COMPREHENSIVE SEQUENCING PROPOSAL FOR PLASMODIUM

PREPARED BY THE PLASMODIUM WRITING GROUP*

Executive Summary

The malaria parasite is one of the most widespread eukaryotic pathogens in the world today. Malaria remains a major global health threat with an estimated 300 – 500 million cases per year and deaths of over a million children in Africa. Resistance to currently effective drugs—including drugs given in combinations to decrease the spread of resistance—is rapidly expanding. The next five years will see billions of dollars spent implementing interventions, particularly in Africa. Past experience suggests that large scale deployment of control measures will lead to selection for drug resistance in the malaria parasite and insecticide resistance in the mosquito vector. The combined need to have genomic tools to monitor the evolutionary consequences of these large scale interventions and to discover new therapies and diagnostics provides a powerful argument to support the overall goal of this white paper to prepare for malaria disease surveillance and control in the 21st century.

We propose here a comprehensive sequencing program with the sequencing of a total of approximately 16 billion base pairs (equivalent to a 5X coverage of the human genome) including significant additional sequencing of *Plasmodium falciparum*, the most deadly form of human malaria, additional sequencing of *Plasmodium vivax*, the second most prevalent of the human malarias, and sequencing of selected non-human primate, other mammalian, avian and reptile malarias and related apicomplexan parasites.

This sequencing effort will complement both the previous work and the ongoing efforts at the Wellcome Trust Sanger Institute (WTSI) including the sequencing of the P. malariae and P. ovale. This white paper was prepared in full consultation with WTSI. In addition to genomic sequencing we propose EST and full-length cDNA sequencing for each of the major species to be sequenced with the express goal of improving the annotation of the genomes. Compared to model organisms, our knowledge of the biology of *Plasmodium* and its nearly 200 named species is limited. Over 50% of the predicted genes from the few available genomes are of unknown function and key processes are yet to be defined at the molecular level.

Species or Group	Number of Genomes at 8X Coverage	Number of Genomes at 3X Coverage	Genome length (Mb)	Total Bases (billions) of Whole Genome Sequence Requested	Number of ESTs
P. falciparum	24	50	25	8.55	200,000
P.vivax—Approved	6*				40,000*
P. vivax—This Proposal		9	27	0.73	
NonHuman Primate	4	2	27	1.03	160,000
Rodent	6	3	27	1.54	80,000
Other (Deep Branch Sequencing)	6		27	1.30	240,000

**denotes sequencing already approved (see Coordinating Committee minutes August 26, 2007)*

This proposal has received broad input from the malaria and infectious diseases research community including open discussions at recent meetings and conference calls and e-mail correspondence with over 50 leading researchers in the field. The advances in technology coupled with the increased public emphasis on addressing the major infectious diseases of HIV, Malaria and TB makes this a timely proposal and one that has broad community support.

Background

Significance of Disease

The malaria parasite is one of the most widespread eukaryotic pathogens in the world today. Malaria remains a major global health threat with an estimated 300 – 500 million cases per year and deaths of over a million children in Africa. The infection and disease have been major determinants in human genetics evidenced by increased prevalence of protective mutations such as sickle cell hemoglobin in African populations. Though it was eliminated in the last century from previous strongholds in North America and Europe, it continues to serve as a major source of morbidity and mortality in tropical and sub-tropical regions, and its medical and productivity costs are a major impediment to economic development in the mostly poor countries where it remains. Resistance to currently effective drugs—including drugs given in combinations to decrease the spread of resistance—is rapidly expanding. Though promising vaccine candidates are being developed, each of them has had important limitations, and once effective vaccines are available it is expected that the parasite's rich capacity for genetic variation will contribute significantly to vaccine failure.

Modern methods of genomics take advantage of genetic variation to identify single nucleotide polymorphisms (SNPs) that are genetically linked to clinically important mutations. The most effective source of SNP discovery is genomic sequencing; the most effective approaches to association studies make use of SNPs that are polymorphic in multiple populations throughout the world. Association studies using SNPs from strains representing a broad geographic catalog will be important for identifying previously unrecognized mutations leading to drug resistance, compensatory mutations that may restore fitness to organisms that have acquired drug resistance, and polymorphic genes associated with transmissibility, immunogenicity, and virulence of the parasite in the field.

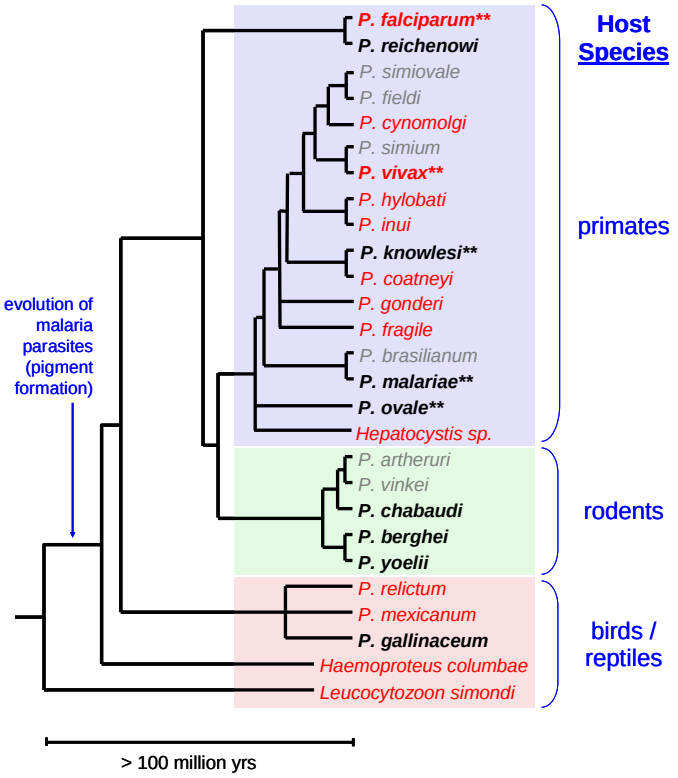
Ambitious goals are being set to reduce childhood mortality due to malaria in response to the Millennium development goals [1]. Significant funding is now available to roll out proven cost effective interventions such as indoor residual spraying (IRS), insecticide treated bednets (ITN's) and artemisinin combination therapy (ACT). The US government is contributing significantly to these programs via President's Malaria Initiative, USAID, UNICEF Child Survival Funding. The Gates Foundation is exploring the even more ambitious goal of malaria eradication. The next five years will see billions of dollars spent implementing these interventions, particularly in Africa. Past experience suggests that large scale deployment of control measures will lead to selection for drug resistance in the malaria parasite and insecticide resistance in the mosquito vector. The combined need to have genomic tools to monitor the evolutionary consequences of these large scale interventions and to design new therapies provides a powerful argument to support the overall goal of this white paper to prepare for malaria disease surveillance and control in the 21st century.

Overall vision

We propose here a comprehensive sequencing program (see Table 2) including significant additional sequencing of *P. falciparum*, the most deadly form of human malaria, additional sequencing of *P. vivax*, the second most prevalent of the human malarias and sequencing of selected non-human primate, other mammalian, avian and reptile malarias and related apicomplexan parasites. This effort will fit into the context of previous work and complement ongoing work at the WTSI. The proposed work will focus on the important human pathogens, in particular on the detection of diversity within the world populations, and will complement that body of knowledge with a comprehensive analysis of the most closely related parasites, those that infect non-human primates for comparative genomic analysis. Sequencing of the rodent malarias will provide not only important information for comparative genomic analysis features but also critical information for drug and vaccine development studies as these are the commonly used small animal model systems for drug and vaccine testing.

In addition to genomic sequencing we propose EST and full-length cDNA for each of the major species to be sequenced with the express goal of improving the annotation of the genomes. Compared to model organisms, our knowledge of malaria biology is limited. Over 50% of the predicted genes are of unknown function and key processes are yet to be defined at the molecular level. A major feature of malaria infection in humans is that a eukaryotic cell lives within a cell. There are few other examples of this in nature and this must have critical implications for core processes such as metabolism, signaling, nutrient acquisition and replication. In addition a set of unique processes such as invasion, host cell membrane remodeling, protein trafficking and sexual differentiation have evolved for effective survival and transmission. Adaptation to a completely new environment in the mosquito requires changes in fundamental machinery such as that required for protein synthesis and activation of alternative metabolic processes.

The detailed delineation of the proposed sequencing is outlined in Table 2 and is supported by the phylogenetic tree presented in Figure 1. *Plasmodium* is a diverse and ancient genus within the Apicomplexa, closely related to the other Apicomplexan parasitic genera *Haemoproteus* and *Leukocytozoon*. *Plasmodium* parasites infect a taxonomically wide range of vertebrate hosts, include primates, rodents, birds, and reptiles. Though capable of switching hosts, phylogenetic evidence suggests that most host transfers occur between similar hosts. Figure 1 shows the phylogenetic history of the genus as inferred from a combination of mitochondrial and nuclear datasets. Of the *Plasmodium* parasites causing human infection, *P. vivax*, *P. ovale*, and *P. malariae* seem to be relatively recently derived from non-human parasitic species. *P. falciparum*, however, appears to be an ancient human parasite. *P. falciparum* is separated phylogenetically from most other primate parasites by a very long branch, and shows close kinship only with the chimpanzee parasite *P. reichenowi*, suggesting that the common ancestor of *P. falciparum* and *P. reichenowi* infected the human/chimp common ancestor 5-6 million years ago.



** Denotes species that infect humans
Bold denotes species already sequenced or approved for sequencing
Red indicates species for which sequencing is requested

We here propose sequencing of __24__ new strains of *P. falciparum* to 8 fold coverage and of __50__ geographical isolates to 3-4 fold coverage, and complete sequencing and closure of _15__ critical model organisms. The extensive sequencing of *P. falciparum* strains is expected to give coverage sufficient to

increase the number of discovered SNPs from the 100,000 currently available to 350,000. The model system work will bring the genetics and biochemistry of important processes including invasion and intracellular survival into reach.

This proposal has received broad input from the malaria and infectious disease community including open discussions at recent meetings (Malaria Diversity, Cambridge MA June 2007; Malaria Gordon Conference, Oxford UK September 2007; Comparative Biology of Plasmodium, New York NY September 2007 and at the broadly attended annual meeting of the American Society for Tropical Medicine and Hygiene in Philadelphia PA, November 2007), conference calls, and e-mail correspondence with over 50 leading researchers in the field. Advances in technology and increased public emphasis on addressing the major infectious diseases of HIV, malaria, and tuberculosis both make this a timely proposal. It has broad community support. Recent successes in mechanical disruption of the malaria cycle with bed nets, promising new vaccine candidates, and drugs in the pipeline make action even more critical: *Plasmodium* is good at surviving, and these promising advances are likely to be followed by observation of compensating changes in parasite and vector that allow malaria to continue. The proposed new sequencing will provide new data critical for linking parasite genotypes to phenotypes and understanding the current distribution of parasite traits. It will establish a baseline for understanding how parasites change over time when faced with these coming new pressures. The proposed work focuses on three objectives:

- generating new sequence that will allow discovery of more SNPs in the human malarias *P. falciparum* and *P. vivax*, greatly expanding capacity for association mapping and providing better representation of sequences from organisms collected across the world's most malarious regions;
- comparative genomics of well established animal models of the major human malarias
- sequencing for the first time examples of several of the less well studied malarial parasites, which are associated with a range of non-human organisms from lizards and birds to small mammals.

The new information generated will be rapidly integrated into existing bioinformatic resources that are well used by the malaria research community.

Plasmodium falciparum **sequencing priorities**

P. falciparum is the most important human malaria in terms of morbidity and mortality. Genetic diversity among parasites contributes directly to drug resistance and antigenic variation, which allow the parasite to evade chemotherapy and immune pressure. Knowledge of this genetic variation is critical for the development of intervention strategies including drugs and vaccines. Although there is evidence of a more recent expansion of the parasite and gene flow among regions, the original dispersal of *falciparum* malaria into the human population occurred sufficiently long ago to permit accumulation of significant genetic differences between *falciparum* strains afflicting human populations in different regions of the world.

A limited number of *P. falciparum* genomes have been sequenced to date (Table 2). These include full assembly-level sequence coverage of 5 strains and skim sequencing (initially 0.25x, now 1.25x) of 9 worldwide strains, though most of this sequence information is derived from parasites representing only a subset of regions in Africa and Asia. Analysis of these limited sequences already indicates that the *P. falciparum* genome is more diverse than previously believed, that most of the variation detected thus far reflects rare variants in the populations tested, and that there is substantial genetic differentiation among populations, both in the polymorphisms present and the extent of linkage disequilibrium (LD) between polymorphisms.

