Genome sequencing white paper for the black fly disease vectors *Simulium* sp.

submitted by C. Brockhouse, A.Papanicolaou, R. Post, D. Boakye , E.W. Cupp, M. Pfrender and J.K. Colbourne on behalf of the Simuliidae community

1 Executive summary

Black flies (Diptera: Simuliidae) are the second most medically-important group of arthropod pest species affecting human health and are now the most important group of disease vectors and pest species lacking a representative genome project. The family also occupies a critical taxonomic placement in the sub-order Nematocera, making it an important out-group for genomic studies in mosquitoes and other important hematophagous Diptera associated with leishmaniasis and viral encephalitis. *Simulium* species are the sole vectors for the human filarial parasite *Onchocerca volvulus*, the causative agent of onchocerciasis, or river blindness. River blindness is a scourge of some of the poorest regions of the world, affecting 39 million people worldwide. Historically, it has been the second most important infectious cause of blindness (after trachoma) and the second most important disease in terms of overall socio-economic impact, eclipsed only by polio.

The socio-economic impact of onchocerciasis has been recognized by the international community, which has supported a number of large control programs to reduce the impact of the disease on the afflicted populations. For example, the Onchocerciasis Control Programme in West Africa, a program that focused primarily upon the control of *S. damnosum* in 13 countries of Africa was active from 1975-2002, and spent more than $565 million over its lifetime. The tradition of interest in onchocerciasis by the international community continues to this day, with ongoing programs in Africa (the African Programme for Onchocerciasis Control (APOC)) and Latin America (the Onchocerciasis Elimination Program in the Americas (OEPA)) spending millions to control the disease.

Until recently, it was felt that onchocerciasis could not be eliminated using the currently available tools, which rely primarily upon mass distribution of the anti-helminthic drug ivermectin. However, recent studies have suggested that long term treatment with ivermectin has dramatic effects on fertility of the adult female parasites. This finding has resulted in a paradigm shift in the field of onchocerciasis control, moving from a focus upon control in Africa towards elimination. The strategy to accomplish elimination involves delineating isolated foci of the infection, and eliminating the infection sequentially in these isolated foci, eventually resulting in continent wide elimination of the parasite. For this strategy to be successful, it is first necessary to delineate these isolated foci or transmission zones. In Africa, the major vector *Simulium damnosum s.l.*, is known to be capable of long distance migration. The extent of the transmission zones will thus likely be determined by vector movement throughout most of Africa. To delineate these transmission zones, polymorphic markers capable of distinguishing populations and of determining effective population sizes are critically needed. Genome sequences of the African vectors for onchocerciasis will provide a rich source of such markers.

This document proposes the sequencing of 11 *Simulium* genomes (genome size smaller than 187 Mb), including both disease vectors and non-vector species. For critical vector species and populations, we simultaneously propose re-sequencing of 50 individuals per species, to identify polymorphisms for vitally needed population genetic studies, and to identify genes involved in vectorial capacity and the ability to colonize. The availability of multiple black fly genomes will provide an epidemiological and genetic framework delimiting the spatial distribution of vector sibling species in areas where onchocerciasis control is on-going, greatly aid the identification of promising species-

specific genetic control targets, provide a frame-work for ecological genomic studies, and give an *entrée* into detailed gene regulation analysis in an otherwise intractable group of insects.

Table 1: Proposed *Simulium* species targeted for genome sequencing.

TIER	Species	Location	Vector	DNA source*	Laboratory
1	*S. vittatum*	North America	-	Colony material	UGA
	S. sirbanum	Burkina Faso (Savanna spp.)	++	Larvae from isofemale progeny	MDSC
	S. damnosum ss	Togo (Savanna spp.)	++	"	MDSC
	S. ochraceum	Mexico	++	"	UANL
	S. squamosum	Ghana (Forest spp.)	+++	"	NMRI
2	*S. ochraceum*	Galapagos Islands	-	Larvae from pure sibling site	Creighton/ CDF
	S. thyolense	Malawi (Colonizing species)	+++	Adults from pure sibling site	NMRI/Creighton
	S. sanctipauli	Ghana (Forest spp.)	+++	Larvae from isofemale progeny	NMRI
	S. woodi	Tanzania (Different subgenus from other vectors)	+++	"	NMRI/LSHTM
	S. exiguum	Ecuador	+++	"	Creighton/CDF
3	*S. yahense*	Ghana (Forest spp.)	+++	"	NMRI

"+++", "++": major *Onchocerca* vector, "+": vector. "-": non-vector model species. UGA, University of Georgia (Athens); MDSC, Multi-Disease Surveillance Centre, WHO; UANL, University of Nuevo Leone, Mexico ; CDF, Charles Darwin Foundation (Ecuador); LSHTM, Post lab London School of Hygiene & Tropical Medicine; NMRI, Noguchi Medical Research Institute, Ghana; Creighton, Creighton University, USA.

2 Background

2.1 Vector profile

The Simuliidae (black flies) are generally regarded as the second most medically important group of insects that impact the health and economic well-being of humans (Adler et al., 2004). They are by now the most important vector group that is not yet represented by large-scale genomics. The blood-feeding activity of adult females transmits several pathogens to humans, most notably *Onchocerca volvulus*, the causative agent of river blindness, as well as *Mansonella ozzardi*; significant veterinary pathogens including species of *Leucocytozoon*, *Trypanosoma*, and *Dirofilaria,* as well as vesicular stomatitis virus (Adler 2005). *Onchocerca volvulus* alone infects approximately 39 million people in Africa, the Arabian Peninsula, and Central and South America (Crainey and Post, 2010), and has been the focus of one of the World's largest insect control programs (Onchocerciasis Control Programme).

