee outbreak. In: Proceedings of the 1994 Conference of the American Water Works Association – water quality: 517-523.

87. Harrus S, Baneth G. 2005. Drivers for the emergence and re-emergence of vector-borne protozoal and bacterial diseases. International J. for Parasitology 35(11-12):1309-1318.

88. Hanski I, Gaggiotti O. 2004. Ecology, Genetics, and Evolutions of Metapopulations. Elsevier Academic Press.

89. Hay S, Guerra C, Tatem A, Noor A, Snow R. 2004. The global distribution and population at risk of malaria: past, present and future. The Lancet, Infectious Disease 4:327-336.

90. Health Canada. 2001. Waterborne Cryptosporidiosis Outbreak, North Battleford, Saskatchewan, Spring 2001. Canadian Communicable Disease Report 27(22): 185-192.

91. Heuser V, Kuenzi P, Rottenberg S. 2001. Inhibition of apoptosis by intracellular protozoan parasites. International J. for Parasitology 31(11):1166-1172.

92. Hey J. 1999. Parasite populations: The puzzle of *Plasmodium*. Current Biology 9(15): R565-R566.

93. Hoar B, Atwill E, Elmi C, Farver T. 2001. An examination of risk factors associated with beef cattle shedding pathogens of potential zoonotic concern. Epidemiology Infection 127:147-155.

94. Hojlyng N, Holten-Anderson W, Jepsen S. 1987. Cryptosporidiosis: a case of airborne transmission. Lancet 2: 271-272.

95. Hooda P, Edwards A, Miller A. 2000. A review of water quality concerns in livestock farming areas. Sci. Total Environment 250: 143-167.

96. Hopp, T. P., K. R. Woods. 1981. Prediction of protein antigenic determinants from amino acid sequences. Proc. Natl. Acad. Sci. USA 78:3824.

97. Horton R, Moran L, Ochs R, Rawn D, Scrimgeour G. 1996. Principles of Biochemistry; 2nd Edition. Prentice Hall.

98. Hunter P, Nichols G. 2002. Epidemiology and Clinical Features of *Cryptosporidium* Infection in Immunocompromised Patients. Clinical Microbiology Reviews 15(1): 145-154.

99. Hunter P, Quigly C. 1998. Investigation of an outbreak of cryptosporidiosis associated with treated surface water finds limits to the value of case control studies. Communicable Disease and Public Health 1(4): 234-238.

100. Isaac-Renton J, Blatherwick J, Robertson W, et al. 1999. Epidemic and endemic seroprevalance of antibodies to *Cryptosporidium* and *Giardia* in residents of three communities with different drinking water supplies. American J. of Tropical Medicine Hygiene 60(4): 578-583.

101. Jakubowski W. 1995. *Crypto* and *Giardia*: the details. Safe drinking water seminar, US EPA.

102. Jackson J, Tinsley R. 2005. Geographic and within population structure in variable resistance to parasite species and strains in a vertebrate host. International J. for Parasitology 35:29-37.

103. Jameson, BA and Wolf, H. 1988. The antigenic index: a novel algorithm for predicting antigenic determinants. Bioinformatics 4:181-186.

104. Joce R, Bruce J, Kiely D, et al. 1991. An outbreak of Cryptosporidiosis associated with a swimming pool. Epidemiology Infection 107:497-508.

105. Joe A, Verdon R, Ward H, et al. 1998. Attachment of *Cryptosporidium parvum* Sporozoites to Human Intestinal Epithelial Cells. Infection & Immunity 66(7): 3429-3432.

106. Jolly C, Vourch C, Robert-Nicoud M, Morimoto R. 1999. Intron-dependent association of splicing factors with active genes. J. of Cell Biology 145(6):1133-1143.

107. Juranek D. 1995. Cryptosporidiosis: sources of infection and guidelines for prevention. Clinical Infectious Disease 21 Supplement 1: S57-S61.

108. Kaiser A, Gottwald A, Maier W, Seitz, H. 2003. Targeting enzymes involved in spermidine metabolism of parasitic protozoa: a possible new strategy for anti-parasitic treatment. Parasitology Research 91:508-516.

109. Keeling, P. 2004. Reduction and compaction in the Genome of the Apicomplexan Parasite *Cryptosporidium parvum.* Developmental Cell: 614-616.

110. Keithly J, Zhu G, Upton S, et al. 1997. Polyamine biosynthesis in *C. parvum* and its implications for chemotherapy. Molecular and Biochemical Parasitology 88:35-42.

111. Keusch G, Joe A, Hamer D, Ward H, et al. 1995. *Cryptosporidia* – who is at risk? Schweiz Med Wochenschr 125(18):899-908.

112. Khan, O. 2003. A review of cryptosporidiosis. Johns Hopkins University.

113. Kimura M. 1983. The Neutral Theory of Molecular Evolution. Cambridge University Press.

114. Kimura M. 1980. A simple method for estimating the evolutionary rates of base substitutions through comparative studies of sequence analysis. J. of Molecular Evolution 16(2):111-120.

115. Koji Lum J, Kaneko A, Tanabe K, Takahashi N, Bjorkman A, Kobayakawa T. 2003. Malaria dispersal among islands: human mediated *Plasmodium falciparum* gene flow in Vanuatu, Melanesia. Acta Tropica 90:181-185.

116. Kyte, J. and Doolittle, R. 1982. A simple method for displaying the hydropathic character of a protein. J. Molecular Biology 157:105-132.

117. Landfear S, Ullman B, Carter N, Sanchez M. 2004. Nucleoside and nucleobase transporters in parasitic protozoa. Eukaryotic Cell 3(2):245-254.

118. LaGier M, Keithly J, Zhu G. 2002. Characterization of a novel transporter from *C. parvum.* International J. for Parasitology 32:877-887.

119. Langer R, Riggs M. 1999. C. *parvum* apical complex CSL contains a sporozoites ligand for intestinal epithelial cells. Infection & Immunity 67(10):5282-5291.

120. Leav B, Mackay M, Ward H, et al. 2002. Analysis of Sequence Diversity at the Highly Polymorphic Cpgp40/15 Locus among *Cryptosporidium* Isolates from Human Immunodeficiency Virus-Infected Children in South Africa. Infection & Immunity 70(7): 3881-3890.

121. LeChevallier, M.W. et al., Occurrence of *Giardia* and *Cryptosporidium spp.* in surface water supplies. Applied and Environmental Microbiology 57(9): 2610-2616 (1991).

122. LeChevallier M.W. et al., *Giardia* and *Cryptosporidium spp.* in filtered drinking water supplies. Applied and Environmental Microbiology 57(9):2617-2621 (1991).

123. Lederberg J. 1998. Emerging Infections: An Evolutionary Perspective. Emerging Infectious Disease 4(3):366-370.

124. Leng X, Mosier D, Oberst R. 1996. Differentiation of *Cryptosporidium parvum, C. muris* and *C. baileyi* by PCR-RFLP analysis of the 18s rRNA gene. Veterinary Parasitology 62 (1 and 2):1-7.

125. Leoni F, Mallon M, Smith H, Tait A, McLauchlin J. 2007. Multilocus Analysis of *Cryptosporidium hominis* and *Cryptosporidium parvum* Isolates from Sporadic and Outbreak-Related Human Cases and *C. parvum* Isolates from Sporadic Livestock Cases in the United Kingdom. J. of Clinical Microbiology 45(10):3286-3294.

126. Lewis, P. O., and Zaykin, D. 2001. Genetic Data Analysis: Computer program for the analysis of allelic data. Version 1.0 (d16c). Free program distributed by the authors over the internet from http://lewis.eeb.uconn.edu/lewishome/software.

127. Li J, Collins W, McCutchan T, et al. 2001. Geographic Subdivision of the Range of the Malaria Parasite, *Plasmodium vivax*. Emerging Infectious Diseases 7(1): Synopsis. 6123.

128. Liberles D, Wayne. 2002. Tracking adaptive evolutionary events in genomic sequences. Genome Biology 3(6):1018.1-1018.4.

129. Lynch M. 2002. Intron evolution as a population genetic process. Proceedings National Academy Science USA 99(9):6118-6124

130. MacKenzie W, Hoxie N, Davis J, et al. 1994. A Massive Outbreak in Milwaukee of *Cryptosporidium* Infection Transmitted through the Public Water Supply. New England J. of Medicine 331(3): 161-167.

131. MacPherson C. 2005. Human behavior and the epidemiology of parasitic zoonoses. International J. for Parasitology 35(11-12):1319-1331.

132. Madern D, Cai X, Abrahamsen M, Zhu G. 2003. Evolution of *C. parvum* lactate dehydrogenase from malate dehydrogenase by a very recent event of gene duplication. Molecular Biology Evolution 21(3):489-497.

133. Madigan M, Martinko J, Parker J. 2003. Brock Biology of Microorganisms. Prentice Hall.

134. Mallon, M., MacLeod A, Wastling J, Smith H, Reilly B, Tait A. 2003. Population structures and the role of genetic exchange in the zoonotic pathogen *Cryptosporidium parvum*. J. Molecular Evolution 56:407–417.

135. Mallon M, MacLeod A, Tait A, et al. 2003. Multi-locus genotyping of *Cryptosporidium parvum* Type 2: population genetics and sub-structuring. Infect. Genetic Evolution 3: 207-218.

136. Marr J, Nilsen T, Komuniecki R. 2003. Molecular Medical Parasitology. Elsevier Academic Press.

137. Marshall M, Naumovitz D, Ortega Y, Sterling C. 1997. Waterborne Protozoan Pathogens. Clinical Microbiology Reviews, 10:67-85.