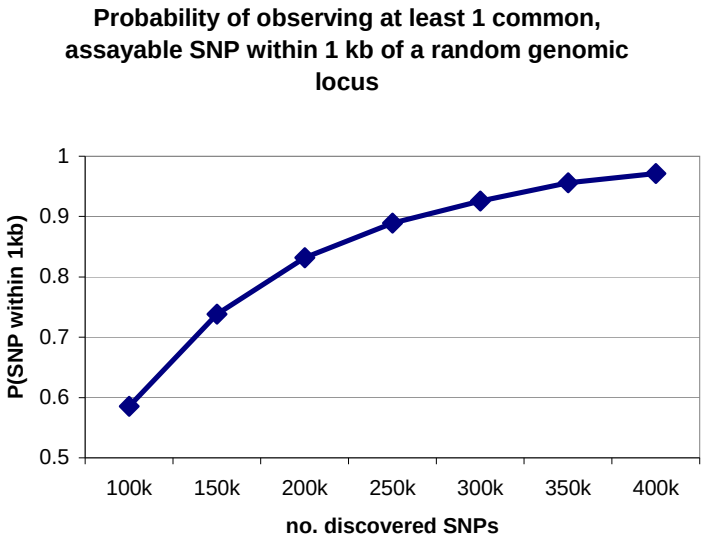
While the existing set of known SNPs may be sufficient to begin mapping strong selective sweeps and analyzing patterns of LD, association mapping will require a denser SNP map. For an association study to be successful, a genomic locus underlying a phenotype of interest must be in strong LD with at least one 'common' (MAF > 5%) SNP marker. We can approximate the expected distance between a random genomic locus and a common SNP by making a few assumptions. If SNPs are distributed through the genome in a Poisson fashion, the distance

between a random locus and a common, assayable SNP will be geometrically distributed according to *p*, where *p* is the frequency of common, assayable SNPs. We can estimate *p* as (no. SNPs/genome size)*(*c*)*(*a*), where *c* represents the proportion of common SNPs and *a* represents the proportion of SNPs assayable using high throughput genotyping. Assuming *c* and *a* are both one third, with our current tally of approximately 100,000 SNPs, *p* is 0.0005, and under the geometric distribution the probability that a common assayable SNP will occur within 1 kb of a random genomic locus is 59% (Figure 2).

Not all SNP associations will rise to significance at the genome-wide level, however, especially if the SNPs are low in frequency. So, it would be preferable to find multiple SNPs in strong LD with a locus of interest. Given that LD in African parasites extends approximately 2 kb, this suggests that we will need more comprehensive discovery of common African SNPs to enable association studies there (Figure 2). Detectable LD in Asian parasite populations extends much further, to 16 kb [2], but overall diversity is lower, so further SNP discovery will also be required to facilitate association studies in Asia and in other populations.

We can use these calculations to estimate the minimum amount of additional genome sequencing required to enable association studies. A reasonable minimum prerequisite to ensure practicality for association studies would be a 95% probability of observing at least one common, assayable SNP within 1 kb of a random genomic locus. As Figure 2 illustrates, approximately 350,000 discovered SNPs will be required to achieve this goal. Full coverage (8X) sequencing of 3 genomes (HB3, Dd2, and PFCLIN) has yielded on average approximately 20,000 SNPs per genome. As more genomes are sequenced and common SNPs are identified in multiple genomes, however, fewer novel SNPs will be discovered, so we can expect this yield to drop somewhat. Therefore, we estimate that full coverage genome sequences of an additional 12-18 isolates may be required to raise the tally of discovered SNPs from 100,000 to 350,000.

Figure 2. This plot illustrates the relationship between the total number of malaria SNPs discovered and the likelihood that a common, assayable SNP will be found within 1 kb of a random genomic locus. For an association study to have power, multiple such SNPs would be most likely be required.



The results of our initial analysis suggest that our SNP discovery effort is far from a complete catalog of genetic diversity in global *P. falciparum* populations, and is limited on multiple fronts: 1) it has only examined a subset of populations described by the CDC and WHO as having significant burden of malaria 2) initial analysis of sequence data indicates that there is significant population structure, and so diversity in one population may not fully predict diversity in another population, and 3) initial analysis of previously generated sequence data indicates that additional sequencing of partially sequenced strains, or sequencing of additional strains from previously characterized regions, will continue to add to the polymorphisms captured at a nearly linear rate.

Global Distribution of sampled isolates

P. falciparum parasites that have been sequenced to date represent only a subset of regions with high malaria burden. These include West Africa (Ghana, Sierra Leone, Senegal), Asia (China, Papua New Guinea, Thailand) and Central America (El Salvador, Honduras) and South America (Brazil). Their origins are shown in red in Figure 3. Many key regions are completely unrepresented as of yet, including Eastern and Southern Africa, India, and the Middle East. Deep sequencing for SNP discovery relies on high quality DNA samples, preferably derived from culture-adapted lines. Culture-adapted parasites are currently available from several sites that represent some of these key undercharacterized or uncharacterized areas. These regions include Peru and Brazil in America, Thailand, Cambodia, Laos and Papua New Guinea (PNG) in Asia, and Mali, Senegal, Nigeria, Zambia, Mozambique, Malawi, Uganda, Kenya and Tanzania from Africa. These strains would greatly enhance the global coverage of *P. falciparum* strains sequenced, and increase depth of coverage in key areas.

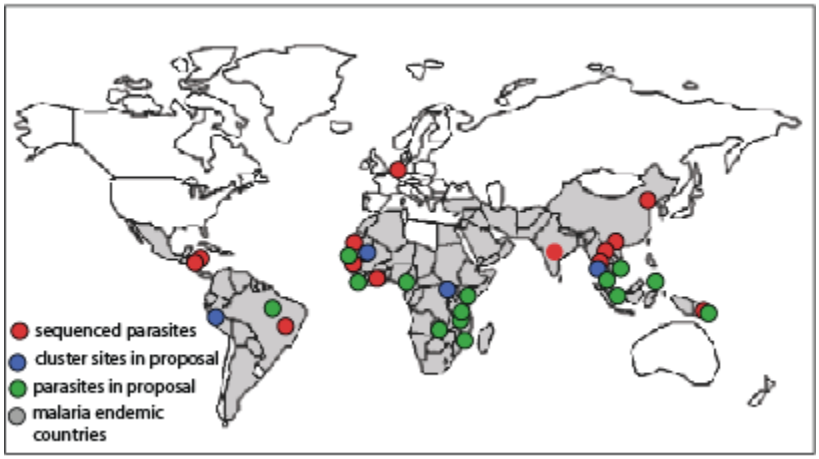


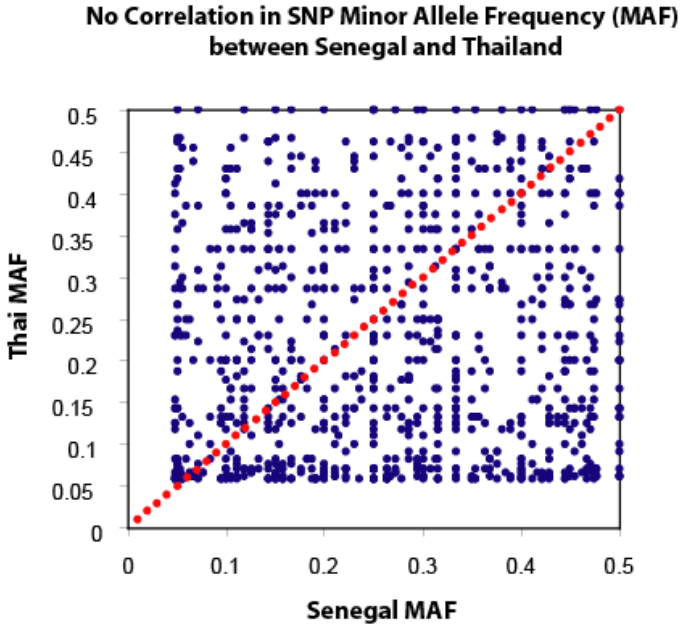
Figure 3. Global distribution of *P. falciparum* parasites that have been sequenced (red), proposed for sequencing (green), as well as countries with significant malaria burden (CDC measures) many of which are unrepresented in current sequencing efforts (gray).

Significant genetic differences exist between malaria populations

The *P. falciparum* parasite exhibits a high degree of population structure, suggesting that further SNP discovery across diverse geographic isolates will be necessary to ensure the utility of a SNP diversity map in all regions. Joy et al. [3] reported strong global population structure from approximately 100 mitochondrial genome sequences, with diversity indicating an ancient (50-100 million years ago) dispersal of the parasite to different geographic regions. Though ongoing expansion of parasite strains' ranges and gene flow between regions is important, the original dispersal of *P. falciparum* occurred sufficiently long ago to permit accumulation of significant genetic differences between and across regions. Mu et al. [4] reported a similar signal of strong population structure from nuclear SNPs on chromosome 3, and we found evidence of strong population differentiation between a sample of Senegal and Thai parasite isolates genotyped on the Broad/HSPH pilot Affymetrix chip (Fst = 0.37, $p<0.0001$).

This strong population structure implies that common SNPs discovered in one geographic region may exhibit a lower minor allele frequency (MAF), or may even be monomorphic, in other geographic regions. This could potentially compromise the power of association studies or epidemiologic analyses of strain relatedness in areas where SNP ascertainment was not performed. For the SNPs genotyped on the Broad/HSPH pilot chip, there is very little correlation between MAF in Senegal and MAF in Thailand ($r^2 = 0.006$). This suggests that a greater overall number of SNPs, ascertained from many populations, will be necessary to recover sufficient common variation to enable downstream analyses in all geographic locales.

Figure 4. No correlation is evident between allele frequencies observed in Senegal and Thailand. For each SNP, we plotted the minor allele frequency (MAF) in Senegal against that in Thailand. Very few SNPs lie close to the diagonal line, which indicates perfect correlation in SNP frequency between populations.



Prioritizing parasites to be sequenced

To generate a map of genetic diversity in *P. falciparum* with a marker density useful for association studies, we propose to sequence 24 *P. falciparum* isolates at 6-8X coverage and an additional 50 *P. falciparum* isolates at 2-3X coverage (Table 2). The 24 isolates represent both contemporary and historical parasites and the additional coverage will be focused in 4 geographic regions with ongoing studies evaluating important clinical phenotypes. A subset of these companion strains would be focused on molecular markers for emerging drug resistance to artemisinin combination therapies (ACTs) including artemisinin compounds and partner drugs.

Based upon sequence data thus far it is estimated that approximately 20 genomes at 8X coverage at the current ascertainment rate would yield a high probability of a common (MAF >5%) assayable SNP within 1 kb of a random genomic locus (Figure 2). Priorities include broadening geographic representation, including parasites from the Americas, Asia and East Africa. Two types of parasites are proposed for sequencing—a contemporary collection comprised of recent patient-derived isolates representing the extant populations in Africa, America and Asia; and, an historical collection including strains used for ongoing vaccine trials and parents of a third genetic cross.

Sequencing of the contemporary collection will provide SNP discovery toward building a marker set for association studies, and companion strain sequencing will allow for assessment of allele frequency of these discovered SNPs to identify useful markers for these studies. Sequencing of the historic collection will provide an important reference to evaluate vaccine trial failures and sequence from the parents of a third genetic cross will allow investigation of traits among the progeny of the cross. In addition, these strains provide broad geographic representation for SNP discovery and provide a valuable reference for the evaluation of important phenotypes such as drug resistance. For example, this set includes a rare chloroquine sensitive isolate from Uganda. [See Table 2 for listing of “historic” and “contemporary” parasites.]

Among the contemporary strains, all parasites will be culture-adapted to enable study of important *in vitro* phenotypes, and will be recently obtained from patients to represent the current picture of global malaria. Samples will be evaluated for their identity using molecular barcoding and genotyping techniques (see below).

