

Sequencing Plan for *Ixodes scapularis*

Broad MSC PI and Primary Contact: Bruce W. Birren

TIGR MSC PI: Claire Fraser

Project PI: Catherine A. Hill, Purdue University

Additional Collaborator: Vishvanath M. Nene, TIGR, and Stephen K. Wikel, UCHC

Executive Summary

- We will generate genomic resources to support research on *Ixodes scapularis*, including BAC-end sequences, ESTs and finished sequence for 60 BAC clones. EST sequencing will be carried out at the Broad MSC, BAC-end sequencing will be carried out at the TIGR MSC and BAC clone sequencing will be shared between the MSCs. These tasks will be performed in the next year.
- Additionally, the Broad and TIGR MSCs will jointly produce a 6X draft assembly of the *I. scapularis* genome with each MSC generating approximately 3X sequence coverage. Genomic sequencing is expected to take place over several years, with the rate being subject to regular review and adjustment.

Introduction

In the United States, *I. scapularis* is the most important tick species from a human health perspective. *Ixodes scapularis* transmits Lyme disease in the northeastern and north-central US as well as human granulocytic anaplasmosis and babesiosis and possibly the flaviviral agent of Powassan encephalitis which is related to West Nile virus. Recent studies by Anderson *et al.*, (2003) also suggest that West Nile virus can be transmitted trans-stadially by *I. scapularis* although vector competence has yet to be established. Lyme disease is arguably one of the most important vector borne diseases in the US, Europe and Asia. Over 17,000 positive Lyme disease cases were reported to the US in 2000 (Centers for Disease Control and Prevention, 2002). Lyme disease and other tick-borne diseases have important long term health consequences. Of further concern is the fact that the incidence and geographic spread of Lyme disease and other tick-borne disease are increasing and many cases are suspected to be vastly under-reported or misdiagnosed (Walker, 1988). *Ixodes scapularis* is a member of the Prostriata, an evolutionarily primitive phyletic line of the Acari that includes a number of medically significant tick species. *Ixodes scapularis* genome data will be widely applicable to studies of other prostriates including *I. pacificus*, the vector of Lyme disease on the US Pacific Coast, *I. ricinus* and *I. persulcatus*, the Eurasian *Ixodes spp.* vectors of Lyme disease and tick borne encephalitis and *I. holocyclus*, an Australian ixodid responsible for transmission of *Rickettsia* and *Borrelia* and human cases of tick paralysis. These factors, particularly the wide range of human diseases that it transmits, make *I. scapularis* an excellent candidate for a genome project that seeks to have an ultimate impact on human welfare through development of novel vector suppression measures, therapeutics and vaccines.

Isolate to be sequenced

Following extensive consultation with the tick community, the *I. scapularis* colony maintained by Dr. S. Wikel at the University of Connecticut Health Center has been selected as the most suitable source of material for sequencing. This colony was established in 1996 using field collected material from New York, Oklahoma and a LD endemic area of Connecticut. This colony has been continuously in-bred since establishment and has not been supplemented with field collected material. The colony is commonly used and has been well characterized. The colony is known to be a competent vector of various *Borrelia* and *Babesia* isolates and has been used to produce a variety of cDNA libraries and EST sequences that are the basis of a number of ongoing genomics studies (see Section V). Dr. Wikel's laboratory is able to supply sufficient material for sequencing and to meet community requests for the reference strain long term. We propose the establishment of satellite colonies at Old Dominion University (Dr. D. Sonenshine) and North Carolina State University (Dr. M. Roe) for long term maintenance and preservation of the

reference strain. Dr. C. Hill has developed a method to obtain large amounts of high molecular weight genomic DNA from ixodid ticks (Hill and Gutierrez, 2003). In anticipation of a genome effort, Dr. Hill is currently extracting high quality genomic DNA from *I. scapularis* embryos for production of BAC libraries. We do not propose the production of endosymbiont free *I. scapularis* for sequencing. Ixodid ticks are obligate hematophagous organisms; many studies suggest that *I. scapularis* is associated with endosymbionts that are thought to play a critical role in the lifecycle of this organism (Sonenshine, 1991). Sequencing of these organisms may help to unravel aspects of this dynamic relationship.

