## ABSTRACT

HARTZOG, MOLLY SUE. Inventing Mosquitoes: Digital Organisms as Rhetorical Boundary Objects in Genetic Pest Management for Dengue and Malaria Control. (Under the direction of Carolyn R. Miller).

Genome databases are widely used tools for facilitating interdisciplinary, inter-institutional, and international communication among scientists conducting research in genetic engineering. Historians and philosophers of science have shown that these databases have become a central driving force in the scientific community, identifying appropriate use of specimens and essentially defining the community standards of experimental biology (Leonelli & Ankeny, 2012). By arguing that model organism databases serve a central role in defining a scientific community, Leonelli and Ankeny, perhaps unintentionally, open genome databases as a site of rhetorical activity, that is, as a site that persuades scientists towards certain actions and certain beliefs about the natural world. In short, it is clear *that* genome databases act as sites of rhetorical invention, but it is unclear *how* they do so. In organizing these databases and generating data and metadata about different species, scientists are essentially debating how to best organize a digital analogy to the natural world to serve as a reliable communication tool and reflect theory, generating scientific thinking. This dissertation brings together these questions regarding the use of digital genome databases and debates about classification to explore the role of these databases in the laboratory as tools for rhetorical invention. In conclusion, I argue that researchers in genetic engineering for dengue and malaria control are taking temporary, stable definitions at the level of species in order to enable questions at genetic level. Mosquitoes are being rhetorically constructed and reconstructed by these researchers, through genome databases, in order to facilitate invention.

# Inventing Mosquitoes: Digital Organisms as Rhetorical Boundary Objects in Genetic Pest Management for Dengue and Malaria Control

by
Molly Sue Hartzog

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Communication, Rhetoric, and Digital Media

Raleigh, North Carolina

2016

APPROVED BY:

| | |
|---|---|
| _____ | _____ |
| Carolyn R. Miller<br>Committee Chair | Huiling Ding |
| _____ | _____ |
| William Kinsella | William Kimler |
| _____ | |
| Fred Gould | |

## DEDICATION

To my family, for their blind support.

To Ruby, for the best kind of inter-species companionship.

## BIOGRAPHY

Molly Hartzog completed her BA in English Literature and Language in her home state at Mississippi State University. She then moved to Raleigh to pursue her MA in English at NC State before joining the Communication, Rhetoric, and Digital Media PhD program and the first cohort of the Integrative Graduate Education and Research Traineeship (IGERT) program in Genetic Engineering and Society. Her interdisciplinary research interests span rhetoric of science, rhetoric of health and medicine, environmental communication, history of biology, and Science, Technology, and Society (STS). Her work has been published in *Environmental Communication: A Journal of Nature and Culture, Science & Technology Studies, Genetic Control of Malaria and Dengue,* and *Proceedings of the Third Iowa State University Summer Symposium on Science Communication*.

## ACKNOWLEDGMENTS

This dissertation is both an argument for and demonstration of the collaborative nature of invention. My acknowledgements could be classified into financial, professional, and personal, but many individuals who helped shape this dissertation provided support in more than one area.

First I must thank the faculty committee for IGERT program at NC State for their generous financial support as well as the travel and networking opportunities that were provided to me through this program. Because of this financial support and professional connections I received through this program, I truly feel like an interdisciplinary scholar, having homes in both the genetic engineering community as well as rhetoric and communication studies.

Perhaps the single most influential person is Carolyn R. Miller. I must credit her for first introducing me to the field of rhetoric of science, for mentoring me as I figured out how to contribute rhetorical scholarship to an interdisciplinary program in genetic engineering, and for giving endless amounts of feedback and guidance, even beyond retirement. I hope she improves at being retired soon.

The rest of my committee has had considerable influence, pushing me into different directions and different areas of scholarship, and giving me many opportunities to practice addressing different disciplines while maintaining my identity as a rhetorical scholar. Thank you, Huiling Ding, Fred Gould, Will Kimler, and Bill Kinsella. Additionally, thank you to Jason Swarts and Stacey Pigg, who have taken a marked interest in my work and helped me to apply this to my teaching.

The reference section of this dissertation does not do justice to the intellectual support that I received from the broader community of rhetoricians. Special thanks goes to Lynda Walsh, Steven Katz, Blake Scott, and all the participants in the "Rhetoric and Science" seminar and "Theory Building in Rhetoric of Health and Medicine" workshop at the 2015 Rhetoric Society of America Research Institute.

Perhaps the largest category of thanks is to my peers at NC State: Tim, Amanda, Sophia, Will, Gabe, Larissa (double thanks for ad hoc Portuguese-English translations), Chelsea, Keon, Emily, Elizabeth, Eli, Hector, Alex, Chris, and Rene (who also gets double thanks for help with coding). And to Ashley Kelly, who has been an unexpected, incredible peer mentor throughout my entire tenure at graduate school, from recruiting me to the IGERT program, to helping me refine my dissertation, and finally to helping me navigate the job market.

Finally, thank you, Zach, for appearing at just the right time.

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

# CHAPTER 1: INTRODUCTION

Collection and categorization have been central practices of the scientific enterprise for many centuries. These activities have resulted in impressive artifacts, such as botanical gardens, seed banks, cabinets of curiosities, museums, menageries, taxidermy collections, and, after the introduction of the computer, genome sequence databases. While it is common to think of only digital databases as information technologies, these other types of collections also serve as information technologies as they assist scientists, doctors, students, and private collectors in furthering their understanding of the natural world. Like other information technologies, they serve as repositories of specimens for research, and, through their organization and metadata, as communication tools for other scientists or students. There is, however, a critical difference between today's genome sequence databases and the other examples listed above, that is, the scientific context in which these artifacts were originally created. Botanical gardens, cabinets of curiosities, and other physical collections were originally developed under the tradition of natural history, which valued identification and ordering for their own sake, with emphasis on the value of common ownership. By the time the first genome sequence databases were developed in the mid to late twentieth century, biology had been operating under an experimental tradition, which values understanding biological mechanisms, not simply collection and categorization. Genome databases merged these two traditions of natural history and experimental biology to best suit so-called "data-driven" science (Strasser, 2008).

Physical collections created by natural historians required long expeditions over great distances and careful transportation and preservation of the collected specimens (as well as

the human crew). Digital genome databases, on the other hand, receive submissions on an hourly basis from laboratories all over the world, drastically increasing in size over just a short period of time. The ease with which data can be collected, analyzed, stored, and retrieved for experimental use has encouraged some scholars to identify the current scientific context as "data-driven" or focused on "big data" (Leonelli, 2012). For many disciplines this is certainly true, especially with the incorporation of computer scientists and biomathematicians in the biology laboratory, who develop programs that enable the research team to analyze a much greater amount of data in a much shorter amount of time than what has traditionally been done (Stevens, 2013).

Like their older physical counterparts, these databases serve as repositories and as resources for scientists conducting research. That is, biologists use data retrieved from the databases in their experiments, and then contribute new data prior to publication of the research. These data include gene sequences and relevant taxonomic information about the organism and metadata on the submission. Scientists, of course, can disagree in significant ways regarding which metadata (i.e. species nomenclature) should be applied to which sequence data, in effect disagreeing on how each species is defined and classified. With the use of these metadata to sort and categorize sequence data that is contributed to the databases, these disagreements are reproduced in the structure of the database itself. These structures of knowledge production have led scholars in science studies to understand databases as centralizing objects of practice within the biology laboratory, defining what counts as biological knowledge (Leonelli & Ankeny, 2012; Stevens, 2013).

While classification and categorization are so often described as “natural” human activities, we know that any system can potentially carry serious social and moral implications, especially when applied to human beings. Bowker and Star (1999) argue that classification systems operate invisibly and are actively kept invisible, thereby operating as powerful social forces. These systems can essentially drive what is highlighted and diminished in a particular community. Bowker and Star (1999) define a classification system as:

> *a spatial, temporal, or spatio-temporal segmentation of the world*. A "classification system" is a set of boxes (metaphorical or literal) into which things can be put to then do some kind of work—bureaucratic or knowledge production. In an abstract, ideal sense, a classification system exhibits the following properties: 1. There are consistent, unique classificatory principles in operation...2. The categories are mutually exclusive...3. The system is complete. (Chapter 1, Section 4, Paragraphs 1-3)

No classification system, they note, will meet all these criteria, but they are the ideal goals of any system.

Classification, particularly in the context of science, can serve useful purposes. A universal, standardized system for categorizing and classifying organisms of the natural world can provide a means for scientists to communicate across spatial and temporal boundaries to reproduce experiments and build scientific knowledge. Such a system was developed by Carl Linnaeus in the eighteenth century and is still used today. In effect, Linnaeus’ categorizations of species, genus, family, order, etc. serve as what Bowker and

Star (1999) call boundary objects, defined as "those objects that both inhabit several communities of practice and satisfy the informational requirements of each of them" (Chapter 1, Section 5, Paragraph 6). These objects enable scientists to accurately point to a particular specimen of interest as well as infer relationships among different specimens based on their proximity to one another in an ordering system.

Defining and classifying species have been long-standing scientific endeavors that have only become more controversial over time, especially so after Darwin. Pre-Darwinian species concepts rested on the idea of species as stable, divinely created entities. Early classification systems were based on perceived complexity of the organism, then ordered based on affinity to a divine entity. This period is perhaps best illustrated by the Great Chain of Being that uses "ladder thinking" to illustrate a logical progression from minerals to plants, animals, humankind, and finally divine figures, with a single deity depicted at the apex (Pietsch, 2012). The transition to species as entities that change over time began during the 18th century, and can be witnessed in the life work of Linnaeus, who converted from a creationist mindset to an evolutionary mindset within his lifetime. His early thinking on species was largely informed by Platonic essentialism, which is the idea that organisms carry an essence (or *eidos*) that is identifiable and classifiable. Essentialism, unlike evolutionary theory, only considers "ideal" organisms and ignores areas where organisms seem to exist on a continuum (Ghiselin, 1969). This way of thinking, already being challenged in Linnaeus' time, had to be dismantled before a theory such as Darwin's natural selection could be accepted into mainstream scientific thought.

The publication of Darwin's *Origin of Species* in the mid-nineteenth century is widely viewed as the climax of the development of evolutionary theory. Darwin was notably uncomfortable with the species concept, as the prevailing concept of the time (species as stable, created entities) did not fit with his evolutionary paradigm (which understood species as inherently unstable and evolving) (Beatty, 1992). Interestingly, even after its wide acceptance, Darwin's work did not overturn the Linnaean classification system, which had been in use for about one hundred years prior. The Linnaean system of classification orders species based on a nested ranking system that grouped species into genus, then the genus into family, then family to order, then class, phylum, and finally kingdom. These groupings were based on physical characteristics of the organisms and did not necessarily reflect genealogical relationships. Debate persists among biologists concerning whether taxonomic systems should be based purely on phylogenetics or on the Linnaean system, which uses morphological methods for classification. Proponents of the phylogenetic system argue that such a system would provide a means of classification that incorporates evolutionary theory and reflect a more accurate understanding of the biological world (de Queiroz & Gauthier, 1994). Defenders of the Linnaean system, on the other hand, argue that evolutionary relationships are always working hypotheses and therefore cannot be a reliable and consistent communication tool (Benton, 2000). While this may, at first blush, appear to be a two-sided issue, it is more of a debate of degree. *How much* should our classification system reflect the natural world? *How stable* should classifications be to function as a communication system? Questions regarding how we define species and understand them in relation to their relatives

cannot be addressed until the community reaches a consensus regarding the central purposes and goals of classification.

This dissertation brings together these questions regarding the use of digital genome databases and debates about classification to explore the role of these databases in the laboratory as persuasive objects and tools for evolutionary thinking. In organizing these databases and generating data and metadata, contributors are debating how to best organize a digital analogy to the natural world in order to serve as a reliable communication tool and generate evolutionary theory. These decisions are not simply an issue of "semantics" or "mere rhetoric." Rather, these activities are a central driving force in the scientific community, identifying appropriate use of specimens and essentially defining the community standards of experimental biology (Leonelli & Ankeny, 2012). By arguing that model organism databases serve a central role in defining a scientific community, Leonelli and Ankeny, perhaps unintentionally, open biological databases as a site of rhetorical activity, that is, as a site that persuades scientists towards certain actions and certain beliefs about the natural world. I adopt a rhetorical approach as I find it to be the most informative theoretical basis for such a study, given that rhetoric traditionally addresses how our symbol-making systems work within specific communities, with special attention to the constraints and consequences of these systems. This exploration will provide a fuller understanding of how species concepts operate in a digital environment to facilitate "data-driven" science.

## Inventing Species: Rhetoric and Taxonomy

### *Classification and Categorization*

There has been a steady scholarly interest in species classification debates (Hull, 1988; Mayr, 1982) as well as in the development of model organisms (Creager, 2001; Endersby, 2007; Kohler, 1994; Rader, 2004), which can be defined as "a specific subgroup of organisms that have been standardized to fit an integrative and comparative mode of research" (Ankeny & Leonelli, 2011, p. 313). These organisms are "standardized" to the point that they are no longer identical to their wild counterparts, but are rather a unique strain bred specifically for laboratory research. At the same time, model organisms are understood to be sufficiently analogous to other biological systems in order to reliably function as models for biological research. The standardization processes of these organisms have contributed to growing genome databases (Leonelli & Ankeny, 2012). Much of the work in science studies on genome databases and computing technologies, more broadly (Stevens, 2013), has focused on the changing social dynamics, labor politics, and value systems in the laboratory as a result of the influx of "big data" and computing technologies. While much of this work has made formidable contributions to science studies as a whole, little is known about the rhetorical impact of such technologies, or how genome databases operate persuasively.