2.2 Onchocerciasis: a debilitating disease

Human onchocerciasis is a severely debilitating disease, which is caused by infection with the parasitic nematode *Onchocerca volvulus* (Onchocercidae: Filarioidea). Ninety-nine percent of cases occur in sub-Saharan Africa, where it causes blindness and skin disease, which are together responsible for the loss of over one million Disability Adjusted Life Years (DALYS) every year (WHO-TDR, 2008). Affected persons spend an additional 15% of their annual income on health, children are more likely to drop out of school and farmers have about 30% less land under cultivation (Fischer & Büttner, 2002), thus directly contributing to poverty. The most important pathologies are blindness and skin disease. Onchocerciasis is the third leading cause of preventable blindness in the tropics (Narita & Taylor, 1993), but skin disease is responsible for 60% of lost DALYs. There is also some evidence that onchocerciasis may be associated with epilepsy and dwarfism (Basáñez et al., 2006), and perhaps it increases susceptibility to malaria and reduces efficacy of vaccinations (Druilhe et al. 2005). The biting nuisance caused by the vectors also has a real but unmeasured effect on economic activities and this impact extends to northern latitudes (Hougard et al. 1998).

2.3 A "translational" genome project: The importance of genomic sequence to Onchocerciasis transmission and control

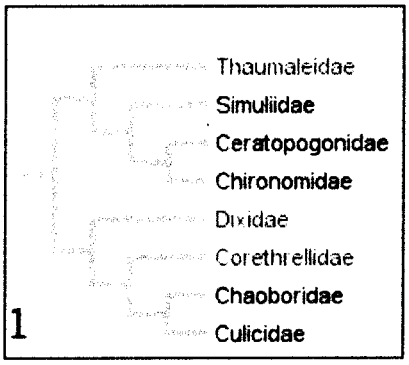
Throughout most of its geographical range in Africa, the parasite is transmitted between people by biting black flies belonging to the *Simulium damnosum* complex. Therefore, larviciding was the main strategy adopted by the World Health Organization Onchocerciasis Control Programme (OCP) in West Africa, which treated vector breeding sites in 11 countries from 1974 to 2002, to interrupt transmission. It was estimated that ≥ 600,000 cases of blindness were prevented at a cost of $556 million (WHO, 2002), and the parasite reservoir was considerably diminished. When the drug ivermectin was approved for use against onchocerciasis in 1987, it was adopted by OCP for annual distribution to communities to give immediate clinical benefit. With the support of the WHO African Programme for Onchocerciasis Control (APOC), the drug has become the mainstay of current onchocerciasis disease control in those 19 endemic countries, which had not been part of the old OCP. Annual Community Directed Treatment with Ivermectin (CDTI) has conferred enormous clinical benefits, and with more than 40 million people currently being treated it is helping to avert the loss of 500,000 DALYS every year (Amazigo & Boatin, 2006). Ivermectin was shown to be a potent microfilaricide when given annually, meaning that little or no effect was seen upon adult parasites in the initial clinical trails conducted to evaluate the effect of treatment on *O. volvulus* (Devaney and Howells, 1984). This necessitated that ivermectin be given at least annually to control onchocerciasis as a public health problem, and that complete elimination of the parasite population using ivermectin was probably not feasible in Africa, due to the scope of the problem and the intensity of transmission there (Borsboom et al., 2003; Winnen et al., 2002). However, more recent studies have suggested that long term annual treatment with ivermectin dramatically reduces the fertility of the adult female parasites (*cf.* Cupp et al., 2010). Furthermore, a recent report has suggested that after 15 years of annual and semi-annual ivermectin distribution, transmission has been interrupted in parts of Senegal and Mali (Diawara et al., 2009). This finding has led to a paradigm shift in the thinking about onchocerciasis in Africa, where the emphasis is now shifting from control of the disease as a public health problem to elimination of the parasite population altogether.

The strategic plan to eliminate onchocerciasis in Africa can be summarized by the title of an informal consultancy convened by the WHO and APOC to develop a plan for onchocerciasis elimination in Africa - "Shrinking the Map". The strategic plan calls for delineating isolated foci of the infection and targeting the isolated foci for elimination using mass treatment with ivermectin. Foci will be successively targeted until elimination is complete. To implement such a strategy, knowledge of the limits of the foci to be targeted is critical. Throughout much of sub-Saharan Africa, the primary vectors

of onchocerciasis are *S. damnosum sensu stricto* and *S. sirbanum*, two savannah dwelling members of the *S. damnosum* species complex. Both of these species are known to be capable of migrating long distances so that the spatial features of the foci in Africa will be determined primarily by the movement of the vector flies. The most effective way to monitor this will be to measure effective vector population sizes. Thus, population level genetic markers are critically needed to refine the estimates of the effective sizes of the onchocerciasis foci throughout Africa, as the current programs move from control to elimination. The most effective way to generate such a suite of markers will be from complete genome sequence data, supplemented by transcriptome data derived from RNA-Seq runs. *Simulium* epidemiology will be supported by the generation of a suite of molecular markers. Since different classes of markers are more appropriate for different applications, we can use the genomic and transcriptomic sequences to generate multiple types of markers including single-nucleotide polymorphisms, microsatellites, and exon-primed markers. Such markers can now be rapidly genotyped, and will assist the community to better understand population structure, gene flow and can also be used as markers for monitoring epidimiological traits. The black fly genome is favorable for detailed scoring of populations with molecular markers, due to the low chromosome number (n=3) and low recombination rate (0.7 chiasma per chromosome arm per meiosis, as judged from detailed cytogenetic studies; Rothels and Nambiar, 1975).

2.4 Beyond parasite transmission

2.4.1. Value of black flies for comparative genomic studies



Not all black flies transmit pathogens, but their taxonomic placement and phylogenetic relationships in a superfamily of vectors is critical. The Simuliidae is a basal family (Nematocera: Culicimorpha) in the order Diptera ("true flies") composed of over 2000 species. It is a sister family to the Chironomidae-Ceraptogonidae clade (non-biting and biting midges, respectively). Using morphological characters, these three families, together with the Thaumaleidae, in turn, form a sister group within the Culicimorpha as a superfamily cluster that includes the Culicidae ("mosquitoes") (Figure 1; from "Tree of Life", after Wood and Borkent, 1989). Understanding the genetics governing life-history traits, which lead to a species evolving into a vector of human pathogens, can be greatly facilitated by comparing vector and non-vector sibling species within this family and between vector species in closely related families such as the Ceratopogonidae ("biting midges").