138. McCole D, Eckman L, Laurent F, Kagnoff M. 2000. Intestinal epithelial cell apoptosis following *Cryptosporidium parvum* infection. Infection & Immunity 68(3): 1710-1713.

139. Mclaughlin J, Amar C, Pedraza-Diaz S, Nichols G. 2000. Molecular Epidemiological Analysis of *Cryptosporidium spp*. In the United Kingdom: Results of Genotyping *Cryptosporidium spp*. In 1,705 Fecal Samples from humans and 105 Fecal Samples from Livestock Animals. J. of Clinical Microbiology 38(11): 3984-3990.

140. McLaughlin J, Pedraza-Diaz S, Amar-Hoetzeneder, C, Nichols G. 1999. Genetic Characterization of *Cryptosporidium* Strains from 218 Patients with Diarrhea Diagnosed as Having Sporadic Cryptosporidiosis. J. of Clinical Microbiology 37(10): 3153-3158.

141. McPherson M, Moller S. 2000. PCR. BIOS Scientific Publishers.

142. Mele R, Morales M, Tosini F, Pozio E. 2004. *C. parvum* at different developmental stages modulates host cell apoptosis in vitro. Infection & Immunity 72(10):6061-6067.

143. Millard P, Gensheimer K, Addiss D, et al. 1994. An outbreak of Cryptosporidiosis from fresh-pressed apple cider. JAMA 272:1592-1596.

144. Morgan U. 2000. Detection and characterization of parasites causing emerging zoonoses. International J. for Parasitology 30:1407-1421.

145. Morgan-Ryan U, Fall A, Xiao L, et al. *Cryptosporidium hominis* n. sp. (Apicomplexa: *Cryptosporidiidae*) from *Homo sapiens*. J. of Eukaryotic Microbiology 49(6): 433-440.

146. Morgan U, Sargent K, Thompson R, et al. 1998. Molecular characterization of *Cryptosporidium* from various hosts. Parasitology 117: 31-37.

147. Morgan U, Weber R, Deplazes P, et al. Molecular characterization of *Cryptosporidium* Isolates Obtained from Human Immunodeficiency Virus-Infected Individuals Living in Switzerland, Kenya and the United States. J. of Clinical Microbiology 38(3): 1180-1183.

148. Morgan U, Xiao L, Thompson A, et al. 1999. Variation in *Cryptosporidium*: towards a taxonomic revision of the genus. International J. for Parasitology 29: 1733-1751.

149. Mu J, Awadalla P, Su X, et al. 2007. Genome-wide variation and identification of vaccine targets in the *Plasmodium falciparum* genome. Nature Genetics 39(1):126-130.

150. Musser J. 1996. Molecular population genetic analysis of emerged bacterial pathogens: selected insights. Emerging Infectious Disease 2:1-17.

151. Navin TR, Juranek DD. 1984. Cryptosporidiosis: clinical, epidemiologic, and parasitological review. Reviews Infectious Disease 6(3):313-27.

152. Nei M. 1987. Molecular Evolutionary Genetics. Columbia University Press.

153. Nime F, Burek D, Yardley J, et al. 1976. Acute enterocolitis in a human being infected with the protozoan *Cryptosporidium*. Gastroenterology 70: 592-598.

154. Nelson K, Williams C, Graham N. 2001. Infectious Disease Epidemiology; Theory and Practice. Aspen Publications.

155. Ngouanesavanh, T, Guyot K, Banuls A, et al. 2006. *Cryptosporidium* population genetics: evidence of clonality in isolates from France and Haiti. J. Eukaryotic Microbiology 53:S33–S36.

156. O'Donoghue P. 1995. *Cryptosporidium* and cryptosporidiosis in man and animals. International J. of Parasitology 25: 139-195.

157. Okhuysen P, Chappell C, Crabb J, Sterling C, DuPont H. 1999. Virulence of three distinct *Cryptosporidium parvum* isolates for healthy adults. J. Infectious Disease 180:1275–1281.

158. Okhuysen P, Chappell C, Sterling C, Jakubowski W, DuPont H. 1998. Susceptibility and serologic response of healthy adults to re-infection with *Cryptosporidium parvum*. Infection & Immunity 66(2):441-3.

159. Oleksiak M, Churchill G, Crawford D. 2002. Variation in gene expression within and among natural populations. Nature Genetics 32:261-266.

160. Ohno, S. 1984. Birth of a unique enzyme from an alternative reading frame of the preexisted, internally repetitious coding sequence. Proceedings National Academy Science USA, 81: 2421–2425.

161. O'Neil R, Lilien R, Donald B, et al. 2003. The crystal structure of DHFR-thymidylate synthase from *C. hominis* reveals a novel architecture for the bi-functional enzyme. J. of Eukaryotic Microbiology, 50(s1): 555-556.

162. Ong C, Eisler D, Isaac-Renton J, et al. 2002. Novel *Cryptosporidium* Genotypes in Sporadic Cryptosporidiosis Cases: First Report of Human Infections with a Cervine Genotype. Emerging Infectious Diseases 8(3): 263-268.

163. Ong C, Isaac-Renton J, Fyfe M, et al. 1999. Molecular epidemiology of Cryptosporidiosis Outbreaks and Transmission in British Columbia, Canada. American J. Tropical Medicine Hygiene 61(1): 63-69.

164. Oura C, Asiimwe B, Weir W, Lubega G, Tait A. 2005. Population genetic analysis and sub-structuring of *Theileria parva* in Uganda. Mol. and Biochemical Parasitology 140(2): 229-239.

165. Patel S, Pedraza-Diaz S, Mclaughlin J, Casemore D. 1997. Molecular Characterization of *Cryptosporidium parvum* from two large suspected waterborne outbreaks. Communicable Disease and Public Health 1(4): 231-233.

166. Pearson T, Busch J, Keim P. 2004. Phylogenetic discovery bias in *Bacillus anthracis* using single-nucleotide polymorphisms from whole-genome sequencing. Proceedings National Academy Science 101(37):13536-13541.

167. Patz J, Gracyzk t, Geller N, Vittor A. 2000. Effects of environmental change of emerging parasitic diseases. International J. for Parasitology 30:1395-1405.

168. Pedraza-Diaz S, Amar C, Nichols G, McLaughlin J. 2001. Nested Polymerase Chain Reaction for Amplification of the *Cryptosporidium* Oocyst Wall Protein Gene. Emerging Infectious Diseases 7(1): 49-56.

169. Peng M, Matos O, Gatei W, Xiao L. 2001. A Comparison of *Cryptosporidium* Sub-genotypes from several Geographic Regions. J. Eukaryotic Microbiology Supp.:28-29.

170. Peng M, Xiao L, Beard C, et al. 1997. Genetic Polymorphism among *Cryptosporidium parvum* Isolates: Evidence of Two Distinct Human Transmission Cycles. Emerging Infectious Diseases 3(4):567-573.

171. Pereira S, Ramirez N, Xiao L, Ward L. 2002. Pathogenesis of Human and Bovine *Cryptosporidium parvum* in Gnotobiotic Pigs. J. of Infectious Diseases 186: 715-718.

172. Perryman L, Jasmar D, Riggs M, et al. 1996. A cloned gene of the *Cryptosporidium parvum* encodes neutralization-sensitive epitopes. Molecular and Biochemical Parasitology 80:137-147.

173. Perz J, Le Blancq S. 2001. *Cryptosporidium parvum* infection involving novel genotypes in wildlife from lower New York State. App. and Environmental Microbiology 67: 1154-1162.

174. Petersen C. 1992. Cryptosporidiosis in patients infected with the human immunodeficiency virus. Clinical Infectious Disease 15: 903-909.

175. Petersen C, Gut J, Leech J. 1992. Characterization of a >900,000-Mr *C. parvum* sporozoites glycoprotein recognized by protective hyper immune bovine colostral immunoglobulin. Infection & Immunity 60(12):5132-5138.

176. Polley L. 2005. Navigating parasite webs and parasite flow: Emerging and re-emerging parasitic zoonoses of wildlife origin. International J. for Parasitology 35(11-12):1279-1294.

177. Pozio E, Gomez M, Barbieri F, La Rosa G. 1992. *Cryptosporidium*: different behavior in calves of isolates of human origin. Transactions Royal Society, Tropical Medicine Hygiene 86: 636-638.

178. Priest J, Li A, Khan M, Arrowood M, Lammie P, Ong C, Roberts J, Isaac-Renton J. 2001. Enzyme immunoassay detection of antigen-specific immunoglobulin g antibodies in longitudinal serum samples from patients with cryptosporidiosis. Clinical Diagnostic Lab Immunology (2):415-23.

179. Priest J, Kwon J, Lammie P, et al. 1999. Detection by Enzyme Immunoassay of Serum Immunoglobulin G Antibodies That Recognize Specific *Cryptosporidium parvum* Antigens. J. of Clinical Microbiology 37(5): 1385-1392.

180. Quiroz E, Bern J, Lal a, et al. 2000. An outbreak of cryptosporidiosis linked to a food handler. J. of Infectious Disease 181: 695-700.

181. Ramirez N, Ward L, Sreevatsan S. 2004. A review of the biology and epidemiology of cryptosporidiosis in humans and animals. Microbes and Infection 6: 773-785.

182. Reid S, Hoe N, Smoot M, Musser J. 2001. Group A streptococcus: allelic variation, population genetics, and host pathogen interactions. J. Clinical Investigations 107:393-399.

183. Rich S, Hudson R, Ayala F. 1997. *Plasmodium falciparum* antigenic diversity: Evidence of clonal population structure. Proceedings National Acadamy Science 94:13040-13045.