Sequencing and analysis strategy

We propose to include the following activities in the *I. scapularis* genome project.

- Sequencing of normalized cDNA libraries, up to a maximum of approximately 100,000 clones from two different libraries, each from pooled material
- Sequencing the ends of all clones in a genomic DNA BAC library offering ~10X physical coverage of the genome (approximately 167,000 clones assuming 120kb average insert size and a 2 Gb genome)
- Complete sequencing and finishing of 40 randomly selected BAC clones (average insert size approximately 120 kb)
- Complete sequencing and finishing of 20 BAC clones (average insert size approximately 120 kbp) selected by the community from the BAC library as encoding defined genes.
- Sequencing whole genome shotgun libraries to a total depth of 6X sequence coverage.

Sequencing of normalized cDNA libraries

Based on the interests and research priorities identified by the tick community, two normalized libraries composed of pooled tick tissues will be constructed:

- Library 1 - Pooled *Ixodes scapularis* Non-infected Library: Composed of non blood fed and various blood fed and replete whole *I. scapularis* and *I. scapularis* life-cycle stages.
- Library 2 – Pooled *Ixodes scapularis* Infected Library: Composed of blood-fed *I. scapularis* and *I. scapularis* infected with *Borrelia burgdorferi*, *Babesia microti* and possibly *Anaplasma phagocytophilum*.

Tissue samples will be collected from participating tick research labs and pooled for library production using a standardized protocol. cDNA libraries will be produced at Express Genomics (Frederick, Maryland).

To ensure a high rate of gene discovery, Library 1 will be sequenced and analyzed by the MSC in batches of 20,000 clones up to a maximum of 80,000 clones. One round of subtraction may be carried out as needed to enrich for rarer cDNAs. Library 2 will be subtracted using abundant ESTs from Library 1 and sequenced up to a maximum of 20,000 clones.

Both the 5' and 3' ends of these clones will be sequenced. In addition to providing the greatest value to the users when these sequences are aligned to the genome, having both 5' and 3' sequences will provide vital information about gene structure that will be used to train the gene prediction software that will be applied to the genome sequence. In fact, in the case of most genes, we expect the 5' and 3' reads to overlap, providing complete sequence for the coding regions of these genes.

Besides their utility in full length gene cloning, genome annotation and the development of micro-arrays, the ESTs will underpin ongoing proteomics studies, inform on the repertoire of expressed genes and gene families and permit evolutionary comparisons between invertebrate genes. ESTs will provide the community with information on the types of genes and gene families that are expressed during tick

developmental stages and within tissues and will also enable the identification of genes involved in blood feeding, host finding and pathogen transmission.

BAC-End Sequencing (BES) and Complete BAC Sequencing

A 10X genomic DNA BAC library (average insert size of approximately 120kb) will be generated from *I. scapularis* embryos. Genomic DNA extraction will be performed and BAC libraries will be produced at the Clemson University Genomics Center in collaboration with Dr. J. Tomkins. We will sequence 40 randomly selected BAC clones and 20 BAC clones selected by the community as encoding defined genes. A committee of tick researchers will solicit and prioritize requests for clones to be sequenced. Additionally, both ends of all clones in the library will be sequenced. Assuming an average insert size 120 kb and a genome size of 2 Gb the BAC library should contain approximately 167,000 clones.

Finished sequence for random and selected BACs will provide preliminary but valuable information on genome organization, including repetitive sequences, gene structure and density as well as serving as a resource for validation of the whole genome assembly. Further, these completely sequenced BACs and BAC-end sequences will provide sequence data for the community to begin analyzing the *Ixodes* genome in advance of obtaining a whole genome assembly. BES will generate important architectural information on the genome and will enable the development of a preliminary physical map for *I. scapularis*. Paired BAC end sequence data are also important as sequence-tagged connectors to assist in scaffold assembly.

Whole-genome Shotgun Sequencing

The shotgun sequencing phase of this project will be carried out jointly by the Broad and TIGR MSCs. Each center will generate approximately 3X sequence coverage of the *I. scapularis* genome. The sequences from both centers will be combined to generate a final assembly of approximately 6X sequence coverage.