The availability of black fly genomes will also be of great value to the mosquito community. The black fly genome will provide a critical outgroup in a sister clade within the Culicimorpha, but outside the mosquito/chaoborid/dixid clade (see Figure 1). Comparative genomics between the two groups should prove enlightening, especially given the suite of critical biological similarities (haematophagy, etc) and differences (disparate host location mechanisms, blood feeding behavior differences, distinct ecological niches, etc) of the two families. Furthermore, other research communities will benefit from the availability of black fly genomes, including investigators studying species belonging to the Chironomidae and the Corethrellidae. Both families are placed within the Nematocera (lower dipterans). The Chironomidae is included in the superfamily Culicomorpha and while not a blood feeding taxon, it is the focus of over 500 research laboratories as a model system for molecular cytogenetics, transcription studies, nuclear export, etc., including such luminaries as Beerman, Daneholt and Edstrom. The black fly genome will provide a closely related entrée into a broader range of *Chironomus* genes than would otherwise be possible. The Corethrellidae is believed to be

ancient family by comparison with the Simuliidae and also contains species that were the first in the superfamily to evolve the ability to blood-feed.

The availability of sequence from this project and other vector species will therefore provide a unique opportunity to study how the genomes of disease-transmitting vectors have evolved. Using sequence similarity- and Hidden Markov Model-driven orthology, we will identify potential genes which are essential for a successful vector or pathogen. First, a robust phylogenetic framework will be established. This will allow us to utilize a phylogenetic approach to identify genes which are diverging at different rates for a particular phenotype such as vector competency. Secondly, the availability of metagenomes of the infected black flies will allow a comparison of pathogen genomic profiles between disease- and non-disease-causing populations within a single species. A complete genome sequence of reference species will also allow us to determine the extent of microsynteny conservation between the Simuliidae. This will assist with the genome assembly efforts. Finally, microsynteny conservation will facilitate the transfer of functional annotations from a well-characterized species to a non-characterized one. These predictions will drive subsequent research for biochemical investigations.

2.4.2. Genetics of host location

Studies of odor receptors (ORs) and odor binding proteins (OBPs) have been identified as promising areas for the development of tools to interrupt transmission by interfering with prey location. Such studies are well advanced in both the model *Drosophila* system (e.g., Robertson et al., 2003; Xu, 2009) and in vector taxa such as mosquitoes (e.g., Fox, 2001; Xu et al., 2003). Comparative genomics examination of black fly odor sensing genes would not only be informative to simudologists but, given the differences between black fly and other dipteran host seeking behavior, should prove illuminating for other vector communities. In addition, identification of such OBPs may lead to the identification of attractants for black flies, which could serve as the basis for the development of traps for their flies. Such traps would have great utility as a surveillance tool to monitor *O. volvulus* infection levels in flies in areas subject to elimination efforts, and may also serve as a supplemental method to achieve elimination. Finally, such traps may have a commercial application as a way to reduce nuisance populations in North America.

2.4.4. Host feeding behavior

Simuliids exhibit a wide variation of host choice, with some species being almost exclusively anthropophilic, while others feed exclusively upon large animals and birds. Host choice is a critical component of vectoral capacity. Importantly, this difference in host feeding behavior has also resulted in apparent biochemical adaptations as well. For example, the bite of the zoophilic North American species (such as *S. vittatum*) is notoriously painful, while the bite of the anthrophilic African species (e.g. *S. yahense*) are painless. This adaptation probably results from the fact that humans exhibit a very effective defensive behavior to being bitten (slapping the offending insect) while the ruminants that are the normal hosts of zoophilic species such as *S. vittatum* exhibit much less effective defensive behaviors. An obvious hypothesis to explain this adaptation is that the salivary secretions of the anthropophilic species contain powerful anesthetic molecules which are lacking in the saliva of the zoophilic species. A genome aided study to identify such differences in the salivary profile of these species would be an effective way to identify such anesthetic molecules, which might in turn have important medical applications.

2.4.5. Black Fly saliva as a source of drugs and vaccines

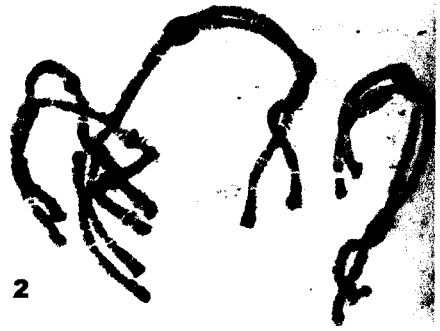
As alluded to in the previous Section, black fly salivary secretions contain an array of novel, pharmacologically active proteins that are capable of (1) inhibiting the anti-hemostatic pathway of vertebrates,, (2) causing maximum and prolonged vasodilation of peripheral blood vessels, and (3) suppressing immune responses. Because of their activity which has been selectively directed by evolution, some of these molecules hold promise as drugs in treatment of clotting disorders and immune diseases. One – a recombinant vasoactive protein – has already been demonstrated as a potential therapeutic agent in promoting wound healing by increasing transdermal vascular flow and accelerating wound repair (Cupp et al., 1998). As such, this (and possibly other) salivary molecules may hold great potential for stimulating the healing of chronic wounds associated with extensive cellular necrosis and compromised blood supply such as decubital ulcers or diabetic foot wounds, as well as helping assure perfusion of severely traumatized wound tissues.