184. Rickard L, Siefker C, Boyle C, Gentz E. 1999. The prevalence of *Cryptosporidium* and *Giardia spp*. in fecal samples from free-ranging white-tailed deer *(Odocoileus virginianus*) in the southeastern United States. J. Veterinary Diagn. Investigations 11:65-72.

185. Riley L. 2004. Molecular Epidemiology of Infectious Disease. American Society for Microbiology Publishing. Washington, DC, U.S.A.

186. Rochelle P, Jutras E, Atwill E, De Leon R, Stewart M. 1999. Polymorphisms in the beta-tubulin gene of *Cryptosporidium parvum* differentiate between isolates based on animal host but not geographic origin. J. of Parasitology 85: 986-989.

187. Roderic D. 2005. TreeViewX Version 0.5.0. www.darwin.zoology.gla.ac.uk/~rpage/treeviewx

188. Roe A, Sperling F. 2007. Population structure and species boundary delimitation of cryptic *Dioryctria* moths: an integrative approach. Molecular Ecology 16:3617-3633.

189. Ryan M, Sundberg J, Sauerschell R, Todd K. 1986. *Cryptosporidium* in wild cottontail rabbit (*Sylvilagus floridanus*). J. Wildlife Disease 22:267.

190. Sasahara T, Maruyama H, Inoue M, et al. 2003. Apoptosis of intestinal crypt epithelium after *C. parvum* infection. J. of Infectious Chemotherapy 9:278-281.

191. Schaechter M, Engleberg C, Eisenstein B, Medoff G. 1999. Mechanisms of Microbial Disease; 3rd Edition. Lippincott, Williams and Wilkins.

192. Schork N, Fallin D, Lanchbury J. 2000. Single nucleotide polymorphism and the future of genetic epidemiology. Clinical Genetics 58:250-264.

193. Sestak K, Ward L, Sheoran A, Feng X, Akiyoshi D, Tzipori S. 2002. Variability among *Cryptosporidium parvum* genotype 1 and 2 immunodominant surface glycoproteins. Parasite Immunology 24:213-219.

194. Shankhar S, Park Y. 2002. Genetic Structure of Six Korean Tea Populations as Revealed by RAPD-PCR Markers. Plant Genetic Resources 42:594-601.

195. Shirley M, Harvey D. 2000. A Genetic Linkage Map of the Apicomplexan Protozoan Parasite *Eimeria tenella*. Genome research, www.genome.org.

196. Simpson V. 2002. Wild animals as reservoirs of infectious diseases. Veterinary Journal 163:128-146.

197. Sischo W. Atwill E, Lanyon L, George J. 2000. *Cryptosporidia* on dairy farms and the role these farms may have in contaminating surface water supplies in the northeastern United States. Prev. Veterinary Medicine 43:253-2667.

198. Slifko T, Smith H, Rose J. 2000. Emerging parasite zoonoses associated with water and food. International J. for Parasitology 30:1379-1393.

199. Smith LM, Priest JW, Lammie PJ, Mead JR. 2001. Human T and B cell immunoreactivity to a recombinant 23-kDa *Cryptosporidium parvum* antigen. J. Parasitol. 87(3):704-7.

200. Sorvillo F, Fujioka K, Mascola R, et al. 1992. Swimming-associated Cryptosporidiosis. American J. Public Health 82(5):742-744.

201. Spano F, Casemore D, Cristani A, et al. 1997. PCR-RFLP analysis of the *Cryptosporidium* oocyst wall protein (COWP) gene discriminates between *C. parvum* isolates of human and animal origin. FEMS Microbiology Letters 150: 209-217.

202. Spano F, Putignani L, Widmer G, et al. 1998. Multilocus Genotypic Analysis of *Cryptosporidium parvum* Isolates from Different Hosts and Geographical Origins. J. of Clinical Microbiology 36(11): 3255-3259.

203. Sreter T, Kovacs G, Varga I, et al. 2000. Morphologic, Host Specificity, and Molecular Characterization of a Hungarian *Cryptosporidium meleagridis* Isolate. App. and Environmental Microbiology 66(2): 735-738.

204. Straub T, Daly D, Chandler D, et al. 2002. Genotyping *Cryptosporidium parvum* with an hsp70 Single-Nucleotide Polymorphism Microarray. App. and Environmental Microbiology 68(4): 1817-1826.

205. Striepen B, Pruijssers A, Kissinger J, et al. 2004. Gene transfer in the evolution of parasite nucleotide biosynthesis. Proceedings National Academy Science 101(9):3154-3159.

206. Strong W, Gut J, Nelson R. 2000. Cloning and Sequence Analysis of a Highly Polymorphic *Cryptosporidium parvum* Gene Encoding a 60-Kilodalton Glycoprotein and Characterization of Its 15- and 45-Kilodalton Zoite Surface Antigen Products. Infection & Immunity 68(7): 4117-4134.

207. Sturbaum G, Jost H, Sterling C. 2003. Nucleotide changes within three *Cryptosporidium parvum* surface protein encoding genes differentiate genotype 1 from genotype 2 isolates. Molecular and Biological Parasitology 128: 87-90.

208. Sturdee A, Bodley-Tickell, Archer A, Chalmers R. 1993. Long-term study of *Cryptosporidium* prevalence on a lowland farm in the United Kingdom. Vet. Parasitology 45: 209-213.

209. Sturdee A, Chalmers R, Bull S. 1999. Detection of *Cryptosporidium* oocysts in wild mammals of mainland Britain. Veterinary Parasitology. 80:273-280.

210. Sulaiman I, Morgan U, Xiao L, et al. 2000. Phylogenetic Relationships of *Cryptosporidium* Parasites Based on the 70-Kilodalton Heat Shock Protein (HSP70) Gene. App. and Environmental Microbiology 66(6): 2385-2391.

211. Sulaiman I, Xiao L, Lal A. 1999. Evaluation of *Cryptosporidium parvum* genotyping techniques. App. and Environmental Microbiology 65: 4431-4435.

212. Sulaiman I, Xiao L, Lal A, et al. 1998. Differentiating human from animal isolates of *Cryptosporidium parvum*. Emerging Infectious Disease 4: 681-685.

213. Sundberg J, Hill D, Ryan M. 1982. Cryptosporidiosis in a gray squirrel. J. American Veterinary Med. Association 181:1420-1422.

214. Tamura K, Nei M, Kumar S. 2004. Prospects for inferring very large phylogenies by using the neighbor-joining method. Proceedings National Academy Science 101(30):11030-11035.

215. Tanabe K, Sakihama N, Kaneko A. 2004. Stable SNPs in Malaria Antigen Genes in Isolated Populations. Science 303: 493.

216. Tanriverdi S, Arslan M, Akiyoshi D, Tzipori S, Widmer G. 2003. Identification of genotypically mixed *Cryptosporidium parvum* populations in humans and calves. Molecular Biochemical Parasitology 130:13–22.

217. Tanriverdi, S, Blain J, Deng B, Ferdig M, Widmer G. 2007.Genetic crosses in the apicomplexan parasite *Cryptosporidium parvum* defines recombination parameters. Molecular Microbiol. 63:1432–1439.

218. Tanriverdi S, Grinberg A, Chalmers M, Widmer G, et al. 2008. Inferences about the Global Population Structures of *Cryptosporidium parvum* and *Cryptosporidium hominis.* App. And Environmental Microbiology 74(23):7227–7234.

219. Tanriverdi, S, Markovics A, Arslan M, Itik A, ShkapV, Widmer G. 2006. Emergence of distinct genotypes of *Cryptosporidium parvum* in structured host populations. App. and Environmental Microbiology 72:2507–2513.

220. Templeton T, Lancto C, Abrahamsen M, et al. The *Cryptosporidium* oocysts wall protein is a member of a multigene family and has homology in *Toxoplasma.* Infection & Immunity 72(2): 980-987.

221. Teodorovic S, Braverman J, Elmendorf H. 2007. Unusually Low Levels of Genetic Variation among *Giardia lamblia* Isolates. Eukaryotic Cell 6(8):1421-1430.

222. Thompson A. 2004. The zoonotic significance and molecular epidemiology of *Giardia* and giardiasis. Veterinary Parasitology 126(1-2):15-35.

223. Tibayrenc M, Kjellbeg F, Ayala F. 1990. A clonal theory of parasitic protozoa: The population structures of *Entamoeba, Giardia, Leishmania, Naegleria, Plasmodium, Trichomonas*, and *Trypanosoma* and their medical and taxonomical consequences. Proceedings National Academy Science 87: 2414-2418.

224. Traub R, Monis P, Robertson I. 2005. Molecular epidemiology: a multidisciplinary approach to understanding parasitic zoonoses. International J. for Parasitology 35(11-12):1295-1307.

225. Turakulav R, Easteal S. 2003. Number of SNPs Loci Needed to Detect Population Structure. Human Hereditary 55:37-45.

226. Tyzzer E. 1907. A sporozoan found in the peptide glands of the common mouse. Proceedings Soc. Experimental Biology Medicine: 12-13.

227. Tzipori S. 1988. Cryptosporidiosis in perspective. Advances in Parasitology 27:63-119.

228. Umejiegon N, Li C, Riera T, et al. 2004. *C. parvum* IMP dehydrogenase. J. of Biological Chemistry 279(39): 40320-40327.

229. Ungar B, Ward D, Fayer R, Quinn C. 1990. Cessation of *Cryptosporidium*-associated diarrhea in an acquired immunodeficiency syndrome patient after treatment with hyper immune bovine colostrum. Gastroenterology 98(2):486-9.