One way to get at this question is to explore how classification and categorization debates are reproduced in genome databases and drive biological knowledge production. Bowker and Star (1999) argue that classification systems, by necessity, highlight certain aspects of reality while diminishing others. In this sense, classification systems act as what

Kenneth Burke (1966) called "terministic screens" that, unavoidably, reflect, select, and deflect different aspects of reality, and, by necessity, shape our understanding of nature. What this means, according to Burke, is that what we often think of as "direct observations" are actually "*implications of the particular terminology in terms of which the observations are made.* In brief, much that we take as observations about 'reality' may be but the spinning out of possibilities implicit in our particular choice of terms" (Burke, 1966, p. 46 emphasis in original). If classification systems operate in this way, as Bowker and Star argue, it is imperative that we understand what is being reflected, selected, and deflected in the classification of digital specimens in genome databases, and how those choices affect biological knowledge production.

This impetus becomes even more critical when the research is focused on controlling any particular species in a natural environment for human health benefits. Most obviously, this type of research has direct consequences to the health and well-being of human populations, as well as implications for environmental health. Understanding transmission cycles of debilitating diseases such as malaria and dengue in order to design a preventative intervention requires a high level of understanding of the pathogen, the transmitting organism, and the effects of the disease on human populations. This breadth and depth of knowledge can only be accomplished through highly collaborative, interdisciplinary, large-scale research that addresses questions concerning the evolution of the pathogen and transmitting organism, reproductive biology of each, genetic mechanisms of transmission, and general biological behavior of each. In short, this requires highly sophisticated evolutionary thinking on the part of interdisciplinary research teams in order to decide the

best method for intervening in the complex transmission cycle in a way that will keep up with the evolution of either the vector or pathogen and eradicate the disease. Scientists must hold a stable, reliable definition of the species they are interested in controlling, but must also keep close watch on the inherently unstable nature of the organism in order to maintain that control.

*Rhetorical Invention*

Classification and definition are rhetorical activities falling in the first of the five canons of rhetoric: invention, arrangement, style, memory, and delivery. The five canons of rhetoric serve as a pedagogical heuristic for teaching students the art of persuasion. The canon of invention has been taught a number of different ways. For much of the twentieth century, writers were trained from an individualist perspective; students were taught to focus internally to find their "inner voice" to develop arguments. LeFevre (1987) calls this tradition the Platonic mode of invention, and heavily critiques this mode as a completely inaccurate understanding of invention. The major theoretical issues she identifies in this Platonic mode are: 1) it favors individualistic studies over studies of writers in social contexts, 2) it "depicts invention as a closed, one-way system," (p. 24), 3) it abstracts the writer from society, 4) it assumes an asocial, isolated self as an inventor, and 5) it does not acknowledge collaboration. These criteria, of course, in no way apply to science, as science is a highly collaborative endeavor. While some scientists may see their work as being fundamentally separate from society, scholars in science studies broadly recognize the scientific enterprise as having its own normative structure that is not entirely immune to society.

While she does not explicitly discuss science, LeFevre (1987) argues that invention is a social activity. She includes this Platonic mode of invention as one extreme on a continuum of perspectives of invention that situate the writer (or scientist) within a community of other writers. This continuum of perspectives

> asks us to look at the inventing writer as part of a community, a socioculture, a sphere of overlapping (and sometimes conflicting) collectives. It draws our attention to social contexts, discourse communities, political aims. It reminds us that writers invent not only in the study but also in the smoke-filled chamber; not only alone but with others with whom they must work; or with whom they choose to think; and not in utter isolation even when they are alone, but by means of inner conversations carried on with internalized others. One invents in part because of others, because one thinks fruitfully in the company of a great many others, who are both possible and real. (p. 93)

One could easily replace LeFevre's argument about writers here with scientists, the smoke-filled chamber with (healthier and more gender inclusive) science cafes and conferences. Scientists, by necessity, work in teams of principal investigators, postdocs, graduate students, and technicians in the laboratory. They produce a great number of collaboratively authored documents including grant proposals, research publications, lab notebooks, and conference presentations. Even when scientists are alone at the bench, they are thinking (whether consciously or unconsciously) of their work as part of a greater community, considering the collectively established norms and values of the scientific method as they complete the tasks in front of them. Once experimental work is complete and written into the research article

format, it is judged through the inherently collaborative peer-review system. As part of the review process (and also in the grant proposal stage), the research is assessed in terms of the current needs of the community as a whole. The scientists, in a sense, "sell" their work by arguing how it will serve the community in furthering its collective goals. It is in this way that scientific communities move through trends where research across many laboratories focuses on a particular set of questions until a general consensus is achieved.

Associating rhetorical invention and science was not always this easy. To begin with, the phrase "rhetoric of science" often seems like an oxymoron to a non-rhetorician's ears (Ceccarelli, 2014). As a field, rhetoric of science has struggled to persuade scientists of the rhetorical nature of science beyond the view of rhetoric as "added" *a posteriori* to knowledge produced in a lab. After work like that of Graves (2005), we know that rhetorical figures and tropes are used in interpreting and understanding data and then communicating them among a particular technical community. She explores how metaphors are used in a physics laboratory to develop theories, analogy is used to interpret the images created in the lab, and metonymy is used to facilitate efficient and effective communication among collaborators as they work towards building theory. These three rhetorical figures have achieved "epistemic status" in the physics laboratory, being used to develop and test hypotheses and construct arguments using the physical data that is produced. Once successful, these figures dissolve and *become* the theory, thus informing the direction of normal science.

Especially when talking about science, it is important to recognize the two senses of "invention" captured by this canon of rhetoric. First, there is the sense of "invention" as the development of something entirely new, something that did not exist before. Second, there is

the sense of "invention" that is synonymous with "discovery" (Miller, 2000). In the typical operations of normal science, scientists seek to make "discoveries" about the natural world. In genetic engineering in particular, these "discoveries" are then exploited in order to achieve a desired effect (e.g. controlling disease) in the natural world. These effects can be achieved through a discrete "invention" (e.g. a genetically modified male insect that is rendered sexually sterile by exposure to radiation), or through an "invented" process (e.g. gene drive techniques). This is where science crosses the boundary between discovery and invention, in their most strict senses, and scientific knowledge/discoveries/inventions become patentable.

Classification and categorization practices are essentially questions of definition: What do we call this organism? What type of organism is it? How is *this* organism not *that* organism? How are they the same? This makes classification and categorization activities of rhetorical invention, as the systematist is essentially *inventing* the species through identifying its boundaries. The practice of definition is one part of stasis theory, which follows four categorical sets of questions called "stases": 1) fact, 2) definition, 3) quality, 4) policy. Questions of fact (e.g. What exists?) are usually addressed and agreed upon before questions of definition (What do we call it?), and so on. Generally speaking, scientific literature deals primarily with the first two stases, fact and definition, with science popularization often moving into the third and fourth stases (Fahnestock & Secor, 1988). Each stasis provides an arena for debate, not telling the speaker *what* to think, but *where* to think (Fahnestock & Secor, 1988).

Classification and categorization do not necessarily remain solely within the second stasis. Rather, in order to classify an organism, one must first discover and obtain the

organism or fossils (stasis of fact), and before one can subscribe to a particular classification system, one must decide the criteria for assessing the value of a classification system (stasis of quality). Exploring the questions surrounding the stasis of definition would provide insight into what remains unstable, and what *topoi* are used in the attempt to stabilize these questions. Sorting out how we define each of these formations will help to inform the third and fourth stasis, which would include questions in the *ought* dimension, that is, how we ought to go about controlling this disease through controlling the mosquito, or whether we should at all.

Once they are stabilized for a specific community and audience, then the stases become *topoi* (or "places") of rhetorical invention that enable normal science to continue (Graham & Herndl, 2011). A speaker (in this case a scientist) assumes a specific audience depending on the *topoi* that are used and the stases that are assumed to already be resolved (Fahnestock & Secor, 1988). Since science is such a community-driven enterprise, assuming a specific audience is simultaneously situating oneself within a community. Addressing a specific audience requires accepting the established stases, taking these stases as assumed, and adopting the acceptable *topoi* that are accepted as persuasive to this particular community. An exploration of the *topoi* used by a specific community, then, provides a way of understanding how this community collectively thinks and generates new knowledge and technologies.

The Aristotelian *topoi* have been a significant pillar in scholarship in rhetorical invention. In 1971, the Committee on the Nature of Rhetorical Invention, part of the National Development Project on Rhetoric, called for rhetoricians to develop a generative theory of

rhetorical invention that would provide a means of explaining how new ideas come into existence (Scott et al., 1971). The Aristotelian *topos* was attractive to many as a starting place for developing this generative theory of invention. However, the *topos* has been a slippery concept that is difficult to pin down. The theoretical undertaking ignited by this 1971 report largely focuses on defining the *topoi*, a process which involved negotiating a number of contrasting dualisms including discovery and invention (in the sense of developing something that has not been thought before), Cartesian and hermeneutic ontologies, the familiar and unfamiliar, generative and managerial understandings of rhetoric, and *topoi* and metaphor. In exploring these contrasting dualisms, I argue that the *topoi* are points of departure for reasoning that both create and reinforce commonly held beliefs, norms, and values in a given rhetorical community. This means that *topoi* are persuasive in a community when they draw on preexisting beliefs, norms, and values, and by departing into new territories of reasoning, generate new beliefs, norms, and/or values for the same community. Most importantly, *topoi* are capable of continually generating discourse that binds rhetors and audiences together to form a community.

In one of the earliest responses to the committee, Karl Wallace defines a system of *topoi* as "an orderly way of searching for meaningful utterances" (Wallace, 1972, p. 395). Despite its simplicity, however, this definition works explicitly for the genome databases discussed in this dissertation. Being a standardized way of organizing all known knowledge on dengue, malaria, and their respective mosquito vectors, the database provides an explicit way of searching for meaningful data and arguments. This definition follows the traditional sense of "invention" in rhetorical studies, that of "coming upon what already exists," despite

the common usage of "invention" in English as "contriving something that never existed before" (Miller, 2000, p. 130).

In this understanding *topoi* operate as folders in a filing cabinet where one goes to select arguments. While this may seem suitable, given the common English translation of *topos* as "place" and the venatic metaphors that are commonly used when discussing activities of rhetorical invention (Miller 2000), this conceptualization of invention points strongly to the Cartesian philosophy of invention that leads to an understanding of "truths" and "facts" to (objectively) exist "out there" (Nothstine, 1988). I would add that in the case of databases, this understanding could render invisible the activities involved in collecting and curating data, reifying the decisions that are made during the process. An opposing philosophy discussed by Nothstine is the hermeneutic ontology, where "place" is understood as one's position, the perspective from which one views a situation: "I understand myself not simply as being, but as *being in a particular set of circumstances, even as those circumstances are understood only by reference to my being within them*" (p. 155, emphasis in original). This understanding of invention is strikingly similar to Burke's "terministic screens," where we understand our world as deflections of the terminology that we use to describe it.

The Cartesian philosophy that construes "place" as a location to search for stock arguments, divorced from the way in which they are argued, creates an interesting division between invention and imagination that has been more-or-less upheld in rhetoric, with the latter being primarily associated with the "creative" and the canon of style. The distinction here would say that "Topics orient us rhetorically in and through time" by operating as

sequential premises, and “metaphors orient us rhetorically in and through space” by evoking images (Leff, 1983, p. 216). Metaphor, which is typically associated with the canon of style rather than the canon of invention, provides a means of associating concepts that would typically be seen as dissimilar, and, Leff argues, would encourage more imaginative thinking when combined with the *topoi.* Likewise, Miller (2000) advocates for a complementary association between decorum and novelty. Associating both decorum and novelty with invention brings attention not only to the substance of arguments, but also of the style, as successful rhetors must "call attention to the *right thing* in the *right way*” (Crosswhite, 2008, p. 174, emphasis added). Additionally, this association creates a “de-radicalization” of novelty that situates it on “the borderland between the familiar and the unfamiliar, the known and the unknown” (Miller, 2000, p. 142). This understanding of novelty removes the strong distinction between invention and discovery. What one searches for is partly what one expects to find, but also something unexpected.

This conceptualization of invention as concerning both substance and style, both familiar and unfamiliar, and both invention and discovery, applies well to invention in the sciences, especially genetic engineering. These scientists are both looking to discover genetic material, to better understand the function and evolution of genes, and also to invent new ways of manipulating genetic material to achieve specific goals. These scientists are concerned with both *what* they find (the substance), but also the *way* they make discoveries, or the *way* they develop new genetic technologies. Graves’ (2005) study in a physics laboratory demonstrates this point well, showing how metaphor, analogy, and metonymy are often first evoked as a short-hand for communicating results among colleagues in a lab.

Later, the tropes and figures that the researchers find most useful for explaining phenomena are adapted into theory and are used to communicate results to the broader scientific community. At this point, style becomes part of the substance. Science can then be critiqued for its methods and results, and also critiqued for the integrity of the tropes and figures used to interpret those results. They are no longer "mere metaphors," but they are part of the substance of the argument.