Black fly saliva also contains one or more chemoattractants that diffuse through the skin and direct *Onchocerca* spp. microfilariae to the bite site during vector blood-feeding, thereby insuring infection. These molecules therefore serve as the fundamental mechanism to control the vector-parasite link and eventually involve the human host after larval development in the fly and subsequent transmission of the L_3. If identified, such molecules could prove extremely useful as transmission-blocking targets to uncouple the vector-parasite link. Because a similar phenomenon has been described for *Wuchereria bancrofti* and *Culex* spp., identification of black fly salivary microfilarial attractants could prove useful for development of a transmission-blocking vaccine for lymphatic filariasis as well.

3 Genomic Resources

3.1 Genome structure

Typically, black flies have n = 3 submetacentric chromosomes, which are greatly amplified in the larval silk glands (Figure 2; the polytene chromosomes of *S. vittatum*). The chromosome complement has been reduced to n=2 independently in several simuliid taxa, including the subgenus *Eusimulium*, and in certain members of the genus *Cnephia* (Leonhardt 1985, Procunier 1982).


2

The only known size of a black fly genome is that of *Prosimulium multidentatum*; it is estimated to be 187 Mbp in length (Sohn et al. 1975). However, Rothfels and Nambiar (1975) indicated that this species has unusually large meiotic/mitotic chromosomes, so the genomes consider here are somewhat smaller.

Crainey et al., (2010) describe the *S. squamosum* BAC library constructed by the Post laboratory. Hybridization results with putatively single copy genes are consistent with a *S. squamosum* genome size slightly smaller than that of *P. multidentatum*. All the species examined by Sohn et al. (1975) had non-trivial amounts of highly repetitive DNA (~30-40 %) and an AT-rich genome (~70 % AT)

3.2 Cytogenetics

Natural populations of black flies have been the focus of intensive cytogenomic studies for over 60 years (reviewed by Rothfels, 1979, 1981, Adler et al., 2005, Adler et al., 2010). Detailed polytene chromosome maps and annotated idiograms of approximately 500 species have been published, with considerably more data available from various laboratories (*cf.* Adler et al., 2004). Many species in the genus *Simulium* have been linked into large chromosome synteny-based phylogenies. To date, very few genes have been mapped to the polytenes by *in situ* hybridization, largely because so few cloned genes are available that no significant coverage can be achieved. However, anti-hemostatic proteins in *Simulium vittatum* (Procunier et al., 2005) and insecticide resistance loci in *Simulium sanctipauli* (Boakye et al., 2000) have been placed in cytogenetic maps. Furthermore, an array of cytogenomic data is also available for *S. damnosum* siblings (Adler et al., 2010; Kreuger, 2006; Post et al., 2007), which should prove invaluable for assembly of the genomes.

3.3 *S. squamosum* BAC library

A *S. squamosum* BAC library, that also contains a complete *Wolbachia* genome and extensive coversage of a merithid nematode parasite of the simuliid, has been desribed by Crainey et al., (2010). This library is available for the genome project, for example in our BAC-FISH genome mapping experiment. The DNA for the library was derived from larvae collected at River Pawnpawn, Ghana (a cytogenomically well characterized population).

3.4 Existing EST resources

The adult female salivary glands of *S. vittatum* and *S. nigrimanum* are represented by EST collections of 1748 and 2147 sequences each, respectively (Andersen et al., 2009; Ribeiro et al., 2010). A carefully normalized transcriptome (using methods described in Meyer et al., 2009) representing all life stages of *S. vittatum* consisting of 16,384 contigs and 61,333 singletons was constructed by the Simulium Genomics Consortium. The contigs have been run through a preliminary annotation. Of these, 16,482 have no significant sequence similarities to proteomes of sequenced organisms and thus represent "black fly unique" sequences (i.e., rapidly evolving or lineage-specific genes that are strong candidates for functions unique to black fly biology). We anticipate submitting discoveries made from the transcriptome sequencing for publication and opening the database (through InsectaCentral.org) to the general scientific community in late summer 2011.

3.5 DNA immediately available

The Brockhouse laboratory (Creighton) currently holds approximately 1000 individuals of *S. vittatum* (UGA colony), 100 of *S. squamosum* (R. Pawnpawn), 20 *S. ochraceum* "Galapagos", 20 *S. thyolense* (Malawi), and 20 *S. sanctipauli* (R. Akran). DNA from any or all of these will be available for sequencing should any delays be experienced by the institutions making purpose-collections of any species listed in Table 1 or 2.

4 Target species

We propose to determine the sequence of isolates of 11 *Simulium* species (Table 1 and Section 5). These species represent a wide range of characteristics within the genus, including documented

onchocerciasis vectors from both Africa and Latin America. The species have also been chosen based upon our ability to obtain material that will be as homogeneous as possible. *S. vittatum* has been included in this list as it is the only *Simulium* species to have been successfully colonized. This colony was initially established at Cornell University roughly 30 years ago and has been maintained continuously since then, without the introduction of any wild material. We feel that material from this colony is likely to exhibit a higher degree of homozygosity than any of the field collected species, and should prove a useful prototype genome to assist in the assembly of the sequences obtained from the field collected species.

Table 2: Proposed sequencing strategy.

Tier	Species	DNA source	Laboratory	Genome Size	Proposed Genome Coverage	Proposed Transcriptome Sequencing
1	*S. vittatum*	Colony material	UGA	≥187Mbp	100X (n=1)	Completed; analysis in progress.
	S. sirbanum	Larvae from isofemale progeny	MDSC	≥187Mbp	100X (n=1) 40X (n=50)	Yes
	S. damnosum ss	"	MDSC	≥187Mbp	100X (n=1) 40X (n=50)	Yes
	S. ochraceum	"	UANL	≥187Mbp	100X (n=1)	Yes
	S. squamosum	"	NMRI	≥187Mbp	100X (n=1)	Yes
2	*S. ochraceum "Galapagos"*	Larvae from pure sibling site	Creighton /CDF	≥187Mbp	100x (n=1) 40X (n=50)	Yes
	S. thyolense	Adults from pure sibling populations in Malawi	NMRI/ Creighton	≥187Mbp	100X (n=1)	Yes
	S. sanctipauli	"	NMRI	≥187Mbp	100X (n=1)	Yes
	S. woodi	Tanzania	NMRI/ LSTMH		100X(n=1)	Yes
	S. exiguum	"	Creighton /CDF	≥187Mbp	100X (n=1)	Yes
3	*S. yahense*	"	NMRI	≥187Mbp	100X (n=1)	Yes