230. Upton S. 2003. Basic biology of *Cryptosporidium.* Parasitology Laboratory, Kansa State University.

231. VanLin L, Pace T, et al. 2001. Interspecies conservation of gene order and intron-exon structure in a genomic locus of high gene density and complexity in *Plasmodium*. Nucleic Acids Research 29(10): 2059-2068.

232. Vasquez J, Gooze L, Nelson C, et al. 1996. Potential antifolate resistance determinants and genotypic variation in the biofunctional dihydrofolate reductase-thymidylate synthase gene from human and bovine isolates of *Cryptosporidium parvum*. Molecular and Bio. Parasitology 79: 153-165.

233. Volkman S, Hartl D, Nilesen K, Winzeler E, et al. 2002. Excess Polymorphisms in Genes for Membrane Proteins in *Plasmodium falciparum*. Science 298:216-218.

234. www.waterunderfire.com University of Lethbridge, Alberta, British Columbia.

235. Weir B. 1990. Genetic Data Analysis. Sinauer Associates Press.

236. Whitaker R, Grogan D, Taylor J. 2003. Geographic Barriers Isolate Endemic Populations of Hyperthermophilic Archaea. Science 301(5635):976-978.

237. Widmer G, Tchack L, Chappell C, Tzipori S. 1998. Sequence Polymorphism in the β-tubulin Gene Reveals Heterogeneous and Variable Population Structures in *Cryptosporidium parvum*. App. and Environmental Microbiology 64(11): 4477-4481.

238. Widmer G, Lin L, Kapur, Feng X, Abrahamsen M. 2002. Genomics and genetics of *Cryptosporidium parvum*: the key to understanding cryptosporidiosis. Microbes and Infection 4:1081-1090.

239. Willocks, Crampin A, Lightfoot N, et al. 1998. A large outbreak of cryptosporidiosis associated with a public water supply from a deep chalk borehole. Communicable Disease and Public Health 1(4): 239-243.

240. World Health Organization (WHO). www.who.org

241. Xiao L, Sulaiman IM, Ryan UM, Zhou L, Atwill ER, Tischler ML, Zhang X, Fayer R, Lal A. Host adaptation and host-parasite co-evolution in *Cryptosporidium*: implications for taxonomy and public health. International J. for Parasitology 32(14):1773-85.

242. Xiao L, Bern H, Checkley J, et al. 2001. Identification of 5 types of *Cryptosporidium* parasites in children in Lima, Peru. J. of Infectious Disease 183: 492-497.

243. Xiao L, Morgan U, Altaf L, et al. 1999. Genetic diversity within *Cryptosporidium parvum* and Related *Cryptosporidium* Species. App. and Environmental Microbiology 65(8): 3386-3391.

244. Xiao L, Morgan U, Lal A, et al. 2000. *Cryptosporidium* systematics and implications for public health. Parasitology Today 16: 287-292.

245. Xu, P. *et al.* 2004. The genome of *Cryptosporidium hominis*. Nature 431, 1107-1112 (2004): Letters to Nature.

246. Zhu G, Marchewka M, Keithly J. 2000. *Cryptosporidium parvum* appears to lack a plastid genome. Microbiology 146: 315-321.

APPENDICES

Appendix 1. *Cryptosporidium hominis* genome characterized in comparison to that of *C. parvum* and *P. falciparum*[1, 245].

Figure A.1

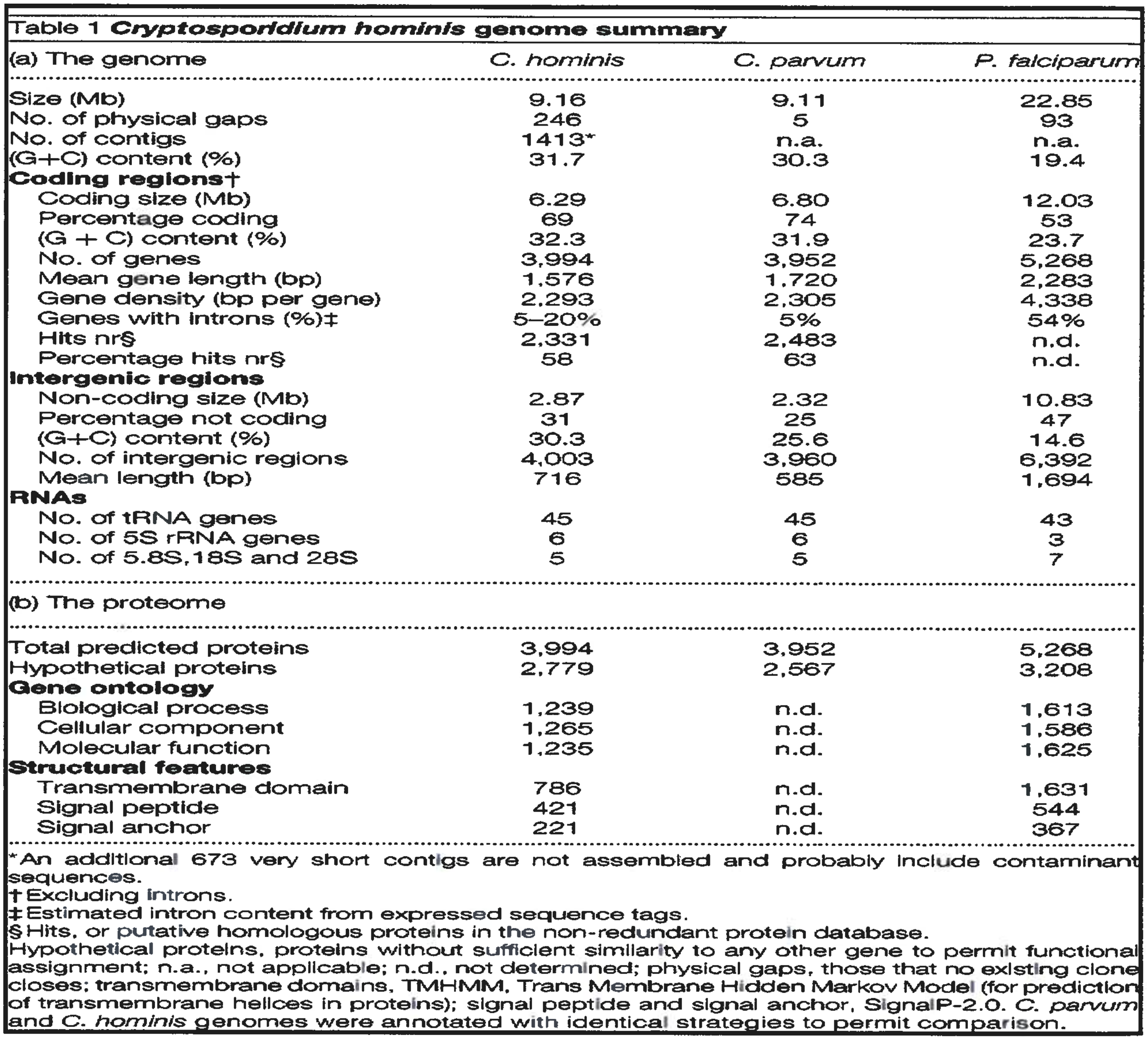


Table 1 ***Cryptosporidium hominis* genome summary**

(a) The genome	*C. hominis*	*C. parvum*	*P. falciparum*
Size (Mb)	9.16	9.11	22.85
No. of physical gaps	246	5	93
No. of contigs	1413*	n.a.	n.a.
(G+C) content (%)	31.7	30.3	19.4
Coding regions†			
Coding size (Mb)	6.29	6.80	12.03
Percentage coding	69	74	53
(G + C) content (%)	32.3	31.9	23.7
No. of genes	3,994	3,952	5,268
Mean gene length (bp)	1,576	1,720	2,283
Gene density (bp per gene)	2,293	2,305	4,338
Genes with introns (%)‡	5–20%	5%	54%
Hits nr§	2,331	2,483	n.d.
Percentage hits nr§	58	63	n.d.
Intergenic regions			
Non-coding size (Mb)	2.87	2.32	10.83
Percentage not coding	31	25	47
(G+C) content (%)	30.3	25.6	14.6
No. of intergenic regions	4,003	3,960	6,392
Mean length (bp)	716	585	1,694
RNAs			
No. of tRNA genes	45	45	43
No. of 5S rRNA genes	6	6	3
No. of 5.8S, 18S and 28S	5	5	7
(b) The proteome			
Total predicted proteins	3,994	3,952	5,268
Hypothetical proteins	2,779	2,567	3,208
Gene ontology			
Biological process	1,239	n.d.	1,613
Cellular component	1,265	n.d.	1,586
Molecular function	1,235	n.d.	1,625
Structural features			
Transmembrane domain	786	n.d.	1,631
Signal peptide	421	n.d.	544
Signal anchor	221	n.d.	367

*An additional 673 very short contigs are not assembled and probably include contaminant sequences.

†Excluding introns.

‡Estimated intron content from expressed sequence tags.

§Hits, or putative homologous proteins in the non-redundant protein database.

Hypothetical proteins, proteins without sufficient similarity to any other gene to permit functional assignment; n.a., not applicable; n.d., not determined; physical gaps, those that no existing clone closes; transmembrane domains, TMHMM, Trans Membrane Hidden Markov Model (for prediction of transmembrane helices in proteins); signal peptide and signal anchor, SignalP-2.0. *C. parvum* and *C. hominis* genomes were annotated with identical strategies to permit comparison.

