THE GENEASE ACTIVITY OF MUNG BEAN NUCLEASE: FACT OR FICTION?

NOTHEMBA KULA

A minithesis submitted in partial fulfillment of the requirements for the degree of Magister Scientiae in Bioinformatics at the South African National Bioinformatics Institute, Department of Biotechnology, University of the Western Cape

Examination copy

Supervisor: Professor Winston Hide

December 2004

KEYWORDS

Alignment

Algorithm

Coding Sequences

Contiguous Sequences

Exon

Genome survey sequence

Intron

Mung Bean Nuclease

Nuclease cleavage site

Plasmodium falciparum

Plasmodium berghei

Sequence Alignment

ABSTRACT

The action of Mung Bean Nuclease (MBN) on DNA makes it possible to clone intact gene fragments from genes of the malaria parasite, *Plasmodium*. This "genease" activity has provided a foundation for further investigation of the coding elements of the *Plasmodium* genome. MBN has been reported to cleave genomic DNA of *Plasmodium* preferentially at positions before and after genes, but not within gene coding regions. This mechanism has overcome the difficulty encountered in obtaining genes with low expression levels because the cleavage mechanism of the enzyme yields sequences of genes from genomic DNA rather than mRNA. However, as potentially useful as MBN may be, evidence to support its genease activity comes from analysis of a limited number of genes. It is not clear whether this mechanism is specific to certain genes or species of *Plasmodia* or whether it is a general cleavage mechanism for *Plasmodium* DNA .There have also been some projects (Nomura *et al*., 2001;van Lin, Janse, and Waters, 2000) which have identified MBN generated fragments which contain fragments of genes with both introns and exons, rather than the intact genes expected from MBN-digestion of genomic DNA, which raises concerns about the efficiency of the MBN mechanism in generating complete genes.

Using a large-scale, whole genome mapping approach, 7242 MBN generated genome survey sequences (GSSs) have been mapped to determine their position relative to coding sequences within the complete genome sequences of the human malaria parasite *Plasmodium falciparum* and the incomplete genome of a rodent malaria

parasite *Plasmodium berghei*. The location of MBN cleavage sites was determined with respect to coding regions in orthologous genes, non-coding /intergenic regions and exon-intron boundaries in these two species of *Plasmodium.* The survey illustrates that for *P. falciparum* 79% of GSSs had at least one terminal mapping within an ortholog coding sequence and 85% of GSSs which overlapped coding sequence boundaries mapped within 50 bp of the start or end of the gene. Similarly, despite the partial nature of *P.berghei* genome sequence information**,** 73% of *P.berghei* GSSs had at least one terminal mapping within an ortholog coding sequence and 37% of these mapped between 0-50 bp of the start or end of the gene. This indicates that a larger percentage of cleavage sites in both *P.falciparum* and *P.berghei* were found proximal to coding regions. Furthermore, 86% of *P.falciparum* GSSs had at least one terminal mapping within a coding exon and 85% of GSSs which overlapped exon-intron boundaries mapped within 50bp of the exon start and end site. The fact that 11% of GSSs mapped completely to intronic regions, suggests that some introns contain specific cleavage sites sensitive to cleavage and this also indicates that MBN cleavage of *Plasmodium* DNA does not always yield complete exons.

Finally, the results presented herein were obtained from analysis of several thousand *Plasmodium* genes which have different coding sequences, in different locations on individual chromosomes/contigs in two different species of *Plasmodium.* Therefore it appears that the MBN mechanism is neither species specific nor is it limited to specific genes.

DECLARATION

I declare that “Genease activity of Mung Bean Nuclease: fact or fiction?” Is my own work, that it has not been submitted for any degree or examination in any other university, and that all the sources I have used or quoted have been indicated and acknowledged by complete references**.**

Nothemba Kula **November 2004**

Signed: _____

DEDICATION

“Who hath despised the day of small beginnings? … God hath chosen the foolish things of the world to confound the wise. God hath chosen the weak things to confound the things that are mighty, and the base things of the world, and the things that are despised hath God chosen, yea and the things, which are not, to bring to naught things that are.

-1cor1: 27-28 (Holy Bible).

This is dedicated to all those who have been made to believe that their humble beginnings would limit what they can achieve. This thesis affirms that there is no limit to those who believe.

ACKNOWLEDGMENTS

I wish to express my heartfelt gratitude to my supervisor and mentor Professor Winston Hide for his patience, academic insight and personal motivation. “Prof, without your involvement in this project, it would have never seen the light of day”, thank you. I am especially thankful to Dr Jane Carlton of the Institute for genomic Research (TIGR) in Rockville Maryland, USA, for having initiated this research project and for further driving the project towards the publication of a research article. “The exposure you gave me broadened my horizons to heights I never thought possible, thank you so much”.

I also wish thank the Medical Research Council’s (MRC) Research and Development unit for funding this project and giving me the opportunity to train, experience and explore bioinformatics research. “You have given me a platform to contribute towards the development of my country and my people, for that I am forever indebted”. I am also grateful to the wonderful people at SANBI who have made a notable contribution to this work. “You guys were such a crazy bunch, thanks for making the hard times fun”. My heart goes out to my mother, my husband and the whole family “You were the ones who gave me a pat on the back and told me to move on, thank you for believing in the purpose of God about my life”. Lastly and most importantly, to my heavenly Father “I know this is only a stepping stone to greater things, SANBI was the right place, at the right time, there is no doubt in my mind that you were the architect behind it all, I owe it all to you” .

CONTENTS

List of Figures and Tables

CHAPTER1

1. INTRODUCTION

The action of Mung Bean Nuclease (MBN) on DNA makes it possible to clone intact gene fragments from genes of the malaria parasite, *Plasmodium*. This “genease” activity has provided a foundation for further investigation of the coding elements of the *Plasmodium* genome. MBN has been reported to cleave genomic DNA of *Plasmodium* preferentially at positions before and after genes, but not within gene coding regions(McCutchan *et al.*, 1984). This mechanism has potential to overcome the difficulty encountered in obtaining genes with low expression levels because the cleavage mechanism of the enzyme yields sequences of genes from genomic DNA rather than mRNA. However, as potentially useful as MBN may be, evidence to support its genease activity comes from analysis of a limited number of genes.

It is not clear whether this mechanism is specific to certain genes or species of *Plasmodia* or whether it is a general cleavage mechanism for *Plasmodium* DNA. Some projects (Nomura *et al.*, 2001;van Lin, Janse, and Waters, 2000) have identified MBN fragments which appeared to contain fragments of genes with both introns and exons, rather than the intact genes expected from MBN-digestion of genomic DNA , which raises concerns about the efficiency of the MBN mechanism in generating complete genes. This study was conducted to provide a final determination of whether the “genease” activity of mung bean nuclease is fact or fiction.

1.1 BACKGROUND

Mung Bean Nuclease (MBN) is a single-strand specific nuclease purified from sprouts of the mung bean *Vigna radiata*. This enzyme has typically been used for its single strand nuclease activity (Kowalski, Kroeker, and Laskowski, Sr., 1976). It has been shown that under a different set of reaction conditions, the enzyme precisely cleaves purified duplex DNA at sites which are outside the coding region of genes from the malaria parasite *Plasmodium falciparum* (McCutchan *et al.*, 1984).Recognition and cleavage of DNA does not seem to be related to any primary sequence but is thought to be related to structural features of the DNA duplex that demarcates genes.

Subsequent to these findings, a novel set of reaction conditions for mung bean nuclease in which *Plasmodium* genes were specifically excised as intact fragments from purified DNA was reported (Vernick, Imberski, and McCutchan, 1988). New conditions where MBN cleaved precisely at sites outside the coding region of every *P.falciparum* gene were detailed. This study also suggested the involvement of an altered DNA structure near gene boundaries in determining the recognition sites for the activity of the enzyme.

Findings revealed that some cleavage sites were within dAT rich regions and that other cleavage sequence, however, were not dAT rich, which suggested that unknown structural features of the overall sequence may be important (Brown, Brentano, and Donelson, 1986) and the sequences of the cleavage sites themselves are not obviously related (Brown *et al*, 1986).

The key aspect of understanding the MBN mechanism is in the nature of the nuclease recognition sites and their relationship to genes (Vernick and McCutchan, 1998). Available data suggests a novel class of DNA site with distinct structural properties which encode biological information by marking the boundaries of at least some gene expression units in organisms as diverse as *Plasmodium* and *Drosophila*. The MBN mechanism has also been exploited in numerous other protozoans, including Giardia (Adam, Nash, and Wellems, 1988), Toxoplasma (Johnson *et al.*, 1987), Leshmania (Muhich and Simpson, 1986) and Babesia (Tetzlaff, McMurray, and Rice-Ficht, 1990;Tripp, Wagner, and Rice-Ficht, 1989).

To study the MBN mechanism further, mung bean nuclease library clones were compared to protein and nucleotide sequences in the databases to judge whether the clones represented complete genes (Reddy *et al*, 1993). It was reported that thirty percent (30%) of the putatively identified clones appeared complete since they showed similarity extending from the initiation codon through to the termination codon. However, results showed that 62% of the remaining clones had one or more exons but appeared incomplete.

Nearly one third of the clones contained intact genes, the majority of the other two thirds appeared to be clones of one or more exons, which suggested that the genes contained introns that were sensitive to MBN digestion. The fact that 62% of the clones studied by Reddy *et al*, appeared to have incomplete genes raises a concern about the efficiency of the mung bean nuclease mechanism in generating complete genes.

A recent study (Carlton *et al.*, 2001) identified some projects (Nomura *et al.*, 2001;van Lin, Janse, and Waters, 2000) in which GSS fragments appeared to contain fragments of genes with both introns and exons, rather than the intact genes expected from MBN-digestion of genomic DNA. These unexpected discoveries prompt a critical examination of the nature of mung bean nuclease recognition sites and their relationship to coding regions. Upon this background, this study conducts a large genome scale analysis (*in silico*) of the MBN mechanism on the *Plasmodium* genome to determine the genease nature of MBN.

1.2 Relevance of mung bean nuclease mechanism in malaria research

Significant fundamental information on the genome organization and molecular biology of the malaria parasite is rapidly increasing as a result of the efforts of the genome-sequencing project of *Plasmodium falciparum.* Completion of *Plasmodium* genome sequences has provided new avenues and opportunities to understand parasite biology in detail and explore new control measures (Gardner *et al.*, 2002).

The new paradigm in biology is to have all the genes of an organism recorded in databases, available to be used as the starting point for further investigations (Gilbert, 1991). Identification, sequencing and mapping of the structural genes provides a foundational database from which to launch applied research programs for anti-malarial drug and vaccine development.