For the additional companion sequencing, we propose to focus on four distinct geographic sites. We propose to survey parasites from sites where both ***good biologic questions*** are studied—emerging drug resistance, vaccine trial setting, important clinical phenotypes—and ***good sample collection*** has been obtained—material preserved for culture-adaptation, numerous samples collected across time and with good patient information. All parasites will be obtained under the international guidelines for human subjects with Institutional Review Board approval. The sites should be broadly distributed, representing America, West Africa, East Africa and Asia and are sites with ongoing drug and vaccine trials. The specific sites selected include Iquitos, Peru; Bandiagara, Mali; Kampala, Uganda; and Thailand. This additional sequencing will provide information about the frequency of mutations within each population, allowing assignment of high or low frequency classes to assist SNP prioritization for association studies. The recommendation is to survey at least 10 companion strains to 3X coverage, which will afford a ~99% chance of getting genotype data for a random SNP in the reference strain from at least 8 out of 10 companion strains. We also propose to sequence 5 parasites from both Laos and Cambodia for the purpose of identification of new emerging molecular markers for ACT failures. Southeast Asia historically provides a source of emerging drug resistance and reports in Western Cambodia from recent clinical trials suggest emergence of resistance to artemisinins. For comparison, Laos is a location where ACTs have been deployed only recently in a few limited areas. Investigation of these parasite populations would allow for early detection of important molecular markers for emerging resistance to artemisinins as well as partner drugs.

- **Iquitos, Peru** represents an American endemic site within the broader Amazon region that has been carrying out epidemiological and biochemical surveillance of drug resistance.
- **Thailand** represents an Asian site where historically drug resistance has emerged and where there is an excellent collection of parasites connected to clinical responses.
- **Western Cambodia** represents an Asian endemic site where reports suggest emergence of drug resistance to artemisinin compounds.
- **Laos** represents an Asian endemic site where limited exposure to artemisinin compounds has occurred.
- **Bandiagara, Mali** represents a West African site where a large number of samples are available in the context of vaccine trials at a well-characterized site, to offer insight into the impact of diversity on vaccine efficacy.
- **Kampala, Uganda** represents an East African site where a large number of parasites linked to well-characterized outcomes after therapy with ACTs are available.

Parasites will be selected to best sample the global distribution of the parasite, and to ensure that they represent broad genetic diversity we will genotype the parasites proposed for sequencing before library construction using molecular barcoding methods that sample for a limited number of SNPs. We will also characterize each candidate sample on SNP arrays that are designed to survey ~3,000 SNPs across the genome to ensure that each parasite genome represents an independent isolate.

Parasite Criteria:

Historic parasites would satisfy at least one of the following criteria:

- Represent a parasite used for ongoing vaccine trials
- Represent historic parasites before drug pressure was applied
- Represent geographically distinct regions to sample global diversity
- Produce gametocytes to represent natural infection

Contemporary parasites should satisfy the following selection criteria:

- Derived from recent natural infections
- Culture-adapted
- Represent geographically distinct regions to sample global diversity
- Represent all lifecycle stages (i.e. produce gametocytes)

- Derived from ongoing studies that represent vaccine or drug trial sites, or settings where severe disease outcomes are being evaluated.

Culture-adapted parasites will be derived from patients with the following patient information:

- Delayed parasite clearance or frank drug resistance in ACT trials and controls
- Vaccine trial escapees and population controls
- Severe versus mild disease (severe malaria anemia, cerebral malaria, etc)
- Pregnancy-related malaria

Culture-adapted parasites will be assayed for the following *in vitro* phenotypes

- Growth rate—look for a range of growth phenotypes
- Invasion characteristics—look for parasites that represent the range of invasion phenotypes
- Drug responsiveness—look for parasites that represent both sensitive and resistant phenotypes with regard to various drugs along a spectrum of reduced clearance times and frank resistance. Specifically, parasites "failing" the three leading ACTs, artesunate/amodiaquine (AS/AQ), artemether/lumefantrine (AL), and dihydroartemisinin/piperaquine (DP) would be evaluated for resistance to both the artemisinin component of the ACT and the artemisinin partner drugs. Continued evaluation of chloroquine, amodiaquine, mefloquine, sulfadoxine, pyrimethamine and other resistances will occur depending upon geographic origin and drug pressure.

Barcoding and Pilot Array Analysis

SNP genotyping has provided us the means to uniquely identify any *P. falciparum* parasite based on a small sample of DNA. Such an assay can be used in clinical samples to distinguishing re-infection from recrudescence in drug trials, to monitor the frequency and distribution of specific parasites in a patient population undergoing drug or vaccine-induced selective pressure. In the lab it provides a rapid, convenient and accurate method for tracking samples and determining purity of both isolates during culture adaptation and cloning as well as DNA preparations. Parallel analysis of human DNA ensures that samples for sequencing are pure parasite DNA.

Based on whole genome sequencing and Affymetrix chip-based genotyping of a worldwide collection of strains we have identified a panel of SNP markers, each with a high minor allele frequency (>30%), for which we could construct a robust TaqMan genotyping assay. Using 21 such markers no two parasites known to be of independent origin have yet been found to have the same allele signature. Thus barcoding is a useful approach to tracking strains. The TaqMan genotyping assays can be performed on DNA from cultured parasites, white blood cell-depleted frozen whole blood, frozen whole blood, or dried filter paper spotted whole blood with a success rate >99%, providing unambiguous identification of the sample at each point in the chain of custody. Less than 25 ng of parasite DNA is needed to complete a panel of 21 markers. This assay permits us to permanently identify and track *P. falciparum* parasites in the lab and to ensure that isolates are pure and unique prior to submission for whole-genome sequencing or genotyping.

Also available is an Affymetrix array containing ~3,000 SNPs derived from chromosome 7 (~2,200 SNPs) and from across the genome (~800 SNPs) that will be used to make sure that each parasite represents an independent isolate and would contribute to our understanding of genetic diversity in the *P. falciparum* malaria population.

Determining depth of coverage

We examined the cost and number of SNPs discovered at increasing genomic coverage of the Dd2 strain, and present it in Figure 5 relative to full coverage sequence. The best cost to SNPs captured ratios occur between 2 and 3x genomic coverage. While shotgun sequencing to 3x is likely optimal for SNP discovery and allele frequency analysis, there is a value in sequencing a

subset of strains to full 8x coverage. With this level of coverage, assemblies can be created and *var* genes analyzed.
It is important to note that the genome sequencing proposed in this white paper largely ignores the extraordinary genomic diversity of the variant antigen genes located at the telomeres [5-9]. The sequencing of approximately 20 genomes cannot begin to capture the genetic variation in these multigene families, and greater depth of coverage would be required to delineate members of this important gene family. To date analysis of laboratory isolates and field samples have shown that no two isolates have the same repertoire of *var* genes.

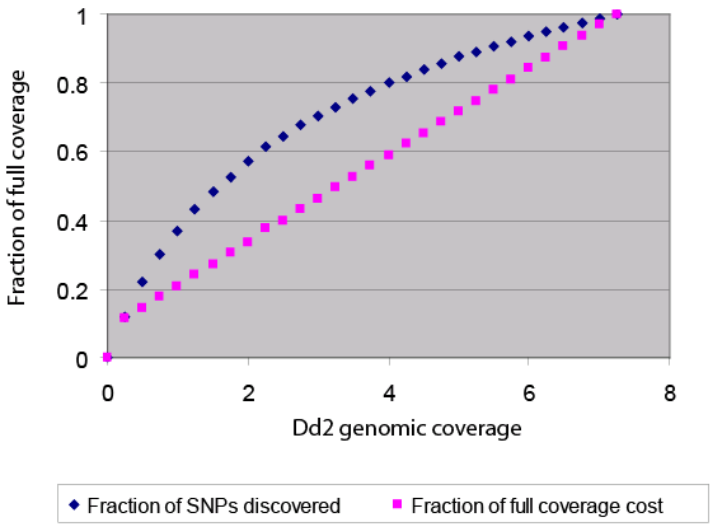


Figure 5. The fraction of SNPs discovered and cost, relative to full coverage, at each level of Dd2 genomic coverage. The best cost to SNP ratios are at between 2x and 3x.

Sequencing Priorities for *P. vivax* and Key Simian Species
Although the majority of malaria-related deaths are caused by *P. falciparum*, *P. vivax* is the most widely distributed species and the major cause of the disease outside Africa, especially in countries of Asia and the Americas. Since *P. vivax* kills less frequently and is not amenable to some experimental methods, including routine continuous *in vitro* cell culture, this malaria parasite remains a relatively neglected disease in the shadow of *P. falciparum*. Currently the ~27 Mb sequence of a single reference strain of *P. vivax* is available for the malaria research community. The lack of sequence data from additional *P. vivax* genomes undoubtedly hampers whole genome analyses and elucidation of the unique biology associated with the species.

Here we propose a strategy for sequencing the genomes of several geographically-diverse isolates of *P. vivax* and the genomes of four key simian species used as models for both human malaria from the related 'monkey malaria clade'. The overall goal is to provide the malaria research community with extensive sequence data from *P. vivax* isolates and related monkey malaria species that will stimulate further research and in part overcome some of the limitations imposed by limited access to laboratory derived parasite material. Specifically, the availability of sequence data from these isolates and species will provide valuable information concerning the rate and mode of evolution of proteins involved in biological diversity, provide the context for understanding where sequence conservation or divergence is critical, and provide insights into sequence/function relationships, with the ultimate aim of generating better methods of disease surveillance and control.

Genome sequencing of six *P. vivax* isolates (two highly drug resistant isolates, two with different relapse phenotypes, and two from untapped geographical regions) were recently approved though this white paper process (Table 2). This proposal incorporates a request for the sequencing of further isolates at low coverage for genetic diversity studies, in addition to several important monkey malaria species used as models for the study of human malaria. Specifically, we are requesting the sequencing of nine *P. vivax* isolate genomes to 3X coverage, which covers 729 million basis and six monkey genomes (four at 8X coverage and two at 3X coverage) along with 160,000 ESTs for a total of 1.15 billion bases.

I. Introduction and rationale for sequencing additional *P. vivax* isolates

Although the majority of malaria-related deaths are caused by *P. falciparum*, *P. vivax* is the most widely distributed species and the major cause of the disease outside Africa, especially in countries of Asia and the Americas [10]. Indeed, more people are at risk from infection with *P. vivax* than with *P. falciparum*. A disease of poor people living on the margins of developing economies, vivax malaria traps many societies in a relentless cycle of poverty. Protective immunity against *P. vivax* is infrequent due to intermittent transmission, and the disease occurs at all ages, though especially in young adult men. Morbidity results from repeated acute febrile episodes of a debilitating intensity that can persist for months. Drug resistance in *P. vivax* is becoming more widespread, hindering management of clinical cases. Since *P. vivax* kills much less frequently and is not amenable as yet to continuous *in vitro* culture, the species remains comparatively neglected in the shadow of *P. falciparum*.

A. Unique aspects of P. vivax biology

Phylogenetically and biologically, *P. vivax* and *P. falciparum* are very distant and very different from each other. *P. vivax* parasites predominantly infect reticulocytes, which results in long-term chronic infections and anemia, but less mortality. In contrast, *P. falciparum* parasites invade cells of various ages and cause acute anemia and frequently death. Moreover, *P. vivax* cannot infect Duffy negative reticulocytes, and so is exceedingly rare in West Africa where Duffy negativity predominates in the human population. Depending on the geographic isolate, *P. vivax* sporozoites differ in their potential for development to maturity in the liver. Some may initiate a primary blood stage infection, while others may remain in the liver as a resting stage (the hypnozoite) until conditions are more favorable for mosquito infection. Thus, in *P. vivax*, there are relapses of malaria due to reactivation of the hypnozoite stage in the liver, a specific adaptation to diminished mosquito populations during the winter months that increases the likelihood of transmission when environmental conditions are more favorable. This contrasts with recrudescences, the patent reappearance of antigenically distinct populations of parasites in the blood that are produced during the course of an existing infection, and which occur with both *P. falciparum* and *P. vivax*. *P. vivax* populations in different geographical regions and environmental conditions have different time courses of relapse, ranging from very delayed in the northern latitudes (*e.g*, Korea), to very frequent in the tropics of the Pacific and Indonesia. The biological phenomenon of relapse presents serious challenges to, and is responsible for, vivax malaria being significantly resistant to elimination and control efforts.

B. Drug resistance

Current treatment for *vivax* malaria primarily relies upon two antimalarial drugs, chloroquine and primaquine, with the latter reducing the number of subsequent relapses through killing of the hypnozoite stage. Reports of chloroquine resistant *P. vivax* strains being found frequently in have several regions of Asia and also South America have been published [11], and likewise, recent reports of primaquine resistance are worrisome [12], although the epidemiology and molecular mechanisms of resistance remain to be determined. Relapse of *vivax* malaria often confounds drug efficacy tests, since the relapse interval can coincide with the time when recrudescence of a drug resistant parasite would occur following drug treatment. The emergence of drug resistance in *P. vivax*—particularly to the only class of compounds available for killing the dormant liver stage—is alarming and of high priority for research.