5 Proposed sequencing strategy

We propose high-coverage sequence for 1 individual (= 2 hapolid genomes) per species. 2 to 5 ug of DNA per individual are routinely obtained in DNA extractions from larvae. Isofemale egg rearing will be conducted to ensure sibling progeny for the genome sequencing experiments, reducing the overall level of genetic variation in the template DNA of the samples to be sequenced. We also propose to conduct in-depth transcriptome analysis of all species included in this project. The proposed target species exhibit different vector competencies and host choice preferences. Having transcriptome data will be essential to carry out the studies described in section 2.4 above.

In addition to the *de novo* sequencing efforts on the nine species outlined in Table 1, we also propose to conduct re-sequencing of 50 individuals from the two most important vectors of onchocerciasis in Africa, *S. damnosum s.s.* and *S. sirbanum* (Table 2). The purpose of this re-sequencing effort will be to identify molecular markers that might be used to delineate endemic onchocerciasis foci throughout Africa, assisting in the elimination programs currently being planned. The 50 individuals will be drawn from geographically distant, cytogenomically characterized populations (approximately 5 individuals from each location), to maximize polymorphism capture. Sites have been identified in Ghana, Nigeria, Togo, Burkina Faso, and Uganda. One silk gland from each larva will be removed, for sibling identification by D. Boakye and R. Post, and the rest of the body used for DNA extractions by established methods (Brockhouse et al., 1993). The *S. ochraceum* population of the Galapagos Islands is also proposed for resequencing. It is an unique example of a recent colonization event, or rapid expansion of a previously relic population, that poses a new threat to human health (*cf.* Nelder et al., 2004). The Galapagos resequencing effort will yield important new data on the biology of black fly population expansion.

The sequencing of natural populations is an active field of research at the Broad Institute and they are generating new sequencing and assembly strategies which will be instrumental in successfully assembling the genomes. In our discussions with VectoBase, Dr. Scott Emrich has expressed confidence that their experience with the mosquito and tsetse genome projects will allow them to hone their ability to assemble "wild" genomes.

Assembly efforts will be aided by the extensive collection of available cytogenetic maps. Dr. Brockhouse's group will work on BAC-FISH to generate a physical map. In addition, the extensive synteny between *Simulium* species will guide our assembly of the low-coverage genomes. (There are no inter-chromosome arm rearrangements within the selected groups. Intra-arm inversions have "scrambled" the synteny among the species, but large regions are conserved and easily recognizable on polytene chromosome maps).

6 Community involvement and impact

The black fly community is represented by several regional groups: the North American Black Fly Association (https://www.clemson.edu/cafls/departments/esps/research/adler/nabfa/index.html), the British Simulium Group (www.blackfly.org.uk) and the European Simuliidae Association. The North American and British groups hold annual meetings, and the European Association hosts a biannual international symposium. Approximately 200 labs word-wide have simuliids as a primary or significant research focus.

Members of the community directly interested in this project have long-standing involvements in exploring the genetics, phylogeny, systematics and vector biology of medically-important *Simulium* species. The collective experience of the group amounts to centuries of laboratory and field work focused on the basic biology and epidemiology of the Simuliidae and its relationship to human and animal diseases. Sequencing of the genomes listed in Table 1 will undoubtedly act as a catalyst for further genetic exploration of this very important vector group. In addition, a number of public health professionals working to control onchocerciasis are keenly interested in the translational aspects of this project as it addresses a crucial need identified by the onchocerciasis control program, namely the identification of foci of transmission.

The black fly community, although comparatively small, has produced the most comprehensive cytogenomic analysis of population chromosome synteny maps and chromosome synteny based phylogenies of any family of living organism at the natural population level. Complete genome sequences will greatly enhance the value of this extensive cytogenomic groundwork, allowing the community to move to the next level, enabling the mapping of specific inversion breakpoints, and the layering of genetic annotations onto well-developed physical maps. The production of this white

paper has already attracted a growing pool of new talent to the field; the full genomes will unquestionably serve as a focus for the further revitalization of the simuliid community.

6.1. Genome sequence annotation.

Data will be released following the procedures established by previous NHGR/NIAID-funded vector genomes as follows. Sequence data generated in the course of the project will be deposited at Eugenes Arthropod database (http://arthropods.eugenes.org/arthropods/data/blackfly/) and VectorBase as soon as it is generated. We anticipate no delay between generation and deposition of the data. We are committed to the immediate release of all RNA-Seq data to the community via InsectaCentral.org and VectorBase, since the insect and vector genomics communities will obtain the most benefit from the data. In addition, we will aim to submit the raw sequence data to the NCBI Trace Archive (http://www.ncbi.nlm.nih.gov/Traces/trace.cgi). We also intend to make all our RNA-Seq MINSEQE-compliant, that is providing the "Minimum Information" about a high-throughput SeQuencing Experiment (see http://www.mged.org/minseqe/), so that the data can be deposited at GEO (see http://www.ncbi.nlm.nih.gov/geo/infor/seq.html).

The Center for Genomics and Bioinformatics (CGB) and CSIRO (A. Papanicalaou) will provide the initial (automated) annotation of the assembled genomes using the Ensembl pipeline (Curwen, V., et al., 2004). More detailed community annotation will be done, and managed according the experience of the *Daphnia* Genome Consortium (John Colbourne, and Michael Pfrender).

