Heartworm Genomics: Unprecedented Opportunities for Fundamental Molecular Insights and New Intervention Strategies

Robin B. Gasser and Cinzia Cantacessi

Vector-borne diseases, including canine heartworm disease (CHWD), are of major socioeconomic and canine health importance worldwide. Although many studies have provided insights into CHWD, to date there has been limited study of fundamental molecular aspects of *Dirofilaria immitis* itself, its relationship with the canine host, its vectors, as well as the potential of drug resistance to emerge, using advanced -omic technologies. This article takes a prospective view of the benefits that advanced -omics technologies will have toward understanding *D. immitis* and CHWD. Tackling key biological questions using these technologies will provide a "systems biology" context and could lead to radically new intervention and management strategies against heartworm.

Keywords: canine heartworm disease (CHWD), *Dirofilaria immitis*, heartworm, drug resistance, genomics, genetics, molecular biology, transcriptomics

Many vector-borne diseases have a major adverse impact on animal and human health worldwide.[1-3] A particularly important example of veterinary importance is canine heartworm disease (CHWD), caused by the filarial nematode *Dirofilaria immitis*.[4,5] Despite current knowledge and developments in advanced technologies, there are still major gaps in our knowledge of many areas, including the fundamental molecular biology, molecular epidemiology, ecology and population genetics of the causative agent and its vectors, the disease itself, and also drug resistance development in *D. immitis*.[6,7] Moreover, there are some limitations in diagnosis and intervention, which represent critical obstacles to the effective control of CHWD.[4,8] Although sustained research and funding have contributed significantly to an improved understanding of vector-borne diseases of humans,[9-12] this is not the case for those of veterinary importance, which are presently neglected in terms of research and development.

The revolution in molecular and computer technologies provides substantial prospects for investigating important pathogens and their vectors, providing insights into their epidemiology, ecology, evolution, and cellular processes. However, the relatively high cost and laborious nature of molecular biological research has sometimes been an impediment to progress, particularly in the veterinary field. Revolutionary developments in a range of genomic technologies[13] provide unprecedented opportunities to explore CHWD and other diseases, at a rate and on a scale unimaginable just a couple of years ago, providing enormous prospects to tackle critically important areas of research for the first time. Future research should harness such technologies to address major knowledge gaps in CHWD. Elucidating *D. immitis*, its relationship with its vectors and its definitive hosts, the disease itself, and the epidemiology and ecology of the parasite will have substantial implications for improving the diagnosis, treatment, prevention, and control of CHWD in years to come. This article takes a prospective view of the impact that -omic technologies could have on our knowledge and understanding of *D. immitis* and CHWD on the molecular level. We briefly explain the principles of transcriptomic and genomic sequencing as well as bioinformatic technologies, and describe the exciting implications of these technologies in both fundamental and applied areas.

Background on Genomic and Bioinformatic Technologies

Genomics is the study of the entire complement of genetic material (DNA or genome) in an individual organism (e.g., parasite). Transcriptomics is the molecular science of examining the transcription of all genes at the level of the cell,

From the Faculty of Veterinary Science, The University of Melbourne, Parkville, Victoria 3010, Australia.
Current research in the Gasser Lab is funded mainly through the Australian Research Council (ARC), the National Health & Medical Research Council (NHMRC), and Melbourne Water Corporation.
Address reprint requests to: Robin B. Gasser, DVM, PhD, DVSc, Faculty of Veterinary Science, The University of Melbourne, Parkville, Victoria 3010, Australia. E-mail: robinbg@unimelb.edu.au.

1527-3369/06/0604-0171\.00/0
doi:10.1053/j.tcam.2011.09.003

tissue, and/or whole organism, allowing inferences regarding cellular functions and mechanisms. The ability to explore and measure the transcription of thousands of genes simultaneously has led to major advances in all biomedical fields, from understanding basic functions in model organisms, such as *Caenorhabditis elegans* (a free-living nematode) and *Drosophila melanogaster* (vinegar fly),[14-18] to investigating molecular processes linked to growth, development, and reproduction, to the study of the mechanisms of survival and drug resistance. Until recently, transcriptomes have been characterized qualitatively by sequencing expressed sequence tags using conventional technologies,[19,20] whereas levels of transcription have been assessed using complementary DNA (cDNA) microarrays[21] or real-time polymerase chain reaction (PCR).[22] In the last few years, there has been an increased demand for practical computer tools for the efficient annotation of nucleotide sequence datasets, particularly within the framework of large-scale expressed sequence tag projects.[23]