The predominant method of obtaining preliminary data on coding sequences of an organism has been the construction of cDNA (complementary DNA) libraries. *P.falciparum* cDNA libraries are limited, however, to the genes expressed in the lifecycle stage used to prepare the mRNA, and the probability of obtaining a given cDNA sequence depends on the level of expression of the gene. Obtaining rare cDNA clones at random, including those that encode regulatory enzymes or proteins is problematic (Reddy *et al.,* 1993). Utilizing the "genease activity of mung bean nuclease has the potential to overcome these problems because the cleavage mechanism of the enzyme yields sequences of genes from genomic DNA rather than mRNA.

The mung bean nuclease mechanism is important in malaria research because it has potential to capture the genes and gene products derived from sequencing the malaria parasite genome. The MBN mechanism may generate a better understanding of the secondary structure of *Plasmodium* DNA and it will ultimately advance our understanding of the biology of the parasite (McCutchan *et al.*, 1984;Dame *et al.*, 1984). The specific manner in which the enzyme cleaves *Plasmodium* DNA generates gene sequence tags which contain complete *Plasmodium* genes (McCutchan *et al,* 1984). Therefore, this "genease" activity provides a foundation for further investigation of the coding elements of the *Plasmodium* genome.

2. *Plasmodium* genomics

2.1 *Plasmodium* genome project

Plasmodium falciparum is the causative agent of cerebral malaria, which afflicts much of the world's population. A minimum of between 700,000 and 2.7 million people die yearly from malaria, over 75% of them African children (Breman, 2001). One of the hallmark features of the malaria parasite is the complex life cycle that includes both mosquito and human hosts (Breman, 2001; Lasonder *et al.*, 2002).

In 1996 an international genome-sequencing consortium was established to sequence and annotate the entire genome of *P.falciparum* with the expectation that the genome sequence would open new avenues for research (Hoffman *et al.*, 1997). Six years later, a complete reference genome for the parasite was published (Gardner *et al.*, 2002). The availability of complete genomic sequence data, use of sequence- based high- throughput technologies and advances in bioinformatics to analyze and interpret genomic data will ultimately provide an integrated picture of malaria parasite biology, pathogenesis and epidemiology (Hoffman *et al.*, 2002).

Genomics is a critical component of 21st century biomedical research. It provides an incredible opportunity for increased insight into the biology of *Plasmodium* parasites, and their *Anopheles* vectors, and the pathogenesis of malaria in humans (Hoffman *et al.*, 2002). However, the genome sequences alone provide little or no relief to those suffering from malaria.

The complete genome sequence needs to be accompanied by larger efforts to develop new methods of control including new drugs and vaccines, improved diagnostics and effective vector control techniques. Much remains to be done. Clearly, research and investments to develop and implement new control measures are needed desperately if the social and economic impacts of malaria are to be relieved.

2.2 The *Plasmodium* genome database

The success of the *Plasmodium* genome project will ultimately be determined by how rapidly and effectively the information it produces is utilized by the research community to advance our understanding of malaria. In particular, effective dissemination of genomic information should accelerate the development of new therapeutics and vaccines.

The *Plasmodium* genome database, PlasmoDB (http://PlasmoDB.org) contains information from multiple sources. The information available includes DNA sequence data and curated annotations, automated gene model predictions, predicted proteins and protein motifs, cross-species comparisons, optical and genetic mapping data, information on population polymorphisms, expression data generated by a variety of complementary strategies, and proteomics data (Kissinger *et al.*, 2002).

The *Plasmodium* genome database seeks to incorporate data and annotation emerging from the *P.falciparum* genome sequencing centers into a searchable resource, with extensive links to relevant resources and databases. Most importantly, the *plasmodium* genome database was designed to establish a database that can incorporate

sequence information for other *Plasmodium* species and related apicomplexan parasites. It also provides full access to the *Plasmodium* genome and related resources for all interested parties regardless of geographic location (Kissinger *et al.*, 2002).

Over the past year PlasmoDB has expanded to include new data as well as changes to the site. The highlights of recent developments are: automated gene predictions for all genomic *P. falciparum* sequences, gene expression data consisting of oligo- and cDNA-based micro array and SAGE (Serial Analysis of Gene Expression) experiments, draft genomic, GSS (Genome Survey Sequence), and/or EST (Expressed Sequence Tag) sequences for four additional *Plasmodium* species: *P. berghei, P. chabaudi, P. vivax, and P. yoelii* and cross-species comparison between *P. falciparum* and *P. yoelii* is also available (Bahl *et al.*, 2003).

2.3 Genome sequence of the human malaria parasite *Plasmodium falciparum*

(i) Genome structure and organization

The *P.falciparum* nuclear genome is composed of 22.8 megabases (Mb) distributed among 14 chromosomes ranging in size from approximately 0.63 to 3.29Mb.Thus the *P.falciparum* genome is almost twice the size of the genome of the yeast *Schizosaccharomyces pombe*. The overall (A+T) composition is 80.6% and rises to ~90% in introns and intergenic regions. *The P.falciparum* genome is the most (A+T) rich genome sequenced to date and it encodes about 5,300 protein encoding genes. The number of genes suggests an average gene density of 1 gene per 4,338 base pairs (bp). Introns were predicted in 54% of *P.falciparum* genes, a proportion roughly similar to *S.pombe* and.*Dictyostelium discoideum,* but much higher than observed in

Saccharomyces cereviciae. Excluding introns, the mean length of *P.falciparum genes* is 2.3 kb, substantially larger than most organisms (e.g. *S.cerevinsia* and *S. pombe*) in which the average gene lengths range from 1.3-1.6 kb. The explanation for increased gene length in *P.falciparum* is not clear (Gardner *et al.*, 2002). The high (A+T) content of the *P.falciparum* genome made gap closure extremely difficult, efforts to close these gaps are still continuing.

(ii) The proteome

Of the 5,300 predicted proteins, about 60% do not have sufficient similarity to proteins in other organisms to justify provision of functional assignments. Thus, almost two- thirds of the proteins appear to be unique to this organism, a proportion much higher than observed in other eukaryotes. Another five percent (5%) of proteins have significant similarity to hypothetical proteins in other organisms. Thirty one percent of the predicted proteins have one or more trans-membrane domains, and 17.3% of the proteins possess signal peptides or signal anchors(Gardner *et al.*, 2002).

2.4 Genome sequence of the rodent malaria parasite *Plasmodium berghei*

P.berghei belongs to the group of four malaria species that infect African murine rodents. Various species of rodent, avian and non- human malaria parasites are used as laboratory models to develop strategies for the eradication of human malaria (Janse and Waters, 1995). Rodent parasites are also valuable in the investigation of the developmental biology of malaria parasites, parasite-host interactions, vaccine development and drug testing. The conservation of housekeeping genes, biochemical and genetic processes between mammalian parasites (and hence similarities and drug susceptibility and

mechanisms of resistance) provide the first justification of the use of these rodent models in malaria research.

(i) Genome organization

P.berghei has an estimated genome size of 25-27Mb, with 14 chromosomes in the size range of 0.6 Mb to 3.8 Mb (http://www.sanger.ac.uk/Projects/P_berghei/, May 2004). *P.berghei* has two extra- nuclear DNA elements comparable to *P.falciparum:* the mitochondrial and plastid DNA -organellar genomes (Yap *et al.*, 1997). The number and size range of the *P. berghei* chromosomes is comparable to *P. falciparum* (Table1), resulting in a comparable genome size of about 25Mbp. Like *P. falciparum*, the nuclear DNA of *P. berghei* has an extremely high overall A+T composition (Table1). This (A+T) rich bias is unevenly distributed between protein coding and non-coding regions. All open reading frames are relatively (G+C) rich while (A+T) composition of the majority of intergenic regions and intragenic introns can rise to more than 90%.

3. Conservation of genome organization between rodent and human malaria parasites

Studies demonstrate a high level of conservation of genome organization between rodent and human parasites (Carlton *et al.*, 2002;van Lin, Janse, and Waters, 2000). The nuclear genome of both *P.falciparum* and the three rodent malaria parasites (*P.berghei*, *P.yoelii*, *P.chabaudi*) are organized into 14 linear chromosomes ranging in size from 0.5 – 3.5 Mb.

Comparative mapping of genes located in the central regions of the chromosomes has shown that both linkage and gene order appear to be well conserved between human and rodent parasites, resulting in a significant level of synteny (gene location and order on chromosomes). Arrangement of the genes (rRNA, rpo-B and tRNA) on the *P. berghei* plastid genome is similar to that found in *P. falciparum.* Partial DNA sequence analysis revealed 69.95 – 95.5% homology to sequenced *P.falciparum.*

Moreover, the conservation of gene domains, regulatory control elements, organization of genomic loci and the presence of multi gene families in rodent and human parasites all emphasizes the high similarity of genome organization, gene content and gene regulation (Janse *et al.*, 1994;Carlton *et al.*, 1998;Thompson, Janse, and Waters, 2001). Although the level of synteny of genes is lower when the genomes of rodent and human species of *Plasmodium* are compared, significant conservation of genome organization has been observed (Carlton *et al*, 1998; Carlton *et al*, 2002).

Table 1: Genome summary statistics for *P. falciparum* and *P.berghei*

Data	*P. Berghei*	*P. Falciparum*
Genome		
Size	17996878	22853764
No. of contigs	7497	93
Av. contig size (bp)	2400	213586
Max. contig size (bp)	37075	2,271,477
No. scaffolds	4700	N/a
Coverage	4x	14.5x
Transcriptome		
No. of ESTs	12277	21371
Av. singleton length (bp)	541	444
No. of EST clusters	1939	7956
Av cluster length (bp)	737	424.2
Genome content		
GC content	23.7	19.4
Total no. of predicted protein coding genes (* with orthologs)	12208 (*5864)	5268
Total no. full length genes	4617	5260
Total no. partial genes	7591	8
No. of tRNAs	65	43
No. of 5.8s,18s,28s rRNA units	4	7
No. of 5s rRNA genes	3	3
Gene density	1476	4338
% coding	56.7	52.6
% genes with introns	40	52
Genes with EST hits	1365	3301
Gene products detected by proteome	1847	2415
Mean no. exons per gene	1.74	2.4
GC content of exons	24.7	23.7
Mean length of exons (bp)	421.6	943.7
GC content of introns	22.8	13.5
Mean length of introns (bp)	135.6	179
Av length of intergenic regions (bp)	735.5	1654.1
GC content of intergenic regions	21.15	13.44

Table reproduced from Hall *et.al*. 2004. *Science*, in press.

4. Properties of Mung Bean Nuclease (MBN)

4.1 Biology and activity

Mung bean nuclease was found to cleave the genomic DNA of the malaria parasite *Plasmodium* at positions before and after genes but not within gene coding –regions.

Under defined reaction conditions, which include altered solvation and elevated temperature; hypersensitive sites surrounding coding regions become sensitive to nuclease cleavage (McCutchan *et al.*, 1984), while sequences within coding regions remain insensitive. A library prepared with fragments generated by mung bean digests is therefore expected to represent an enrichment of coding sequences and a very significant increase in the proportions of complete and intact coding regions (Rathore and McCutchan, 2002). Due to its single strand nuclease activity, MBN has also been used to cut restriction overhangs into blunt ends prior to ligation (Fig 1).