Figure A.1. *C. hominis* genome summary, obtained from: Ping Xu, Giovanni Widmer, Yingping Wang, Luiz S. Ozaki, Joao M. Alves, Myrna G. Serrano, Daniela Puiu, Patricio Manque, Donna Akiyoshi, Aaron J. Mackey, William R. Pearson, Paul H. Dear, Alan T. Bankier, Darrell L. Peterson, Mitchell S. Abrahamsen, Vivek Kapur, Saul Tzipori and Gregory A. Buck. The genome of *Cryptosporidium hominis* *Nature* 431, 1107-1112(28 October 2004) doi:10.1038/nature02977

Appendix 2. Simplified sketch of spatially structured species.

Figure A.2

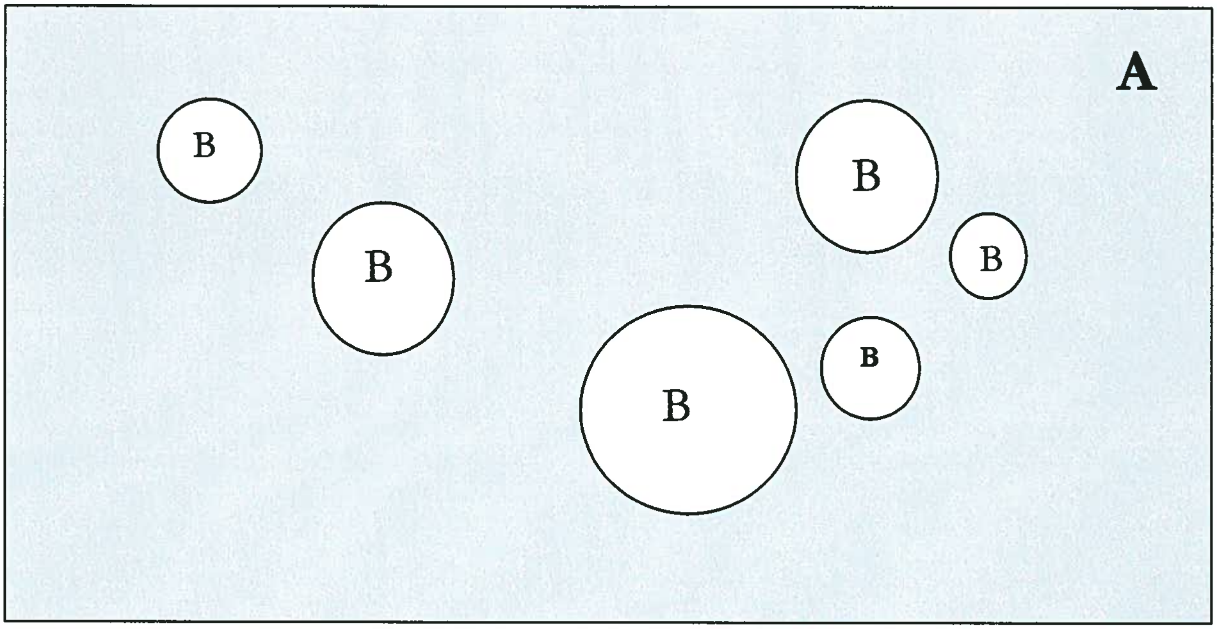


Figure A.2. Simplified sketch of spatially structured species, adapted from Connor and Hartl[47]. Shaded areas, denoted B, are where the organisms live and are considered subpopulation of the metapopulation. They area within which some gene flow occurs is denoted by A and can range greatly in size.

Appendix 3. Applied Biosystems 3130xl automated sequence analyzer.

Figure A.3

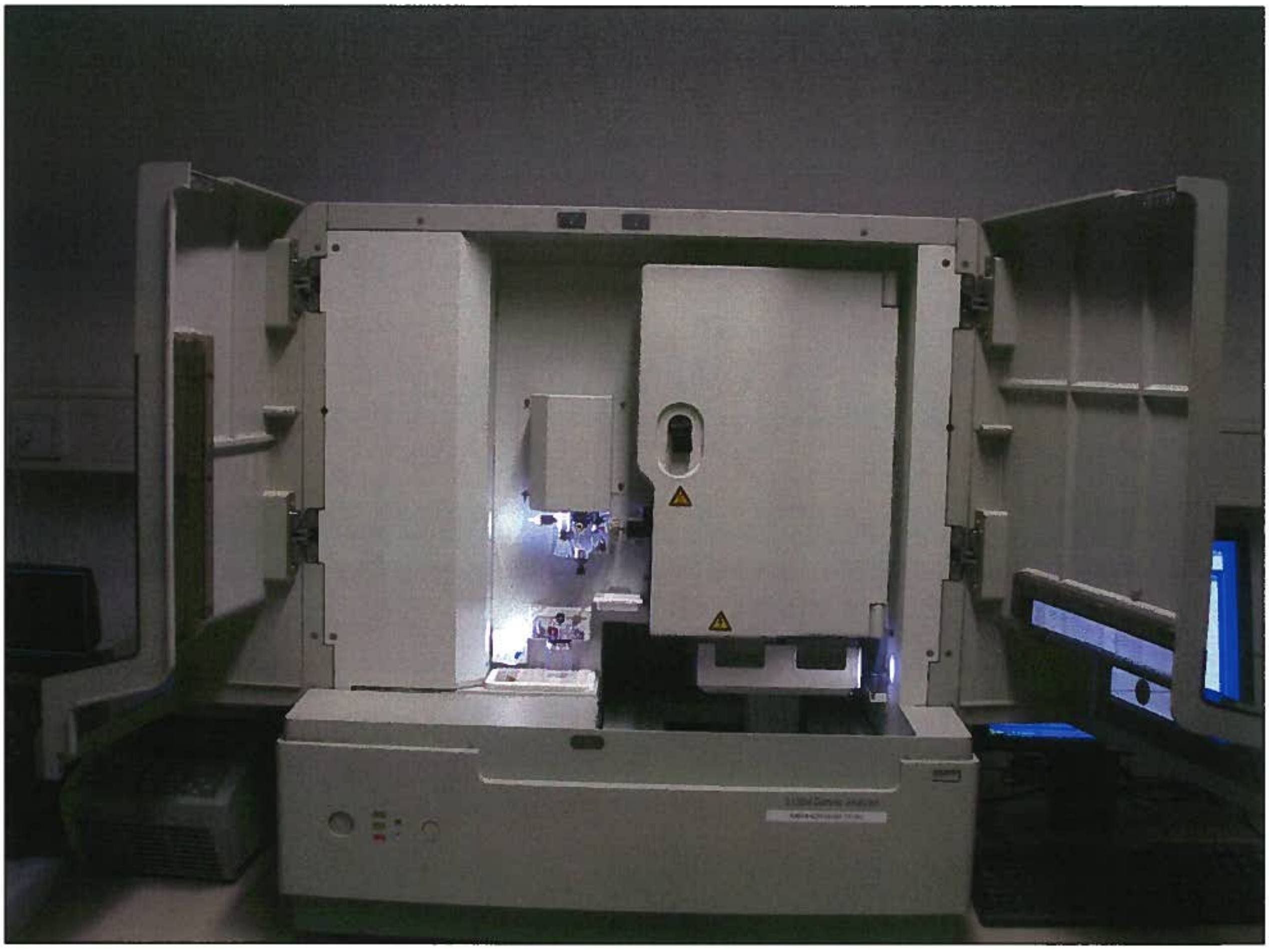
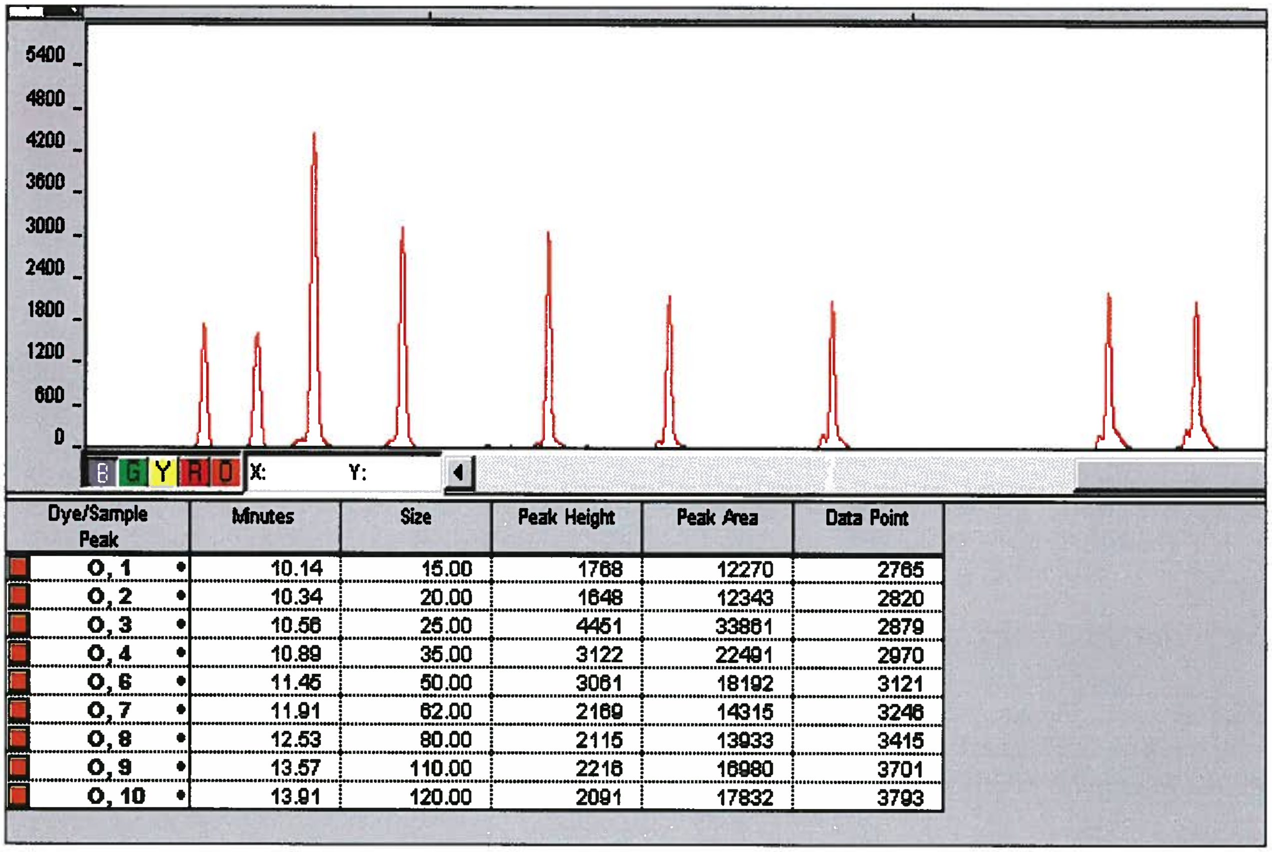