C. Geographical variation

Despite the essential role that assessing the genetic variation of malaria parasites plays in developing, testing and deploying control interventions, investigation of genetic variation in human *Plasmodium* species to date has been limited and biased toward *P. falciparum.* There have been a few population-based studies of *P. vivax* (*e.g.,* [13-15]), mostly of genes encoding parasite surface proteins or neutral markers such as microsatellites and mitochondrial loci. Significantly, even these limited studies indicate high levels of genetic polymorphism, large numbers of gene duplication events in the species' evolutionary past, and rapid evolution of particularly repetitive tandem repeats in *P. vivax* proteins. They also indicate a distinct geographic structure with at least two major *P. vivax* "clusters", one in Southeast Asia and another in Melanesia [16]. Thus *P. vivax* is an ancient and diverse species for which the range of diversity has yet to be defined. To understand this diversity and its consequences will require extensive sampling of a wide variety of geographical isolates.

D. Refractoriness of P. vivax to experimental manipulation

There are several reasons why genome sequencing presents one of the easiest and most cost effective methods of obtaining genetic data about *P. vivax*. The parasite cannot be grown in long term *in vitro* culture, although short term culture has been reported. While several groups within the research community have ongoing projects to develop a culture system -- for example, through using stem cell technology to establish and maintain human erythrocytic cell lines for the generation of reticulocytes which can support *P. vivax* growth -- genetic crosses, which have identified candidate loci for several phenotypes in *P. falciparum*, currently are not feasible in *P. vivax* since the cloning and phenotyping of progeny clones would have to be carried out *in vivo*. Long-oligo arrays have been developed through the NIAID-funded Pathogen Genomics Resource Center for the community, but it is unclear at this stage whether gene expression studies or CGH studies will be possible using these arrays due to inherent problems of obtaining enough parasite DNA and RNA that is free from contaminating host material; *P. vivax* infects reticulocytes that contain host DNA and RNA, and parasitemias rarely exceed ~1% due to restriction to this class of erythrocyte. However, association studies that identify links between genetic markers and a particular phenotype in populations of parasites are one of the few methods available to connect phenotype to genotype in *P. vivax*. Using microsatellites developed from the genome project, the first association mapping studies are underway in *P. vivax*. However, such studies would benefit tremendously from the development of a haplotype map, which requires sequencing of more *P. vivax* isolates. Once candidate loci have been identified, recent advances in transient *P. vivax* transfection [17] and heterologous transfection [18] ensure that the next steps to determine candidate gene function and molecular mechanisms are possible.

E. Secrets revealed by the first P. vivax genome sequence

The *P. vivax* genome sequencing project began at The Institute for Genomic Research (TIGR) in 2002 with the goal of producing a finished sequence at least as good as the sequence of *P. falciparum*. Surplus funds from the US Department of Defense and NIAID, which supported part of the *P. falciparum* genome sequencing project, were used to finance the project [19]. After a nine month halt in 2004 due to diminution of funds, the project was rescued by additional funding from the Burroughs Wellcome Fund and NIAID. A paper describing the genome is under pre-submission consideration at *Nature* (Carlton *et al*. 2007) in conjunction with the paper describing the genome of the monkey malaria parasite and close relative, *P. knowlesi*, undertaken at the WTSI.

Throughout the project, *P. vivax* genome sequence data have been made available to the community and have been used in multiple ways to understand the basic biology of this pathogen. Some of the major findings are:

- Screening of ~330 microsatellites in the *P. vivax* genome has identified ~150 that are polymorphic among eight strains. Many researchers have used these for world-wide population studies of *P. vivax* that have confirmed the tremendous genetic variation exhibited by the parasite (see for example [14, 20, 21]). Another study has used the microsatellites to determine that *P. vivax* populations emerging from hypnozoites routinely differ from the populations that caused the acute episode, and that activation of heterologous hypnozoite populations is the most common cause of first relapse in patients with vivax malaria [21]. The first association studies using these microsatellites are now underway.
- *P. vivax*-specific genes involved in biologically interesting phenotypes such as red blood cell invasion, antigenic variation, and host-cell interactions have been identified (*e.g.* [22]).
- Crystal structures and homology models of proteins implicated in drug interactions and drug resistance have been compared between *P. vivax* and *P. falciparum*, in order to predict the *P. vivax* parasite's active sites involved in drug interaction (Carlton *et al*. 2007).
- A network of *P. vivax* protein interactions from yeast-two hybrid experiments, Rosetta stone analysis and phylogenetic profiling, has been generated and compared to the interactome of *P. falciparum* (Carlton *et al*. 2007).

In addition, comparative analysis with completed *Plasmodium* genomes *P. falciparum*, *P. yoelii* (and other rodent malaria species) and *P. knowlesi*, presented in Carlton *et al*. 2007, provided:

- the fundamental characteristics of mamalian *Plasmodium* genomes (genome size range, number of genes, gene length, GC content, codon bias etc).
- the major *Plasmodium* gene families shared across species, which could be targeted to generate a genus-specific vaccine.
- four-way synteny maps, used to infer the chromosomal rearrangements that have occurred since divergence of the species from a common ancestor and the evolution of the *Plasmodium* genus (Figure 6).
- the number of synonymous substitutions per synonymous site (dS) and non-synonymous substitutions per non-synonymous site (dN) between orthologous pairs of *P. vivax* and *P. knowlesi* genes, and used these values to visualize how the mutation rate varies over the length of chromosomes, and identify genes under possible positive selection pressure that may represent infected erythrocyte surface proteins that could be targeted for vaccine development.

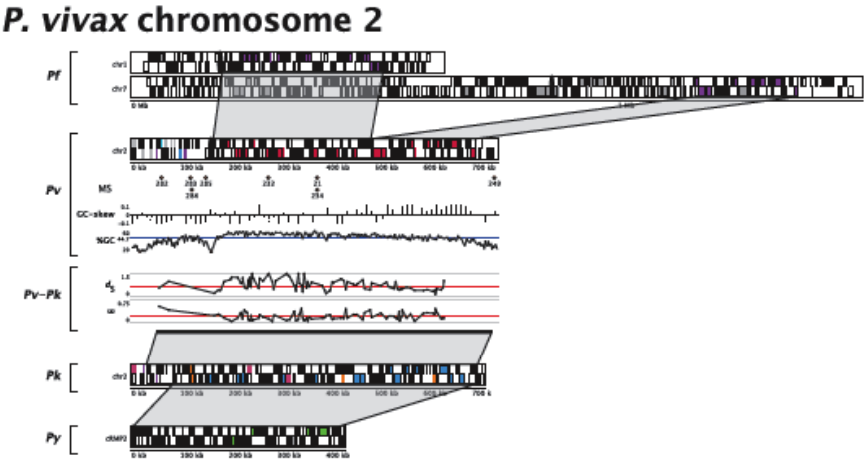


Figure 6. Map of *P. vivax* chromosome 2 showing syntenic regions and orthologous genes between four species of malaria parasite (cRMP refers to a composite rodent malaria chromosome generated from partial sequence data of three rodent malaria species). The plots between *P. vivax* and *P. knowlesi* show: (i) the position of polymorphic microsatellite markers (MS) in the genome; (ii) GC skew along the *P. vivax* chromosome; and (iii) plots of molecular evolution statistics such as dS (a rough approximation to the mutation rate) and dN/dS (areas of the genome under possible positive selection) determined from analysis of orthologs between *P. vivax* and *P. knowlesi.* Taken from the *P. vivax* genome paper (Carlton *et al*. 2007).

A wealth of information has been generated from the first genome sequence of *P. vivax*. However, further analyses, in particular studying the natural variation of *P. vivax* genes such as those involved in invasion have been limited by the lack of sequence data from additional isolates. We hope to ameliorate this situation via the plans outlined here.

F. Additional P. vivax isolates: the generation of a diversity map and linkage mapping

To extend our current knowledge of genetic diversity and population structure of *P. vivax*, we propose sequencing seven additional isolates to 3X coverage which, in combination with the six isolates from Melanesia, Brazil, North Korea, India and West Africa recently funded and the reference genome sequence of the Salvador I isolate from El Salvador, will drive further genome diversity studies (Table 2). With the genome sequences of a total of 16 *P. vivax* isolates, the vivax research community will be able to (i) identify SNPs, indels and gross chromosomal rearrangements and construct a map of genetic diversity ('haplotype map') of the species; (ii) calculate allele frequencies and identify patterns of linkage

disequilibrium; (iii) estimate recombination rates and patterns; (iv) identify correlates of protein evolution; and (v) identify potential vaccine candidates through identification of genes under selective pressure.

We propose sequencing one isolate from Sri Lanka (Sri Lanka) to compare with the isolate from India (India VII, recently funded); isolates from the west coast of South America (Ecuador) and from Nicaragua (NICA) to compare with the reference genome isolate Salvador I from El Salvador; an isolate from Vietnam (Vietnam Palo Alto, highly virulent) and an isolate from Thailand (Thai III) to compare together; an isolate from Melanesia (Chesson, also used as a standard strain for drug studies in humans) to compare with the isolates from Melanesia, Indonesia XIX and AMRU I (recently funded); and a second isolate from Brazil (Brazil VII) to compare with Brazil I (recently funded). India, S.E. Asia and Brazil represent three regions of the world where vivax malaria predominates, and obtaining sequence data from more than one isolate in these key areas is of high importance. Moreover, obtaining low-coverage sequence data from strains located in close geographical proximity to the six approved isolates will provide data as to whether any of the six isolates appear to be atypical.

How will obtaining sequence data from *P. vivax* isolates advance studies on population structure and genetic diversity? Briefly: (i) we will be able to make preliminary statements concerning the gene repertoires (for example the range of antigen genes such as *vir*, *msp*, *rbp, trag*) in parasites from seven different countries; (ii) we will obtain a first glimpse of what a typical *P. vivax* parasite from India, West Africa and Brazil looks like in comparison to parasites from other countries, allowing us to answer such questions as: Do African parasites have additional copies of invasion genes, enabling it to use alternative pathways of invasion that do not use the Duffy antigen receptor, as has been postulated recently in a study of Duffy negative patients from western Kenya infected with *P. vivax* [23]; (iii) we will be able to determine the rate and mode of evolution of *P. vivax*-specific proteins, allowing identification of groups of genes that may be evolving fast within *P. vivax* compared to other species; and (iv) we will be able to compare genome architectures, including the isochore structure in *P. vivax*; is the isochore structure the same or does it vary geographically? What implications does this have for the evolution of genes found within these regions?

The sequencing of multiple *P. vivax* isolates will also provides the SNPs necessary for location of drug resistance loci using linkage analysis methods. Traditionally, linkage mapping involves analysis of segregation of drug resistance phenotypes with genetic markers in the progeny of genetic crosses. Linkage analysis has been extremely successful for locating drug resistance loci in *P. falciparum*. However, this method has been difficult to apply to non-culturable *Plasmodium* species such as *P. vivax*, because the progeny of genetic crosses cannot be cloned. Recent methods developed for use with rodent malaria parasites now allow linkage mapping of loci underlying drug resistance without cloning of progeny [24]. This is done by selecting uncloned progeny using the drug of interest, and analyzing SNP representation in uncloned progeny using quantitative methods. These methods could be adapted to *P. vivax* and other malaria species that grow in primate models in order to examine genes underlying drug resistance, and they could also be used to map other important phenotypes such as host specificity and strain specific immunity [25]. In view of the difficulty in accurately measuring drug resistance and other phenotypes in *P. vivax*, these methods are expected to be particularly valuable alternative approach to association mapping for this parasite, and provide strong incentive for the sequencing effort proposed here.

Finally, we have also included a place-holder (Table 2) for two additional isolates of *P. vivax* to be acquired from documented parasite failures in ongoing clinical trials of 8-aminoquinoline drugs. The only means available to ensure radical cure of relapsing *P. vivax* malaria is treatment with primaquine, an 8-aminoquinoline, that eliminates the dormant hypnozoite forms in liver. Over the past decade there have been numerous clinical observations of primaquine failures, some with high levels of drug be administered, which in some cases have been authenticated by genotype analysis. At present, we have no understanding of the mode of action for primaquine, nor do we know the mechanism by which resistance to it may arise. Genomic sequencing of an authenticated primaquine resistant *P. vivax* patient isolate therefore, would be highly valuable in providing insights into this important antimalarial drug. Furthermore, the next generation of hypnozoite-eliminating drugs under development also belong to the

8-aminoquinoline class, and thus information gained from studying primaquine-resistant *P. vivax* will be widely applicable.