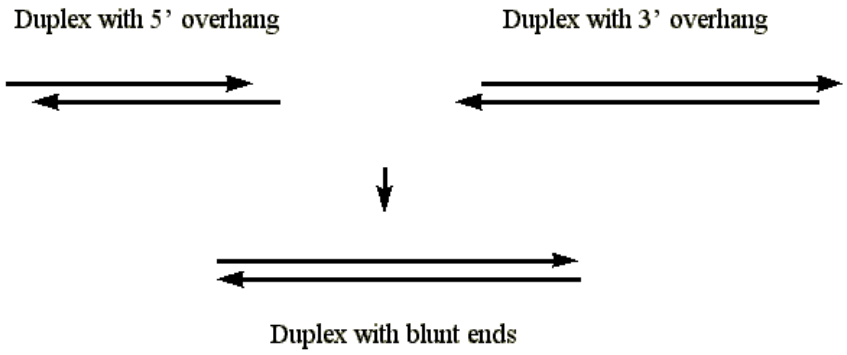

Fig1: MBN degrades single strand extensions from the termini of double stranded DNA and RNA generating blunt ended duplex DNA.

The diagrams below give insight into the biology and activity of MBN as reported in previous studies, in each case the enzyme was found to cut genomic DNA outside the coding region of each gene analyzed. It was observed that introns are not recognized as distinct structural features under specified reaction conditions, although they can contain within them specific cleavage sites (Vernick *et.al*, 1988).

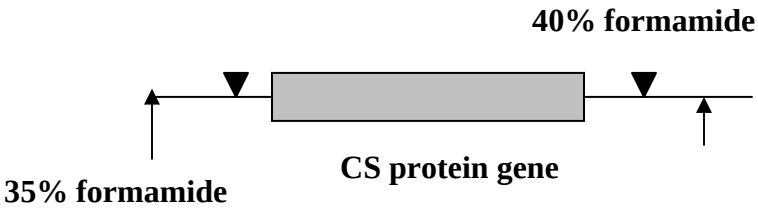

Fig2: Mung bean cleavage sites in the *P.falciparum* circumsporozoite protein (CSP). The coding region is boxed; non-coding regions are represented by a line. Arrows show cleavage in 35% formamide**.** Large arrowheads show the predominant cleavage sites in 40% formamide. Figure obtained from (McCutchan *et al.*, 1984).

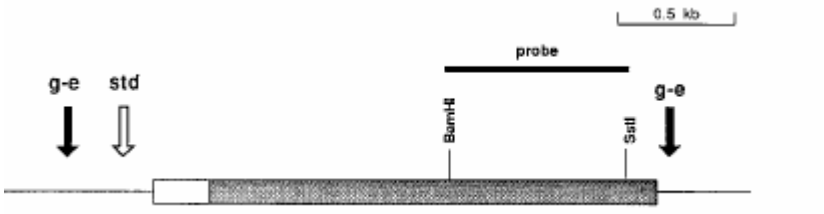

Fig3: Positions of 5' and 3' sequences recognized by MBN in Drosophila DNA. Gene excision (g-e) conditions are indicated by filled arrows. The 5' sequence region recognized by MBN under standard conditions is indicated by an open arrow (std). The coding region of the gene is shown by a shaded rectangle, 5' UTR by open rectangle and location of a probe fragment is indicated. Figure reproduced from Vernick *et al*, 1998.

4.2 Recognition sequences

It was previously suggested that MBN prefers to cut 3'of A and T residues in single stranded DNA (Johnson and Laskowski, Sr., 1970). However, several cases were identified where the enzyme did not favor what was expected to be a preferred site, (Kabotyanski *et al.*, 1995). It was concluded that there is no apparent sequence identity either 5' or 3' to the MBN cleavage cuts (McCutchan, 1984). Although the cut sites are rich in dA.dT they have no more dA.dT than surrounding areas both inside and outside the gene that are not cut. Recognition and cleavage does not seem to be related to any primary sequence but may be related to structural features of the DNA duplex that demarcate genes (McCutchan *et al.*, 1984).

It also emerged from some findings that MBN does not recognize introns as distinct structural features although introns can contain within them specific cleavage sites. It was concluded that the cleavages, which do occur in introns, are not general events (Vernick, Imberski, and McCutchan, 1988) and that the presence of introns does not hinder the genease approach (Reddy *et al.*, 1993).

Available data indicates that sites of cleavage depend on the structure of naked DNA (McCutchan *et. al*, 1984). Since the enzyme cleaves single stranded DNA, cleavage might be directly related to denaturation of DNA and cutting of single stranded areas. The fact that a point of cleavage that is totally spared in one concentration of formamide is completely cleaved in another also indicates that conformational transitions are

induced by formamide and that the cleavage is not a result of primary sequence recognition (McCutchan *et al.,* 1984). Although several studies have shown that single strand specific nucleases recognize structural features in the vicinity of hairpin termini in DNA, the nature of these features has not been fully defined (Xodo *et al.,* 1991).

5. STUDY AIMS AND OBJECTIVES

Genomic DNA libraries represent the total complement of the genetic information of an organism's DNA, as opposed to cDNA libraries, which contain only the protein encoding sequences expressed at a particular stage of the life cycle. The action of mung bean nuclease on *Plasmodium* DNA makes it possible to capture complete coding sequences which can be used as the starting point for further investigation into the pathogenicity of the malaria parasite.

Therefore, the overall objectives of this thesis are to:

- Analyze the mechanism by which the enzyme, mung bean nuclease cleaves genomic DNA of rodent and human malaria parasites.
- Investigate, using large -scale genome data, the relationship between MBN cleavage sites and coding sequences in the malaria parasite genome.
- Make a determination of the bias with which MBN cleaves *Plasmodium* DNA

CHAPTER2

GENERATING A GENOME MAP OF MUNG BEAN NUCLEASE CLEAVAGE SITES ON THE *PLASMODIUM* GENOME

1. INTRODUCTION

In order to provide a final determination of whether the genease activity of mung bean nuclease is 'fact or fiction', several important questions must be addressed:

(i) Do all or only a subset of genes in *Plasmodium* contain flanking mung bean nuclease recognition signals that will allow their precise excision under certain reaction conditions?

(ii) Does MBN digestion of *Plasmodium* DNA generate complete exons?

(iii) How general is the MBN phenomenon? is it limited to specific genes or species of *Plasmodia* ?

To answer the questions specified above, it is necessary to closely examine the location of MBN cleavage sites proximal to coding elements within the *P.falciparum* and *P.berghei* genome. Several strategies were used to determine whether the entire coding region of a gene was included in a MBN generated fragment. Generating a map of mung bean nuclease cleavage sites across the genome facilitates a genome wide comparison of MBN cut site locations relative to gene locations and coding exons in *P.falciparum* and *P.berghei*. To generate the map of MBN cleavage sites, genomic DNA (gDNA) sequencing libraries prepared from randomly sheared gDNA sequenced from both ends

were used -such sequences are termed genome survey sequences (El Sayed *et al.*, 1997;Smith, 1998;Liu *et.al* 1999).Once the genome map was generated we were able to determine, using our data, the percentage of:

(i) Genome survey sequences (GSSs) mapping to non-coding /intergenic regions

(ii) GSSs mapping completely to coding regions

(iii) GSSs overlapping exon- intron boundaries / gene-intergenic region boundaries.

2. Tools and datasets selected for generating a genome map of MBN cleavage sites.

2.1 Biological Sequence alignment

Sequence alignment is the process of comparing two (pair-wise) or more (multiple sequence alignment) sequences by searching for a series of individual characters or character patterns that are in the same order in the sequences, (Mount, 2001). It provides an explicit mapping between sequences. The primary goal of sequence alignment is to determine whether sequences display sufficient similarity such that an inference of homology can be made. Sequence alignment is useful for discovering structural, functional and evolutionary information in biological sequences.

In global alignment an attempt is made to align the entire sequence using as many characters as possible, up to both ends of each sequence. Sequences that are similar and approximately the same length are suitable candidates for global sequence alignment. In local alignment, stretches of sequences with the highest density of matches are aligned, thus generating one or more islands of matches or sub-alignments in the aligned

sequences. Local sequence alignments are suitable for aligning sequences that are similar along some of their lengths, but dissimilar in others and sequences that differ in length or sequences that share conserved domains (Mount, 2001).

2.1.1Sequence alignment algorithms used

2.1.1.1 MEGABLAST

MEGABLAST is a greedy local sequence alignment algorithm optimized for aligning sequences that share significant similarity and differ only by a small fraction. It handles large input sequences, incorporates a model for sequencing errors and low-cost affine gap qualities, matches larger word sizes than traditional BLAST and it runs considerably faster than most of the other heuristic sequence comparison tools in its class (Zhang *et al.*, 2000). MEGABLAST performs a nucleotide sequence alignment search and concatenates many queries to save time spent scanning the database. This program is optimized for aligning sequences that differ slightly as a result of sequencing or other similar "errors". It is up to 10 times faster than more common sequence similarity programs and therefore can be used to swiftly compare two large sets of sequences against each other.

MEGABLAST differs from other blast programs in that it uses a 28mer word size for database searching and is thus able to efficiently handle much longer DNA sequences. Megablast takes as input a set of FASTA formatted DNA query sequences. Since Megablast can only work with DNA sequences, the only program it supports is BlastN. MEGABLAST is most efficient in both speed and memory requirements.

2.1.1.2 BLASTN PROTOCOL

BLAST (Basic Local Alignment Search Tool) is the most basic and common tool used for determining sequence similarity. The program looks for " words" that are present in both sequences and then extends these at either end to find the longest segments that are present in both sequences. The statistical significance of these high – scoring segment pairs (HSP) is evaluated to determine whether the matches are random or not. Thus, the scores assigned in a BLAST search have a well- defined statistical interpretation (Gentner and Gunn, 2001). The results are reported in a form of a ranked list followed by a series of individual sequence alignments, plus various statistics and scores, but an important question is whether or not the score is high enough to provide evidence of homology. To address this, it is helpful to have some notion of how high a score is expected to be due to chance alone. Several statistical models are available to make these approximations**.**

The BlastN algorithm is used to compare a nucleotide query sequence against a nucleotide sequence database. BLAST programs have been designed for speed to find high scoring local alignments. BLAST uses a strategy based on matching sequence fragments by employing a powerful statistical model, developed by Samuel Karlin and Stephen Altschul (Altschul *et al.*, 1990). The program uses a heuristic algorithm which seeks local as opposed to global alignments and is therefore able to detect relationships among sequences which share only isolated regions of similarity (Altschul et al., 1990)**.** Because of its design for speed, there may be a minimal loss of sensitivity to distant sequence relationships.