Figure A.3. Automated DNA sequencer, model 3130xl, used for fragment analysis of *Cryptosporidium hominis* subpopulations. With the SNaPshot protocol, running a full 96-well plate, consisting of 96 individual samples, the potential for each sample to be genetically typed for up to 8-12 SNPs could be accomplished in less than 90 minutes.

Appendix 4. Liz120 size standard profile for sizing fragments using SNaPshot single base extension chemistry.

Figure A.4



Dye/Sample Peak		Minutes	Size	Peak Height	Peak Area	Data Point
🟥	O, 1 ●	10.14	15.00	1768	12270	2765
🟥	O, 2 ●	10.34	20.00	1648	12343	2820
🟥	O, 3 ●	10.56	25.00	4451	33861	2879
🟥	O, 4 ●	10.89	35.00	3122	22491	2970
🟥	O, 6 ●	11.46	50.00	3061	18192	3121
🟥	O, 7 ●	11.91	62.00	2169	14315	3246
🟥	O, 8 ●	12.53	80.00	2115	13933	3415
🟥	O, 9 ●	13.57	110.00	2216	16980	3701
🟥	O, 10 ●	13.91	120.00	2091	17832	3793

Figure A.4. Electropherogram of LIZ120 size standard, Dye Set E, with predefined peaks ranging in size from 15nt to 120nt (shown in table below electropherogram). Acted as the reference against which *Cryptosporidium* fragments were sized.

Appendix 5. Example protean profiles: Cp23 and Gp60 loci.

Figure A.5 A,Cp23 profile and B, Gp60 profile.

A. Cp23 ORF bio-physical profile.

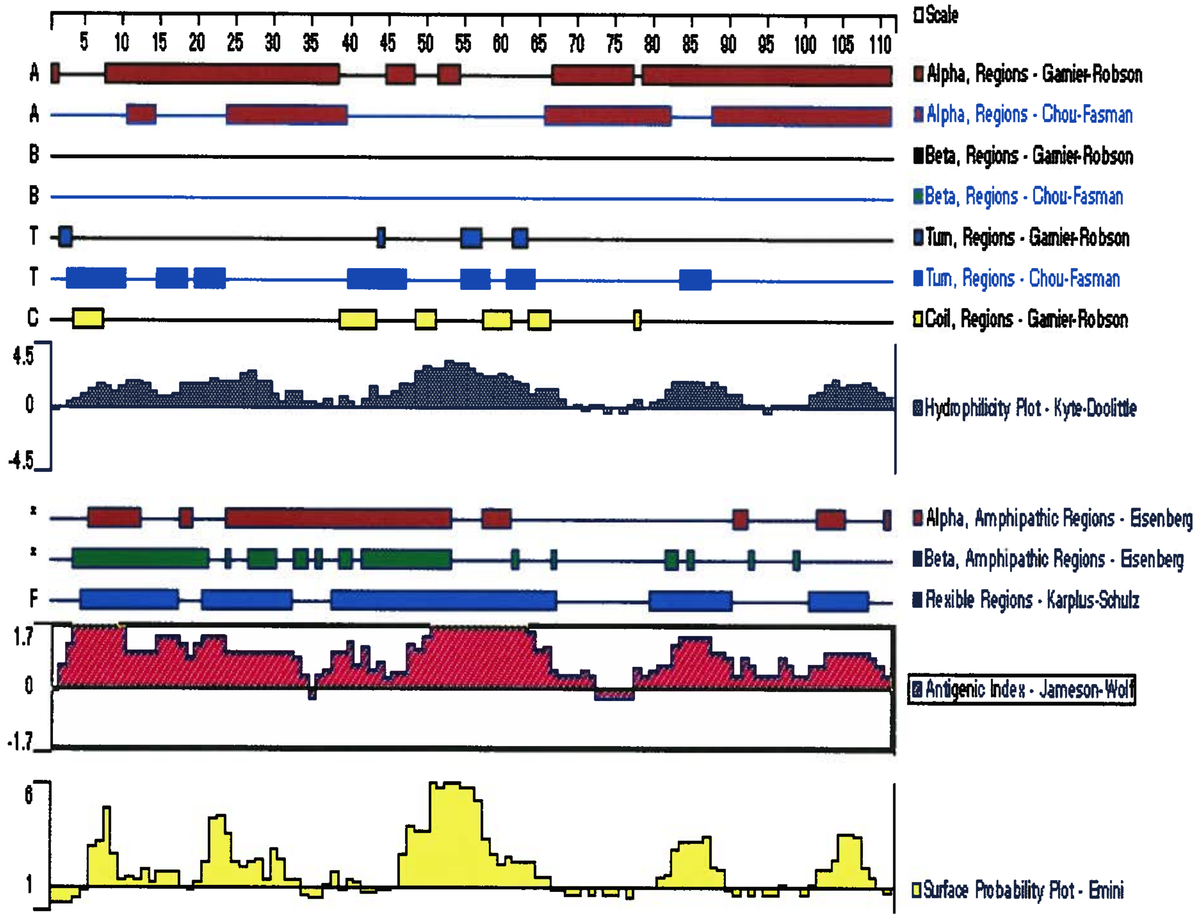


Figure A.5, A. Complete depiction of Cp23 open reading frame in regards to biophysical properties including Jameson Wolf antigenic index (pink), Emini surface probability (bottom: yellow), Kyte Dolittle hydrophilicity plot (blue, middle), and multiple secondary structure predictors (top: red, blue, yellow). Within the Protean program an individual SNP locus could be targeted generating a mathematical output for its position.

Appendix 5, Figure A.5 continued.

B. Gp60 ORF bio-physical profile.

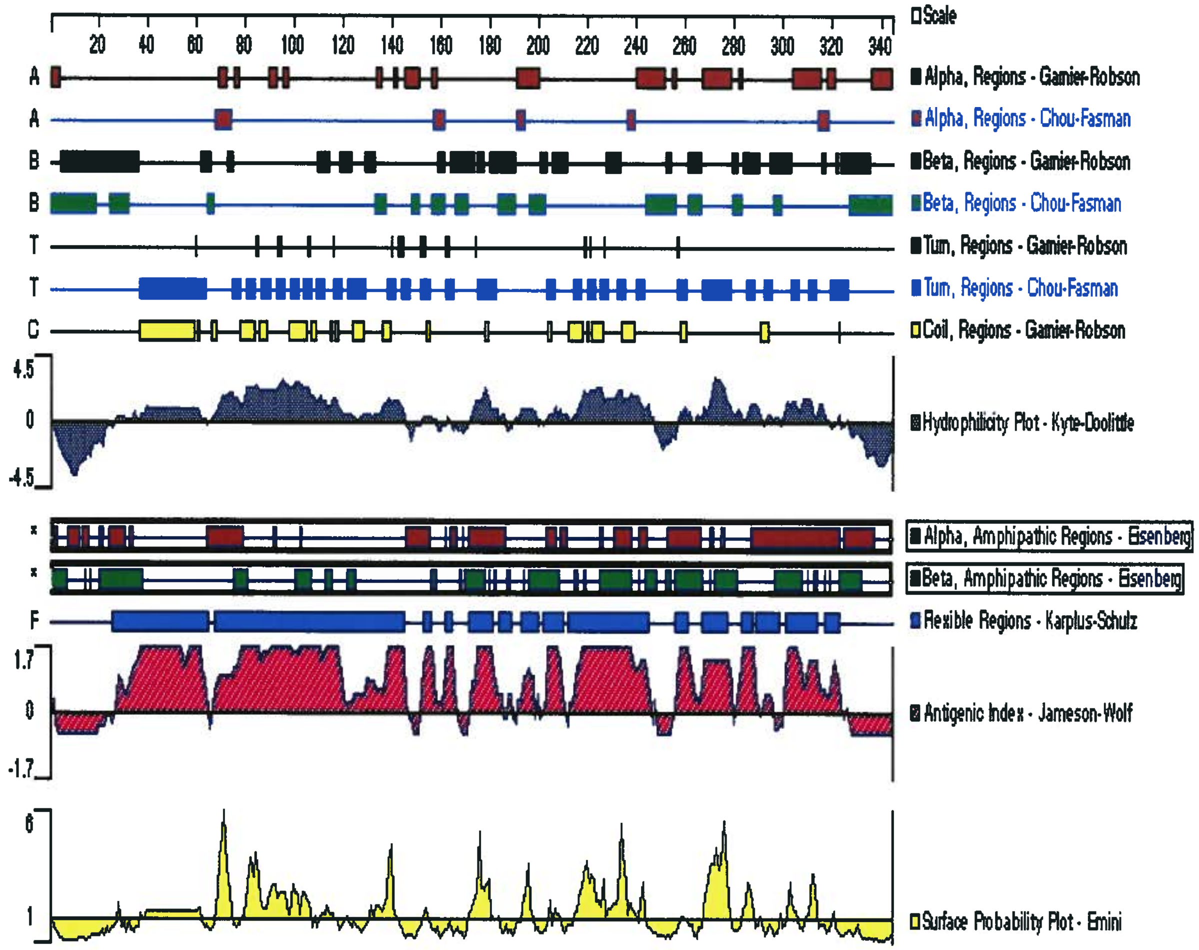


Figure A.5, B. Complete depiction of Gp60 open reading frame in regards to biophysical properties including Jameson Wolf antigenic index (pink), Emini surface probability (bottom: yellow), Kyte Dolittle hydrophilicity plot (blue, middle), and multiple secondary structure predictors (top: red, blue, yellow). Within the Protean program an individual SNP locus could be targeted generating a mathematical output for its position.