In order to explore the social activities of invention in the sciences, this dissertation focuses on one research community in genetic engineering for disease control, specifically dengue fever and malaria. This focus is advantageous for several reasons. First, exploring a specific community with common goals provides a defined scope for a rhetorical study. A specific community adheres to common research and writing practices, attends conferences together, participates in the peer review process together, writes and evaluates grant proposals together, exchanges organisms and other laboratory materials, and socializes new scientists under their common norms and values, all while working and residing at different institutions around the world. All of these practices inform what is considered to be appropriate rhetoric within the community, so researching a specific research community enables me to closely scrutinize one community's rhetorical nuances that are more-or-less successful. Second, applying boundaries to the study enables me to question those boundaries. I begin with a community that is clearly defined on an institutional level—having its own centers for research, its own conferences, databases, and journals—and then explore how that community defines itself internally, and how knowledge and materials are exchanged within that community to work towards a commonly-held goal.

Lastly, the specific community that I have chosen to explore provides several advantages in terms of substance. Within this research community of genetic engineering for disease control, I have chosen to focus on *Aedes aegypti,* which transmits dengue fever, and *Anopheles,* a mosquito genus that includes several species that transmit malaria. These two examples provide two highly contrastive case studies. In the case of *Aedes aegypti*, the species has a well-understood, fully sequenced genome and is well established as the primary carrier of the dengue virus. The *Anopheles* genus, on the other hand, includes a number of species that may or may not transmit malaria, may or may not have a fully sequenced genome, may or may not reproduce with one another, and may or may not cohabitate in the same regions of the world. It is a much messier species complex to work with. In fact, some researchers are attracted to the idea of genetically modifying *Aedes aegypti* in order to get a better handle on some of the potential techniques and then apply them to one or more of the *Anopheles* mosquitoes to control malaria. In the sections that follow, I provide an overview of these two case studies, the diseases they transmit, and one database that houses information related to these organisms and is widely used by researchers in this community.

## Case Study Materials

### *Dengue and Aedes aegypti*

Dengue fever is considered one of sixteen neglected tropical diseases (NTD), that, when taken together, hold a higher global disease burden than any one of the "big 3" (malaria, tuberculosis, HIV/AIDS) individually (Hotez, 2010). Dengue is growing in prevalence in tropical and sub-tropical areas of the world, especially in Latin America and

Asia. It is primarily transmitted by the *Aedes aegypti* mosquito, a highly anthropophilic (favoring human blood) and urban-dwelling mosquito. Due to global warming and evolutionary adaptation, the *Aedes aegypti* mosquito is growing in population and spreading into new areas where it was once thought to be eradicated, or had never existed before.

In 2011 the world saw the first release of genetically modified mosquitoes in the Cayman Islands to control the dengue fever virus. Since then, releases have been conducted in Brazil, Malaysia, and Florida. Dengue fever is growing in concern around the world due to the increasing prevalence of the disease and the movement of the mosquito into new areas. Development of a vaccine for dengue has proven difficult given the presence of multiple dengue serotypes (or strains), the need for an inexpensive intervention, and the lack of an appropriate animal model for dengue. The discovery of a fifth serotype will likely slow this development even further (Normile, 2013). As a result, mosquito control has proven to be a more promising direction for dengue control. Additionally, developing transgenic mosquitoes for dengue control is appealing to the broader scientific community because it serves as a model for genetically modifying *Anopheles* mosquitoes, some of which transmit malaria.

In 2007, the genome of the *Aedes aegypti* mosquito was sequenced, revealing the genome to be much larger and more complex than both *Drosophilia melanogaster*, the common fruit fly widely used as a model organism in the laboratory, and *Anopheles gambiae,* one major vector of malaria (Nene et al., 2007). This sequencing project has enabled further research in identifying genes involved in the uptake and transmission of the dengue virus, and thus furthered understanding of possible genetic control techniques.

*Malaria and Anopheles*

Like dengue, malaria is transmitted by the bite of an infected female mosquito. Unlike dengue, which is a viral disease, malaria is caused by parasitic protozoans, specifically *Plasmodium*. Current control measures include bed nets, insect repellent, and controlling mosquito population with insecticides. In addition, some preventative anti-malarial drugs are available for travellers in endemic countries; however, resistance is developing to both insecticides and anti-malarial drugs. As for dengue, vaccine research is ongoing but no effective vaccine yet exists.

The full sequencing of *Anopheles gambiae*, *Plasmodium*, and the human genomes complete the "malaria triad" and enable genetic research at all levels of malaria transmission and infection (Aultman, Gottlieb, Giovanni, & Fauci, 2002). However, the *Anopheles gambiae* mosquito is notoriously difficult to research on the molecular scale due to high genetic variance within the species. While it is currently still classified as one species, it is believed that *Anopheles gambiae* is currently undergoing a speciation event, meaning the species is splitting into two genetically distinct species (Lawniczak, Emrich, Holloway, Regier, et al., 2010). Additionally, while *An. gambiae* is one of the more significant vectors for malaria, it is only one of approximately five hundred species of *Anopheles*, and only one of approximately two dozen known significant vectors of malaria, all of which are *Anopheles* (Besansky et al., 2008). As I will discuss in detail later, the relationships among different known malaria vectors has been a major constraint to this research community.

The wide genetic variance within this species means that researching genetic control options is more difficult than genetic control for dengue, despite its having a smaller and less

complex genome than the *Aedes aegypti* mosquito. On a surface level, in the case of malaria control research, how the species is defined is central to developing genetically modified strains. This affects not only how the species is modified, but also how the modified genes are spread throughout the ecosystem and through the mosquito population as a whole.

*VectorBase: An Database for Invertebrate Vectors of Human Pathogens*

Genome sequence databases are widely used in the scientific community. The most influential database is GenBank, which is supported by the National Center of Biotechnology Information (NCBI). GenBank is an open-access database with data on over 250,000 species as of 2012 (Benson et al., 2013). GenBank shares information globally with the European Molecular Biology Laboratory Nucleotide Sequence Database (EMBL) in Europe and the DNA Data Bank of Japan. Sequence data are generally submitted by individual laboratories. Sharing sequence data through a database such as GenBank is a requirement for publication by many academic journals.


**Figure 1 VectorBase Home Page**

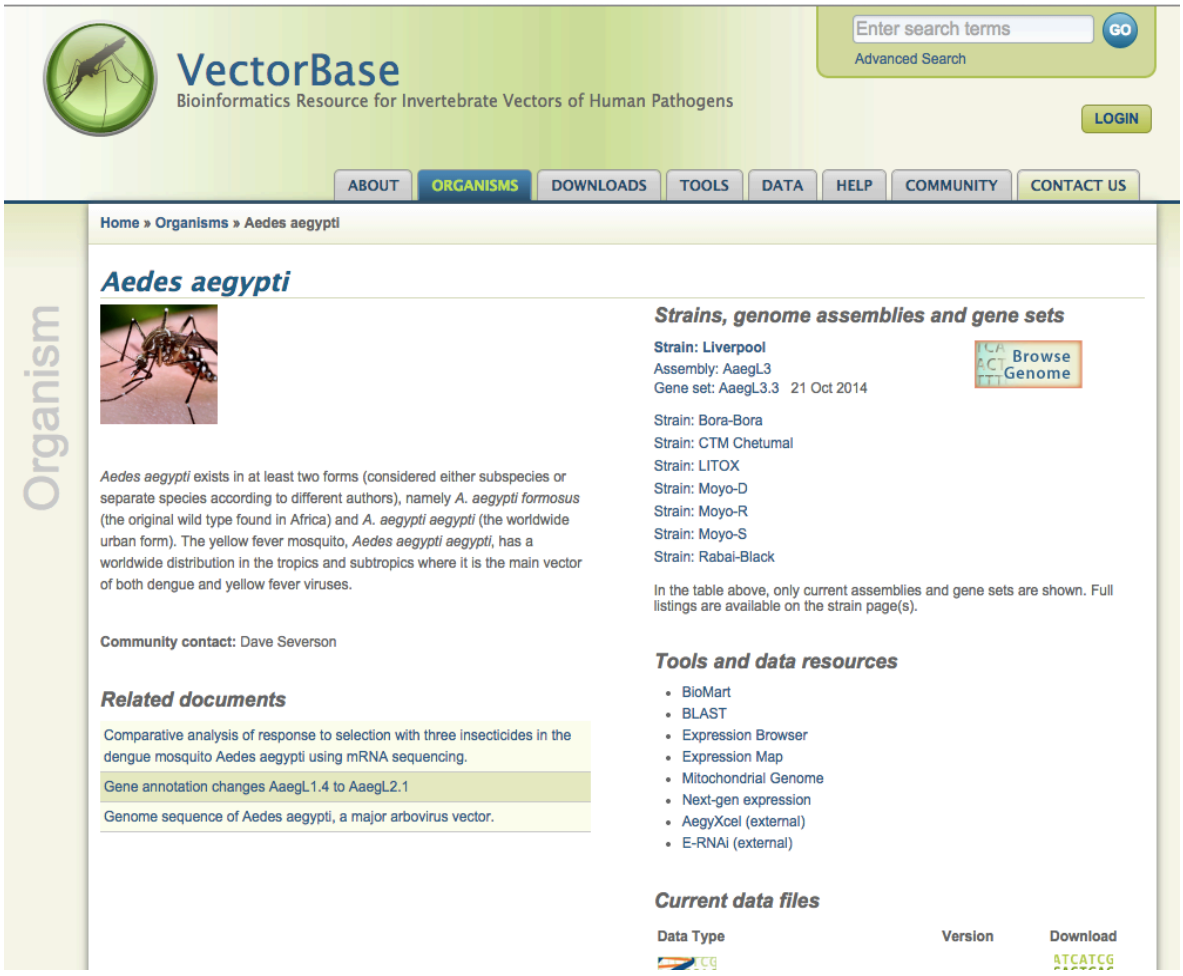



**Figure 2 Aedes Aegypti Organism Page on VectorBase**

There are many genome databases that pull information from larger databases like GenBank and EMBL to create smaller, organism-specific databases. These include Wormbase, Flybase, Zebrafish Information Network (ZFIN), Xenbase (for Xenopus tropicalis and Xenopus laevis, two species of frogs), among many others. One database of specific relevance to this project is VectorBase, which collects sequence data on invertebrate vectors of human pathogens (see Figure 1). This database includes sequence data for 51

different species, including several species of mosquito, tick, fly, louse, snail, and others. All species are vectors for human diseases such as dengue, malaria, Lyme, and others. Users can "browse the genome" of each species or search for specific genes or gene function. Figure 1 shows a screenshot of the *Aedes aegypti* organism page, which is the primary vector for dengue fever.

**Chapter Outline**

This dissertation is concerned with the topical method in two senses: as a method of invention for rhetorical production, and as a method of rhetorical analysis. The former is the business of science, and the latter is the business of rhetoric. I am concerned with both methods in this dissertation, and ultimately show how both are interrelated—the strategies of invention used by this community seem to work *because* they reflect certain beliefs, norms, and values that are widely accepted in the community, and beliefs, norms, and values of a community are reinforced, strengthened, and perhaps modified through continued use.

In order to accomplish this, it is useful at some points to employ the very strategies and methods that I am theorizing. For instance, it is helpful to categorize the general questions about genome databases into three sub-categories: First, those concerning the design; second, those concerning content; and lastly, those concerning use. Chapters two, three, and four address each of these categories in sequence.

In chapter two, I conduct two analyses that explore database design. Together, these analyses show the beliefs, norms, and values that guide the database developers and the types of arguments that are privileged in the database design. This gives an idea about the *intended* use of this genome database. The first analysis explores four reports published by the

VectorBase developers from 2007-2015 in *Nucleic Acids Research.* Tracing the *topoi* that are present in these reports tells us what is valued by the developers, how they are envisioning their target audience, how they situate themselves as part of a community, and how these values have shifted in emphasis over the four published reports. The second analysis of this chapter focuses on two so-called "mosquito ontologies" that serve as an organizational vocabulary for the database itself. This analysis tells us, in greater detail, what assumptions are made by this particular community, how the stasis of definition is stabilized (if at all) for work on mosquito vectors, and how the stases are used as *topoi* for invention.

In chapter three, I explore the rhetorical problems of taxonomy in further detail, focusing on the *Anopheles* mosquito genus. This particular genus of mosquitoes provides an interesting case of a very problematic taxonomic situation. This chapter offers a close reading of papers linked to the VectorBase organism pages for *Anopheles* mosquitoes. One paper, identified as a "white paper," calls for the sequencing of several *Anopheles* mosquitoes related to the transmission of malaria. An additional three papers, linked to VectorBase but also published in *Parasites and Vectors,* describe the distribution of key *Anopheles* populations around the globe. I intend to show how malaria researchers approach this taxonomy and employ alternative rhetorical strategies for justifying research on different malaria-transmitting species.

In chapter four, I take a break from published texts and turn to interviews with practicing scientists in genetic engineering. I explore one community of genome database users, one target audience of VectorBase. In order to understand how genome databases are integrated into the day-to-day workings of one laboratory, I conducted a series of interviews

with graduate students, technicians, and principal investigators in one laboratory at a research university in the United States. While not all are users of VectorBase specifically (although many of them are), these interviewees use one or more genome databases on a regular basis, whether a little or a lot. An analysis of these interviews reveals what users consider to be valuable with genome databases, what they see as constraints within genome databases, and what phases of the research process databases are most helpful to them. While many of these databases they consult are built around one specific set of organisms, many of these researchers, whether graduate student or faculty, consult a portfolio of different databases as a way to fill gaps in information, retrieve different types of data, or to compare data across species. Additionally, while many of these researchers pointed out inconsistencies and gaps in genomic data, many indicated that they rely on this data to give themselves a "starting point" for research.