3. Sequences

3.1 Genome survey sequence library construction

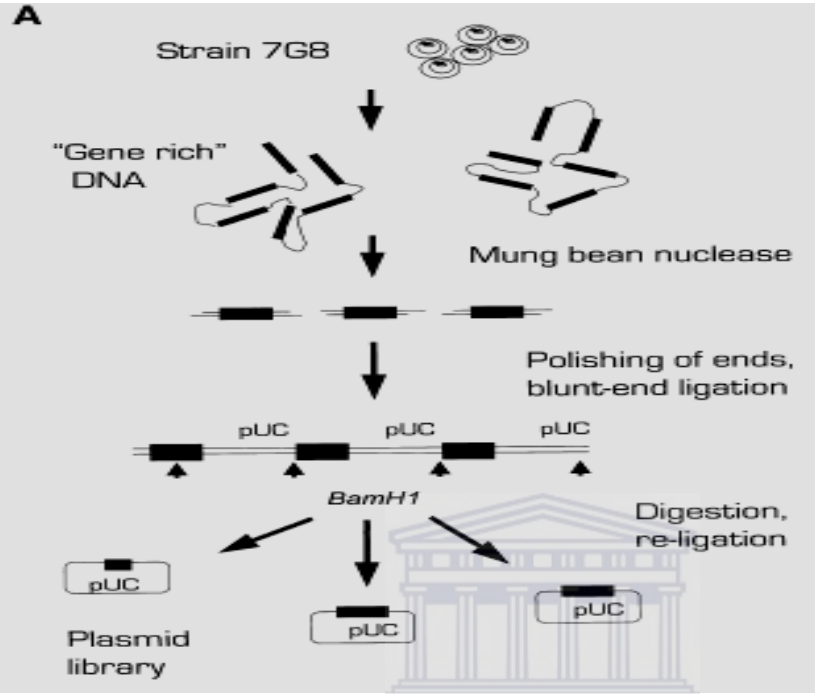

Fig4: Preparation of mung bean nuclease digested *Plasmodium* genomic DNA. Diagram from (Hayward *et al.*, 2000).

To obtain a representative sampling of the genetic coding capacity of parasite genomes, several gene discovery expeditions have utilized genomic DNA (gDNA) sequencing libraries prepared from randomly sheared gDNA and have sequenced both ends of the gDNA inserts of random clones isolated from these libraries; such sequences are termed genome survey sequences –GSSs (El Sayed *et al.*, 1997;Smith, 1998; Liu *et al*, 1999). Genome survey sequences are similar to expressed sequence tags- ESTs (DNA sequences derived from cDNA clones) (Adams MD *et al*, 1991), with the exception that these

sequences are genomic in origin, rather than cDNA (mRNA). GSSs are random "single pass read" genome survey sequences, which have not been mapped in the genome.

Random genome survey sequencing (GSS) of organisms with relatively small genomes has recently been shown to be very effective in gene discovery (Strong and Nelson, 2000; El Sayed *et al.*, 1997). Since most parasitic protozoa are composed of several asexual and sexual stages, focusing gene discovery exclusively on genes expressed during particular life cycle stages would place serious limitations upon the type and extent of sequence information obtained.

The GSS approach to gene discovery is a powerful compliment to EST profiling as it samples and identifies genes irrespective of the life cycle stage in which they are expressed. Some studies have even suggested that genome survey sequencing is a more productive and efficient method of gene discovery than EST profiling (Strong and Nelson, 2000).

Table2: Summary data for *Plasmodium* dataset

	P. berghei	*P.falciparum*
1. Total number of GSSs	5476	1766
2. Clone Libraries	Pb MBN#21	gmPfHB3.1, G.R Reddy
3.Strain	ANKA 15cy1	HB3 clone
4. GSS A/T(%)	74.00	77.00
5. A/T of genome (%)	74.08	76.04
6. Ave. Length of GSS	562.48	412.29
7. Direction of sequencing	3'	3'
8. Total genome size (bp)	17,99*	22,853,764
9. % coding	56.70	52.60

***** This is an under-representation of the ~23 Mb actual genome size, due to the partial 4x coverage of the genome**.**

3.2 DNA sequences analyzed

Nucleotide and deduced amino acid sequences as well as location and transcriptional directions of *P.falciparum* 3D7 were obtained from the *Plasmodium* genome resource – (http://plasmodb.org). Mung bean nuclease generated genome survey sequence clones were obtained from NCBI dbGSS (http://ncbi.nlm.nih.gov). *Plasmodium berghei* shotgun reads with 4X genome coverage, nucleotide and amino acid sequence of automatically predicted and annotated *P.berghei* genes were obtained at the Sanger Institute (ftp://ftp.sanger.ac.uk/pub/pathogens/P_berg).

P.berghei nucleotide sequences of *P.berghei* genes that are orthologous to some *P.falciparum* genes were obtained from Hall *et.al*, 2004.

4. Methods

4.1 Alignments: defining GSS location on the *Plasmodium* genome

4.1.1 *P.falciparum*

A total of 1766 MBN generated genome survey sequence fragments were mapped to *P.falciparum* chromosomes using Megablast (V2.2.4). Megablast was run with default parameters (gap opening and extension penalties: 0.0) and good alignments were classified as those having 95% identity over a length of at least 100 base pairs (bp) at a cut off probability (e- score) of 0.0001. A search was also conducted without the identity threshold to determine whether the chosen cutoff yielded a significant fraction of the data.4039 *P.falciparum* genes with corresponding *P.berghei* orthologs were analyzed to determine their relationship with MBN cleavage sites.

4.1.2 *P.berghei*

5476 genome survey sequences were aligned to *P.berghei* contiguous sequences using Megablast (parameters and cutoffs were the same for both the *P.falciparum* and *P.berghei* datasets). Since the annotation for *P.berghei* is incomplete, there is no information on the location of *P.berghei* genes on the contigs. To establish approximate gene locations on the contigs, a *P.berghei _ P.falciparum* ortholog dataset was used (Hall, *et al*, 2004) was used .A total of 4039 *P.berghei* genes, which had corresponding orthologs in *P.falciparum*, were mapped to the contigs using NCBI BlastN.

High scoring segment pairs were considered significant when they were at least 95% identical over a length of 100 bp at a random probability of at least 0.0001. *P.berghei* hypothetical protein sequences (non-orthologs) were mapped to P.*berghei* contigs using MEGABLAST (parameters used were similar to those used for the ortholog set). The GSS and gene coordinates were extracted and then used to build a database of GSS locations with respect to contigs.

***P.FALCIPARUM* ANALYSIS**

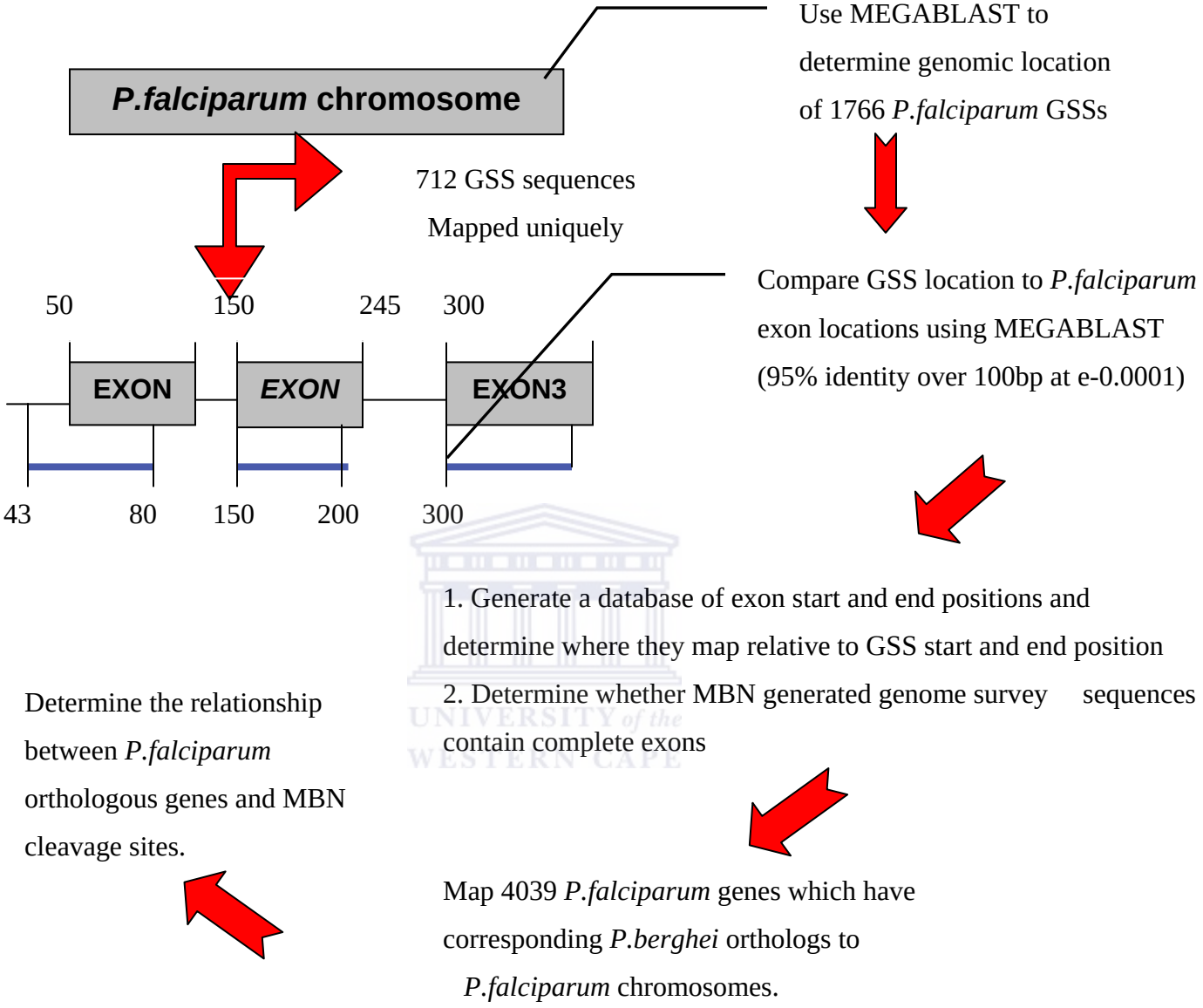

Figure 5: Schematic showing the generation of a map of MBN cleavage sites in the *P. falciparum* genome. GSSs were mapped to the chromosomes. A Database of chromosome and exon coordinates was created and later compared to the database of chromosome and GSS coordinates to determine whether MBN generated GSSs contain complete exons.

4.2 Creating a Database of exon, gene and GSS coordinates

4.2.1 *P.falciparum*

Since the *P.falciparum* genome has been fully annotated, we made use of the location of genes and their exons (whole genome annotated coding sequences in GeneBank format- http://plasmodb.org) to create a database of all exon positions in the genome and the corresponding chromosomes on which they were found. Megablast output was parsed to generate a similar database representing GSS chromosomal coordinate location across all 14 *P.falciparum* chromosomes.