Appendix 6. Alleles scored per SNP marker per subpopulation.

Table A.1

Number of alleles sampled per SNP locus and population.					
SNP locus	**Australia**	**Kenya**	**Peru**	**Scotland**	**Total**
BT1	2	1	1	1	2
BT4	1	1	1	1	1
BT3	2	2	2	2	3
BT5	n.d.	1	1	1	3
BT7	2	2	1	1	3
BT8	2	1	1	2	2
COWP5	2	1	1	1	2
COWP6	2	1	2	1	3
COWP1	2	1	1	1	2
COWP3	2	1	2	1	3
COWP7	1	1	2	1	2
23Cp4	2	1	1	1	2
23Cp3	2	1	1	1	2
23Cp1	2	1	1	1	2
23Cp5	2	1	1	1	2
23Cp6	2	1	1	1	2
18sRNA	1	1	1	1	1
18sRNA	1	1	1	1	1
HSP14	2	1	1	1	2
HSP17	2	2	1	1	2
HSP19	2	1	1	1	2
HSP20	2	1	1	1	2
HSP22	2	1	1	1	2
60Gp80	2	2	2	n.d.	2
60Gp108	2	2	2	1	2
60Gp126	1	1	2	1	2
60Gp79	1	2	1	1	2
60Gp98	2	3	3	1	3
60Gp115	2	3	3	3	3
LDH10	2	1	1	1	2
LDH3	2	1	1	1	2
MDH8	2	1	1	1	2
MDH7	2	1	1	1	2
EMAg29	1	1	1	1	1
EMAg27	1	1	1	1	1
UPRTase	2	1	1	1	2
UPRTase	2	1	1	1	2

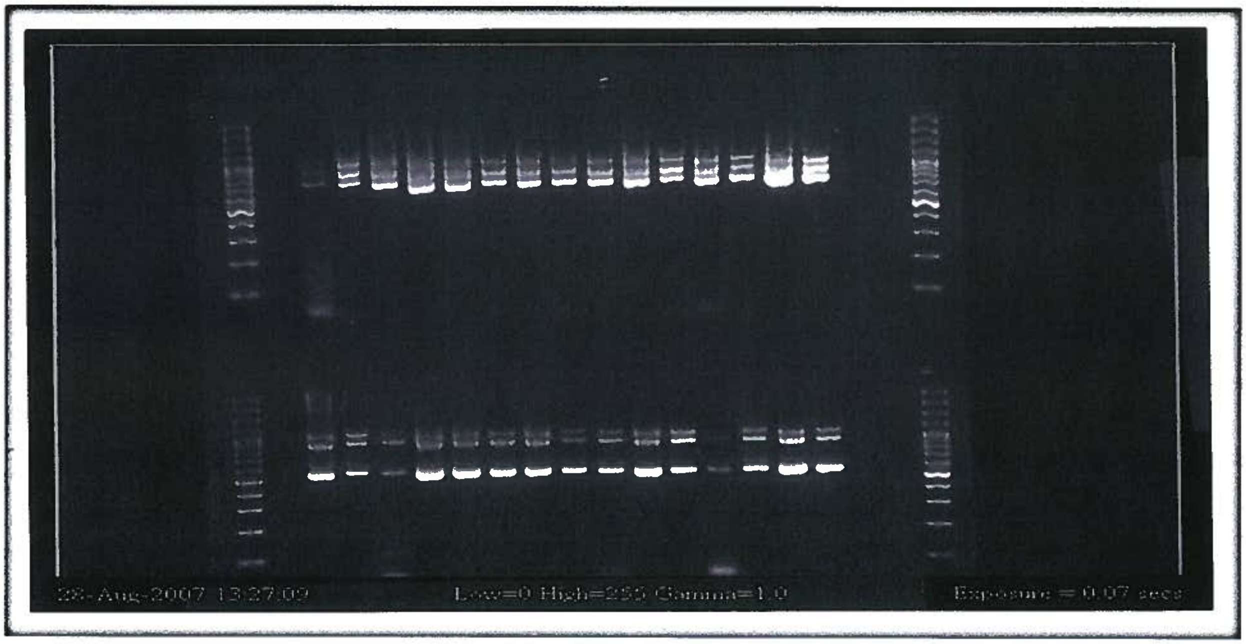
Table A.1. Number of different alleles scored at each SNP molecular marker for each subpopulation as a whole; ranging from 1 to 4 (A, C, T or G). Most scored for a marker was 3.

Appendix 7. Future target protein loci for SNP-typing; multi-plex PCR gel electrophoresis, run against 100bp ladder.

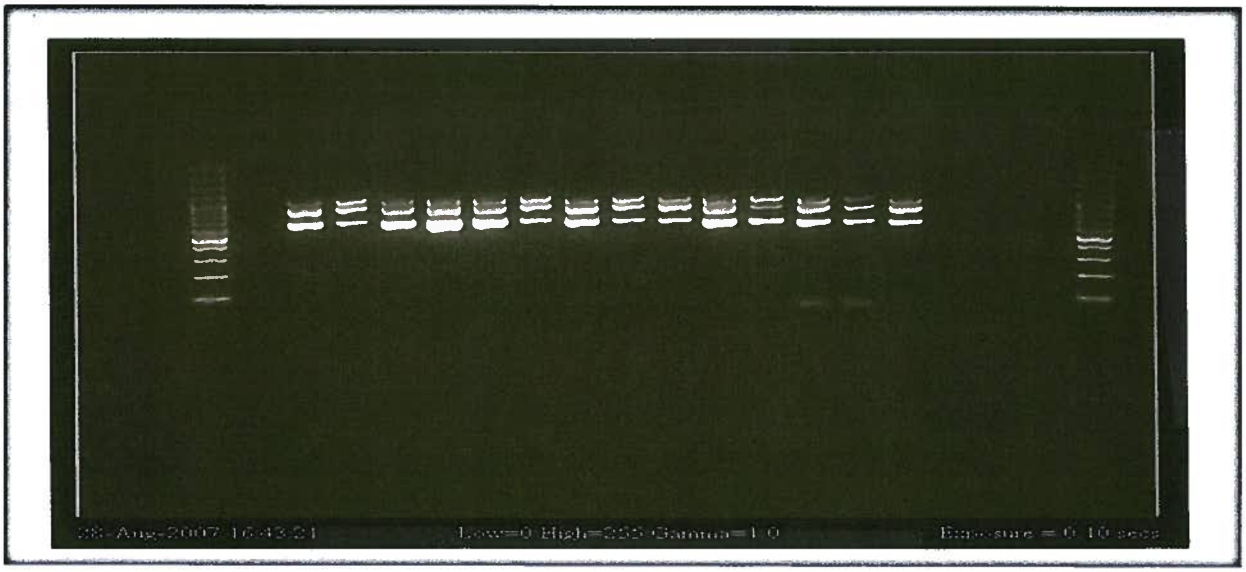
Figure A.6

A.

Top Gel: Isolates K_{1-15}, reaction set 6; CCR, CTCL, AAD at 953, 678, and 817bp respectively
Bottom Gel: Isolates $K_{1-15,}$ reaction set 7; CLL, SSK, FLJ at 983,507, and 849bp respectively.



B. Isolates P_{1-14}, reaction set 9; Exp, SeroAg, RIK at 949, 777, and 573 respectively.



Appendix 8. Genomiphi; whole genome amplification of *Cryptosporidium* DNA from fecal specimens.

An initial objective of the present study was to amplify genomic DNA in its entirety to assess the potential of isolating molecular markers without having to amplify gene specific regions. This would work to increase the throughput of the typing system almost 2-fold and was extremely cost-effective; it negated the need for gene specific primers and accompanying reagents and/or materials. In addition it was designed to generate large amounts of DNA which can be especially useful when the amount of a given sample is small.

The system used was the GenomiPhi kit by GE Healthcare. The GenomiPhi kit utilizes bacteriophage Phi29 DNA polymerase to exponentially amplify single- or double-stranded linear DNA templates via a strand displacement reaction and therefore thermal cycling is not required. The genomic DNA template is combined with a sample buffer containing random hexamer primers. The mixture is heat denatured and cooled to allow random priming of the hexamers. Then, the remaining reaction components— including Phi29 DNA polymerase, deoxynucleotide triphosphates, and buffer components optimized for linear DNA synthesis—are added. This reaction mixture is incubated overnight at 30°C, during which time the available nucleotides are consumed and converted into high molecular weight fragment copies of the template DNA. The DNA replication is extremely accurate because of the proofreading activity of Phi29 DNA polymerase. Once genome amplification is completed various genotyping assays can be undertaken from a large base of synthetic DNA copies.