Additional gene models will be predicted and improved at the CGB using in-house pipelines that include Fgenesh family models (Salamov and Solovyev, 2000), Genewise family models (Birney and Durbin, 2000), SNAP (Korf, 2004) and a newly developed protocol by Dr. Don Gilbert (Indiana University) called Evidence Directed Gene predictions for Eukaryotes (EvidentialGene) (http://arthropods.eugenes.org/genes2/). Colleagues at the NCBI RefSeq Project Group will provide RefSeq transcript alignments (Pruitt et al., 2005) and Gnomon gene prediction (http://www.ncbi.nlm.nih.gov/genome/guide/gnomon.shtml). Finally, ESTs will be used to extend predicted gene models into fuller-length genes by adding to 5' and/or 3' UTRs. The PASA annotation pipeline (Haas et al., 2003) will be used to further refine the gene models by verifying that spliced alignments of ESTs are congruent with the predicted gene structures. The elected gene set will be given putative functional assignments by homology to annotated genes from NCBI non-redundant sets and classified according to Gene Ontology (Harris et al., 2004), eukaryotic orthologous groups (Koonin et al., 2004), and KEGG metabolic pathways (Kanehisa et al., 2004).

One of the most challenging elements of a genome sequencing project is functional annotation, which is essential for extracting biological significance from the vast amounts of newly acquired sequence information (Elsik et al., 2006; Stein, 2001). To aid in this task, Indiana University's CGB will host an annotation training workshop with supporting web-conferencing, and design and implement a community-wide manual annotation project modeled from earlier experiences with the waterflea (*Daphnia;* Colbourne et al., 2011) (http://conferences.cgb.indiana.edu/daphnia2007/index.html), jewel wasp (Werren et al. 2010) and pea aphid (https://dgc.cgb.indiana.edu/display/aphid/Workshop+I) genome projects, which are coordinated through collaboration wikis. The *Simulium* annotation project will employ a hybrid "jamboree" and "cottage industry" model that will bring together the *Simulium* community with bioinformaticians and insect genome biologists in a five-day intensive annotation workshop that will serve to train the community and jump-start a longer-term decentralized annotation effort that will be dispersed throughout the community (Elsik et al., 2006; Stein, 2001).

The annotation workshop will be hosted at the Mount Desert Island Biological Laboratory (MDIBL) located on the coast of Maine. John Colbourne is Adjunct Associate Professor at MDIBL and will obtain funding for this event via contacts made through the Environmental Genomics Summer course that he offers with colleagues every summer. MDIBL is equipped and staffed for hosting such events.

Participants will work hands-on with annotation training modules and will hear from invitees from other genome projects who will highlight the annotation process and speak from experience on ways to improve our annotation efforts. The goal is to educate and excite the community about their role in building this resource, familiarize them with the annotation software and other technical aspects, and facilitate future collaborative research efforts. The workshop will be held during the first academic break (summer or winter) following genome assembly and Ensembl annotation and will be open to all interested researchers. We hope to identify opportunities from the NIH to help provide travel awards for students and post-doctoral researchers.

The *Simulium* genome database will be housed at VectorBase (see letter of support), built with common Generic Model Organism Database (GMOD; Stein et al., (2002)) components and open source software shared with other genome databases. The computationally intense analysis we propose will benefit from the TeraGrid project (www.teragrid.org), which is part of a shared cyber infrastructure for sciences, funded primarily by NSF. We have used TeraGrid to annotate and validate the assembly of a *Daphnia* genome, where results included homologies to nine eukaryote proteomes, gene predictions, marker genes, and EST locations.

7 References

Adler, P. H. 2005. black flies, the Simuliidae. Pp. 127-140. *In* W. C. Marquardt (ed.). Biology of Disease Vectors, 2nd edition. Elsevier Academic Press, San Diego, CA.

Adler, P. H., D. C. Currie and D. M. Wood. 2004. The black flies (Simuliidae) of North America. Cornell University Press, Ithaca, NY

Adler PH, Cheke RA, and Post RJ. 2010. Evolution, epidemiology, and population genetics of black flies (Diptera: Simuliidae). Infect Genet Evol. 2010 Oct;10(7):846-65

Amazigo U & Boatin B (2006) The future of onchocerciasis control in Africa. *Lancet* 368, 1946-1947.

Andersen JF, Pham VM, Meng Z, Champagne DE, Ribeiro JM. 2009. Insight into the sialome of the Black Fly, Simulium vittatum. J Proteome Res. 2009 Mar;8(3):1474-88.

Baker RHA, Guillet P, Seketeli A, Poudiougo P, Boakye D, Wilson MD & Bissan Y (1990) Progress in controlling the reinvasion of windborne vectors into the western area of the Onchocerciasis Control Programme in West Africa. *Philosophical Transactions of the Royal Society of London* B **328**, 731-750.

Basáñez M-G, Pion SDS, Churcher T, Breitling LP, Little MP & Boussinesq M (2006) River blindness: a success story under threat? *PLoS Medicine* **3**, e371.

Borsboom GJJM, Boatin BA, Nico JD, Nagelkerke NJD, Agoua H, Akpoboua KLB, Soumbey Alley EW, Bissan Y, Renz A, Yameogo L, Remme JHF & Habbema DF (2003) Impact of ivermectin on onchocerciasis transmission: assessing the empirical evidence that repeated ivermectin mass treatments may lead to elimination/eradication in West Africa. *Filaria Journal* 2:8

Beaumont, M.A. and D.J. Balding, *Identifying adaptive genetic divergence among populations from genome scans.* Molecular Ecology, 2004. **13**(4): p. 969-980.

Birney, E. and R. Durbin, *Using GeneWise in the Drosophila annotation experiment.* Genome Research, 2000. **10**(4): p. 547-548.

Boakye, DA., Cornel, AJ., Merideth, SE., Brakefield, PM., and Collins, FH. 2000. DNA in situ hybridization DNA in situ hybridization on polytene chromosomes of Simulium sanctipauli at loci relevant to insecticide resistance. Med. Vet. Entomo. 14: 217-222 .