4.2.2 *P.berghei*

Although sequence data was available (for 4039 *P.berghei* genes which had corresponding orthologs in *P.falciparum*), the location of these genes on the *P.berghei* genome was unknown. NCBI BlastN was used to map these orthologous genes to contigs, output was parsed and gene coordinates were extracted to create a database of contig IDs and contig positions wherein each of the genes mapped. Similarly, Megablast was used to map *P.berghei* GSSs to contigs and output was parsed to create a database of contig IDs and contig positions wherein GSSs mapped on the *P.berghei* genome.

4.3 Distance Analysis: Gene vs. GSS location on the genome

GSS**,** Gene and exon coordinate databases were compared against each other to generate a map of mung bean nuclease cleavage sites across the *P.falciparum* and *P.berghei* genome. The map of MBN cut sites was used to determine whether MBN clones contained complete exons. Perl scripts (perl version 5.005_03) were used to identify the position of GSSs relative to coding exons and coding sequences in the *P.falciparum*

genome and relative to gene positions in *P.berghei*. GSS sequences were classified according to their distance (in base pairs) with respect to the start and end of the coding sequences in the *plasmodium* genome. MBN cut sites were further categorized with respect to their proximity to coding sequences in the genomes of both the human and rodent malaria parasite.

4.4 Classification of alignments

A number of perl scripts were written to classify alignments according to:

Category A: Genome survey sequences (GSSs) mapping to intergenic /intronic regions

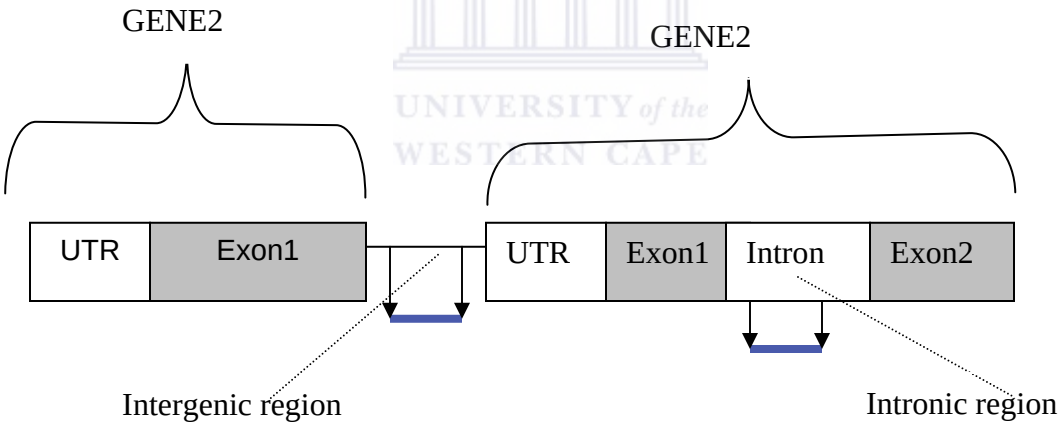

Fig6: Previous studies suggest that GSSs generated by MBN contain complete coding sequences. However, it has also been reported that MBN does cleave within some introns. Genome survey sequences that map completely to non-coding regions would illustrate that MBN digestion of *Plasmodium* DNA does not consistently (in all possible reaction conditions) yield fragments with complete coding sequences.

Category B: GSS mapping completely in coding regions

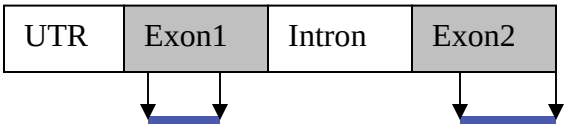

Fig7: Some projects (Nomura *et al.*, 2001;van Lin, Janse, and Waters, 2000) identified GSS fragments which appeared to contain fragments of genes with both introns and exons, rather than the intact genes expected from MBN-digestion of genomic DNA**.** This category comprises GSSs which are enriched for coding exon sequences.

Category C: GSS overlapping exon intron boundaries

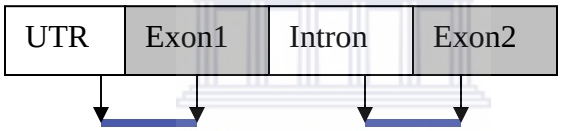

Fig8: A genome survey sequence is the sequence of one end of a single MBN clone, and is only 500-600 nucleotides, therefore, the GSS is not the complete sequence of the MBN insert. A GSS that overlaps an exon-intron boundary gives partial coverage of the coding sequence, although the original full-length MBN insert could have contained the entire coding sequence.

CHAPTER3

RESULTS

1. *P.falciparum* analysis

1.1 Genome survey sequence mapping

Despite filtering for low complexity regions, 33% of the *P.falciparum* GSS dataset was redundant. This means that these GSSs mapped to two or more separate locations either on the same or on different chromosomes, therefore they were not considered for further analysis.

1.2 Relationship between MBN generated GSSs and exons

Forty one percent (41%) of *P.falciparum* GSS fragments analyzed mapped completely within coding exon sequences. For this GSS set, the 5' terminal mapped downstream of the exon start site and the 3'terminal of the GSS mapped upstream of the exon end site (Table3). 45% of GSSs overlapped an exon-intron boundary (i.e., one end mapped within an exon sequence whilst the other mapped outside). Therefore, 86% of *P.falciparum* GSSs had at least one end mapping within a coding exon.

Furthermore, 85% of GSSs which overlapped exon-intron boundaries mapped within 50bp of the exon start and end site (Fig9). In order to distinguish between those GSSs which mapped to introns and those mapping to intergenic regions, introns and intergenic regions were distinctly defined. Results showed that of the 13% of GSSs which mapped

to non-coding regions 2% mapped to intergenic regions whilst the remaining 11% mapped to intronic regions (Table3).

Table 3 Classification of GSS location with respect to coding regions

	P.falciparum		*P.berghei*	
Total number of GSSs	1766		5476	
Total hits satisfying cutoffs	1074		4176	
Total non-redundant GSSs	702 (65%)		2450(59.4%)	
No. GSS within coding regions	294 (41.9%)		363(14.8%)	
No. GSSs overlapping 3' exon	126 (17.9%)	44.7%	184(7.5%)	16.1%
No. GSSs overlapping 5' exon	188 (26.7%)		211(8.6%)	
Total No. of GSS in non-coding regions	94 (13.4%)		40 (1.6%)	
(i) GSSs in intergenic regions	16 (2.3%)	13.4%		
(ii) GSSs in intronic regions	78 (11.1%)			

86% (608/702) of *P. falciparum* GSSs have at least one terminal mapping within an exon compared to 31% (758/2450) of *P. berghei* GSSs which have one terminal within a coding sequence. This might be attributed to the partial nature of the *P.berghei* annotation and the unavailability of *P.berghei* exon sequences.

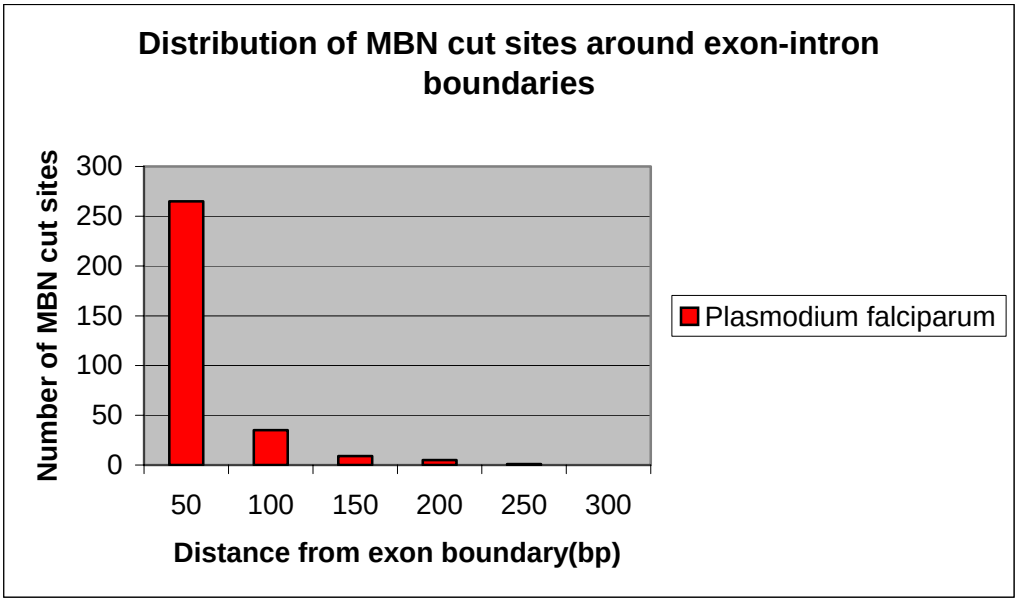

Figure9: Distribution of MBN cut sites around exon boundaries in *Plasmodium falciparum*. 85% of MBN cut sites in The *P.falciparum* genome mapped within 50 bp 5' and 3' of exon boundaries.

1.3 Location of cleavage sites proximal to *P.falciparum* orthologous genes

Analysis of the ortholog dataset revealed that 47% of GSSs mapped completely within coding regions (Fig12) whilst 32% overlapped a coding sequence boundary (Fig11) and 20% mapped to an intergenic region. In total, 79% of *P.falciparum* GSSs had at least one terminal mapping within an ortholog coding sequence. 72% of GSSs which overlapped coding sequence boundaries mapped within 50 bp of the start or end of the gene (Fig10).

2.*P.berghei* analysis

2. **1 Genome survey sequence mapping**

40% of *P.berghei* GSSs mapped to multiple locations within the *P.berghei* contiguous sequences and these were discarded from the dataset and were not considered for further

analysis. Multiple hits were noted as greater than two alignments for one consensus sequence on the same or on different contigs.

2.2 Relationship between MBN generated GSSs and automatically predicted and annotated *P.berghei* genes.

No information was available with regards to the location of exons within the *P.berghei* genome, therefore genome survey sequences could not be analyzed with respect to the location of coding exons in the *P.berghei* genome.

Analysis of the relationship between *P.berghei* genes (non-orthologs) and GSSs revealed a pattern that deviated significantly from the *P.falciparum* dataset. Only15% of GSSs mapped completely within coding regions, as opposed to 41% in *P.falciparum*. 17% of GSSs overlapped coding sequence boundaries and 2% aligned completely in intergenic regions (Table 3). A total 31% of genome survey sequences had at least one terminal (3' or 5') mapping to a coding region. This is a very low percentage when compared to 86% in *P.falciparum*.