2. Non-human primate *Plasmodium* sequencing priorities

P. vivax is intimately related to a large clade of malaria parasite species that infect the cercopithicoid (Old World) monkeys and the lesser apes (gibbons) of South Asia, from Taiwan to Sri Lanka and India. The Southeast Asian non-human primate malaria parasites form a monophyletic group containing *P. fragile, P. knowlesi*, *P. coatneyi, P. inui, P. hylobati, P. simiovale*, *P. fieldi,* and *P. cynomolgi*, together with *P. vivax*. These species in turn form a divergent sister group with *P. gonderi* (and other species) from African monkeys (Figure 1). An important feature of this group of parasites is that several biological traits, such as periodicity, morphology and virulence, have little value for inferring their phylogenetic relationships [26, 27]. For instance, the length of periodicity is a convergent characteristic: the 'quartan' (72-hour erythrocytic cycle) parasites *P. inui* and *P. malariae* do not form a monophyletic group, and *P. knowlesi*, the only 'quotidian' (24-hour erythrocytic cycle) primate parasite shares this characteristic with several species parasitic to rodents. Not all species of the monkey malaria clade are capable of relapse; *P. vivax, P. cynomolgi, P. simiovale* and *P. fieldi* are among those that exhibit this trait. Based on the phylogeny depicted in Figure 1, it appears that the capacity to relapse appeared *de novo* in *P. vivax* and the clade including *P. cynomolgi*. Three of the non-human primate malaria species, *P. cynomolgi*, *P. coatneyi* and *P. fragile* are excellent and faithful models for the biology, immunology and pathology of human species with the same phenotypic characteristics, *i.e., P. vivax* for *P. cynomolgi* and *P. falciparum* for *P. coatneyi* and *P. fragile*. A fourth, *P. inui*, is the only other known quartan parasite aside from its human equivalent *P. malariae* (the latter is currently being sequenced at the WTSI).

Of the seven malaria parasite species currently known to naturally infect Asian macaque monkeys, three (*P. cynomolgi*, *P. inui*, and *P. knowlesi*) are capable of infecting humans [28]. The ability of macaques to thrive in human-altered environments has made them the most common non-human primate species in Asia [29]. In addition, several Asian cultures honor centuries-old traditions of human - non-human - primate commensalism (interactions associated with habitually sharing a space); such close interaction may facilitate the emergence of zoonotic malaria parasites. In fact, there have been increasing reports of naturally-acquired human infections with *P. knowlesi* [30, 31]. In line with this apparent frequent host-switching undertaken by members of the monkey malaria clade, *P. vivax* is thought to have arisen in humans through several host switches from Asian macaque monkeys [26, 32]**.**

A. *Plasmodium cynomolgi*: a robust monkey model for *P. vivax*

In the absence of continuous culture, and until it has been developed, there is an urgent need to have a good model system for *P. vivax*. *Plasmodium cynomolgi* possesses most if not all of the phenotypic, biologic and genetic characteristics of *P. vivax.* In fact, it is the closest relative of *P. vivax* and resides in the same subclade in the ancestral tree of the simian malaria parasites (Figure 1). It infects both Old World and New World monkeys as well as humans. There is a great deal of experimental information about this species, and it has the same dormant liver stage (the hypnozoite) as *P. vivax*. This parasite is found to infect a variety of macaque species stretches from the Philippines to Sri Lanka, and it infects a variety of anopheline mosquitoes. *P. cynomolgi* has been used extensively as a model for *P. vivax* in drug efficacy and biological studies.

Obtaining the genome sequence of *P. cynomolgi* would enhance the experimental capability for work in this model many fold. With *P. cynomolgi*, infections of both New World and Old World monkeys (which are more closely related to humans) can be used to study the hypnozoite stage, to transfect *P. vivax* vaccine candidates for challenge experiments, for homologous gene knock-out studies, and for mosquito transmission studies. In addition, *P. vivax* parasite material such as DNA, RNA and protein can be generated either from monkey infections or long term *in vitro* cultures of transgenic *P. cynomolgi* and used in molecular and biochemical assays. The genome sequence of *P. cynomolgi* will be required if one is to determine the precise genetic relatedness of the two species, and to compare the gene complement. For example, construction of a transgenic line of *P. cynomolgi* carrying a copy of a *P. vivax* gene would be senseless unless the copy number of the *P. cynomolgi* homolog(s) was known. In addition, between-species comparisons of *P. cynomolgi* and *P. vivax* genomes will shed light on (i) genome architecture, such as the degree of synteny and number of orthologous genes, and whether the isochore structure is

shared in location and contains orthologous genes in the two species; (ii) gene families that are shared or have undergone expansion in one lineage; and (iii) identify genes which are under positive selection pressure, which can be compared to the genes identified as under positive selection pressure between *P. vivax* and *P. knowlesi* (Figure 1).

A powerful approach for investigating the rate and mode of evolution of particular regions of the genome is the comparison of genetic polymorphism (within species) versus the divergence of closely related species. These comparative approaches allow the identification of putative adaptive polymorphisms (maintained by positive selection), regions of the genome evolving under evolutionary constrains (negative selection), and a better understanding of the evolutionary processes taking place in regions with low complexity. *P. cynomolgi* exhibits high genetic polymorphism (relative to *P. vivax*) and is found in multiple non-human primates. Although a comparison of multiple strains of *P. vivax* versus only one strain of *P. cynomolgi* is a first valuable approximation, we will overestimate the divergence between *P. vivax* and its closely related species by ignoring the genetic polymorphism within *P. cynomolgi*. The inclusion of two strains from different geographic origins per non-human malaria species increases the probability of detecting divergent alleles within each of these non-human malarial parasites. This information provides an estimate of their genetic polymorphism improving our capacity of identifying fix divergences between these species and *P. vivax*. A better estimate of the fixed divergences increases the power of comparative approaches and other evolutionary genetic analyses that aim to understand the rate and mode of evolution of the *P. vivax* genome.

The two *P. cynomologi* isolates proposed for sequencing are: (1) the Berok strain (to 8X coverage), because it can also be cultured *in vitro* [33]. It is the only other monkey malaria species known to harbor an isochore genome structure; and (2) Ceylonensis (to 3X coverage), which was isolated from a toque monkey, *Macaca sinica* in Sri Lanka, and has been studied extensively in rhesus monkeys.

B. *Plasmodium inui*: a monkey model for *P. malariae*

Plasmodium inui infects Old World monkeys and has been experimentally transmitted to humans and *Aotus* monkeys. *P. inui* infections, like those of *P. malariae* in humans, are characterized by being extremely long-term, often lasting for the lifetime of the host, despite lacking a relapse mechanism. Infections frequently last up to 13 or 14 years [34], whereas human malaria parasites (with the exception of *P. malariae*) last one to three years on average. The asexual stages of both *P. malariae* and *P. inui* require approximately 72 hours for development. In addition, chronic infections of *P. inui* cause renal disease in rhesus monkeys, much as *P. malariae* is known to cause chronic nephrosis in human and nonhuman primates [35]. The parasite is readily transmitted by a variety of anopheline mosquitoes from Taiwan throughout Southeast Asia and on to India. It is quite possible that humans in Southeast Asia could be naturally infected with *P. inui,* but be misdiagnosed as having *P. malariae* because of their morphological similarity. Many different isolates of *P. inui* have been adapted to *M. mulatta* monkeys with over a dozen different strains available. In addition, the OS strain has been experimentally transmitted to humans and to *Aotus* monkeys, and adapted to *in vitro* culture [36]. Sequencing and subsequent investigation of the *P. inui* genome have the potential to explain how a malaria parasite such as *P. malariae* establishes a lifelong relationship with the host, whereas *P. falciparum* and *P. vivax* infections terminate within several years.

Similar to *P. cynomolgi*, *P. inui* exhibits high genetic polymorphism and is found in multiple non-human primates. Thus the same concerns exist should only one isolate of *P. inui* be sequenced *i.e* we will overestimate the divergence between *P. vivax* and its closely related species by ignoring the genetic polymorphism within *P. inui*. The two isolates proposed for sequencing are: (1) the OS strain (to 8X coverage) isolated from a monkey in India; and (2) Taiwan I (to 3X coverage), isolated from a *Macaca cyclopis* monkey from Taiwan and has been passaged to rhesus monkeys; however, no attempts have been made to infect New World monkeys or humans. There are many isolates of *P. inui* from which to choose for sequencing; the selection of these two from different hosts and different geographic locations is a logical basis for selection.

C. *Plasmodium coatneyi*: a monkey model for cerebral malaria

Plasmodium coatneyi grows in Old World monkeys and is characterized by marked sequestration of the mature parasite forms in the deep vascular tissues of various organs and tissues—sites which are very similar to the location of sequestration for *P. falciparum. P. coatneyi* sequestration affects the brain in particular, and symptoms of cerebral malaria occur in *Macaca mulatta* and *M. fuscata* monkeys. *P. coatneyi* characteristically produces severe acute and long-term chronic infections, with waves of parasitemia that represent the appearance of new variant gene antigens on the surface of the infected erythrocytes (antigenic variation), highly similar to *P. falciparum* infections. The parasite has variant antigen genes similar to the *SICAvars* found in *P. knowlesi*, which are also functional paralogs of the *var* gene family of *P. falciparum. P. coatneyi* is also morphologically similar to *P. falciparum* (excluding the gametocyte stages). *P. coatneyi* is genetically close to and falls in the same subclade as *P. knowlesi*, a parasite with quite different phenotypic and biologic characteristics, such as variable sequestration and a 24 hour asexual growth cycle (*P. knowlesi* has been sequenced at the WTSI).

Thus, *P. coatneyi* represents a robust and faithful model species for *P. falciparum* in terms of sequestration, cerebral malaria, antigenic variation, severe pathology (such as acute nephritis, major organ dysfunction and anemia), immunobiology and determination of the genetic characteristics responsible for wide phenotypic differences amongst closely-related sibling species. Arguments for obtaining the genome sequence follow those presented above for *P. cynomolgi*, but in addition identifying the complete complement of genes in the species will enable insight into the evolution of the phenotypes specific to *P. falciparum* and *P. coatneyi* such as sequestration and severe pathology, *i.e* whether they are the result of convergent evolution or loss of the phenotypes in other *Plasmodium* species.

D. *Plasmodium fragile*: a monkey model for sequestration and severe disease

Plasmodium fragile grows in both Old World and New World monkeys and is characterized by marked sequestration of the mature parasite forms in the deep vascular tissues of various organs and tissues—sites that are very similar to the location of sequestration for *P. falciparum*. *P. fragile* sequestration affects the heart in particular. *P. fragile* characteristically produces severe acute and long-term chronic infections, with waves of parasitemia that represent the appearance of new variant gene antigens on the surface of the infected erythrocytes (antigenic variation), similar to *P. falciparum* and *P. coatneyi*. The parasite can also be grown continuously in culture. *P. fragile* is genetically close to the Asian simian malaria parasites but is not in the subclade of *P. knowlesi* and *P. coatneyi* or the other two subclades of these simian parasites and thus perhaps represents a closer evolutionary link with African simian malaria parasites. The parasite is transmitted experimentally by *Anopheles dirus* and has been found in Sri Lanka and southern India. Other common laboratory-raised anopheline mosquitoes fail to support sporozoite infections. Attempts to infect humans were unsuccessful. The Sri Lankan isolate produces infective gametocytes and is the one recommended for sequencing over the Nilgiri isolate which does not. *P. fragile*, like *P. coatneyi* represents a faithful model for *P. falciparum* in terms of sequestration, antigenic variation, severe pathology, immunobiology.

To summarize, having the genome sequence of these four monkey malaria models at hand will provide the means and incentives for scientists to participate in genetic and biological research on human species of malaria parasite. Monkeys and non-human primates are more appropriate animal models than rodents because of their close phylogenetic relationship to humans. Establishment of such animal models is an essential prerequisite of the research on human diseases such as malaria. Such disease animal models can be used effectively as common research resources. Over 100,000 primates are used in US-based biomedical research each year, because of their utility in medical science.