Brockhouse, C.L., Vajime, C.G., Marin, R., and Tanguay, R.M. 1993. Molecular identification of

onchocerciasis vector sibling species in black flies (Diptera: Simuliidae). *Biochem. Biophys. Res. Commun.* **194**: 628-634.

Butlin, R.K., *Population genomics and speciation.* Genetica, 2010. 138(4): p. 409-418.

Colbourne, J.K., et al. (2011) The ecoresponsive genome of *Daphnia pulex*. Science 331: 555-561.

Crainey JL, Hurst J, Wilson MD, Hall A, Post RJ.. (2010). Construction and characterisation of a BAC library made from field specimens of the onchocerciasis vector Simulium squamosum (Diptera: Simuliidae). Genomics. 2010 Oct;96(4):251-7.

Crosskey, RW. 1990. The Natural History of Blackflies. John Wiley & Sons. Chichester.

Cupp, EW., Sauerbrey, M, and Richards, F. 2010. Elimination of human onchocerciasis: History of progress and current feasibility using invermectin (Mectizan®) monotherapy. Acta Trop. Aug. 2010. Eprint ahead of publication.

Cupp MS, Ribeiro JM, Champagne DE, and Cupp EW. (1998). Analyses of cDNA and recombinant protein for a potent vasoactive protein in saliva of a blood-feeding black fly, Simulium vittatum. J. Exp. Biol. 201: 1553-1561.

Curwen, V., et al., (2004) *The Ensembl automatic gene annotation system.* Genome Research, **14**(5): p. 942-950.

Devaney, E., and Howells, RE. 1984. The microfiliaricidal activity of invermectin in vitro and in vivo. Tropenmed Parasitol 35: 47-49.

Diawara L, Traoré MO, Badji A, Bissan Y, Doumbia K, Goita SF, Konaté L, Mounkoro K, Sarr MD, Seck AF, Toé L, Tourée S & Remme JHF (2009) Feasibility of Onchocerciasis Elimination with Ivermectin Treatment in Endemic Foci in Africa: First Evidence from Studies in Mali and Senegal. *PLoS Neglected Tropical Diseases* 3(7): e497.

Druilhe P, Tall A & Sokhna C (2005) Worms can worsen malaria: towards a new means to roll back malaria? *Trends in Parasitology* **21**, 359-362.

Elsik, C.G., et al., , (2006) *Community annotation: Procedures, protocols, and supporting tools.* Genome Research. **16**(11): p. 1329-1333.

Fischer P & Büttner DW (2002) The epidemiology of onchocerciasis and the long term impact of existing control strategies on this infection. In: *The Filaria.* (ed. TR Klei & TV Rajan) Kluwer Academic Publishers, London, pp. 43-57.

Fox AN, Pitts RJ, Robertson HM, Carlson JR, and Zwiebel LJ. (2001) Candidate odorant receptors from the malaria vector mosquito *Anopheles gambiae* and evidence of down-regulation in response to blood feeding. *Proc Natl Acad Sci U S A.* 98: 14693–14697.

Haas, B.J., et al., (2004) *The Gene Ontology (GO) database and informatics resource.* Nucleic Acids Research, **32**: p. D258-D261.

Hougard J-M, Agoua H, Yaméogo L, Akpoboua KLB, Sékétéli A & Dadzie KY (1998) Black fly control: what choices after onchocerciasis? *World Health Forum* **19**, 281-284.

Kanehisa, M., et al., (2004)*The KEGG resource for deciphering the genome.* Nucleic Acids Research, **32**: p. D277-D280.

Koonin, E.V., et al. (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biology, **5**(2): R7. Epub April 16.

Korf, I., (2004) Gene finding in novel genomes. BMC Bioinformatics. **5**: p. 59.

Krueger A (2006) Guide to black flies of the *Simulium damnosum* complex in eastern and southern Africa. *Medical and Veterinary Entomology* **20**, 60-75.

Krueger A, Mustapha M, Kalinga AK, Tambala PAJ, Post RJ & Maegga BTA (2006) Revision of the Ketaketa subcomplex of black flies of the *Simulium damnosum* complex. *Medical and Veterinary Entomology* **20**, 76-92.

Leonhardt, K. G. (1985) A cytological study of species in the Eusimulium aureum group (Diptera: Simuliidae). Can. J. Zool. 63: 2043-2061.

Maegga BTA & Cupp EW (1994) Cytotaxonomy of the *Simulium damnosum* complex and description of new cytotypes in the Tukuyu focus, southwest Tanzania. *Tropical Medicine & Parasitology* **45**, 125-129.

Nelder, M., McCreadie, J.W.,. Coscaron, C. and Brockhouse, C.L 2004. The first report of a trichomycete fungus (Zygomycota: Trichomycetes) inhabiting larvae of *Simulium ochraceum sensu lato* Walker (Diptera: Simuliidae) from the Galapagos Islands, Ecuador. *Journal of Invert. Path.* 87: 39-44

Narita AS & Taylor HR (1993) Blindness in the tropics. *Medical Journal of Australia* **159**, 416-420.

Post RJ, Mustapha M & Krueger A (2007) An inventory of the cytospecies and cytotypes of the *Simulium damnosum* complex (Diptera: Simuliidae). *Tropical Medicine & International Health* 12, 1342-1353.

Procunier, WS. (1982) A cytological study of species in Cnephia s.str. (Diptera: Simuliidae). Can.J.Zool. 60: 2866-2878.

Procunier W, Zhang D, Cupp MS, Miller M, and Cupp EW. 2005. Chromosomal localization of two antihemostatic salivary factors in Simulium vittatum (Diptera: Simuliidae). *J Med Entomol.* 42: 805-811.

Pruitt, K.D., T. Tatusova, and D.R. Maglott. (2005) *NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.* Nucleic Acids Research. **33**: p. D501-D504.