We spent much time with the system attempting to evaluate it, finesse the procedure and obtain reproducibility. While the system design is beautiful in its theory and simplicity and amplification results were positive in those samples that did amplify, there was little confidence that downstream typing results were specific to *Cryptosporidium* genomic DNA. This is largely due to the fact that isolates were collected and processed from fecal specimens of patients under different protocols in facilities around the world so the likelihood of genetic material from other microorganisms, naturally occurring or invasive; being amplified as well was considered too high.

Appendix 8 continued.

Figure A.7 Simplified schematic of Genomiphi protocol.

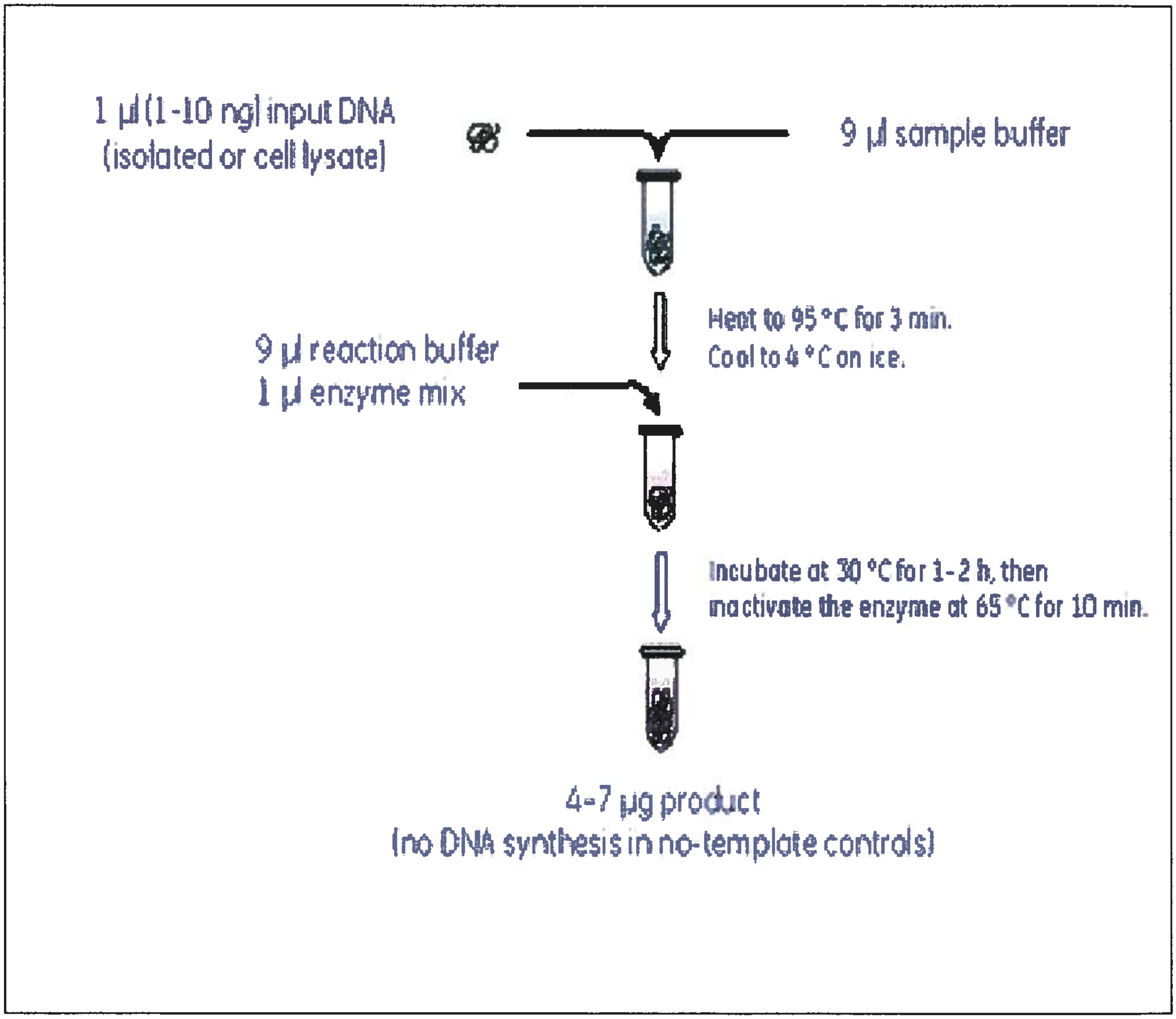

1 µl (1-10 ng) input DNA
(isolated or cell lysate)
9 µl sample buffer
Heat to 95 °C for 3 min.
Cool to 4 °C on ice.
9 µl reaction buffer
1 µl enzyme mix
Incubate at 30 °C for 1-2 h, then
inactivate the enzyme at 65 °C for 10 min.
4-7 µg product
(no DNA synthesis in no-template controls)

Figure A.7. The Genomiphi protocol is engineered on the basis of amplification of genomic DNA material using the whole genome as a template.

Appendix 9. eBURST; inferring patterns of evolutionary descent.

Figure A.8 eBURST representation of evolutionary descent.

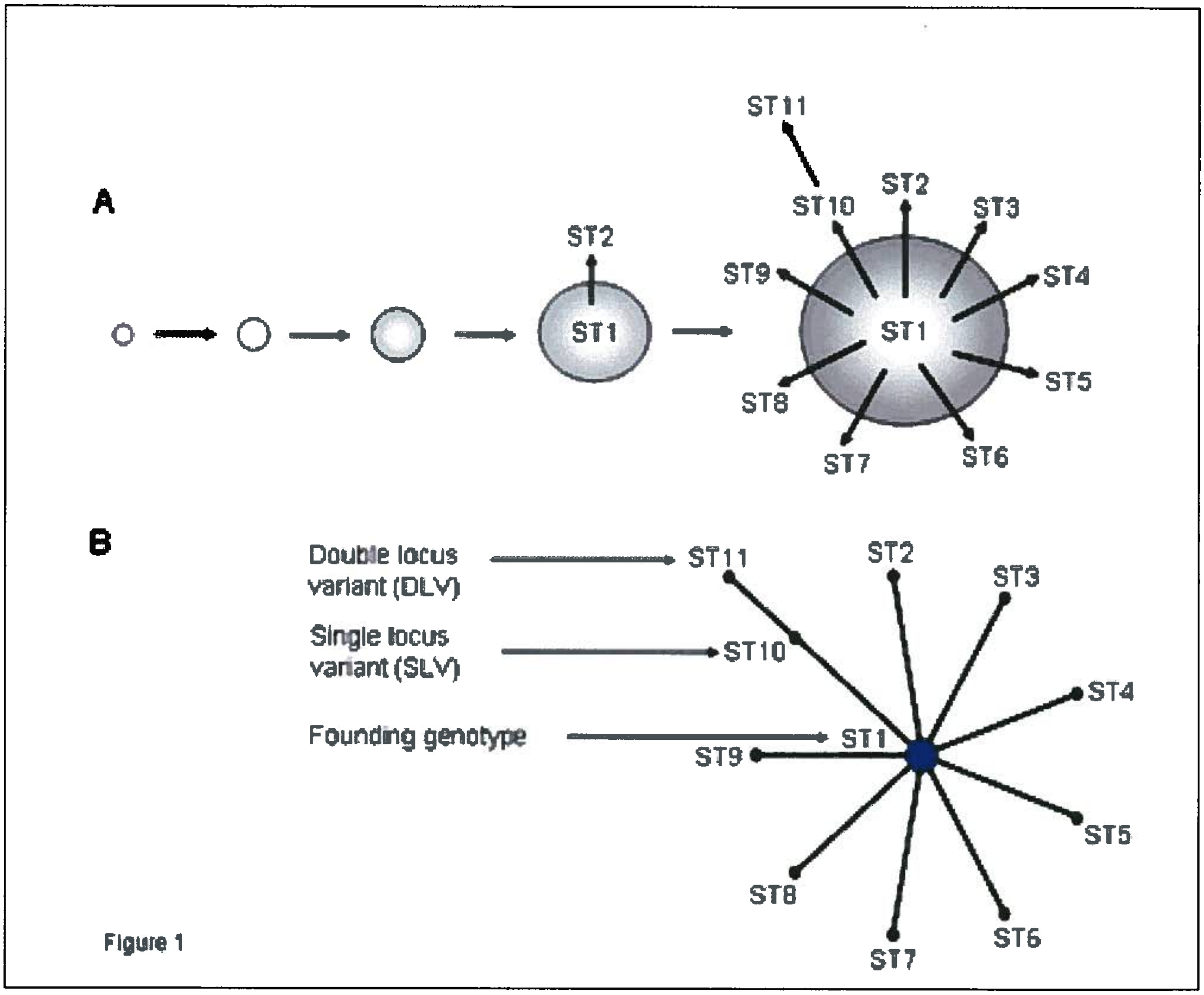


Figure A.8. The primary founder of a group is defined as the ST that differs from the largest number of other STs at only a single locus (i.e. the ST that has the greatest number of single-locus variants; SLVs). This method of assigning the primary founder takes account of the way in which clones emerge and diversify (Figure A); the initial diversification of the founding genotype of a clonal complex will result in variants of the founder that differ at only one of the seven loci (i.e., SLVs of the founder). The eBURST diagrams display the patterns of descent within each group from the predicted founding ST (Figure B). http://eburst.mlst.net

END OF DISSERTATION; JMW, 2009