2.3 Location of cleavage sites proximal to *P.berghei* orthologous genes

The annotation for *P.berghei* is incomplete and no exon coordinate information is available. The comparison in 2.2 above compares dissimilar sequence fragments (genes and their exons in *P.falciparum* and only predicted genes in *P.berghei*), hence the poor quality of the results. To conduct a uniform comparison, orthologous genes (genes in different species that evolved from a common ancestral gene by speciation) were used to

investigate the relationship between MBN cleavage sites in *P.falciparum* and *P.berghei* genes, which have evolved from a common ancestor. This strategy produced a pattern of results nearly consistent with the *P.falciparum* dataset in that 73% of *P.berghei* GSSs had at least one terminal mapping within a coding sequence (79% in *P.falciparum*) and 37% of these mapped between 0-50bp of the start or end of the gene. 32% of GSSs mapped completely within a coding sequence (47% in *P.falciparum*) whilst 41% overlapped a coding sequence boundary (32% in *P.falciparum*) and the remaining 22 % mapped completely within intergenic regions (Table4; Fig 10).

Table4 Classification of GSS sequences with respect to orthologous genes within *the P.berghei* and *P.falciparum*

	P.falciparum		*P.berghei*	
Total number of orthologs	1029		1029	
GSSs mapping to orthologs	112		201	
No. GSS within coding regions	53(47.3%)		63(31.3%)	
No. GSSs overlapping 3' boundary	1 6 (14.3%)	32 .2%	43 (21.4%)	41.3%
No. GSSs overlapping 5' boundary	20(17.9%)		40(19,9%)	
No. of GSSs in non-coding/intergenic regions	23 (20.5%)		55 (27.4%)	

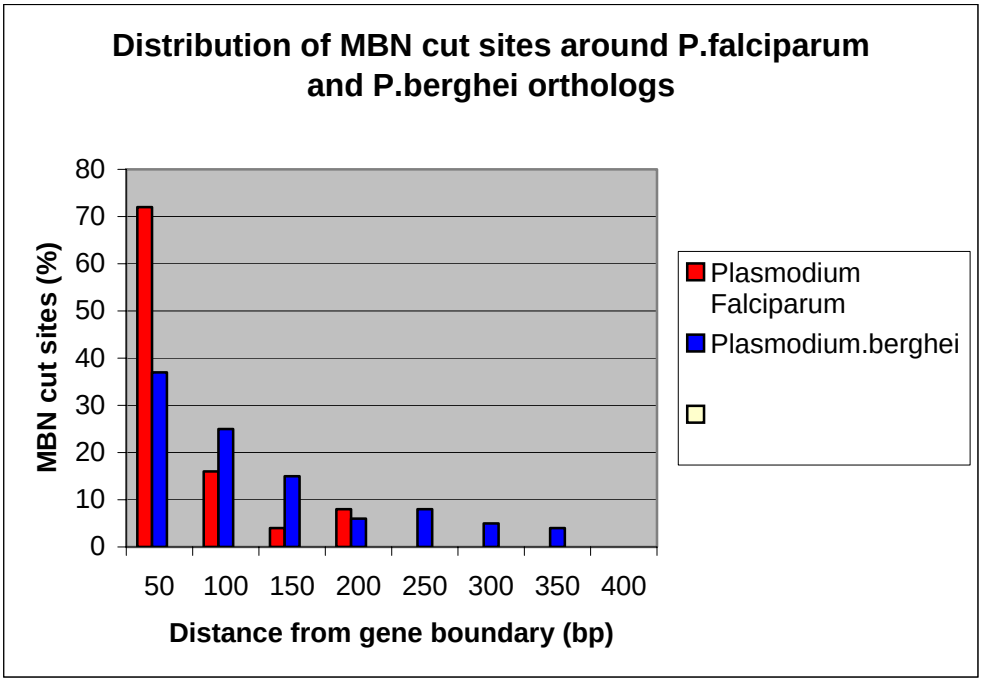

Fig10: In both the human and rodent malaria parasite, the majority of GSSs mapped between 0- 50 bp of the start and end of the gene.

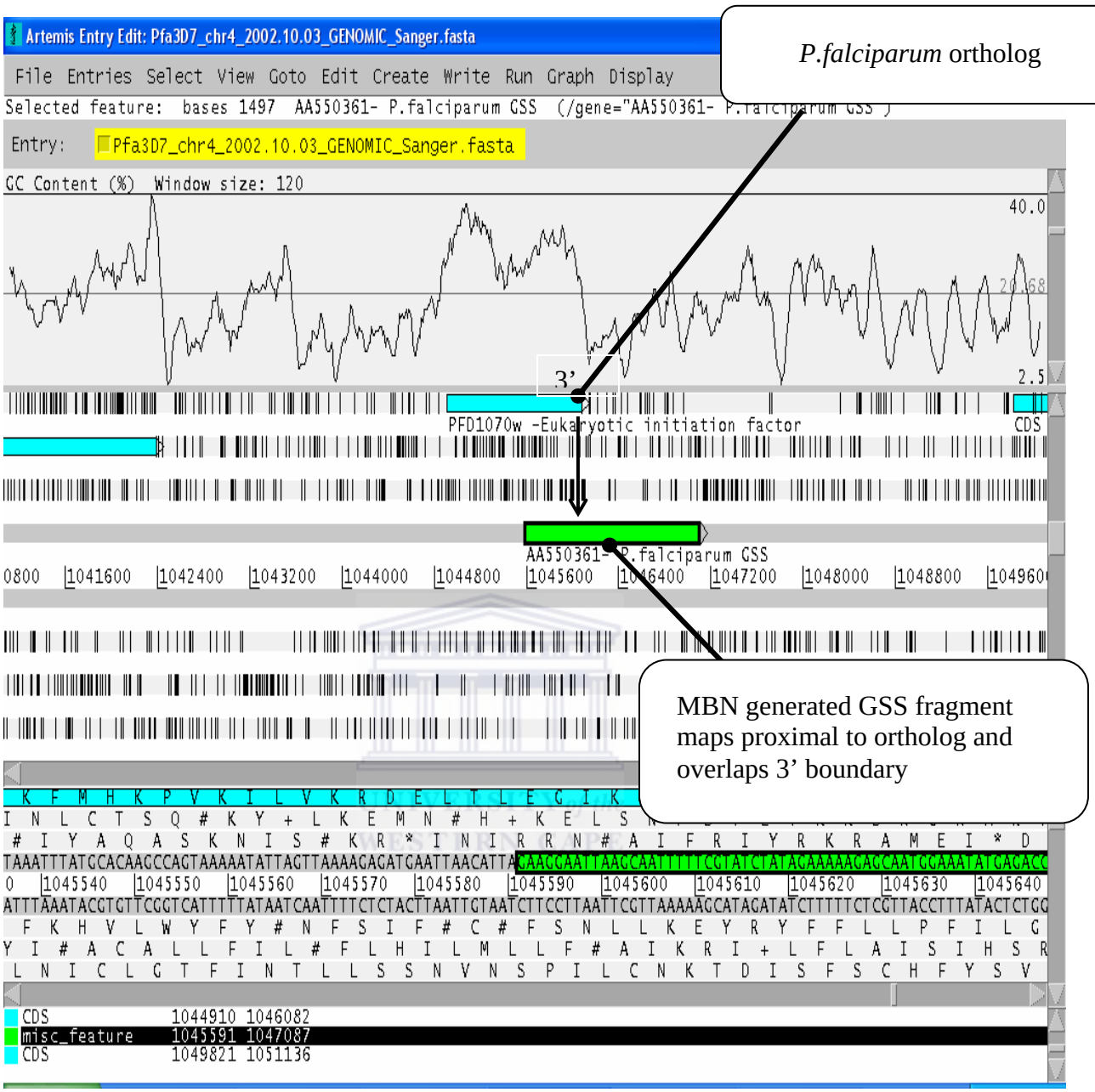

Fig 11: Screenshot of a GSS fragment which contains part of a *P.falciparum* ortholog sequence and overlaps the 3' boundary. A total 32.2% of GSSs analyzed overlapped coding sequence boundaries and 72% of these mapped within 50bp of the boundary**.** Image generated by Artemis sequence viewer (Berriman *et al*, 2000).

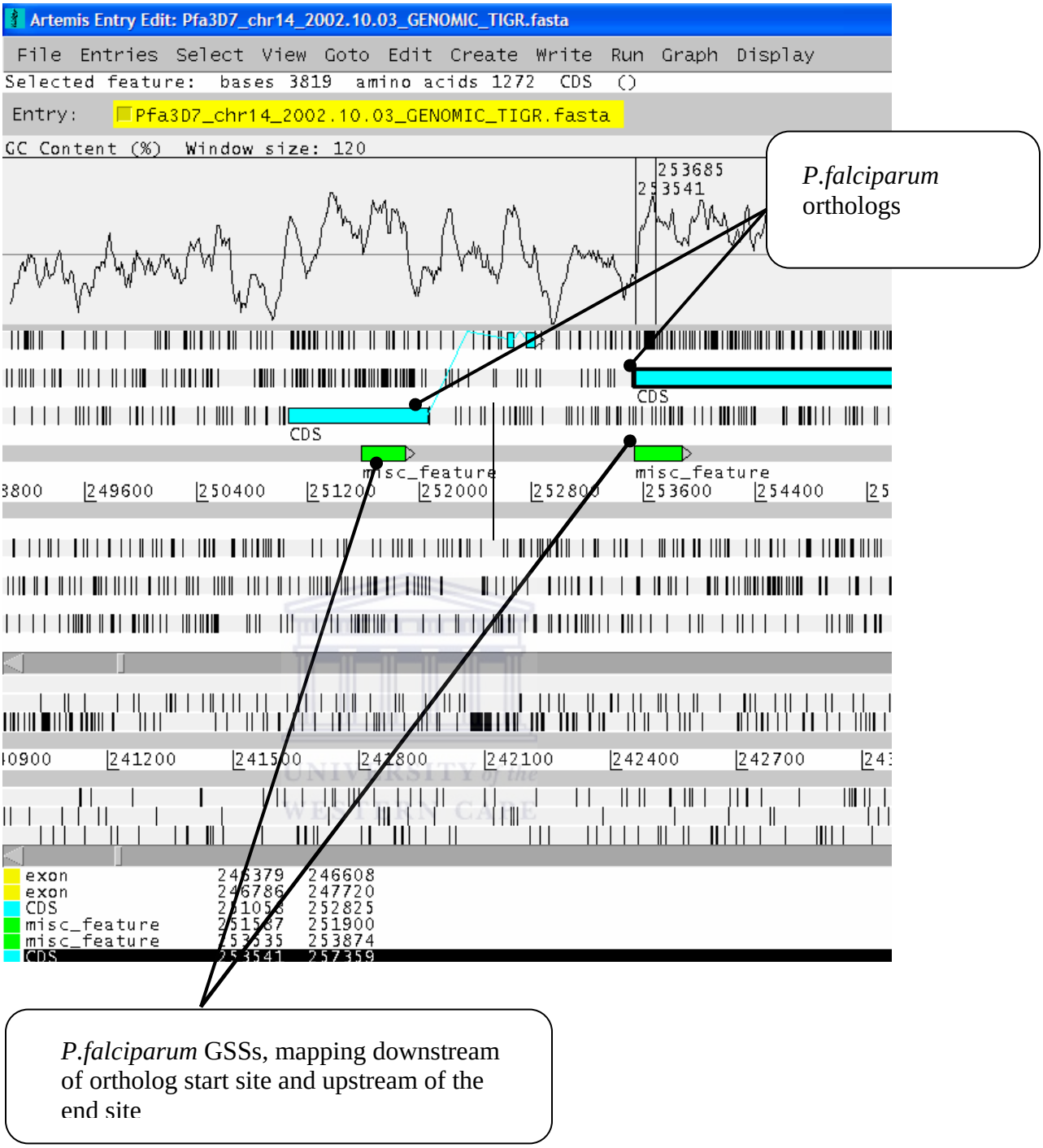

Fig12: Mung bean nuclease generated fragments mapping completely within *P.falciparum* coding sequences.