This sequence of analyses is intended to provide a detailed rhetorical description of how species operate as boundary objects in genetic engineering, facilitating rhetorical invention through their representation in genome sequence databases. The mix of materials I have chosen to include provides a way of looking at this problem through the database design (Chapter 2), through content (Chapter 3), and finally through users of these databases (Chapter 4). This dissertation contributes to ongoing efforts in rhetorical invention by considering how genome databases function as sites that generate invention, borrowing the STS concept of "boundary objects" in order explain how species are conceptualized and exchanged through databases to facilitate rhetorical invention in genetic engineering. In conclusion, I argue that mosquitoes are continually invented and reinvented to serve different

purposes across different research projects. Genome databases provide one medium for this process of invention. Collaboration is both process and product of genome databases, providing a way for researchers to engage inventive practices despite stable definitions of the mosquitoes that are subject of their research.

# CHAPTER 2: BUILDING AN "ONTOLOGY-DRIVEN" DATABASE FOR INVERTEBRATE VECTORS OF HUMAN PATHOGENS

Genome databases can be considered in both senses of invention: as providing familiar, stock arguments (in a quite literal sense), and as providing unfamiliar perspectives to a situation. In the former sense, the data themselves become the *topoi*; in the latter sense, the *topoi* are the organizing principles of a database. In order to develop a generative theory of invention that incorporates databases, I focus on the latter sense in this chapter. This focus on the organizing principles of a genome database provides a way of exploring the database from the perspective of the developers, telling us what needs the developers are attempting to address, and what values they draw upon to inform the overall design. First, I look at a series of reports published in *Nucleic Acids Research* by VectorBase developers. The topical analysis of these reports provides a way of understanding the beliefs, norms, and values that are utilized by the developers themselves. Second, I analyze two ontologies (or standardized languages) that are used to structure VectorBase. This analysis tells us what types of arguments are favored by the database.

As discussed in chapter one, both of these analyses will focus on VectorBase, a genome database designed for invertebrate vectors of human pathogens, or insects that transmit disease to humans. The first analysis uses a series of reports published in *Nucleic Acids Research* that describe development and improvements to VectorBase (Giraldo-Calderon et al., 2015; Lawson et al., 2007, 2009; Megy et al., 2012). Tracing the *topoi* present in VectorBase reports reveals the assumptions the developers hold concerning their target community and how they envision the database to be used. The second analysis

focuses on the Infectious Disease Ontology for Dengue and Malaria (IDODEN and IDOMAL), which provide a structure for VectorBase. Tracing the *topoi* that are present in these disease ontologies provides a means of mapping what types of arguments are more privileged by the database in its very structure.

These analyses provide a point of comparison for the analyses in later chapters that focus on actual users. That is, I am able to discuss here *intended* use, and later I am able to compare that to *actual* use. This chapter also provides a theoretical starting point. By starting at the level of the design, I am able to situate rhetorical invention as a social activity (LeFevre, 1987). Scientific invention is certainly no exception to this rule. Scientific invention responds to exigencies, whether they are offered by funding agencies, past research, anomalous data, critical environmental and medical needs, etc. It connects to past research by filling a gap, raising new questions, reproducing past experiments, etc. Rhetoric can offer explanatory devices for how scientists develop potential responses to pressing exigencies, negotiating what is understood as appropriate and timely within the community, reverberating between data and warrant (Leff, 1983), and connecting novel ideas to past research. These processes are punctuated throughout the entire scientific process, not just in the experimental design phase. Invention occurs anywhere a scientist develops a new or repurposes an old means of persuasion in order to develop "new" scientific knowledge. In other words, "rhetorical *practice* allows reflection and invention to occur" (Farrell, 1991, p. 185 emphasis in original).

In this dissertation, I am seeking a theory of topical invention that incorporates databases as sites of *generative* invention. Databases are commonly thought of as places to

find and retrieve data (and, by extension, arguments). In other words, they are understood in the Cartesian manner, as a place housing "raw data" that exist "out there," or "in the cloud." I would like to complicate this notion to bring databases into the consideration of rhetoric. I hope to find the place where databases are used to not only store information, but to generate new ideas. This would be the place where databases are considered part and parcel of the social activities involved in invention.

**VectorBase Reports**

As I described in chapter one, VectorBase is a genome database developed specifically for invertebrate vectors of human pathogens. It includes sequence data for fifty-one different species, including several species of mosquito, tick, fly, louse, snail, and others. All species are vectors for human diseases such as dengue, malaria, yellow fever, West Nile, Chagas, and Lyme. VectorBase was developed in order to serve a smaller, more specialized community than those of more comprehensive databases like GenBank. The goal of this database is to "provide web-based resources to the scientific community for organisms considered to be causing or transmitting emerging or re-emerging infectious disease" (Lawson et al., 2007, p. D503). VectorBase provides a common platform for community efforts in genome annotations and comparative analysis. Improvements to the database have been documented in reports published in database issues of Nucleic Acids Research (NAR) in 2007, 2009, 2012, and 2015 (Giraldo-Calderon et al., 2015; Lawson et al., 2007, 2009; Megy et al., 2012). These reports are published as part of the "database issue" in NAR, which has been published annually as the first issue of the year for the past twenty-two years. These issues include "brief descriptions of bioinformatics databases," and in 2015 included 176

reports, 56 of which described new databases (Galperin, Rigden, & Fernandez-Suarez, 2015, p. D1). These issues include reports from a number of different databases spanning all areas of interest of *NAR* readership: Chemistry and synthetic biology; Computational biology; Gene regulation, chromatin and epigenetics; Genome integrity, repair and replication; Genomics; Molecular biology; Nucleic acid enzymes; RNA and Structural biology. In later issues, the reports are organized within the issue by the following categories: (i) nucleic acid sequence and structure, transcriptional regulation; (ii) protein sequence and structure, motifs and domains, protein-protein interactions; (iii) metabolic and signalling pathways, metabolites, enzymes, protein modification; (iv) viruses, bacteria, protozoa and fungi; (v) human genome, model organisms, comparative genomics; (vi) genomic variation, diseases and drugs; (vii) plant databases and (viii) other molecular biology databases (Galperin et al., 2015, p. D1). In 2015, the report on VectorBase was categorized under "human genome, model organisms, comparative genomics." These reports are different from what you might find usually published in *Nucleic Acids Research*. They do not follow the IMRAD structure typical of a scientific report. Instead, they offer descriptive reports of technical improvements to the interface and web hosting, updates on the types of data and metadata incorporated into the database, funding sources, collaboration, and outreach, and directions for future development. These reports steadily increase in detail and length with each year, beginning with three pages in 2007 and seven pages in 2015. See  for additional meta-data on each report.

**Table 1 VectorBase Reports**

| Year Published | Authors | Title | Word Count | Page Count |
|---|---|---|---|---|
| 2007 | Lawson, et al. | "VectorBase: a home for invertebrate vectors of human pathogens" | 1301 | 3 |
| 2009 | Lawson, et al. | "VectorBase: a data resource for invertebrate vector genomics" | 2225 | 5 |
| 2012 | Megy, et al. | "VectorBase: improvements to a bioinformatics resource for invertebrate vector genomics" | 2228 | 6 |
| 2015 | Giraldo-Calderon, et al. | "VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases" | 3345 | 7 |

The reported improvements to VectorBase have focused on increasing usability for purposes of facilitating community work in genome annotation and comparative analysis. In this way, these reports show directly how the developers of this database envision their targeted audience. As I discussed in detail in Chapter 1, identifying one's audience is simultaneously situating oneself within a community. Therefore, to explore how the developers understand their audience (in this case, database users), is, to a certain extent, to explore how they are envisioning themselves as members within that community. Identifying the *topoi* employed by these developers, then, will provide an overview of the beliefs, norms, and values that shape this community and drive interpretation and invention (Walsh, 2013).

To accomplish this, I inductively derived a list of expected *topoi* after an initial close reading of these reports, describing what seemed to be recurring special *topoi* throughout the texts, that is, the *topoi* that were specific to this particular rhetorical context. I identified places in each report where I saw these *topoi* operating and defined these *topoi* inductively, through several iterations with a second coder. Portions of the reports omitted from analysis were brief, primarily descriptive, and offered no substantive argument. The vast majority of

the text was included for analysis. After identifying where *topoi* occurred and developing a list of expected topoi, their definitions, and examples, I worked with a second coder to further refine the definitions. See Appendix A for definitions and examples of the *topoi* we identified. The texts were segmented by main idea or end of sentence or paragraph. Segments ranged from 1-4 sentences and were typically under 100 words. The second coder coded a randomized 30% of the segmented text that I identified as employing one or more *topoi.* After working with the second coder for three passes over the data, we were able to achieve 100% negotiated inter-coder reliability. I coded the remainder of the data based on the definitions we developed collaboratively.

The frequencies of the *topoi* I identified in these reports are shown in Table 2. This table shows what was most valued by the writers of the report for each year, as well as how those values have shifted with each report. Considering the reports together, "review," "breadth/scope of data," "integration," and "community" were the most commonly evoked *topoi*. We coded as "review" any mention of the data review process in VectorBase, including annotation and re-annotation processes. For example, "Once an annotation is finalized, *additional analyses are performed* such as our standard orthology/paralogy relationship predictions (6) and *cross-referencing to other resources*. This system was *trialled* for the R. prolixus and G. morsitans genomes." "Review" increased slightly in in the second report (2009), but has decreased in frequency since then.

We coded as "integration" any mention of data being linked, cross-referenced, compared or connected in some way, for example, **"***Integration of these data with existing gene sets* has greatly improved reference gene sets (e.g. An. gambiae) and has led to a new

'patch' build system that uses heuristics to merge manual and automated gene predictions to allow more frequent gene set updates." "Integration" declined in frequency between 2007 and 2012, but then more than doubled in frequency between 2012 and 2015. These results suggest that while the developers may have been initially concerned with reviewing data (in many cases this involved annotation and re-annotation), they have since shifted their focus to the integration of data, which, by my definition, includes "any mention of data being linked, cross-referenced, compared or connected in some way" (see Appendix A for definitions and examples of all coded *topoi*).

We coded as "breadth/scope of data" any mention or *display* of the breadth, scope, or variety of data included in VectorBase, for example, "VectorBase currently hosts *nine genomes* of which the majority are mosquitoes, reflecting their importance in disease agent transmission. The seven corresponding species are *Anopheles gambiae (three genomes, for the PEST, Mali-NIH and Pimperena colonies), Aedes aegypti, Culex quinquefasciatus, Glossina morsitans, Ixodes scapularis, Pediculus humanus and Rhodnius prolixus*." "Breadth/scope of data" remained relatively consistent across all reports, but slightly higher in 2012 than other years.

**Table 2 Rates of Special *Topoi* Occurring in VectorBase Reports**

| | 2007 | 2009 | 2012 | 2015 | Total |
|---|---|---|---|---|---|
| Efficiency | 3% (2) | 3% (3) | 8% (8) | 5% (8) | 5% (21) |
| Collaboration | 6% (4) | 7% (6) | 4% (4) | 4% (6) | 5% (20) |
| Review | 9% (6) | 13% (12) | 7% (7) | 5% (9) | 8% (34) |
| Consistency | 6% (4) | 3% (3) | 7% (7) | 7% (11) | 6% (25) |
| Integration | 12% (8) | 11% (10) | 7% (7) | 17% (28) | 13% (53) |
| Search/Retrieval | 5% (3) | 3% (3) | 5% (5) | 8% (14) | 6% (25) |
| Future Work | 5% (3) | 7% (6) | 7% (7) | 4% (7) | 5% (23) |
| Community | 9% (6) | 11% (10) | 12% (12) | 12% (20) | 11% (48) |
| Past Growth | 3% (2) | 4% (4) | 0% (0) | 13% (21) | 6% (27) |
| Breadth/Scope of Data | 12% (8) | 10% (9) | 16% (16) | 11% (19) | 12% (57) |
| Total *topoi* identified | 66 | 90 | 98 | 168 | 422 |

Since I developed this list of *topoi* inductively, these would be considered "special *topoi*" in the strict Aristotelian sense. The importance of distinguishing these as "special *topoi"* is simply to indicate that these are the *topoi* that are specific to this particular community of VectorBase developers. What is interesting, then, to note here is not only those *topoi* that are used most frequently, but also those that relatively rare. For instance, while references to "community" are particularly frequent, references to "collaboration" have remained relatively rare. We coded as "community" any mention of users, a community of scientists, work being outsourced to a community, etc. "Collaboration," which we defined as

any mention of two entities collaborating or partnering together to improve VectorBase or benefiting from the collaborative affordances of Vectorbase, occurred much less frequently but when it did occur, it typically co-occurred with "community."