Next-generation Sequencing of Transcriptomes and Genomes

There has been a massive expansion in next-generation sequencing (NGS) technologies,[24-27] which now provide unique opportunities to explore de novo the transcriptomes and nuclear genomes of different species, strains and developmental stages of *D. immitis* as well as its vectors and its canine or other definitive hosts. The capacity of NGS techniques to generate millions to hundreds of millions of sequences, in parallel, has put them at the forefront of the scientific research.[28–30]

NGS technologies include 454 (Roche),[24] SOLiD,[27] and Illumina/HiSeq.[25] For example, the 454 technology platform[24] uses a sequencing-by-synthesis approach, by which cDNA is randomly fragmented into 500 to 1000 base pair fragments. During the process of cDNA library construction, an adaptor is ligated to each end of these fragments and then mixed into a population of agarose beads whose surfaces anchor oligonucleotides complementary to the 454-specific adapter sequence, such that each bead is linked to a single fragment. Each of these complexes is transferred into individual oil-water micelles containing amplification reagents and then subjected to an emulsion PCR step, during which ~10 million copies are produced and bound to individual beads. In the sequencing step, the beads anchoring the cDNAs are deposited on a picotitre plate, together with other enzymes required for the pyrophosphate sequencing reaction on a solid support system.[31] The 454 technology (a "long-read" platform) has been used frequently for de novo genomic or transcriptomic studies. However, increasingly, Illumina/HiSeq is being used for this purpose. This latter technology differs significantly from the 454 approach,[25] in that, after fragmentation of cDNA, Illumina-specific adaptors are ligated to each cDNA template and then covalently attached to the surface of a glass slide, allowing individual templates (cDNAs) to form bridge-like structures. During the amplification step (bridge PCR), clonal clusters of ~1000 amplicons are generated and immobilized to a single physical location on the slide. Then the cDNAs are linearized, and the sequencing reagents (including 4 fluorescently labeled nucleotides) are added to the flow cell. After the individual fluorescent bases have been incorporated, the flow cell is interrogated with a laser beam in several locations, allowing several image acquisitions at the end of a single synthesis cycle, and the sequences to be read.[31] This technology is considered ideal for resequencing projects, targeted sequencing, single nucleotide polymorphism analyses, and gene transcription studies.

In the last years, numerous studies have shown the utility of NGS technologies for investigating, for example, aspects of the molecular biology, systematics, and population genetics of parasites.[32-35] In particular, 454 technology has been used recently for the rapid de novo sequencing of the transcriptomes of numerous pathogens of veterinary and human health importance,[33-41] yielding substantial datasets and providing a major step forward in the understanding of the basic molecular biology of these organisms.

Bioinformatic Tools (= Computer Programs) for the Analysis of NGS Datasets

The development of practical and efficient bioinformatic tools has become critical for comprehensive analyses of DNA, RNA and protein sequence data as well as for making biological sense of such datasets. Therefore, there has been relatively rapid progress in the construction of new programs and/or integrated pipelines, some of which are accessible via the Worldwide Web.[42-46] The principles, methods and protocols for the analysis of sequence data, together with currently available bioinformatic tools and pipelines, have been reviewed recently.[47]

In brief, after the acquisition of data, nucleotide sequences are firstly screened for repeats, contaminants and/or adaptor sequences,[47,48] and "clustered" (assembled) into contiguous sequences (of maximum length; called contigs) based on sequence similarity.[47] Long reads (produced by 454 or conventional sequencing technologies) and short reads (e.g., Illumina) are assembled using the algorithms "overlap-layout-consensus"[49] and "de Bruijn graph,"[50,51] respectively. For the former algorithm,[49] pairwise overlaps among reads are computed and stored in a graph; the graphs are used to compute a layout of reads and then a consensus sequence of contigs.[42,52-57] For the de Bruijn graph,[50,51] reads are fragmented into short segments, called "*k*-mers" (number of nucleotides in each segment; overlaps among *k*-mers are captured and stored in graphs, which are then used to produce consensus sequences.[51,56-60]

After assembly, the contigs and single reads (singletons) are compared, using different types of the Basic Local Alignment Software Tool (BLAST),[61] with known sequence data available in public databases, to infer an identity for each query sequence, if significant matches are found.[47] In addition, assembled nucleotide sequences are usually conceptually translated into pre-