3. Sequencing logistics

A. Parasite material

Unlike *P. falciparum*, *P. vivax* and *P. coatneyi* have not been maintained in continuous culture, and all parasite material including DNA for sequencing is most readily obtained from infection of non-human primates (*P. inui* and *P. cynomolgi* can be grown *in vitro*.) All of the *P. vivax* isolates described above have been adapted to growth in New World monkeys, and *P. coatneyi* is maintained as frozen parasite stocks from rhesus monkeys. These frozen parasite stocks, plus detailed genealogies of both primate

passage and mosquito transmission, are available as a unique resource at the Division of Parasitic Diseases, Centers for Disease Control. Most stocks have also been provided to the malaria repository, *MR4*, under a contract with the CDC Foundation. Similarly, the CDC Foundation would be able to provide parasite material for the genome sequencing proposed here via a contract funding mechanism.

Of paramount importance with many parasite genome sequencing projects is the need to limit the amount of contaminating host material in preparations of parasite material for genomic and cDNA library construction. During the *P. vivax* Salvador I sequencing project [19], protocols were perfected which minimized the amount of contaminating monkey DNA, including methods utilizing acid-washed glass beads to remove activated platelets, and cellulose fiber columns, filters and affinity columns to remove monkey white blood cells. These procedures successfully removed greater than 99.7% of host DNA from the Salvador I parasite material, and can be used during preparation of parasite material from the isolates discussed here. In addition, genotyping of monkey-adapted parasite lines has shown that the lines of parasites generated during the adaptation process are genetically homogenous, *i.e.,* they consist of a single clone. The completion of the Salvador I genome sequence has also confirmed these results. Thus, heterozygosity or mixed infections of the isolates discussed here is unlikely to be a problem.

B. Sequencing strategies

The *P. vivax* genome is ~27 Mb, haploid and distributed among 14 chromosomes (all *Plasmodium* species studied so far share genomic characteristics similar to these). The haploidy of *Plasmodium* genomes makes them uniquely amenable to large-scale sequencing. Unlike the *P. falciparum* and rodent malaria parasite genomes, which are extremely biased in their genome content, members of the monkey malaria parasite clade have a more balanced GC content (~35-45%) (Table 3). *P. vivax* and *P. cynomolgi* are unique in containing an 'isochore' chromosome structure, *i.e.,* GC-rich regions interspersed with GC-poor regions; in *P. vivax* the GC-poor regions predominate in subtelomeric regions of the chromosomes, which made it more difficult to assemble the ends of *P. vivax* chromosomes during the sequencing project. Due to their phylogenetic relatedness, it is likely that all the species belonging to the monkey malaria clade have similar genome size and GC content, although only *P. vivax* and *P. cynomolgi* have an isochore structure. For the purpose of this proposal we have assumed that the isolates proposed here have an average genome size of ~27 Mb.

i. Genomic libraries

Since the GC content of the *P. coatneyi* and *P. inui* species is sufficiently high that vector/insert instability issues are minimal, we propose that large insert BAC and/or fosmid libraries be prepared for these species. End sequencing of large clones from these libraries will provide scaffolding information for the genome assemblies and, since the availability of DNA from monkey malaria species is limited, a minimal tiling path of clones across each genome in addition to the BAC/fosmid library itself can be provided as a resource for the malaria community (a similar tiling path of 10-12 kb plasmid clones for *P. vivax* Salvador I is available through *MR4*, reagent #MRA 840). The isochore structure of *P. cynomolgi* chromosomes reduces the usefulness of large insert vectors for cloning, since low GC regions in the genomes will be less stable resulting in possible biased libraries. However, low GC isochore regions constitute less than 18% of the *P. vivax* genome, and if this holds true for *P. cynomolgi* then the majority of the species' genome is likely to be stably cloned in large insert vectors and could still be made available to the community as a source for DNA.

Table 3. Genome characteristics of several *Plasmodium* species. *Approximate GC content derived from a limited number of protein coding genes. ** Tandem repeats and low complexity sequences. NK, not known.

Clade	Species	Genome size (Mb)	No. genes	% GC	Repeat content**	Structure
Monkey	*P. vivax*	26.8	~5,400	42.3	Medium/high	Isochore
Monkey	*P. cynomolgi*	NK	NK	34.5*	NK	Isochore
Monkey	*P. coatneyi*	NK	NK	34.5*	NK	Uniform
Monkey	*P. inui*	NK	NK	40.2*	NK	Uniform
Monkey	*P. fragile*	NK	NK	42.0*	NK	Uniform
Monkey	*P. knowlesi*	23.5	~5,200	37.5	Medium	Uniform
Human	*P. falciparum*	23.3	~5,400	19.4	High	Uniform
Rodent	*P. yoelii*	23.1	~5,800	22.6	High	Uniform

ii. Sequence coverage

We propose sequencing the four non-human primate species and isolates to 8x (full) coverage, and sequencing the additional *P. vivax* isolates to 3x (sample) coverage. Previous low-coverage genome projects (*e.g.* the rodent malaria parasites [37, 38]) have highlighted the inherent problems of analysis of partial sequence data, and thus it is important that at the very least the four simian model parasite genomes be sequenced to full coverage since these will be the reference sequences for each of these species. (Sample sequencing to 3X of the *P. cynomolgi* Ceylonensis strain and *P. inui* Taiwan I strain is sufficient). In addition, we recommend sample sequencing to 3X coverage for the nine *P. vivax* isolates since high-coverage is not required for the generation of a diversity map. The six previously funded *P. vivax* isolates have been approved for whole genome sequencing to 8X coverage and will thus provide the community with a set of gold-standard genome sequences against which low-coverage genome sequence can be compared.

Rodent Malaria Parasite Sequencing Priorities

The dominant status of Rodent Malaria Parasites (RMP) as *in vivo* experimental models for detailed investigations of *Plasmodium* biology and interactions of the parasite with both host and vector is unquestioned. RMP have served as extremely useful surrogates for the investigations of biology of human malaria parasites and parasite interactions with both host and vector since their introduction into the laboratory beginning in the 1950's. The power of such investigations like for all species of *Plasmodium* has only been empowered by the availability of significant genome sequence and the realization that ~80% of the genes in all *Plasmodium* genomes are orthologous. While no one RMP is the perfect model system, between them the three major species used in the laboratory offer the virtually complete palette of modern investigational approaches, forward and reverse genetics, unparalleled access to the complete life cycle and facile *in vitro* culture of many life cycle stages. These features when combined with the exploitation of (post)-genomics technologies have generated significant datasets and advances in our knowledge of *Plasmodium*:

- Discovery of and investigation of numerous vaccine candidates with homologues in human malaria
- Discovery and validation of attenuated sporozoite vaccines (either irradiated or genetically manipulated)
- Elucidation of many novel aspects of pre-erythrocytic biology
- Initial definition of innate immune responses of mosquito vectors to *Plasmodium*
- Defining correlates of immune protective responses that control virulence and clear infections
- Investigations of the immunobiology of cerebral malaria
- Drug discovery: assessing initial in vivo efficacy, pharmacokinetics, ADME, and toxicology for all anti-malarial drug development
- Molecular developmental biology of the parasite in the mosquito vector

Nevertheless, the existing incomplete RMP genome sequence resources are a significant limiting factor for experimental research. For this reason the highest priority is to complete the existing genome projects that were initiated to establish the genome sequences of *P. yoelii 17XL, P. chabaudi* AS and *P. berghei* ANKA. For serious genetic work it is very frustrating, and can verge on the non-productive, to have lots of unjoined contigs, as has been the case for *P. chabaudi* until very recently and still so for the other two projects. Therefore, it is very important that at least two lines from each species are characterized to completion without significant gaps. Whatever additional genomes are chosen they should be sequenced to a level that permits development of tools for comparative genetic analyses among the lines that have the distinct phenotypes. We propose to finish the sequencing and assembly of P. yoelii 17XL, P. chabaudi AS and P. berghei ANKA. In addition, we proposed to sequence to 8X coverage two additional P. yoelli genomes, two additional P. chabaudi genomes and two P. berghei genomes. All parasites will be carefully genotyped to ascertain independent organisms before sequencing will be initiated.

The case for additional RMP genomes is as follows:

P. chabaudi

P. chabaudi chabaudi strains AJ and CB are highly prized models for studying immune responses and virulence in malaria. Therefore, these strains should be sequenced from the perspective of strain-specific immunity, growth rate and drug resistance studies where they would provide the greatest exploitable sequence information that can be linked to parasite phenotypes. Two additional strains—the *P. chabaudi adami* lines DS and DK, which have striking growth rate differences between them—will be held in reserve should the uniqueness of the AJ and CB strains not be validated. These two strains would be of much biologic interest as indeed, linkage group selection (LGS) analysis in Richard Carter's lab suggests the different growth patterns between themselves and all the *P. chabaudi chabaudi* lines (the former, *P. chabaudi adami*, first peaking on day 8 of an infection, regardless of intrinsic growth rate, and the *P. chabaudi chabaudi* all first peaking on day 6 regardless of growth rate) have a genetic basis. Since differences in growth rates are often correlated with virulence in *P. falciparum*, this RMP system is a useful model to study this virulence factor.

P. yoelii

P. yoelii is a valuable model for studying invasion phenotypes of blood-stage parasites and is a leading model for pre-erythrocytic vaccine development. Homologous to alternative invasion pathways of *P. falciparum*, changes in the type of erythrocytes invaded are linked in *P. yoelii* to differential expression of parasite ligands (i.e., phenotypic variation) thereby providing an excellent experimental model for understanding the genetic basis of this switches in vivo. There are two major genotypes of *P. yoelii* and only genotype 1, represented by *P. yoelii* 17XL/17XNL(clone1.1), has been partially sequenced and so is the only represented in the genome database. Moreover, all other 17X lines for which information is available are effectively congenic with 17XNL in the genome database. The "second " *P. yoelii yoelii* 17X genotype represented by the 17X (WHO) line, which is genotypically the same as the 17XA line (used in Edinburgh), is genetically distinct (within the *P. yoelii* species) from the genotype one of the 17XNL. The differences evident between genotypes 1 and 2 are reflective of the sum of differences in all human isolates of *P. falciparum.* All the "virulent" lines of 17X are congenic with 17X clone1.1 with the very interesting qualification that they—or at least the virulent 17XYM - have what appear to be numerous "small" genomic differences with respect to the mild lines of genotype of 17X/17XA. LGS analysis in Edinburgh has identified a long (1Mb) section on chromosome 13 that carries the genes for fast or slow growth rate in addition to understanding mechanisms regulating alternative invasion pathways. Sequencing of these genomes will permit identification and subsequent analysis of the "small" genomic differences between fast and slow growing 17X genotype 1 linking sequence content to important phenotypes *e.g.* virulence. Much could be learnt about growth rate from the generation of the genome sequence of a fast growth rate 17X genotype 1 such as 17XYM. This in itself is further justification for getting some more *P. yoelii* lines genome sequenced, and two strains, 33X along with 17X/17XA, which has been used in most *P. yoelii* genetic crosses, will be held in reserve should the independence of the 17X (WHO) and 17XYM lines be a concern after genotyping.

P. berghei

Plasmodium berghei is the most tractable parasite available for reverse genetics technologies and is an invaluable model for studying all aspects of malaria parasite biology. However, the incompleteness of its genome significantly limits high resolution genetic mapping (e.g., Quantitative Trait Locus (QTL)) and comparative genomic studies. *P. berghei* ANKA line K173 neither makes gametocytes nor cytoadheres. Furthermore it lacks the 2.3kb element which is the bane of the assembly process for the ongoing ANKA project and it is not yet clear that this obstacle is one that can be overcome. A large scale rearrangement between K173 and ANKA has yet to be detected so the sequence of K173 would allow the full assembly of a core *P. berghei* genome as well as reveal some interesting absences that could be exploited by reverse genetics. A full core assembly for *P. berghei* ANKA would be an important tool for application of reverse genomics. NK65 is a standard *P. berghei* strain used for drug testing, therefore it will also be included for full genome sequencing. As back-up strains the 233 and 234 are two *P. berghei* ANKA clones that do and do not make gametocytes and so differences there should be of interest, and the XAT line is the irradiated ANKA (by Waki and colleagues) that is attenuated and again should reveal interesting genetic alterations. Of course, without crosses interesting areas responsible for phenotype cannot be identified so readily but reverse genetics would attempt this allied to the comparative genomics and other global studies such as transcriptional analyses.