The contrast between "collaboration" and "community" is surprising, given that one of the major motivations for building databases is the increased ability to collaborate and share data across space and time. Other values one might connect to the use of databases are also notably absent. These include references to decreased cost of laboratory procedures, accuracy of data, manual vs. automated procedures, and customization. The fact that these *topoi* were only marginally referenced in these reports, while "community" was referenced frequently, suggests a potential point of conflicting values. The developers seem to be concerned primarily with serving a very specific community of users, but are perhaps more concerned with users as consumers rather than as producers of data. Given that "breadth/scope of data" has remained consistently high throughout all reports, and, by my definition, includes "any mention or *display* of the breadth, scope, or variety of data included in VectorBase," it seems that the developers have not indicated that there is a strong need for additional data. Additionally, given that references to "community" have been consistently frequent, the developers must be imagining the contributions of the community in another way than producing more data, since there is no obvious acknowledgment of a *need* for additional data. Then, if we consider that references to "integration" have increased, which involves cross-referencing and linking data for ease of consumption, then the developers seem to be imagining the community as consumers of integrated data rather than producers and contributors of original data. It seems that so-called "big science" may be shifting focus

from churning out more and more data to cleaning up and thoroughly reviewing the data that has already been produced. This would corroborate recent work in science studies (Stevens 2013, Leonelli and Ankeny 2012) that has indicated a shift in science from the “wet bench” to the computer. Moving from the wet bench to the computer and working with “big data” requires highly sophisticated information technologies for organizing and mining massive amounts of data. I turn to one such tool in the following section.

**Mosquito Ontologies**

Ribes and Bowker (2009) define an ontology as "an information technology for representing specialized knowledge in order to facilitate communication across disciplines, share data or enable collaboration. In a nutshell, they describe the sets of entities that make up the world-in-a-computer, and circumscribe the sets of relationships they can have with each other" (p. 199). While not making this argument explicitly, Ribes and Bowker are implying that ontologies create boundary objects, creating entities that facilitate communication and collaboration across boundaries. Observing the development of an ontology in geosciences, Ribes and Bowker describe the strategies used by participants in this group for developing an ontology. They find that in the activity of developing an ontology, participants were required to communicate across domains of expertise to understand the purpose of an ontology. This process, Ribes and Bowker argue, “reconstituted science through the eyes of its data. No longer were data only an individual researchers’ raw materials, rather, they became a community resource. But these data were not yet a community resource until they were interoperated, able to move seamlessly across disciplinary, institutional and technical barriers.” (p. 214).

This reconstitution of science through data can also be seen in the effort to create ontologies for malaria (IDOMAL) and dengue (IDODEN). For instance, the designers of IDOMAL begin an article reporting on its development by describing the transformation in the approach to disease control:

> The failure of the campaign to eradicate malaria about 40 years ago led, among others, to a widespread notion that this disease can simply not be wiped out. This modified the goals of the majority of malaria workers worldwide towards achieving a mitigation of the problem, rather than seeking a final solution. (Topalis et al., 2010, p. 1)

Malaria workers rescaled their goal from total eradication of the disease to more reasonable control measures. Part of what prompted this restructuring of goals was the realization that malaria transmission was far more complex than was originally thought, and required input from many different areas of expertise in order to develop a manageable control strategy. This is where information technologies become useful. Ontologies, the IDOMAL designers note, can be used "as an efficient instrument to enhance the impact of IT tools in vector biology and malaria entomology. This can be achieved by building databases and/or decision support systems driven by wide-ranging ontologies that follow common and established rules" (Topalis et al., 2010, p. 2). In this case, data become integrated, or "interoperated," to use Ribes and Bowker's (2009) term, into a system of support for making decisions on malaria control efforts.

The final step in developing an ontology, according to Ribes and Bowker, is engaging the community for maintenance and use of the database. In this step, participants shift from

questions regarding ontology development to questions about the community of users. They argue that in this step, "'the broader community' became important as part of an outreach project: a community which itself had to be engaged and transformed such that they would use and contribute to ontologies…In order to engage the community in using ontologies and registering their data, members of the community had to understand the value of sharing data, and of ordering them through ontologies" (p. 215)

This emphasis on community that Ribes and Bowker describe is also present in efforts to create an ontology for malaria (IDOMAL) and dengue (IDODEN). The designers of IDOMAL make the case for the use of ontologies explicitly:

> It is apparent that if this kind of data exchange and comprehension by information systems can be achieved, a world-wide malaria eradication campaign would greatly benefit from the adoption of standardized ontologies, which would allow for an extensive data exchange across national boundaries and specific projects. (Topalis et al., 2010, p. 2)

Furthermore, the designers write "The aim was to produce a tool that will be useful to the malaria community working towards effectively reducing the global malaria burden" (Topalis et al., 2010, p. 8). These ontologies re-envision the community that produces this data as a community of users that exchange data. In effect, this ontology re-envisions the "malaria workers" introduced in the opening chapter of the designers' article as the "malaria community" that they discuss in the conclusion. In a sense, through facilitating data integration and exchange, these designers are also designing a community of users.

As mentioned in chapter one, VectorBase's broad target community is researchers of invertebrate vectors of human pathogens. The *Aedes aegypti* mosquito is included for purposes of researching dengue and yellow fever transmission cycles. *Aedes aegypti* data is organized following the standards provided by the Infectious Disease Ontology for Dengue Fever (IDODEN) (Mitraka, Topalis, Dritsou, Dialynas, & Louis, 2015). In addition, as of October 4, 2015, VectorBase houses data on nineteen different species of *Anopheles* for the purpose of researching malaria transmission cycles. Like the data on *Aedes aegypti*, these data are organized following an ontology, in this case the Infectious Disease Ontology for Malaria (IDOMAL).

IDOMAL and IDODEN are formalized languages intended to be readable by both humans and computers, thus improving the overall usability of information technologies (IT) like databases. Improved IT can facilitate data exchange and comprehension, enabling more strategic control programs for complex transmission cycles like those of dengue and malaria. Thus, these ontologies serve as tools of rhetorical invention. By providing a standardized set of terms and relationships, an ontology provides a system for finding and creating new arguments. If an ontology is a tool of rhetorical invention, it is important that we understand what is enabled and constrained by this tool.

Similar to the definition provided by Ribes and Bowker, the designers of IDODEN (who include some of the designers of IDOMAL) state that an ontology consists of "definitions of terms in a given domain, as well as, most importantly, the relations that link these terms to each other. Based on the relationships between terms, the parent-children configuration leads to a tree-like format when an ontology is laid out graphically" (Mitraka et

al., 2015, p. 2). IDODEN includes twelve relations; IDOMAL includes eleven relations. As shown in **Error! Reference source not found.**, only one relation is included in IDOMAL that is not also included in IDODEN; only two relations are included in IDODEN that are not also included in IDOMAL. These relations, given their purpose of creating the links between different terms included in IDODEN and IDOMAL, operate quite literally as *topoi*. As I defined this concept earlier, *topoi* are points of departure for reasoning that both create and reinforce commonly held beliefs, norms, and values in a given rhetorical community. These relations provide points of departure by linking different concepts through an explicit logical structure. The user can then depart down different paths, following these relations/*topoi* into familiar and unfamiliar territory. In a sense, these relations provide warrants for generating an argument that uses the data organized by the ontology. In being constantly evolving entities, these relations/*topoi* are capable of continually generating discourse that binds this community of researchers. Exploring these relations through a rhetorical lens provides a way of looking further at the beliefs that drive this community. The relations/*topoi* that the designers chose to include will, to a certain extent, reflect what they believe to be acceptable warrants within this community. In order for a warrant to be acceptable, it must adequately reflect a particular belief, value, or norm of a community. The following analysis of these relations tells us what types of arguments are favored by these ontologies, and by extension VectorBase. By understanding what is favored, I am able to draw conclusions about what is valued in this community.

In Aristotle's work on the places of invention, he divided his system of *topoi* into two broad categories: *koinoi topoi,* or the "common topics," and the special topics. The former

are topics that apply to all genres and domains of discourse, the latter apply to specialized genres and domains, such as specialized scientific discourse. The common topics are believed to provide "places" to develop arguments or a "method of reasoning" from commonly held beliefs. Another way of thinking of the *topoi* is as warrants connecting data to claim by commonly used reasoning patterns. This system of *topoi* has been widely adapted to studies in STEM discourse (e.g. Miller & Selzer, 1985; Prelli, 1989; Walsh, 2010). Prelli (1989) expands on Aristotle's original *koinoi topoi* by exploring those that are "used again and again in the sciences" (p. 216), essentially designating the *koinoi topoi* of STEM as a set of special *topoi*. Minimizing this distinction is particularly important in such highly interdisciplinary research like disease control, as "specialized" groups often perceive themselves as resorting to "common" language in order to communicate across disciplines. As I discuss above, *topoi* are persuasive because they draw from a community's existing beliefs, norms, and values to justify arguments. Additionally, the persistent use of specific *topoi* help to develop beliefs, norms, and values in a specific community. In other words, there is a reciprocal relationship between *topoi* and the beliefs, norms, and values of a community. Thus, the *topoi* provide an analytical method of understanding a community and how it collectively engages in rhetorical invention.

This relationship between organizational values and texts has been explored by others. Like Miller and Selzer's arguments on special *topoi* in engineering reports (1985), Prelli shows that the *koinoi topoi* of STEM reflect organizational as well as textual relationships. Walsh (building on Prelli, 1989; 2010) added nine additional *koinoi topoi* she found present in science, technology, engineering, and mathematics (STEM) research

articles. Together with Aristotle's original 28, these 37 *koinoi topoi* of STEM research are divided into three (overlapping) reasoning families: causal, dimensional, and comparative (Walsh, 2010). In this analysis, I matched each relation in IDODEN and IDOMAL to its Aristotelian *koino*s *topos* counterpart (see Table 3). A total of thirteen relations are used between the two ontologies. Relations in IDODEN were listed in Mitraka et al. (2015) and relations from IDOMAL I obtained on October 3, 2015 through a simple command + F search for "relationship" in the browser-based ontology at http://anobase.vectorbase.org/idomal/IDOMAL.obo.

**Table 3 Relations in IDOMAL and IDODEN with corresponding *koinoi topoi***

| Relation | Example | *Koinos topos** | Reasoning family* |
|---|---|---|---|
| is_a | “dengue fever” is_a “infectious disease” | 12. parts | causal |
| agent_in | “Aedes albopictus” agent_in “dengue transmission” | 24. cause/effect | causal |
| bearer_of | “Aedes albopictus” bearer_of “dengue virus” | 3. correlation | comparative |
| happens_during | “ascites” happens_during “clinical manifestation of dengue” | 5. time | causal/dimensional |
| has_role | “Aedes albopictus” has_role “infectious agent vector role” | 7. definition and conclusion | causal |
| inheres_in† | “dengue virus seroprevalence” inheres_in “human population” | 9. division | dimensional |
| part_of | “acquired immunity to dengue” part_of “immunity” | 9. division | dimensional |
| participates_in | “dengue C protein” participates_in “dengue virion assembly” | 3. correlation | comparative |
| preceeded_by | “vitellogenic stage” preceeded_by “previtellogenic development” | 5. time | causal/dimensional |
| precedes‡ | “pre-oviposition behavior” precedes “egg laying behavior” | 5. time | causal/dimensional |
| realized_by | “response to visual cue” realized_by “adult vision” | 24. cause/effect | causal |
| realizes | “progression of dengue fever” realizes “dengue shock syndrome” | 24. cause/effect | causal |
| results_in† | “asymptomatic dengue” results_in “convalescence” | 24. cause/effect | causal |

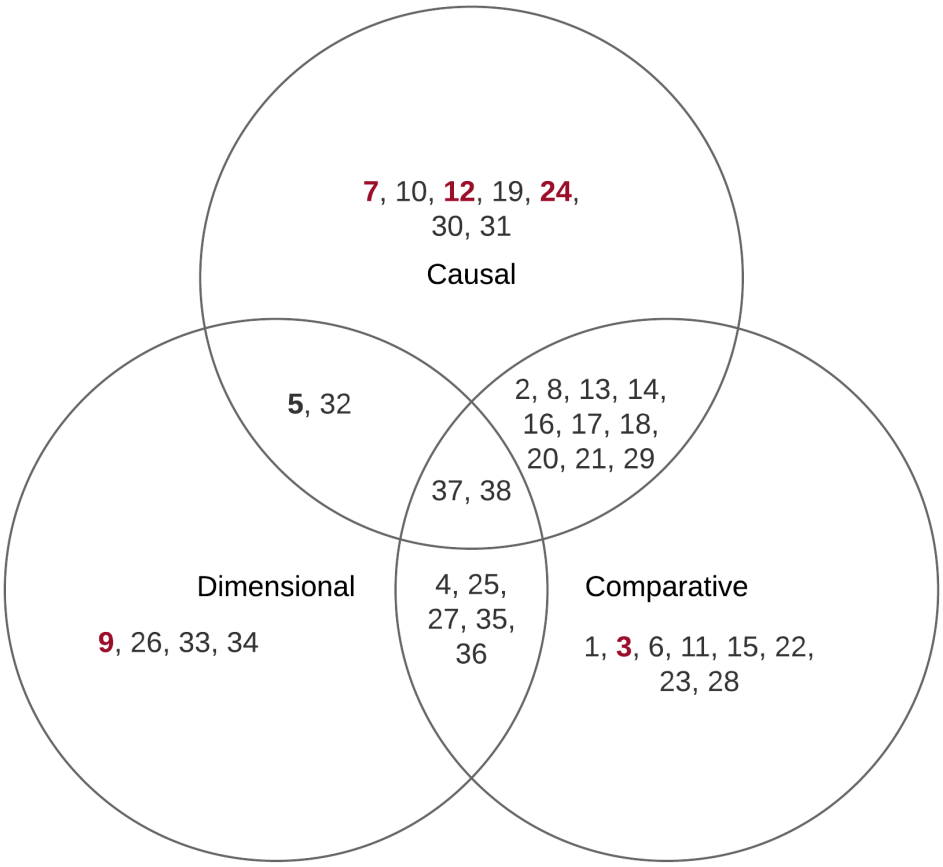
† IDODEN only
‡ IDOMAL only

* Numerical identifiers and reasoning families adapted from Walsh (2010)

All but four of the relations in IDODEN and IDOMAL reside in the "causal" reasoning family, with some overlap in "dimensional" and "comparative" (see Figure 3). This is not surprising given that the primary purpose of these data is to identify (and potentially halt) causes and effects of disease transmission. It is helpful to consider which of these occur most often. Table 4 shows the number of occurrences of each unique *koinos topos*. In order of most to least occurrences, they are cause/effect, time, division, correlation, parts, and definition and conclusion. What this indicates is that these ontologies favor greater use of the *topoi* of causality and time, and less use of parts and definition.