CHAPTER 4

DISCUSSION AND CONCLUSIONS

1. Effect of formamide concentration

The "genease" activity of mung bean nuclease, which has nearly the preciseness of a restriction nuclease, requires controlled conditions in the presence of formamide. It is possible that a point of cleavage that is totally spared in one concentration of formamide may be completely cleaved in another (McCutchan *et al.*, 1984). It has also been reported that, depending on the particular gene, increasing formamide above the effective concentration can result either in cleavage at less sensitive sites within the gene or else in nonspecific degradation of the sequence (Vernick, Imberski, and McCutchan, 1988). It appears that conditions of the digestion may be modified to release genes as intact coding sequences or exons.

In the human malaria parasite, *P.falciparum*, purified DNA was digested with mung bean nuclease in the presence of 30% formamide at 50° C (Vernick, Imberski, and McCutchan, 1988) ; http://www.vetmed.ufl.edu .However, the *P. berghei* genomic clone library analyzed in this study (Pb MBN library #21) was digested with MBN in the presence of 36-38% formamide at 50°C (Vernick, Imberski, and McCutchan, 1988). The varying reaction conditions under which both of these genomic libraries were prepared might account for some of the differences observed in the results for each data set. As a consequence of these differences, Table 3 suggests a non- specific degradation of

P.berghei genomic DNA, since there seems to be no correlation between these results and the expected " genease activity of MBN that was observed in *P.falciparum* DNA. As an explanation for this, it was initially thought that the MBN mechanism was species specific, but the analysis of the genease activity with respect to particular *P.beghei* genes which have orthologs in *P.falciparum* significantly improved the results.

2. Availability of complete genomic DNA sequence data

A complete reference genome for the human malaria parasite, *P.falciparum* is now available in public databases (Gardner *et al.*, 2002)**.** The information available includes DNA sequence data and curated annotations, automated gene model predictions, predicted proteins and protein motifs, cross-species comparisons, expression data generated by a variety of complementary strategies, and proteomics data (Kissinger *et al.*, 2002). This made the analysis of *P.falciparum* genomic DNA libraries much easier than *P.berghei* GSS data since the *P.berghei* annotation is incomplete.

Currently sequence information that is available comes from only 4X coverage of the *P.berghei* genome. The partial nature of *P.berghei* genome data did not provide sufficient genome coverage to make verifiable inferences with regards to the relationship between *P.berghei* genes/exons and mung bean nuclease cleavage sites. Characterization of *P.berghei* clones indicated that the DNA is over digested and contains both full-length genes and shards of genes (Waters *et.al,* 2002). Therefore, some of the difficulties encountered indicate that there is a need to confirm, correct and expand the annotation and ascribe function and location of genes, exons and introns in the *P.berghei* genome.

3. Location of genease cleavage sites near *Plasmodium* genes

It appears from these and other observations that MBN predominantly cleaves genomic DNA in *Plasmodium* proximal to coding sequences. Although the results were different when comparing MBN cleavage sites to *P.falciparum* genes and their exons and to *P.berghei* predicted and annotated genes (non-orthologs-Table 3), using common ortholog genes between these species significantly improved the results (Table 4).

A close examination of the termini of cloned inserts revealed that, in total, 79% of *P.falciparum* GSSs had at least one end mapping within an ortholog coding sequence and 72% of GSSs which overlapped coding sequence boundaries mapped within 50 bp of the start or end of the gene (Fig 10). Similarly, 73% of *P.berghei* GSSs had at least one terminal mapping within an ortholog coding sequence and 37% of these mapped between 0-50 bp of the start or end of the gene. This indicates that a larger percentage of cleavage sites just outside coding regions.

These results are consistent with previous studies which demonstrated a MBN generated fragment which was cut 52bp in front of the gene and 60bp after it (McCutchan *et al.*, 1984). Other fragments, had sites either 10 or 11 bp from the start of the gene and sites either 25 or 35bp from the 3' end of the gene.

In African trypanosomes, variant surface glycoprotein (VSG) genes were flanked by MBN susceptible sites that were closer to coding regions (KH Brown, 1986). DNA from both humans and Drosophila have been cleaved under slightly different conditions with similar success, but the number of analyses performed to date is limited (Vernick and

McCutchan 1998). This indicates that *Plasmodium* DNA is not unique in its susceptibility to this type of cleavage.

However, our observations also show that introns and intergenic regions can contain within them specific MBN cleavage sites. Comparing the sequences at the ends of cloned mung bean fragments and available *P.falciparum* intronic sequences showed that, of the 13% of GSSs which mapped to non-coding regions, 2% mapped to intergenic regions whilst the remaining 11% mapped to intronic regions. The cleavages which do occur within certain introns may be informative by virtue of not being general events.

Finally, to make a final determination of whether the genease activity of mung bean nuclease is fact or fiction we answer the following questions:

(i) Does MBN digestion of *Plasmodium* DNA generate complete exons?

We have found that 41% of *P.falciparum* GSS fragments analyzed mapped completely within a coding exon sequence**.** These results could suggest one of two things:

(i) MBN generated fragments do not necessarily contain complete exon sequences since one terminal should ideally map outside the exon for the fragment to contain the entire exon sequence.

(ii) Since the GSS fragment does not represent the full-length sequence of the MBN insert and only one end of the clone insert is sequenced, it is possible that (with full coverage) the rest of the clone does contain the entire coding exon.

Since these two possibilities do not answer the question, we take into account the fact that 45% of *P.falciparum* GSSs overlapped an exon-intron boundary and in total 86% of GSSs had at least one terminal mapping within a coding exon. Furthermore, 85% of GSSs which overlapped exon-intron boundaries mapped within 50bp of the exon start and end site (fig9). The fact that more than 80% of these cut sites lie within 50bp of the exon boundary strongly suggests that MBN digestion of *P.falciparum* DNA does in part generate complete exons or at least fragments that are enriched for coding exon sequences.

(ii) Do all or only a subset of genes in *Plasmodium* contain flanking mung bean nuclease recognition signals that will allow their precise excision under certain reaction conditions?

The results presented herein were obtained from analysis of several thousand *Plasmodium* genes which have different coding sequences, in different locations on individual chromosomes/contigs. We have demonstrated that mung bean nuclease preferentially cleaves at sites proximal to/ just outside gene boundaries. These results can be explained by the fact that these genes contain hypersensitive sites surrounding coding regions which become sensitive to nuclease cleavage under defined reaction conditions.

As previous studies have suggested, it seems more likely that a discreet and locally-defined structural form in the DNA that flanks genes is revealed to the enzyme by the conditions in the reaction mixture. The structural characteristics of the sites recognized

by MBN are not known, however, it is clear that the cleavage is not the result of primary sequence recognition (McCutchan *et al*, 1984).

(iii) How general is the MBN phenomenon? is it limited to specific species of *Plasmodia* ?

Analysis of the ortholog set of rodent and human malaria parasite genes revealed that in both species, above 70% of GSSs mapped proximal to coding sequences and largely within 50 bp of the coding sequence boundary. This suggests that fragments generated by MBN digestion in both species are enriched for coding regions. This indicates that the MBN phenomenon is a general event that is not limited to specific species of *Plasmodium*. As previously mentioned, DNA from humans, Drosophila, Trypanosoma, Giardia, Leshmania and other protozoans have been cleaved by this enzyme under slightly different conditions with similar success.

4. Overall Conclusions

The results of this study have shown that the MBN mechanism does in part yield fragments which are enriched for coding regions. The enzyme preferentially cleaves proximal to coding regions, although there are intronic/intergenic regions which contain specific cleavage sites sensitive to cleavage. The fact that some GSSs mapped completely to non- coding/ intergenic regions, suggests that MBN cleavage of *Plasmodium* DNA does not always yield complete exons.

The large number of genes surveyed in two different species of *Plasmodium* indicates that MBN mechanism is neither species specific nor is it limited to specific genes.

Reference List

Adam,R.D., Nash,T.E., and Wellems,T.E. (1988) The Giardia lamblia trophozoite contains sets of closely related chromosomes *Nucleic Acids Res.* **16**: 4555-4567.

Altschul,S.F., Gish,W., Miller,W., Myers,E.W., and Lipman,D.J. (1990) Basic local alignment search tool *J.Mol.Biol.* **215**: 403-410.

Bahl,A., Brunk,B., Crabtree,J., Fraunholz,M.J., Gajria,B., Grant,G.R., Ginsburg,H., Gupta,D., Kissinger,J.C., Labo,P., Li,L., Mailman,M.D., Milgram,A.J., Pearson,D.S., Roos,D.S., Schug,J., Stoeckert,C.J., Jr., and Whetzel,P. (2003) PlasmoDB: the Plasmodium genome resource. A database integrating experimental and computational data *Nucleic Acids Res.* **31**: 212-215.

Breman,J.G. (2001) The ears of the hippopotamus: manifestations, determinants, and estimates of the malaria burden *Am.J.Trop.Med.Hyg.* **64**: 1-11.

Brown,K.H., Brentano,S.T., and Donelson,J.E. (1986) Mung bean nuclease cleaves preferentially at the boundaries of variant surface glycoprotein gene transpositions in trypanosome DNA *J.Biol.Chem.* **261**: 10352-10358.

Carlton,J.M., Angiuoli,S.V., Suh,B.B., Kooij,T.W., Pertea,M., Silva,J.C., Ermolaeva,M.D., Allen,J.E., Selengut,J.D., Koo,H.L., Peterson,J.D., Pop,M., Kosack,D.S., Shumway,M.F., Bidwell,S.L., Shallom,S.J., van Aken,S.E., Riedmuller,S.B., Feldblyum,T.V., Cho,J.K., Quackenbush,J., Sedegah,M., Shoaibi,A., Cummings,L.M., Florens,L., Yates,J.R., Raine,J.D., Sinden,R.E., Harris,M.A., Cunningham,D.A., Preiser,P.R., Bergman,L.W., Vaidya,A.B., van Lin,L.H., Janse,C.J., Waters,A.P., Smith,H.O., White,O.R., Salzberg,S.L., Venter,J.C., Fraser,C.M., Hoffman,S.L., Gardner,M.J., and Carucci,D.J. (2002) Genome sequence and comparative analysis of the model rodent malaria parasite Plasmodium yoelii yoelii *Nature* **419**: 512-519.