<u>*Plasmodium* diversity and evolution: Deep-branching malaria parasites</u>

Plasmodium parasites are grouped within the protozoan phylum Apicomplexa, which is comprised of several thousand species [40]. All of these organisms are obligate intracellular parasites, and they are characterized by a distinctive "apical complex" of organelles involved in host cell attachment, invasion, and manipulation of the intracellular environment. Because many of these pathogens cause diseases of clinical and/or veterinary importance, the Apicomplexa are among the most extensively sampled groups eukaryotes: effectively complete genome sequence is currently available for two species of *Babesia,* three species each of *Cryptosporidium, Eimeria, Neospora,* two species each of *Theileria,* and *Toxoplasma* [41], this greatly facilitates comparative genomic analysis of unicellular eukaryotes, biological research into the evolution of parasitism, and studies on the specific diseases caused by these pathogens.

While certain apicomplexans are generalists (a single species of *Toxoplasma can* infect virtually any nucleated animal cell), *Plasmodium* parasites are striking for their extreme adaptation to survival in specific host cells and tissues, and host and vector species. These adaptations—such as the choice of transmission vector, or the evolution of mechanisms for hemoglobin degradation and heme detoxification—are critical for parasite ecology survival and pathogenesis. In order to elucidate these important aspects of parasite biology, and determinants of host range, we propose to sequence five parasite species that are the most accessible albeit with some challenges, but should be particularly informative.

The six species of *Plasmodium* that are the causative agents of human malaria—causing millions of deaths and hundreds of millions of illnesses each year—give only a hint of the substantial systematic and ecological diversity of malaria parasites. Hundreds of species of *Plasmodium* and closely related genera exploit a variety of wildlife mammal species, birds, turtles, and squamate reptiles. These parasites reveal many similarities with the biology of human malaria, but also many differences in life history, presumed nutritional strategy, and vectors infected. Genomic studies of the following taxa are expected open novel windows into the evolutionary history and present life cycle strategies of malaria parasites, including: *Hepatocystis* of monkeys (and bats), *Plasmodium mexicanum* of lizards, *Plasmodium relictum* of birds, and *Haemoproteus* and *Leucocytozoon* of birds.

Rationale for species selected:

Hepatocystis: **morphological vs. molecular phylogeny.** Intriguingly, recent molecular phylogenetic studies show that some parasites with very different morphology and life cycle features (such as the "genus" *Hepatocystis*) fall within the clade of mammal-infecting *Plasmodium,* while some with very similar appearance (*Haemoproteus*) are actually quite divergent (see phylogenetic tree).

P. mexicanum & P. relictum: **Host and vector range (avian/lizard), virulence, accessible liver stages.** In addition to their value for comparative evolutionary studies, these species are of interest because they both infect nucleated erythrocytes, in contrast to the anucleate red blood cells infected by mammal parasites. It is therefore likely that their genomes will be informative as the diversity of strategies used for host cell infection, which may yield insights into such important phenomena as the use of liver vs. erythrocytes by human malaria parasites. The *Shiva* strain of *P. relictum* and *P. mexicanum* are also highly virulent, causing far more harm to their hosts than most other species. They may therefore yield important insights into the ecological devastation wrought by *P. relictum* in native Hawaiian avifauna and unusually virulent nature of *P. falciparum* infections. *P. mexicanum* is also of interest because it is transmitted by sandflies; it is the only *Plasmodium* species known to be transmitted by a vector other than mosquitoes.

Haemoproteus **and** ***Leucocytozoon:*** **The evolution of malaria parasite specialization.** These genera branch more deeply in the phylogeny of malaria parasites [42], and are therefore likely to yield special insights into the evolution of entry into the red blood cells and survival within this hostile environment. Neither species undergo schizogony in red blood cells, but only cast gametocytes into erythrocytes. This appears to be an ancestral strategy for all malaria parasites. Genomic studies may reveal the genetic architecture underlying this important change in the life cycle, and may reflect on changes in placement of schizogony in the human malaria parasites, such as the lack of schizont stages seen in the peripheral blood of *P. falciparum.*

Sample acquisition and quality:
Because all of these isolates are best obtained from the field (although several can grow in the lab; see table), the quantity and quality of DNA are both a matter of some concern. Contamination with host cell DNA is potentially problematic for avian and lizard parasites, due to DNA contamination from the nucleated red cells found in these systems.

<u>*Hepatocystis*</u>

Hepatocystis will be obtained from naturally infected non-human primates (rather than wild bats, which are also infected), and will be coordinated by Dr. John Barnwell, CDC (who purified the *P. gallinaceum* DNA for successful genome sequencing). Material will come either from quarantined rhesus imported from China, or by screening baboons at the Kenya Primate Research Institute facilities outside of Nairobi, with whom the CDC has a good working relationship. This would probably be quite productive as infection rates can be very high in baboons, with *H. simiae* gametocytemias as the best source (although merocysts in the liver are ~2-4 mm and could easily be biopsied. Gametocytes can easily be purified away from contaminating host leucocytes and platelets by first passing the blood through a column containing acid washed glass beads to remove platelets, then Plasmodipur filters and a CF11 cellulose column to remove WBC. The low density gametocytes are then readily purified away from RBC by centrifugation of the blood over density cushions of Percoll or Nycodenz. Being able to bleed about 50 ml of blood twice from several infected rhesus monkeys or baboons when gametocytemias are several thousand or much more per ul would yield ~15 ug or more of DNA.

<u>*P. relictum and P. mexicanum*</u>

The Fleischer laboratory (Smithsonian Institution) has several isolates (and frozen stocks) of *P. relictum* from both Hawaii and Bermuda, and routinely propagates these parasites in canaries in the laboratory, at very high parasitemia. Parasites may also be grown in ducklings, with commensurately higher yields. High quality parasite DNA will be separated from contaminating avian red cell nuclei by Dr. Thomas

McCutchan (NIH), who was responsible for the isolation of *P. gallinaceum* DNA from chickens for the parasite's successful genome project. In the event that material is limiting, DNA will be isolated instead from the related avian parasite *P. juxtanucleare,* which can be raised in the laboratory in chickens .

A population of *S. occidentalis* naturally infected with *P. mexicanum* in Hopland CA has been studied for many years by the Schall laboratory (University of Vermont) and their students and colleagues (including Dr. Susan Perkins, American Museum of Natural History). Prevalence of infection can be as high as 50% of lizards infected in the population, with parasitemias up to 100%. Infections can also be transferred to uninfected animals via injection of infected blood into the peritoneal cavity. The maximum amount of blood obtainable per lizard is only about 70 µl, however. Preliminary calculations suggest that it should be possible to obtain sufficient material for library construction, but this remains uncertain. Assuming that adequate quantities can be obtained, Dr. McCutchan will prepare genomic DNA from these parasites as well. A further possibility would be to dissect the large parasite oocysts from the insect vector.

Haemoproteus and Leucocytozoon

From the genus *Haemoproteus,* the best studied species is *H. columbae* (also the type species for this genus) and from the genus *Leucocytozoon* the equivalent would be *L. simondi* (although not the type species for the genus). Both species are cosmopolitan and well documented in sense of their host and geographic distribution as well as their life cycles and invertebrate host use. Parasites isolated from field isolates will be cultivated in pigeons or ducks in the laboratory, and DNA isolated by Dr. McCutchan. With methodological improvements in sequencing and genotyping, it may be possible to analyze additional representatives that only occur in wild populations and cannot be propagated in the laboratory but that is beyond the scope of the current proposal.

EST and Full-length cDNA Sequencing

EST and full-length cDNA (fl-cDNA) sequencing will provide critical information for the accurate annotation of each of the genomes proposed for sequencing in this proposal. Approximately 50% of the predicted genes in *P. falciparum*, for instance, have no significant similarity to any other known gene, necessitating a completely *de novo* annotation process. The situation is similar across the genus. Further, information on the type or frequency of alternate splicing or other processing is essentially absent for *Plasmodium* spp. Finally, EST and fl-cDNA sequencing will provide key insights into the variant antigen expression repertoire, helping to identify which of the computationally identified potential surface antigens are, in fact, expressed. We propose to sequence at least 40,000 additional ESTs from each species under consideration, focusing on the most readily obtainable stage, and 200,000 from *P. falciparum*, representing all accessible life cycle stages. Existing EST resources, uses and issues have recently been reviewed [43].

There are numerous difficulties in preparing high-quality cDNA libraries for including quantities of material available for non-blood stages, contamination with host RNA, and difficulty in purifying mRNA due to the high AT content of the organism. Additional selection or enrichment of desired clones will doubtless be necessary. It makes sense, therefore, to maximize the information obtained from and usefulness of these libraries. We would therefore propose that the *P. falciparum* project be conducted as a fl-cDNA project, generating all the available information with respect to splicing, allowing better assembly of gene models and creating a clone resource for the community. Due to the presence of high abundance ribosomal RNAs and human hemoglobin RNA, enrichment strategies will be considered to select for messenger RNAs.

Quality Control and Validation of Materials for Sequencing

All parasite DNA will undergo a quality control analysis, including analysis for purity and molecular barcoding or preliminary PCR resequencing to determine that the strains or isolates are indeed unique and suitable for sequencing. It is paramount to ensure that the DNA samples represent parasite genomes of independent origin, and that the material will be as free as possible from host genetic

material. The working group will not only provide these quality samples, but carefully assess them for identity and quality to make available vouchered materials for the sequencing efforts.

Ensuring access to *Plasmodium* genome sequence data (and other relevant genomic-scale datasets), and parasite biologic materials

Sequences emerging from the proposed genome sequencing project will be rapidly released to the public via GenBank and other archival repositories, in accord with NHGRI (http://www.genome.gov/25521732 http://www.genome.gov/25521732) and NIAID (http://www.niaid.nih.gov/dmid/genomes/mscs/data_release.htm) policies. In addition to deposition in the relevant EST, HTS, genome, and SNP databases, *Plasmodium*-related data is also made available to the broader research community via the *Plasmodium* genome database (http://PlasmoDB.org) [44, 45], a component of the larger ApiDB project [41], supported in the context of the broader ApiDB Bioinformatics Resource Center through a contract from NIAID [46]. In addition to information on *Plasmodium,* ApiDB also supports genomic-scale datasets from many other species in the protozoan parasite phylum Apicomplexa, including other pathogens of clinical and/or veterinary interest, and parasites that serve as accessible experimental models for *Plasmodium.*

The PlasmoDB component of ApiDB incorporates a wide range of genomic-scale datasets, including finished and unfinished genome and EST sequence data, manually curated and automatically generated annotations, SNPs and other population genetics data, comparative genomics and synteny analysis, and a variety of functional genomics data (expression profiling studies, proteomics results, interactome data, etc). Powerful queries enable users to identify genes relevant to many studies of biomedical and evolutionary interest, including the identification of targets for drug and vaccine development [47-49].

PlasmoDB is widely used by the malaria research community and others, receiving ~15,000 hits per day, from >100 countries worldwide, and a CD version is available to endemic country researchers without reliable access to high-speed internet connections [50]. Standard protocols have been implemented to facilitate regular data exchange with the various sequencing centers that have been engaged in *Plasmodium* sequencing projects over the years (including WTSI, The Institute for Genomic Research (TIGR) / J. Craig Venter Institute, genome centers at Stanford and Washington University, The Broad Institute of MIT & Harvard, and elsewhere). These protocols also help to ensure synchrony with GenBank, the WTSI's GeneDB [51], and other data providers; GenBank depositions are typically jointly held by the relevant data generators and PlasmoDB, in order to facilitate continued curation and annotation [52].

The NIAID-funded malaria repository *MR4* (http://www.mr4.org/ [39]) provides a convenient and efficient mechanism for archiving and distributing biological material generated by this white paper. Where possible and appropriate, the species and isolates described in this paper will be deposited as frozen parasite stabilates and/or DNA so that they can be made available to the wider malaria research community. In addition, whole genome reagents such as genomic and cDNA libraries, tiling paths and cDNA clones will be deposited where appropriate. Deposition of such material, in particular for *Plasmodium* species which cannot be grown *in vitro* and for which biological material is scarce, will be of prime importance since *P. vivax* biological resources, either patient or monkey derived, are not easily obtained.

Final summary

The proposed work will strengthen existing *P. falciparum* genomic resources and build a platform of new polymorphism data on which current and diversity of *P. falciparum* can be understood. This will be a critical tool for understanding how the parasite's defenses reshape themselves under the pressure of current and coming efforts at malaria control and eradication. At the same time, we propose new sequencing in the difficult to study human malaria *P. vivax*, complementing and extending the limited current data available. Additionally this work will characterize the genomes of non-human parasites that already serve as good models for the human malarias and will extend our knowledge of the biology of the Apicomplexa, a vast group of deadly parasites of species ranging from lizards to small mammals and birds and to the human and non-human primates. Proposed work in rodent malarias will improve current model systems and perhaps enhance the development of new ones. All data generated will be rapidly placed into the context of an established and well-utilized community resource, PlasmoDB, from which

data will easily flow into the broader apicomplexan database ApiDB. More than fifty labs around the world have contributed and commented on the proposed work, which will comprise a remarkable step forward for the malaria and broader apicomplexan fields.