**Table 4 Occurences of Unique *Koinoi Topoi* in IDODEN and IDOMAL**

| *Koinos Topos* | Reasoning Family | Occurrences |
|---|---|---|
| Cause/effect | Causal | 4 |
| Time | Causal/dimensional | 3 |
| Division | Dimensional | 2 |
| Correlation | Comparative | 2 |
| Parts | Causal | 1 |
| Definition and conclusion | Causal | 1 |




1. Opposites
2. Different forms of the same word
3. **Correlatives**
4. Degree
5. **Time**
6. Turning the tables
7. **Definition and conclusions**
8. Semantics
9. **Division**
10. Induction
11. Precedent
12. **Parts**
13. Evaluation by consequence
14. Catch 22
15. Appearance v. reality
16. Consequence by analogy
17. Similar results must have similar causes
18. Ironic choices
19. Purpose is cause
20. Motives to action or inaction
21. Truth is stranger than fiction
22. Compare/contrast
23. Correcting false impressions
24. **Cause and effect**
25. Search for a better plan
26. Question terms of debate
27. Previous mistakes
28. Meaning of names

Additions from Walsh (2010)

29. Corroboration
30. Methods
31. Assumption
32. Scope
33. Future work
34. Significant anomaly
35. Novelty
36. Unification
37. Consensus

**koinoi topoi* in IDOMAL and IDODEN appear in **red**

**Figure 3 *Koinoi Topoi* in IDODEN and IDOMAL**

Walsh (2010) argues that new common STEM *topoi* (#29-37 in Figure 3) indicate a movement in 21st century STEM research towards consensus and collaboration, as opposed to difference and conflict. She argues that this relationship between norms and *topoi* should be viewed both ways: that social norms influence the use of certain *topoi* in texts, and texts influence social norms. It is in this way, she argues, that *topoi* "carry knowledge across disciplinary boundaries" by enabling scientists to communicate findings to different stakeholders (p. 128). This definition of *topoi* strongly resembles the definition of boundary objects. Boundary objects, as defined by Star and Griesemer (1989), are objects that enable collaboration among different social worlds. These are robust enough to maintain a common identity across boundaries, but flexible enough to adapt to local needs. I do not believe it is the *topoi* that carry knowledge across boundaries, but the boundary objects that carry knowledge, and enable a family of *topoi* to emerge around the object. For example, as I will discuss in Chapter 3, the "malaria mosquito" functions as a boundary object among different communities of malaria researchers, enabling special *topoi* such as "vector capacity," or the ability to carry malaria, to emerge and serve as a source of rhetorical invention in disease control research. Without the boundary object of the malaria mosquito, the *topos* of "vector capacity" would not be useful, and perhaps not even exist.

The emphasis on collaboration in Vectorbase, IDOMAL, and IDODEN seems to enable and is enabled by the interdisciplinary nature of dengue and malaria research. Given that there is fewer occurrences and thus less emphasis on the *topoi* of correlation, parts, and definition and conclusion in IDOMAL AND IDODEN, these *topoi* seem to be assumed, stable points of agreement. The *topoi* of cause/effect and time are more frequent in these

ontologies, suggesting that these are the points of disagreement, or at least highly flexible points, providing a space for invention. If science is being pulled towards collaboration and consensus building, the *topoi* of correlation, parts, and definition and conclusion seem to be the stable points of consensus that *allow for* a more detailed, in-depth exploration in cause/effect and time while still enabling collaboration.

While the use of an ontology-driven database can facilitate collaboration and consensus, there may be some loss in deemphasizing the comparative reasoning family, where novel connections can occur. Other work in rhetorical invention suggests that it is through creative use of metaphor (Leff, 1983), one form of comparative thinking, that the most innovative arguments can be made. Prelli (1989) calls this transpositional thinking. He writes: "By bringing X into relation with Y and viewing X from that vantage point, X displays selective features. X is transformed and is given fresh meaning because Y brings to the fore special details and qualities perhaps previously unforeseen. A transposition thus allows new insights by letting unforeseen relationships come into clear view" (Prelli, 1989, p. 66). By governing the specific relationships a scientist is able to use in this database, the database thus limits the option of creating novel comparisons. While this level of standardization may be useful when considering a database as a communication tool, it is less useful when considering a database as a tool of invention.

**Conclusion**

This chapter has focused on the development and structure of one genome database, VectorBase. Here I presented two analyses. The first analysis looks at a set of published reports written by the developers of VectorBase, where they detail the improvements and

updates made to VectorBase from 2007 to 2015. This analysis suggests that developers emphasize the capacity (or potential) of VectorBase to integrate data, and see their intended audience as consumers rather than producers of data. The view of users as consumers rather than producers of data may be a result of the overwhelming amount of data that is undoubtedly housed in VectorBase, shifting the focus from production and collection of additional data to the organization and usability of existing data.

The problem of organization is exactly what drives the development of ontologies, the subject of the second analysis in this chapter. This second analysis considered the Infectious Disease Ontology for Malaria (IDOMAL) and Dengue (IDODEN) that structure part of VectorBase. This analysis reveals the types of arguments that are favored by the database by outlining the lines (or "places," to continue with Aristotle's metaphor) of reasoning that the ontology designers considered to be relevant to the community of users. The *topoi* that were present in the ontologies reside primarily in the causal and dimensional reasoning families. Walsh (2010) argues that this reflects a shift of emphasis in 21st century STEM research to collaboration and consensus rather than dispute and difference. Walsh's observations may be an artifact of an increasing emphasis on collaborative, interdisciplinary research. The first analysis in this chapter seems to support Walsh's argument. Looking at the *topoi* that are present in the developers' reports, we see a clear emphasis on community. However, given the lack of emphasis on collaboration in these reports, I do not believe that 21st century STEM research is necessarily being pushed towards consensus, at least in the case of research on vectors of human pathogens. Collaboration can occur without consensus, and almost certainly doesn't happen without dispute.

The analyses of this chapter suggest that designers of the database and ontologies focus on the integration of data to facilitate causal and dimensional thinking in a community of interdisciplinary researchers. The IDOMAL designers, in particular, envision a "malaria community" as a group of interdisciplinary, collaborative workers that focus on "mitigating" the problem of malaria rather than completely eradicating the disease. "Eradication" proved to be a failure. "Mitigation," according to these designers, requires a more collaborative effort than some "eradication" efforts. In providing pre-established lines of reasoning through VectorBase and its organizing ontologies, these designers create a machine, in a literal sense, for continually generating arguments that are, in theory, widely accepted among this community. However, this machine is dependent on collaboration in order to work, and collaboration is dependent on this machine.

In the next chapters, I explore a significant constraint of invention in dengue and malaria research, that is, species concepts. This constraint is of special significance because research in species concepts, broadly speaking, has been an active area of dispute and division for quite some time. In chapter three, I attempt to explain how collaboration can still occur despite this particularly volatile constraint. In chapter four, I shift to *actual* use of databases like VectorBase in an analysis of interviews with practicing researchers in these communities.

## CHAPTER 3: RHETORICAL PROBLEMS OF *ANOPHELES* PHYLOGENY

Mosquitoes, to the majority of U.S. citizens who come in contact with them, are often thought of as a nuisance—a disruption to backyard summertime barbeques, but (relatively) easy to control with window screens, bug spray, tiki torches, and citronella candles. Of course, in many other areas of the tropical and sub-tropical world, mosquitoes are considered more than a nuisance; they are a threat to one's health and livelihood. Mosquito control efforts, in all areas of the world, have mostly involved pest-deterrent strategies to minimize human-mosquito contact. These include technologies like insecticides and bed nets. However, insecticide resistance in mosquitoes is a growing concern, on top of the known adverse health and environmental effects of insecticide use. Additionally, bed nets are not always successful for reasons of access, human use error, and the presence of daytime-biting mosquitoes. For these reasons, medical experts are turning to alternative pest management techniques, like genetic engineering, to minimize or eliminate the need for pest-deterrent strategies. These techniques, in a nutshell, alter the mosquito on the genetic level to either prohibit the mosquito from carrying and transmitting the target pathogen, or to restrict the mosquito in reproducing, thus decreasing the population of the target species. Reframing mosquito control in this way leads medical researchers to think of the mosquito as a technology of medical intervention, rather than a target for control or elimination. Thinking of the mosquito as something that can be manipulated and exploited to prohibit transmission of disease requires a very different rhetorical framework than thinking of the same organism as something that needs to be eliminated, or at least deterred from human contact. This new approach to pest management requires researchers to develop a more comprehensive

understanding of the life cycle, reproductive cycle, disease transmission cycle, and genetic makeup of specific vectors of the disease. Scientists can then take this more comprehensive knowledge base and identify a place of intervention in order to break the disease transmission cycle. As I discuss in the previous chapter, genome databases are one response to the need for a more comprehensive understanding of these systems in order to *control* rather than *eliminate* these diseases. My analyses in Chapter 2 show how VectorBase developers respond to this exigence and envision their target audience. What I argue in the previous chapter is that the structure of the database, guided by two disease ontologies (IDODEN and IDOMAL for dengue and malaria, respectively), favors collaboration and consensus building, but may be discouraging more imaginative metaphorical thinking.

My purpose in this chapter is to explore how the relationships between different species of mosquitoes, or phylogeny, and the way species are subsequently classified constrain rhetorical invention in malaria. I have chosen to focus on malaria in this chapter because the *Anopheles* genus provides an interesting, complex example of how phylogeny and classification can constrain research. The *Anopheles* genus includes hundreds of different species; approximately two dozen of these are known vectors of malaria. To make matters more complicated, not all of these vectors are closely related. Furthermore, there are so-called "cryptic species" that look identical but are considered different species. Sometimes, the difference between these cryptic species is simply the ability to transmit malaria. These complications in the phylogeny and classification of mosquitoes create at least two rhetorical problems for researchers: First, this significantly increases the burden to justify research into any one of the *Anopheles* species, justifying how they know this species to be relevant to

malaria transmission. Second, this requires the researchers to think about malaria transmitting mosquitoes in a manner other than evolutionary relatedness. Ordinarily, the principle of maximum parsimony dictates that traits only evolve once, meaning that species sharing similar attributes (like the ability to spread malaria, referred to as "vector capacity") are thought to be closely related. This form of thinking is often referred to as "tree thinking" or "evolutionary thinking." However, when looking at vector capacity in mosquitoes, phylogenists would conclude that this trait must have evolved independently more than once, or, conversely, was lost independently more than once, since it is present on many of the branches on the evolutionary tree, but not in every species on the same branch. Figuring out why and how this trait evolves then becomes an important special *topos* in malaria control efforts. Once it is known how this trait evolves, geneticists can then develop a method of intervention to disrupt disease transmission that is applicable to all malaria vectors.

To understand how researchers are working around this problem of phylogeny, I borrow the concept of "boundary objects" from sociology to explore how the malaria mosquito is constructed and re-constructed as different species and species complexes, and how this construction and re-construction impacts thinking in malaria research. In conclusion, I argue that these scientists are using a form of tree thinking that pushes against the assumption that "organisms with similar attributes must be related," and defines the "malaria mosquito" by a specific set of characteristics they believe to play a role in malaria transmission, using *An. gambiae* as an "anchor" (their word), and using "vector capacity" as a special *topos* of invention.

The following section provides an overview of the concept of boundary objects and its application in rhetoric of science, to show how this concept can be adapted for exploring scientific practices of definition and classification, two essential concepts for understanding organisms as species. Next, I give a brief review of literature on "species concepts," which are used to define and delineate species. Following this, I conduct an analysis of technical documents that call for research in specific "dominant vector species" of malaria. Finally, I outline some conclusions on the constraints of the inventive strategies used in this area of genetic engineering and some points of consideration for technical communication and regulation of genetically modified pests.