Ribeiro, JM., Valenzuela, JG., Pham, VM., Kleeman, L., Barbian, KD., Favreau, AJ., Eaton, DP, Aoki, V., Hans-Filho, G., Rivitti, EA., and Diaz, LA. (2010) An insight ino the sialotranscriptome of Simulium nigrimanum, a black fly associated with fogo sevagem in South America. 2010. Am. J. Trop. Med. Hyg. 82: 1060-1075.

Robertson HM, Warr CG, Carlson JR. (2003) Molecular evolution of the insect chemoreceptor gene superfamily in *Drosophila melanogaster.* PNAS 100:14537–42.

Rothfels, K.H. (1979) Cytotaxonomy of black flies (Simuliidae). Annu. Rev. Entomol. 24: 507-539.

Rotfhels, K.H. (1981) Cytotaxonomy: principles and their application to some northern species-complexes in Simulium. Pp 19-29. In: Laird (ed). Blackflies: the future for biological methods in integrated control. Academic Press, New York.

Rothfels, K., and Nambiar, R. (1975) The origin of meiotic bridges by chiasma formation in heterozygous inversions in Prosimulium multidentatum (Diptera: Simuliidae). *Chromosoma* 52: 283-292.

Salamov, A.A. and V.V. Solovyev. (2000) *Ab initio gene finding in Drosophila genomic DNA.* Genome Research: **10**(4): p. 516-522.

Sohn, U.I., Rothfels, KH and Straus, NA. (1975) DNA:DNA hybridization studies in black flies. Jounal of Molecular Evolution. 5: 75-85.

Stein, L., (2001) *Genome annotation: From sequence to biology.* Nature Reviews Genetics: **2**(7): p. 493-503.

Stein, L.D., et al., (2002) *The Generic Genome Browser: A building block for a model organism system database.* Genome Research. **12**(10): p. 1599-1610.

Werren, J.H., S. Richards, C.A. Desjardins, O. Niehuis, J. Gadau, J.K. Colbourne, and the *Nasonia* Genome Working Group. (2010) Functional and evolutionary insights from the genomes of three parasitoid *Nasonia* species. Science 327:343-348.

Winnen, M., Plaisier AP, Alley ES, Nagelkerke NJ, van Oortmarssen G, Boatin BA, and Habbema JD. (2002) Can ivermectin mass treatments eliminate onchocerciasis in Africa? Bull World Health Organ. 2002;80(5):384-91.

W.H.O. (2002) *Success in Africa: The Onchocerciasis Control Programme in West Africa 1974-2002.* W.H.O., Geneva.

W.H.O.-T.D.R. (2008) http://www.who.int/tdr/diseases/oncho/files/oncho-poster.pdf

Wood, D.M, and Borkent A (1989): Phylogeny and classification of the Nematocera. pp. 1333–1370 in:. McAlpine J F (ed). Manual of Nearctic Diptera, Vol. 3.

Xu YL, et al. (2009) Large-scale identification of odorant-binding proteins and chemosensory proteins from expressed sequence tags in insects BMC Genomics: 10:632

Xu PX, Zwiebel LJ, Smith DP (2003) Identification of a distinct family of genes encoding atypical odorant-binding proteins in the malaria vector mosquito, *Anopheles gambiae*. Insect Mol Biol 12: 549–560.

Appendix I. Author, Collaborator and Biological Material Supplier Affiliations

Table 3. Author Affiliations

Author	**Affiliation**	**Dates of involvement with *Simulium* genome proposals**
Charles Brockhouse	Biology, Creighton University	2004-present
Alexie Papanicolaou	CSIRO (Canberra, Australia)	2009-present
Rory Post	London School of Hygiene and Tropical Medicine (formerly with Natural History Museum)	2004-present
Daniel Boakye	Noguchi Memorial Institute for Medical Research, University of Ghana	2004-present
E.W. Cupp	Entomology, Auburn University	2010-present
M. Pfrender	Biology, Notre Dame	2010-present
J.K. Colbourn	Center for Genomics and Bioinformatics, Indiana University (Bloomington)	2004-present

Table 4. Collaborator Affiliations

Collaborator	**Affiliation**	**Dates of involvement with *Simulium* genome proposals**
Daniel Lawson	VectorBase-EBI	2009-present
Scott Emrich	VectorBase-Notre Dame	2009-present
Peter H. Adler	Clemson University	2004-present
John McCreadie	Univ. of South Alabama	2004-present
Soochin Cho	Biology, Creighton University	2008-present
Guishin (Gary) Xiao	Creighton Medical Center, Creighton University	2009-present
Eric Haas	Chemistry, Creighton University	2009-present
Sara Lustigman	New York Blood Center	2004-present
Stavros Hamodrakas	Univ. Athens (Greece)	2010-present
Peter Cherbas	Center for Genomics and	2005-present

	Bioinformatics, Indiana University (Bloomington)	
Thomas Grigliatti	Univ. of British Columbia	2004-2007 (previous white paper)
Robert Tanguay	Université Laval (Canada)	2004-2007 (previous white paper)

Table 5. Suppliers of Biological Samples.

Supplier	Affiliation	Species
Daniel Boakye	Noguchi Memorial Institute for Medical Research, Univ. of Ghana	*S. squamosum, S. sirbanum, S. damnosum, S. sanctipauli, S. thyolense, S. yahense.*
Mario Alberto Rodriguez	University of Nuevo Leone, Mexico	*S. ochraceum.* Mexico.
Mark Gardener, Charlotte Causton, Jessica Gaulter	Charles Darwin Foundation, Quito, Ecuador	*S. ochraceum*, "Galapagos", *S. exiguum.*
Elmer Gray, Ray Noblet	Entomology, Univ. Georgia - Athens (USA)	*S. vittatum* colonized
Charles Brockhouse (Secondary, back-up supplier from frozen, archived material)	Biology, Creighton	*S. ochraceum* "Galapagos" *S. squamosum, S. thylolense, S. yahense, S. vittatum*