1. Bryce, J., et al., *Countdown to 2015: tracking intervention coverage for child survival.* Lancet, 2006. **368**(9541): p. 1067-76.
2. Volkman, S.K., et al., *A genome-wide map of diversity in Plasmodium falciparum.* Nat Genet, 2007. **39**(1): p. 113-9.
3. Joy, D.A., et al., *Early origin and recent expansion of Plasmodium falciparum.* Science, 2003. **300**(5617): p. 318-21.
4. Mu, J., et al., *Recombination hotspots and population structure in Plasmodium falciparum.* PLoS Biol, 2005. **3**(10): p. e335.
5. Albrecht, L., et al., *Extense variant gene family repertoire overlap in Western Amazon Plasmodium falciparum isolates.* Mol Biochem Parasitol, 2006. **150**(2): p. 157-65.
6. Barry, A.E., et al., *Population genomics of the immune evasion (var) genes of Plasmodium falciparum.* PLoS Pathog, 2007. **3**(3): p. e34.
7. Elizabeth V. Fowler, J.M.P., Michelle L. Gatton, Nanhua Chen and Qin Cheng, *Genetic diversity of the DBLα region in Plasmodium falciparum var genes among Asia-Pacific isolates.* Mol Biochem Parasitol, 2002. **120**: p. 117-26.
8. Taylor, H.M., et al., *A study of var gene transcription in vitro using universal var gene primers.* Mol Biochem Parasitol, 2000. **105**(1): p. 13-23.
9. Fowler, E., Peters, JM, Gatton, ML, Chen, N and Cheng, Q, *Genetic diversity of the DBLα region in Plasmodium falciparum var genes among Asia-Pacific isolates.* Mol Biochem Parasitol, 2002. **120**: p. 117-26.
10. Mendis, K., et al., *The neglected burden of Plasmodium vivax malaria.* Am J Trop Med Hyg, 2001. **64**(1-2 Suppl): p. 97-106.
11. Baird, J.K., *Chloroquine resistance in Plasmodium vivax.* Antimicrob Agents Chemother, 2004. **48**(11): p. 4075-83.
12. Collins, W.E. and G.M. Jeffery, *Primaquine resistance in Plasmodium vivax.* Am J Trop Med Hyg, 1996. **55**(3): p. 243-9.
13. Figtree, M., et al., *Plasmodium vivax synonymous substitution frequencies, evolution and population structure deduced from diversity in AMA 1 and MSP 1 genes.* Mol Biochem Parasitol, 2000. **108**(1): p. 53-66.
14. Imwong, M., et al., *Microsatellite variation, repeat array length, and population history of Plasmodium vivax.* Mol Biol Evol, 2006. **23**(5): p. 1016-8.
15. Jongwutiwes, S., et al., *Mitochondrial Genome Sequences Support Ancient Population Expansion in Plasmodium vivax.* Mol Biol Evol, 2005. **22**(8): p. 1733-1739.
16. Cornejo, O.E. and A.A. Escalante, *The origin and age of Plasmodium vivax.* Trends Parasitol, 2006. **22**(12): p. 558-63.
17. Pfahler, J.M., et al., *Transient transfection of Plasmodium vivax blood stage parasites.* Mol Biochem Parasitol, 2006. **149**(1): p. 99-101.
18. Ramjanee, S., et al., *The use of transgenic Plasmodium berghei expressing the Plasmodium vivax antigen P25 to determine the transmission-blocking activity of sera from malaria vaccine trials.* Vaccine, 2007. **25**(5): p. 886-94.
19. Carlton, J., *The Plasmodium vivax genome sequencing project.* Trends Parasitol, 2003. **19**(5): p. 227-31.

20. Ferreira, M.U., et al., *Population structure and transmission dynamics of Plasmodium vivax in rural Amazonia.* J Infect Dis, 2007. **195**(8): p. 1218-26.
21. Imwong, M., et al., *Relapses of Plasmodium vivax Infection Usually Result from Activation of Heterologous Hypnozoites.* J Infect Dis, 2007. **195**(7): p. 927-33.
22. Mongui, A., et al., *Identifying and characterising the Plasmodium falciparum RhopH3 Plasmodium vivax homologue.* Biochem Biophys Res Commun, 2007. **358**(3): p. 861-6.
23. Ryan, J.R., et al., *Evidence for transmission of Plasmodium vivax among a duffy antigen negative population in Western Kenya.* Am J Trop Med Hyg, 2006. **75**(4): p. 575-81.
24. Culleton, R., et al., *Linkage group selection: rapid gene discovery in malaria parasites.* Genome Res, 2005. **15**(1): p. 92-7.
25. Martinelli, A., et al., *A genetic approach to the de novo identification of targets of strain-specific immunity in malaria parasites.* Proc Natl Acad Sci U S A, 2005. **102**(3): p. 814-9.
26. Escalante, A.A., et al., *A monkey's tale: the origin of Plasmodium vivax as a human malaria parasite.* Proc Natl Acad Sci U S A, 2005. **102**(6): p. 1980-5.
27. Escalante, A.A., A.A. Lal, and F.J. Ayala, *Genetic polymorphism and natural selection in the malaria parasite Plasmodium falciparum.* Genetics, 1998. **149**(1): p. 189-202.
28. Coatney, G.R., et al., *The Primate Malarias*. 1971, Washington, D.C.: U.S. Department of Health, Education and Welfare.
29. *Conservation Assessment and Management Plan for the Primates of Indonesia*. Final Report 2001, Conservation International Indonesia, Taman Safari Indonesia.
30. Jongwutiwes, S., et al., *Naturally acquired Plasmodium knowlesi malaria in human, Thailand.* Emerg Infect Dis, 2004. **10**(12): p. 2211-3.
31. Singh, B., et al., *A large focus of naturally acquired Plasmodium knowlesi infections in human beings.* Lancet, 2004. **363**(9414): p. 1017-24.
32. Mu, J., et al., *Host Switch Leads to Emergence of Plasmodium vivax Malaria in Humans.* Mol Biol Evol, 2005. **22**(8): p. 1686-1693.
33. Nguyen-Dinh, P., et al., *Cultivation in vitro of the vivax-type malaria parasite Plasmodium cynomolgi.* Science, 1981. **212**(4499): p. 1146-8.
34. Schmidt, L.H., et al., *The course of untreated Plasmodium inui infections in the rhesus monkey (Macaca mulatta).* Am J Trop Med Hyg, 1980. **29**(2): p. 158-69.
35. Nimri, L.F. and N.H. Lanners, *Immune complexes and nephropathies associated with Plasmodium inui infection in the rhesus monkey.* Am J Trop Med Hyg, 1994. **51**(2): p. 183-9.
36. Nguyen-Dinh, P., C.C. Campbell, and W.E. Collins, *Cultivation in vitro of the quartan malaria parasite Plasmodium inui.* Science, 1980. **209**(4462): p. 1249-51.
37. Carlton, J.M., et al., *Genome sequence and comparative analysis of the model rodent malaria parasite Plasmodium yoelii yoelii.* Nature, 2002. **419**(6906): p. 512-9.
38. Hall, N., et al., *A comprehensive survey of the Plasmodium life cycle by genomic, transcriptomic, and proteomic analyses.* Science, 2005. **307**(5706): p. 82-6.
39. Adams, J.H., Y. Wu, and A. Fairfield, *Malaria Research and Reference Reagent Resource Center.* Parasitol Today, 2000. **16**(3): p. 89.
40. Levine, N.D., *Progress in taxonomy of the Apicomplexan protozoa.* J Protozool, 1988. **35**(4): p. 518-20.

41. Aurrecoechea, C., et al., *ApiDB: integrated resources for the apicomplexan bioinformatics resource center.* Nucleic Acids Res, 2007. **35**(Database issue): p. D427-30.
42. Martinsen, E.S., S. L. Perkins, and J. J. Schall, *A three genome phylogeny of malaria parasites (Plasmodium and related genera): evolution of life history traits and host switches.* . Molecular Phylogenetics and Evolution, 2007. **In Revision**.
43. Carlton, J., *Pilot gene discovery in plasmodial pathogens*. Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics. 2005.
44. Bahl, A., et al., *PlasmoDB: the Plasmodium genome resource. A database integrating experimental and computational data.* Nucleic Acids Res, 2003. **31**(1): p. 212-5.
45. Kissinger, J.C., et al., *The Plasmodium genome database.* Nature, 2002. **419**(6906): p. 490-2.
46. Greene, J.M., et al., *National Institute of Allergy and Infectious Diseases bioinformatics resource centers: new assets for pathogen informatics.* Infect Immun, 2007. **75**(7): p. 3212-9.
47. Chaudhary, K. and D.S. Roos, *Protozoan genomics for drug discovery.* Nat Biotechnol, 2005. **23**(9): p. 1089-91.
48. Roos, D.S., et al., *Mining the Plasmodium genome database to define organellar function: what does the apicoplast do?* Philos Trans R Soc Lond B Biol Sci, 2002. **357**(1417): p. 35-46.
49. Tongren, J.E., et al., *Malaria vaccines: if at first you don't succeed.* Trends Parasitol, 2004. **20**(12): p. 604-10.
50. Milgram, A.J., J.C. Kissinger, B. Gajria, D.S. Pearson, A. Bahl, P. Labo and D.S. Roos. , *Plasmodium falciparum GenePlot: Internet-independent access to the malaria parasite genome.* . Nature, 2003. **422**: p. CD-ROM.
51. Hertz-Fowler, C., et al., *GeneDB: a resource for prokaryotic and eukaryotic organisms.* Nucleic Acids Res, 2004. **32**(Database issue): p. D339-43.
52. Berry, A.E., et al., *Curation of the Plasmodium falciparum genome.* Trends Parasitol, 2004. **20**(12): p. 548-52.

*THE PLASMODIUM WRITING GROUP (In alphabetical order)

John Adams	jadams3@health.usf.edu
Tim Anderson	tanderso@sfbrgenetics.org
John Barnwell	wzb3@CDC.GOV
Mathew Berriman	mb4@sanger.ac.uk
Bruce Birren	bwb@broad.mit.edu
Richard Carter	rcarter@staffmail.ed.ac.uk
William Collins	wec1@cdc.gov
Jane Carlton	jane.carlton@med.nyu.edu,
Alan Cowman	cowman@wehi.edu.au
Karen Day	karen.day@med.nyu.edu
Abdoulaye Djimde	adjimde@yahoo.com
Patrick Duffy	patrick.duffy@sbri.org
Manoj Duraisingh	MDURAISI@hsph.harvard.edu
Ananias Escalante	ananias.escalante@asu.edu
Robert Fleischer	FleischerR@si.edu
Kasturi Haldar	k-haldar@northwestern.edu
Dan Hartl	dhartl@oeb.harvard.edu
L. Hiller	luhiller@gmail.com
Dominic Kwiatkowski	Dominic.Kwiatkowski@paediatrics.ox.ac.uk
Manuel Llinas	manuel@genomics.princeton.edu
Ayo Oduola	oduolaa@who.int,
Victoria McGovern	vmcgovern@bwfund.org
Mathias Marti	MMARTI@hsph.harvard.edu
Ellen Martinsen	Ellen.Martinsen@uvm.edu
Dan Neafsey	neafsey@broad.mit.edu
Chris Neubold	cnewbold@hammer.imm.ox.ac.uk
Christopher Plowe	cplowe@medicine.umaryland.edu
Susan Perkins	perkins@amnh.org
David Roos	droos@sas.upenn.edu
Phil Rosenthal	rosnthl@itsa.ucsf.edu
Pardis Sabeti	pardis@broad.mit.edu,
Joseph Schall	jschall@uvm.edu
Balbir Singh	bskhaira55@gmail.com
Xinxhuan Su	XSU@niaid.nih.gov
Akhil Vaidya	av27@drexel.edu
Sarah Volkman	svolkman@hsph.harvard.edu
Andy Waters	A.P.Waters@lumc.nl
Roger Wiegand	rwiegand@broad.mit.edu
Elizabeth Winzeler	ewinzeler@gnf.org
Dyann Wirth	dfwirth@hsph.harvard.edu