**Boundary objects, boundary work, and science**

Understanding organisms as species, and understanding a species as belonging to genus, a genus to a family, and so on, are practices of definition and classification. In the context of defining and classifying organisms, these are practices of identifying specific boundaries for one organism, or a set of organisms. It is no surprise, then, that the notion of the boundary object was developed out of a study of species and their function in museum research. The notion of boundary objects was first introduced by Star and Griesemer (1989) in their study of the work of Joseph Grinnell and Annie Alexander in Berkeley's Museum of Vertebrate Zoology. They define boundary objects as

> those scientific objects which both inhabit several intersecting social worlds … *and* satisfy the informational requirements of each of them. Boundary objects are objects which are both plastic enough to adapt to local needs and the constraints of the several parties employing them, yet robust enough to maintain a common identity

> across sites. They are weakly structured in common use, and become strongly structured in individual-site use. These objects may be abstract or concrete. They have different meanings in different social worlds but their structure is common enough to more than one world to make them recognizable, a means of translation. The creation and management of boundary objects is a key process in developing and maintaining coherence across intersecting social worlds. (393)

Star and Griesemer offer this idea of boundary objects as a mechanism to explain how cooperation occurs despite the heterogeneous nature of scientific work. They argue that consensus is not a requirement to achieve cooperation. Boundary objects, instead, provide points of stabilization among different social worlds, enabling members of these different social worlds to cooperate and collaborate. Boundary objects are both adaptable and rigid. They reinforce boundaries between different social worlds in their exchange while also bridging these social worlds together. They both originate within different social worlds but are also exchanged between social worlds and adapt as they are exchanged. According to Star and Griesemer, "In natural history work, boundary objects are produced when sponsors, theorists and amateurs collaborate to produce representations of nature" (408).

In the case of genome databases, Star and Griesemer's work has direct relevance, given that they open informational systems up for rhetorical critique, while not being rhetoricians themselves. To put it in rhetorical terms, they question how informational systems adapt for a given rhetor, audience, and situation, and they ask what would be the consequences for managing and exchanging information in a given system in order to achieve cooperation. Boundary objects, like *topoi*, are tools for generating new ideas. While

*topoi* describe the relations between things, boundary objects are the things themselves; *topoi* are used for identifying, defining, and arranging these objects. For example, when mosquitoes are used as boundary objects within a scientific research community, they may evoke the *topos* of comparison to other dipteran species, or flies. The exchange of a particular boundary object will evoke a plethora of inventional *topoi*. Narrowing focus on a few of the *topoi* that the object evokes helps a rhetor define the features of a boundary object relevant to an exigence. This is a recursive process—through identifying the boundaries of a centralizing boundary object, a rhetor is also refining her exigence.

A second, less cited strategy for managing cooperation offered by Star and Griesemer (1989) is methods standardization. They argue that there is an intimate relationship between the standardized system of collection and the substance of scientific claims:

> Grinnell's managerial decisions about the best way to translate the interests of all these disparate worlds not only shaped the character of the institution he built, but also the content of his scientific claims. His elaborate collection and curation guidelines established a management system in which diverse allies could participate concurrently in the heterogeneous work of building a research museum … There was an intimate connection between the management of scientific work as exemplified by these precise standards of collection, duration and description, and the content of the scientific claims made by Grinnell and others at the museum (p. 392)

Both of these strategies, the use of boundary objects and methods standardization, have direct relevance to the case of genome databases. Databases provide both a medium of exchanging information, and also a systematic method for curating and organizing this information. In a

sense, they operate much like museums in that they provide a management system for the heterogeneous research being conducted on a set of organisms. Genome databases help to shape the character of the research community, albeit spread across many institutions and even nations, and helps to shape the content of scientific claims made by this community.

A concept closely related to boundary objects is boundary work. Boundary work, as it was first developed by Gieryn (1983) is the demarcation of science from non-science; this involves defining what science *is* as well as defining what science *is not*. This concept has been widely used in rhetorical studies to describe the rhetorical work involved in demarcating boundaries in order to accomplish work (e.g. Miller, 2005). Wilson and Herndl (2007) take a different approach to boundary work by considering it alongside the concept of boundary objects in a way that "encourages an integrative rather than a demarcation exigence" (132). What they are advocating here is an expansion of the idea of "boundaries" into "shared social, organizational, and discursive spaces" (131). While Star and Griesemer's original idea of boundary objects also emphasizes how cooperation can occur among different social worlds with different interests, what Wilson and Herndl do for boundary objects is refine what can easily be an overused metaphor by opening the space of the "boundary" for rhetorical investigation. Opening this space brings into question how the object is defined and classified, or what *topoi* are used to position the object in a rhetorical situation. It then becomes the work of the rhetorician not only to identify boundary objects and trace their exchange, but also to question what the object *does* for the conversation in different situations.

While Wilson and Herndl are right to caution against the overuse of a perhaps overly simplified theoretical concept, I find the simplicity of the concept of boundary objects part of its theoretical appeal. The process of defining the plastic features of these objects is part of what gives a rhetorical community its common identity, while the robust features enable collaboration between communities. The boundary object, as a theoretical concept, provides a simple construct for analyzing otherwise highly complex rhetorical activities, such as defining and classifying species. Identifying and describing these objects, in terms of their plastic and robust features, and how they facilitate rhetorical activities will help to understand the nature of these activities and the rhetorical communities in which they take place. In short, "boundary objects" as a theoretical construct is also plastic and robust, giving the rhetorician room to adapt the concept for productive use in a specific analytical situation.

**Species concepts and the taxonomy wars**

There is a wealth of literature in the history and philosophy of science on debates concerning species concepts and taxonomy. Much of this debate (dubbed the "taxonomy wars") has involved demarcation and conflict. In the following section I provide a brief overview of the literature in this area and map out the basic arguments. For the remainder of the chapter, I focus on the cooperative activities that occur in genome databases for mosquito vectors of human pathogens. In other words, I seek to answer a question much like that originally posed by Star and Griesemer (1989): How do cooperation and collaboration occur *despite* the contested nature of defining species boundaries?

For many centuries, classification was established as an exploratory means of observing the natural world. Prior to Darwin's work on evolution, this was easy. Species

were primarily understood to be stable, created entities. There was no need to questions what makes a pigeon a pigeon, because it was understood to have always been a pigeon. This understanding lent itself to ideas like the *scala naturae* that put species on a hierarchy based on their affinity to a singular deity (Pietsch, 2012). The goal of this type of classification was to illustrate harmony and creation (Mayr, 1982). It was based on the idea of essentialism, which argues that organisms have an essence (or *eidos*) that is identifiable and classifiable: the pigeon's "pigeoness."

Darwin's evolutionary theory presented a number of rhetorical problems for science. Under this theory, species are understood to be changing, dynamic entities that do not have clear, well-defined boundaries. However, the need for a stable system of communication requires that researchers agree upon, if only temporarily, a universal definition of specific species. In *Origin,* Darwin problematized the issue of species and thus their classification by illustrating how organisms exist on a continuum rather than as discretely compartmentalized (and created) entities (Ghiselin, 1969). Darwin's novel way of looking at species is a result of his novel method for conducting work in natural history. He borrowed from geologists a way of looking at the earth chronologically, looking at past events as explanatory devices for present phenomena (Ghiselin, 1969), for example, earthquakes forming the mountains of Chile, or volcanoes forming atolls. Darwin's theory of evolution by natural selection looked at the history of a "species" to explain its current development.

Darwin's theory of evolution created a difficult rhetorical problem. In order to persuade others of the mechanisms of evolution by natural selection, Darwin had to use the prevailing species concept of the time (species as stable, created entities). However, his

theory debunked this very notion (Beatty, 1992). According to Beatty (1992), Darwin resolved this dilemma by "formulat[ing] his position in terms of the evolution of what naturalists call 'species' and 'varieties,'… but [Darwin] was also able to communicate and defend a position concerning the undefinability of those terms" (p. 239). Darwin did not undo the species concept; rather, he proposed a process for the development of species. This enabled him to show the lack of distinction between what naturalists call "species" and "varieties." In fact, Darwin did not seem that concerned with a definition of the concept at all, according to Beatty (1992): "Darwin felt that natural history would be liberated by abandoning the search for [a definition of 'species']—liberated in particular from the constraints of nonevolutionary thinking built into pre-Darwinian definitions of the term" (p. 243).

What Darwin added was that these classifications were relations of degree and time, not of parts. In other words, he shifted the conversation to draw from different *koinoi topoi.* Despite his desire to keep the definition of species fluid, Darwin did not believe that species did not exist; to make this claim would prevent one from seeing any evidence for evolution. We need species to understand evolution, but we need them to be fluid in order for the theory to hold true (Winsor, 2013). This idea, though, puts classification in a quandary. When species diversify it becomes difficult to impose boundaries without recognizing a certain degree of artificiality. In fact, it even becomes difficult to discuss what we recognize as "species" (Ghiselin, 1969).

A theory that makes the idea of "species" difficult to even discuss seems to go completely against the goal of species classification to begin with, that is, to provide a system

of information storage and retrieval (Mayr, 1982). Classification should be both a way of communicating with the natural science community and serve as a source of invention--it should tell us something about the natural world (Mayr, 1982). The peculiar part of this history of evolutionary theory is the fact that discussions about how to integrate the theory into taxonomy did not begin until the twentieth century. Current debates on "the species problem" and how to integrate evolutionary theory and taxonomy span a number of issues. Essentialism continues to make an appearance despite being long debunked by Darwin (Hull, 1992; R. A. Wilson, 1999). Some argue over the relevance of the Linnaean ranking system, especially the higher taxa (M. Ereshefsky, 1992, 1999). Others have made attempts to distinguish between classification and phylogeny (Benton, 2000), and between nomenclature and taxonomy (de Queiroz, 2006).

*Phylogenetics and Classification*

The Linnaean classification system and binomial nomenclature continue to be widely used, well after the publication of Darwin's *Origin.* In the twentieth century, scientists began to make attempts to integrate taxonomy with evolutionary theory, beginning with the work of Willi Hennig mid-century. Phylogenetics attempts to construct a tree of life that reflects the actual ancestry of organisms through the principle of maximum parsimony, which states that the most plausible tree will include the fewest number of changes. However, like the system of classification by downward division developed by Aristotle, as outlined by Mayr (1982), a system of classification by phylogenetics requires an entire reworking as new species are discovered and more information is gathered. This leads to an instability in names of taxa because phylogenies are always working hypotheses, not stable categories (Benton, 2000).

Proponents of the Linnaean system of classification point out that this system provides a stable means of categorizing species, fulfilling the utilitarian goal of classification, communication (Benton, 2000), even though it may not represent evolutionary history accurately.

Proponents of a phylogenetic basis for classification emphasize the theoretical goal of classification, that it should not only just describe the natural world, but should also tell us something about its history. The phylogenetic approach advocates a definition of a species as one common ancestor and all its descendants. There are three ways to classify species phylogenetically: node-based, stem-based, and apomorphy-based. All three are monophyletic, meaning stemming from a single common ancestor (de Queiroz & Gauthier, 1990). This approach requires that at the moment a species diverges from an ancestor, but the ancestor continues to exists, the ancestor must then be considered a different species. What phylogenetics leaves out is a consideration of evolutionary methods other than natural selection, for example, genetic drift, hybridization, or genetic changes through parasitism. In addition, depending on the particular characteristics (or apomorphies) that are chosen for a phylogenetic analysis, one could come to different conclusions regarding its phylogenetic history. For example, two species may be similar in early life stages but then diverge as adults; characteristics may change in the natural development of an organism.

The Linnaean system continues to be relevant for the biological sciences because of its stability as a system of communication. Perhaps one of the biggest contributions of Linnaeus is his system of categorizing based on genitalia, a characteristic which does not change over the development of the organism. In addition, the Linnaean system does not

seem to be as incompatible with evolutionary theory as phylogenists seem to believe. The Linnaean hierarchy, while it does not follow the principle of maximum parsimony or monophyly, does create a nested, branching system of organization for species, while remaining flexible enough to allow for the addition of new species as new discoveries are made and new information is gathered. The branching of species (an extension of the tree metaphor in biology) exactly follows Darwin's idea of species as being related, yet different.

My reason for including this brief overview of species concepts and taxonomy is to demonstrate the rhetorical problems for classification presented by evolutionary theory. Under this theory, species are understood to be changing, dynamic entities that do not have clear, well-defined boundaries. As demonstrated by this overview, evolutionary theory creates a number of rhetorical problems in the sciences. The debates seem to emerge from the tension between a theory that suggests species are changing and the need for a stable system of communication. In the following section, I explore one specific case of a complex group of species that demonstrates this tension and the rhetorical constraints that come along with it.

**The Consequences of Classification on Mosquitoes and Malaria**

These debates about classification in biology are not merely debates over our choice in terminology or how we choose to organize species like we organize books in a library. These choices have direct consequences for the way research is conducted. The case of vector borne diseases and pest management can provide apt illustrations of the kinds of consequences these decisions bring to bear. The following analysis is based on a white paper that proposed the sequencing and further research on thirteen specific mosquito species

known to transmit malaria, and three research articles (published as a sequence) that provide reports on the distribution of known vectors of malaria. I perform a close reading of these texts in order to understand the warrants for defining specific species of mosquitoes, and classifying them as "malaria mosquitoes." As I demonstrate below, rhetorical invention in malaria research is constrained by the complex taxonomic status of the species that are known to transmit the disease.

Malaria-transmitting mosquitoes provide an interesting example of evolutionary development. There are approximately 500 known species in the *Anopheles* family. All mosquitoes that transmit malaria are in the *Anopheles* genus, but not all *Anopheles* mosquitoes transmit malaria. To be specific, only ~30 species are currently known to transmit malaria. Furthermore, those species that transmit malaria are not necessarily closely related. The branches on this portion of the evolutionary tree are very deep, meaning that the *Anopheles* mosquitoes split from other species of mosquitoes and flies many million years ago. From an evolutionary perspective, this seems counter-intuitive. Generally speaking, species that share similar attributes are generally assumed to be related. However, this line of thought assumes that traits only evolve once. When looking at vector capacity in mosquitoes, phylogenists would conclude that this trait must have evolved more than once. Figuring out why and how this trait evolves then becomes a central question to malaria control efforts. Once it is known how vector capacity evolves, geneticists can then develop a method of intervention to disrupt disease transmission that is applicable to all malaria vectors.