The Broad MSC proposes to generate 3X sequence coverage using a combination of 4kb pOT plasmid and 40 kb Fosmid end sequences in the ratio described in Table 1.

Library type	% Sequence Coverage	Sequence Coverage	# Reads	Clone Coverage
4 kb pOT	90%	2.7	8,275,968	8.3
40 kb Fosmid	10%	0.3	1,283,712	12.8
Total	100%	3.0	9,559,680	21.1

Table 1. Broad MSC coverage strategy for *Ixodes scapularis*

The TIGR MSC proposes to generate 3X sequence coverage for the genome using a 10 and 4 kbp pHOS and 50 kb pHOS-Kan plasmid end sequences in the ratio described in Table 2. The exact ratios will be determined based on the quality of genomic libraries. This strategy complements the sequence data that will be generated by the Broad MSC.

Library type	% Sequence Coverage	Sequence Coverage	# Total Reads	Clone Coverage
4 kbp pHOS	10%	0.3	900,000	0.91
10 kbp pHOS	80%	2.4	7,200,000	18.36
50 kbp pHOS-Kan	10%	0.3	900,000	11.48
Total	100%	3.0	9,000,000	30.75

Table 2. TIGR MSC coverage strategy for *I. scapularis*

Timeline and milestones

In order to best serve the needs of the research community and NIAID, *I. scapularis* sequencing will be scheduled to accommodate NIAID's changing priorities. At any given time, the MSCs will devote sequencing capacity to NIAID projects that have sequencing reagents available according to directions from the NIAID Project Officer. While the MSCs will attempt to complete the *I. scapularis* project as efficiently as possible, any timeline developed at this stage could be altered by instructions from the Project Officer and the arrival of new, high-priority projects. These timelines also depend upon the receipt of reagents (DNA and libraries) compatible with our production pipelines. For example, the receipt of low quality DNA, an insufficient quantity of DNA, or libraries constructed with vectors that are not compatible with our process could impact these time estimates.

Construction of the first cDNA library and the BAC library are expected to be completed by February 2005. Construction of the second cDNA library is expected to be completed by September of 2005. Genomic DNA is expected to be available by February 2005.

Assuming no delays due to the above-mentioned reasons, we expect sequencing of the first 20,000 cDNAs to be completed within 6 months of receiving each library. Decisions about subsequent sequencing will be made after analysis of these data. BES and BAC clone sequencing will be completed within 7 months of receipt of the BAC libraries and clones. BAC assemblies will be generated and released within 45 days of obtaining shotgun sequence assuming no significant errors are detected during the internal validation process.

We expect genome sequencing to be completed over the course of the next three years. An interim assembly will be generated when 3X coverage is achieved. The 6X assembly will be generated within three months of completion of the shotgun sequencing. The assemblies will be released within 45 days of their generation assuming no significant errors are detected during the validation process. An automated annotation of the 6X assembly will be released within 45 days of its generation assuming no significant errors are detected during the validation process.

Costs

The Broad MSC anticipates that sequencing and finishing costs for their component of this project will be approximately \$8.2M (see appendix A). This amount includes all equipment, labor, reagents and overhead associated with library construction, sequencing, detection, and supporting informatics infrastructure for cDNA end sequencing, 3X whole genome shotgun sequencing, and complete sequencing of 30 BAC clones. This amount also includes the costs associated with finishing 30 BAC clones. All other costs, including personnel and equipment costs associated with assembly, annotation, and analysis are charged at a monthly rate and are not included in this cost estimate. Actual costs may

vary from this estimate depending on, for example, actual genome size, aggregate pass rates, read lengths, and the final characteristics of the cDNA libraries currently being constructed.

The TIGR MSC anticipates that sequencing and finishing costs for their component of this project will be approximately \$ 9.32 M (see appendix A). This figure includes all equipment, labor, reagents and overhead associated with library construction, sequencing and supporting bioinformatics infrastructure for BAC end sequencing, sequencing of 30 BAC clones and 3 X whole genome shotgun sequence coverage of the *Ixodes scapularis* genome. Sequence data from the Broad MSC will be downloaded from trace archive at GenBank and incorporated into assembly analyses. Other costs, including personnel and equipment costs associated with assembly, BAC closure, annotation and analyses are charged at a monthly rate. Actual costs may vary from this estimate depending, for example, on aggregate sequence success rate, read length, genome size and randomness of genomic libraries.