Carlton,J.M., Muller,R., Yowell,C.A., Fluegge,M.R., Sturrock,K.A., Pritt,J.R., Vargas-Serrato,E., Galinski,M.R., Barnwell,J.W., Mulder,N., Kanapin,A., Cawley,S.E., Hide,W.A., and Dame,J.B. (2001) Profiling the malaria genome: a gene survey of three species of malaria parasite with comparison to other apicomplexan species *Mol.Biochem.Parasitol.* **118**: 201-210.

Carlton,J.M., Vinkenoog,R., Waters,A.P., and Walliker,D. (1998) Gene synteny in species of Plasmodium *Mol.Biochem.Parasitol.* **93**: 285-294.

Dame,J.B., Williams,J.L., McCutchan,T.F., Weber,J.L., Wirtz,R.A., Hockmeyer,W.T., Maloy,W.L., Haynes,J.D., Schneider,I., and Roberts,D. (1984) Structure of the gene encoding the immunodominant surface antigen on the sporozoite of the human malaria parasite Plasmodium falciparum *Science* **225**: 593-599.

El Sayed,F., Elbadir,S., Ferrere,J., Marguery,M.C., and Bazex,J. (1997) Chronic balanitis: an unusual localisation of necrobiosis lipoidica *Genitourin.Med.* **73**: 579-580.

Florea,L., Hartzell,G., Zhang,Z., Rubin,G.M., and Miller,W. (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence *Genome Res.* **8**: 967-974.

Gardner,M.J., Hall,N., Fung,E., White,O., Berriman,M., Hyman,R.W., Carlton,J.M., Pain,A., Nelson,K.E., Bowman,S., Paulsen,I.T., James,K., Eisen,J.A., Rutherford,K., Salzberg,S.L., Craig,A., Kyes,S., Chan,M.S., Nene,V., Shallom,S.J., Suh,B., Peterson,J., Angiuoli,S., Pertea,M., Allen,J., Selengut,J., Haft,D., Mather,M.W., Vaidya,A.B., Martin,D.M., Fairlamb,A.H., Fraunholz,M.J., Roos,D.S., Ralph,S.A., McFadden,G.I., Cummings,L.M., Subramanian,G.M., Mungall,C., Venter,J.C., Carucci,D.J., Hoffman,S.L., Newbold,C., Davis,R.W., Fraser,C.M., and Barrell,B. (2002) Genome sequence of the human malaria parasite Plasmodium falciparum *Nature* **419**: 498-511.

Gentner,D., Gunn,V. (2001) Structural alignment facilitates the noticing of differences *Mem.Cognit.* **29**: 565-577.

Gilbert,W. (1991) Towards a paradigm shift in biology *Nature* **349**: 99.

Hall Neil, Marianna Karras, J. Dale Raine, Jane M.
Carlton, Taco W.A. Kooij, Matthew Berriman, Laurence Florens Ç,
Christoph S. Janssen, Arnab Pain, Georges K. Christophides, Keith
James, Kim Rutherford, Barbara Harris, David Harris, Carol
Churcher, Michael A. Quail, Doug Ormond, Jon Doggett, Holly E.
Trueman, Jacqui Mendoza, Shelby L. Bidwell, Marie-Adele Rajandream1,
Daniel J. Carucci, John R. Yates III, Fotis C. Kafatos, Chris J.
Janse, Bart Barrell, C. Michael R. Turner, Andrew P. Waters*,Robert E. Sinden (2004) . A comprehensive survey of the malaria parasite life-cycle by integrated genomic, transcriptomic and proteomic analyses. *Science in press.*

Hayward,R.E., DeRisi,J.L., Alfadhli,S., Kaslow,D.C., Brown,P.O., and Rathod,P.K. (2000) Shotgun DNA microarrays and stage-specific gene expression in Plasmodium falciparum malaria *Mol.Microbiol.* **35**: 6-14.

Hoffman,S.L., Bancroft,W.H., Gottlieb,M., James,S.L., Burroughs,E.C., Stephenson,J.R., and Morgan,M.J. (1997) Funding for malaria genome sequencing *Nature* **387**: 647.

Hoffman,S.L., Subramanian,G.M., Collins,F.H., and Venter,J.C. (2002) Plasmodium, human and Anopheles genomics and malaria *Nature* **415**: 702-709.

Janse,C.J., Carlton,J.M., Walliker,D., and Waters,A.P. (1994) Conserved location of genes on polymorphic chromosomes of four species of malaria parasites *Mol.Biochem.Parasitol.* **68**: 285-296.

Janse,C.J., Waters,A.P. (1995) Plasmodium berghei: The application of cultivation and purification techniques to molecular studies of malaria parasites *Parasitol.Today* **11**: 138-143.

Johnson,A.M., Illana,S., Dubey,J.P., and Dame,J.B. (1987) Toxoplasma gondii and Hammondia hammondi: DNA comparison using cloned rRNA gene probes *Exp.Parasitol.* **63**: 272-278.

Johnson,P.H., Laskowski,M., Sr. (1970) Mung bean nuclease I. II. Resistance of double stranded deoxyribonucleic acid and susceptibility of regions rich in adenosine and thymidine to enzymatic hydrolysis *J.Biol.Chem.* **245**: 891-898.

Kabotyanski,E.B., Zhu,C., Kallick,D.A., and Roth,D.B. (1995) Hairpin opening by single-strand-specific nucleases *Nucleic Acids Res.* **23**: 3872-3881.

Kissinger,J.C., Brunk,B.P., Crabtree,J., Fraunholz,M.J., Gajria,B., Milgram,A.J., Pearson,D.S., Schug,J., Bahl,A., Diskin,S.J., Ginsburg,H., Grant,G.R., Gupta,D., Labo,P., Li,L., Mailman,M.D., McWeeney,S.K., Whetzel,P., Stoeckert,C.J., and Roos,D.S. (2002) The Plasmodium genome database *Nature* **419**: 490-492.

Kowalski,D., Kroeker,W.D., and Laskowski,M., Sr. (1976) Mung bean nuclease I. Physical, chemical, and catalytic properties *Biochemistry* **15**: 4457-4463.

Lasonder,E., Ishihama,Y., Andersen,J.S., Vermunt,A.M., Pain,A., Sauerwein,R.W., Eling,W.M., Hall,N., Waters,A.P., Stunnenberg,H.G., and Mann,M. (2002) Analysis of the Plasmodium falciparum proteome by high-accuracy mass spectrometry *Nature* **419**: 537-542.

Liu,J., Zhou,G.Q., and Straby,K.B. (1999) Caenorhabditis elegans ZC376.5 encodes a tRNA (m2/2G(26))dimethyltransferance in which (246)arginine is important for the enzyme activity *Gene* **226**: 73-81.

McCutchan,T.F., Hansen,J.L., Dame,J.B., and Mullins,J.A. (1984) Mung bean nuclease cleaves Plasmodium genomic DNA at sites before and after genes *Science* **225**: 625-628.

Mount, D.W 2001. Bioinformatics: Sequence and Genome analysis.

Muhich,M.L., Simpson,L. (1986) Specific cleavage of kinetoplast minicircle DNA from Leishmania tarentolae by mung bean nuclease and identification of several additional minicircle sequence classes *Nucleic Acids Res.* **14**: 5531-5556.

Nomura,T., Carlton,J.M., Baird,J.K., del Portillo,H.A., Fryauff,D.J., Rathore,D., Fidock,D.A., Su,X., Collins,W.E., McCutchan,T.F., Wootton,J.C., and

Wellems,T.E. (2001) Evidence for different mechanisms of chloroquine resistance in 2 Plasmodium species that cause human malaria *J.Infect.Dis.* **183**: 1653-1661.

Rathore,D., McCutchan,T.F. (2002) Construction of a gene library with mung bean nuclease-treated genomic DNA *Methods Mol.Med.* **72**: 253-263.

Reddy,G.R., Chakrabarti,D., Schuster,S.M., Ferl,R.J., Almira,E.C., and Dame,J.B. (1993) Gene sequence tags from Plasmodium falciparum genomic DNA fragments prepared by the "genease" activity of mung bean nuclease *Proc.Natl.Acad.Sci.U.S.A* **90**: 9867-9871.

Smith,M. (1998) The genome--where to from here? *Can.J.Cardiol.* **14**: 1343-1347.

Strong,W.B., Nelson,R.G. (2000) Gene discovery in Cryptosporidium parvum: expressed sequence tags and genome survey sequences *Contrib.Microbiol.* **6**: 92-115.

Tetzlaff,C.L., McMurray,D.N., and Rice-Ficht,A.C. (1990) Isolation and characterization of a gene associated with a virulent strain of Babesia microti *Mol.Biochem.Parasitol.* **40**: 183-192.

Thompson,J., Janse,C.J., and Waters,A.P. (2001) Comparative genomics in Plasmodium: a tool for the identification of genes and functional analysis *Mol.Biochem.Parasitol.* **118**: 147-154.

Tripp,C.A., Wagner,G.G., and Rice-Ficht,A.C. (1989) Babesia bovis: gene isolation and characterization using a mung bean nuclease-derived expression library *Exp.Parasitol.* **69**: 211-225.

van Lin,L.H., Janse,C.J., and Waters,A.P. (2000) The conserved genome organisation of non-falciparum malaria species: the need to know more *Int.J.Parasitol.* **30**: 357-370.

Vernick,K.D., Imberski,R.B., and McCutchan,T.F. (1988) Mung bean nuclease exhibits a generalized gene-excision activity upon purified Plasmodium falciparum genomic DNA *Nucleic Acids Res.* **16**: 6883-6896.

Vernick,K.D., McCutchan,T.F. (1998) A novel class of supercoil-independent nuclease hypersensitive site is comprised of alternative DNA structures that flank eukaryotic genes *J.Mol.Biol.* **279**: 737-751.

Xodo,L.E., Manzini,G., Quadrifoglio,F., van der,M.G., and van Boom,J. (1991) DNA hairpin loops in solution. Correlation between primary structure, thermostability and reactivity with single-strand-specific nuclease from mung bean *Nucleic Acids Res.* **19**: 1505-1511.

Yap,M.W., Kara,U.A., Heggeler-Bordier,B., Ting,R.C., and Tan,T.M. (1997) Partial nucleotide sequence and organisation of extrachromosomal plastid-like DNA in Plasmodium berghei *Gene* **200**: 91-98.

Zhang,Z., Schwartz,S., Wagner,L., and Miller,W. (2000) A greedy algorithm for aligning DNA sequences *J.Comput.Biol.* **7**: 203-214.