Additional reagent funds in the amount of \$20,000 will be needed for the generation of cDNA libraries by Express Genomics.

Data release

In accordance with NIAID's principles regarding data release, we will publicly release all data generated under this contract as rapidly as possible.

Chromatogram Files: Unless otherwise directed by NIAID, the MSCs will submit all sequence and trace files (chromatograms) generated under this proposal to the Trace Archive at NCBI on at least a weekly basis. These data will also include information on templates, vectors, and quality values for each sequence.

BAC Assemblies: BAC assemblies will be made available via GenBank and the relevant MSC web site after internal validation. Assuming no significant errors are detected during the validation process, the assembly will be released within 45 calendar days of being generated.

Genome Assemblies: 3X and 6X assemblies will be made available via GenBank and the relevant MSC web site after internal validation. Assuming no significant errors are detected during the validation process, the assemblies will be released within 45 calendar days of being generated.

Genome Annotation: Automated annotation data will be made available via GenBank and our web sites after internal and community validation. Assuming no significant errors are detected during the validation process, annotation data will be released within 45 calendar days of being generated.

Publication and Authors

The Broad MSC and TIGR MSC will work with investigators to facilitate the publication of research results as they are discovered. Sequencing, assembly, and analysis of the *I. scapularis* data will be carried out as part of collaboration between the Broad Institute, TIGR and the research community. The project will be coordinated by Dr. Hill. Authorship and author order will be determined immediately prior to submitting manuscripts for publication based on the specific contributions of the individuals involved. All editorial issues associated with manuscript preparation and authorship will be coordinated by Dr. Hill.

Appendix A: Project Costs

Broad MSC Costs			
Total Costs			\$8,207,683
Total Sequencing Reads			9,880,320
Sequencing Costs			
	Reads	Cost per read	Total cost
cDNA sequencing	240,000	\$0.78	\$187,200
BAC clone sequencing	80,640	\$1.28	\$103,219
Whole genome shotgun sequencing (3X)			
4 kb pOT	8,275,968	\$0.78	\$6,455,255
40 kb Fosmid	1,283,712	\$1.08	\$1,386,409
			\$7,841,664
Finishing Costs			
	Finished bases	Cost per base	Total cost
BAC clone finishing	3,600,000	\$0.021	\$75,600
Notes			
* Sequencing costs include all equipment, labor, reagents and overhead associated with library construction, sequencing, detection, and supporting informatics infrastructure.			
* All other costs including personnel and equipment costs associated with assembly, annotation, and analysis are charged at a monthly rate and are not shown in this budget.			
* Actual costs may vary from this estimate depending on, for example, aggregate pass rates, read lengths, final characteristics of the BAC and cDNA libraries currently being constructed, and actual genome size.			

TIGR MSC Costs

Total Costs **\$ 9,322,318**
Total Sequencing Reads 9,571,000

Sequencing Costs

	Reads	Cost per read	Total cost
BES sequencing	334,000	\$ 2.25	\$ 751,500
BAC insert sequencing	57,000	\$ 0.94	\$ 53,580
Library construction and closure reagents			\$ 54,238
			\$ 107,818
Whole genome shotgun sequencing (3X)			
4 kbp pHOS	900,000	\$ 0.94	\$ 846,000
10 kbp pHOS	7,200,000	\$ 0.94	\$ 6,768,000
50 kbp pHOS-Kan	900,000	\$ 0.94	\$ 846,000
Library construction			\$ 3,000
			\$ 8,463,000

Notes

- * Random sequencing costs include all equipment, labor, reagents and overhead associated with sequencing, detection, and supporting informatics infrastructure.
- * All other costs including personnel and equipment costs associated with assembly, annotation and analysis are charged at a monthly rate and are not shown in this budget.
- * Actual costs may vary from this estimate depending on, for example, aggregate pass rates, read lengths, characteristics of the BAC libraries currently being constructed and genome size.

Additional Reagent Costs (Express Genomics)

cDNA library construction **\$20,000**