These constraints of the *Anopheles* family tree serve as the warrant in a white paper that calls for the sequencing and comparative genome analyses of thirteen anopheline

mosquitoes (Besansky et al., 2008). This paper is posted to the *Anopheles* pages at VectorBase, and was compiled by the *Anopheles* Genome Cluster Committee, a group of thirteen scientists from universities at Cambridge, London, Liverpool, Yale, Pennsylvania, and others, chaired by Nora Besansky at Notre Dame. According to an email published in an appendix to the paper, this paper was shared over the VectorBase email list in May 2007 to solicit endorsement from the wider scientific community in order "to assess the size and strength of the community of potential users of these sequence data" proposed in the paper. In just two weeks' time, the authors indicated they received support from over 70 scientists.

The authors divide the thirteen proposed species into three tiers, with decreasing significance to malaria research. Tier 1, those species that are deemed most important for malaria research, includes *An. Arabiensis, An. quadriannulatus, An. merus, and An. epiroticus* (formerly *An. sundaicus* species A), which were chosen because they are considered "the species most closely related to *An. gambiae,*" which is widely considered the most important malaria vector (p. 9). Tier 2 includes seven additional species that represent the "most evolutionary diversity," but still closely related to the species represented in tier 1, with divergence times ranging from 10,000 years ago to 40-50 million years ago. Tier 3 represents species the authors consider to be "outgroups" respective to malaria transmission; these species extend evolutionary divergence up to 100 million years ago. In this proposal, the authors are using a form of tree thinking to make the case for research on these thirteen species—each tier branches out from the previous, with relatedness to *An. gambiae* servingas

the “anchor” for the project (see

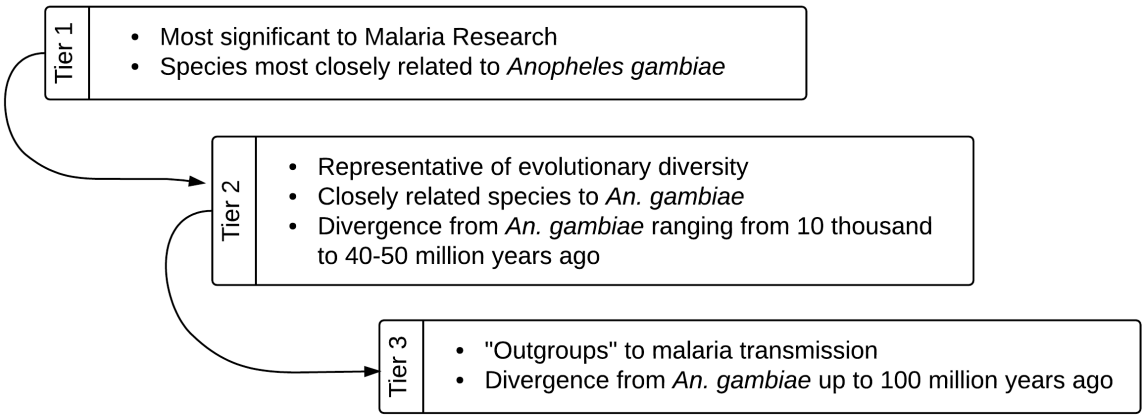



Figure 4 for a diagram of these proposed tiers for research). This particular species is actually referred to as a “species complex” that “comprises seven formally recognized species that vary considerably in vectorial capacity, from the nominal *An. gambiae* considered as the world’s most important malaria vector to its non-vector sibling *An. quadriannulatus*” (p. 2).

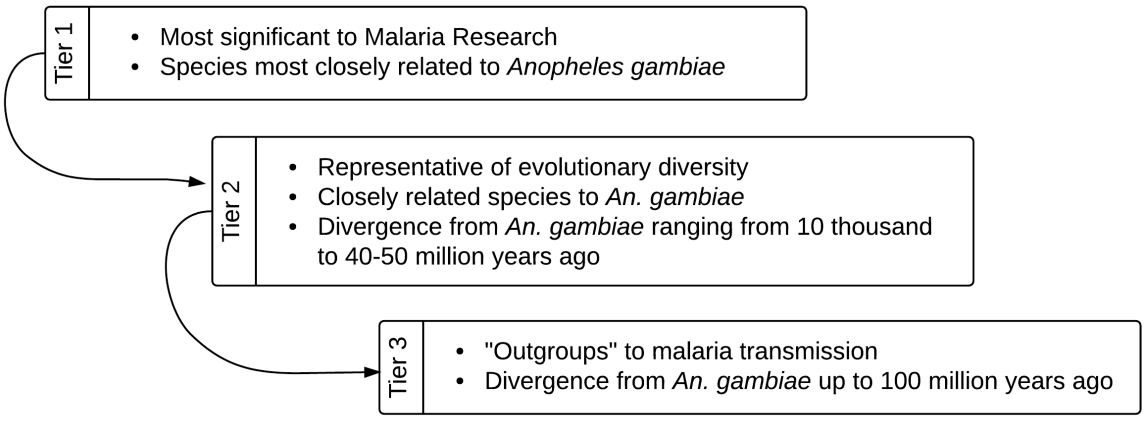



**Figure 4 Tiers for *Anopheles* research proposed by Besansky, et al. (2008)**

Given that malaria-transmitting mosquitoes are not necessarily closely related, vector capacity becomes a central point of concern. The importance of vector capacity in anopheline mosquitoes can be summarized in two important points provided by the authors:

> First, because vectors from different complexes are not closely related, at least some of the underlying vector traits arose independently multiple times in different lineages. Second, the presence of both vector and non-vector species in the same species complex implies either rapid loss or rapid gain of "vector traits". Thus, at least some of the genes that are associated with vectorial capacity—whether involved in immunity, host-preference, or some other physiological or behavioral response—are likely to be rapidly evolving rather than highly conserved over long evolutionary distances. In particular, genes associated with behaviors like preference for human blood meals, selection of anthropogenic breeding sites, or preference to rest inside

> human dwellings, all of which represent 'use of' the human environment, are likely to be very recent evolutionary adaptations that postdate human cultural innovations such as the development of agriculture and animal husbandry that enabled the human populations to reach high and stable (ie. non-nomadic) densities. (p. 3)

While these authors are fighting against the usual line of thought that "species with similar attributes must be related," the underlying assumption in their reasoning for choosing these thirteen species for genomic analysis is that vector capacity must evolve for similar reasons or in a similar manner. This could turn out to be yet another constraint, given that there is clearly no "one size fits all" approach to understanding anopheline mosquitoes. In the paragraph following the above quotation, the authors make this assumption a bit more explicit:

> The traits that impact a mosquito's role in malaria transmission are known, in principle. They include susceptibility of the mosquito to the parasite throughout the entire sporogonic stage, and mosquito population density, longevity, and bloodfeeding behavior. Acquisition of genome assemblies for the mosquito species highlighted in this proposal is critical for understanding the genetic basis for these traits. (p. 3)

The authors here are building to the conclusion that once the evolutionary basis of these traits is understood, we would be able to apply this information across different species, closely related or not. This is quite explicitly stated in the conclusion of the paper:

> The entire community of vector biologists and parasitologists who are seeking novel solutions for controlling malaria will benefit from the availability of additional

> anopheline genome sequence data. *An. gambiae* represents *the* model organism for the study of malaria vector biology and control. Our ultimate goal is to extract from its genome information that will expedite development of new malaria control methods, both chemical and genetic, that will alter vectorial capacity. (p. 10)

To work around the constraint that vector species of malaria are not nicely grouped together on a phylogenetic tree, these authors are proposing to use a different form of tree. The authors are essentially proposing to use *An. gambiae* as the "trunk" and branch out from there using "vector capacity" as the connective thread, rather than common ancestry. While I do not intend to discredit this line of thinking, I do wish to bring attention to the new set of constraints (whether enabling or restricting) that come with this line of thinking. As I mention above, this line of thinking assumes that the trait of vector capacity evolves for the same reasons and in the same way across species, albeit at different times. Should this thinking work well, we could then have a control technique that would be transportable, or applicable to all vectors, thus, theoretically speaking, eliminating the spread of disease. While the broader malaria research community has recognized that *eradication* of the disease may be an unattainable goal, it seems that these researchers are still holding out hope for developing a control strategy that would lead to eradication.

Understanding the important role of vector capacity in malaria control research helps to situate a more recent project that focused on mapping the distribution of known malaria vectors. This project was published as three research papers that focus on malaria vectors in the Americas (Sinka, Rubio-palis, et al., 2010); Africa, Europe, and the Middle East (Sinka,

Bangs, et al., 2010); and the Asia-Pacific region (Sinka et al., 2011). The authors explicitly state the goal of this series of three articles in the final article on the Asia-Pacific region:

> This article concludes a project aimed to *establish the contemporary global distribution of the [dominant vector species] of malaria*. The three articles produced are intended as a detailed reference for scientists continuing research into the aspects of taxonomy, biology and ecology relevant to species-specific vector control. This research is particularly relevant to help unravel the complicated taxonomic status, ecology and epidemiology of the vectors of the Asia-Pacific region. All the occurrence data, predictive maps and EO-shape files generated during the production of these publications will be made available in the public domain. We hope that this will encourage data sharing to improve future iterations of the distribution maps. (Sinka et al., 2011, p. 1, emphasis added).

Several of the *Anopheles* organism pages on the VectorBase website link to this series of articles as directly providing the information reproduced on the page.

In each of these articles, the authors focus on mapping the distribution of known malaria vectors in each of these three regions. Vectors are mapped based on biting habit (inside or outside), feeding habit (human or animal), biting time (day, dusk, night, dawn), pre-feeding resting habit (inside or outside), and post-feeding resting habit (inside or outside), larval sites (light intensity, salinity, turbidity, movement, and vegetation, natural, artificial or manmade. The options listed are not mutually exclusive; for example, *An. albimanus* is both anthropophilic and zoophilic, meaning it feeds on both human and non-human animals. Additionally, some of these species don't differ at all based on these

descriptors; for example, *An. albimanus* shares the exact same characteristics listed as *An. aquasalis*, and is only different from *An. albitarsis* in that *An. albitarsis* has some evidence of endophilic post-feeding resting habit, meaning that the species has been known to rest inside human habitats after feeding.

The authors choose to focus on these characteristics because they encompass the behaviors that they believe affect the disease transmission cycle, and thus provide some insight into "vector capacity." The focus on these specific, definable, behaviors enables scientists to explore how those behaviors can be manipulated or disrupted to break the transmission cycle. What this focus does not consider is the role of humans, non-feeding related habits of the mosquito, or the biology of the malaria pathogen. Given that these reports were published in *Parasites & Vectors*, the intended audience is entomologists and other experts in related fields. The other aspects of the general ecology of malaria transmission would be relegated to experts in other fields such as epidemiology and virology.

Working from the perspective of vector capacity, there are major taxonomic constraints in malaria control. The authors of the mapping projects identify "species sympatry," meaning two species that co-habitate but do not interbreed, as one such constraint. The assumption with species sympatry is that the two species were once a single species and at some point in the past diverged. To help work around this constraint in mapping dominant vector species, the authors propose "an overview of the life history characteristics (bionomics) of vector species pertinent to epidemiology and control" (Sinka, Rubio-palis, et al., 2010, p. 2). Bionomics, the authors explain,

> highlights DVS [dominant vector species] behaviour and life-history characteristics that are relevant for mosquito control, but also clearly indicates *the marked behavioural plasticity* of each species. The influence of human behaviour such as insecticide use, environmental disturbance to a greater or lesser extent, or host activities in the evening and night also drive local variation in species bionomics. Moreover, concerns regarding species identity also add to the uncertainty in categorising species behaviour and thus local, expert knowledge must be consulted when interpreting or acting on the data summarised here. (Sinka et al., 2010, p. 20, emphasis added)

The authors here point to the use of species as boundary objects quite explicitly. The authors are choosing those characteristics that they see as relevant to malaria control for this particular audience of mosquito experts, but indicating flexibility in the transmission-related behavior of each species, thus flexibility in how they are defined by this community. Related to the constraint of species sympatry is the concept of "species complexes" that

> encompass a group of closely related, morphologically indistinguishable species, which may occur in sympatry (but not interbreeding), yet still display behavioural differences that could confound any control efforts that ignore their bionomics and epidemiological importance. Moreover, even amongst those species that are not members of a complex, behavioural differences are common depending upon location, such that a species can be considered a primary vector in one area, but of secondary or no importance elsewhere. The correct identification of any vector

> implicated in malaria transmission is key to successful control. (Sinka et al., 2011, p. 2)

Here we see that these boundary objects remain rigid in their evolutionary relatedness, but flexible in the behavioral characteristics that are related to vector capacity, enabling the authors to demarcate species of interest (those that transmit malaria) and behavioral characteristics of interest (those related to malaria transmission to humans). Given their goal of mapping species distribution based on these characteristics, the authors identify one major constraint that impacts their results. Discussing interventions that limit human-mosquito contact, such as insecticides, repellant, and bed nets, the authors write:

> These interventions are often deployed without a detailed understanding of the distribution, species composition and behaviour of local vectors. This complicates impact monitoring, the appraisal of arguments for more holistic integrated vector control and evaluation of the potential of novel vector control methods. Distribution maps can also be applied to gauge the importance of emerging insecticide resistance among the DVS of Africa. In contrast to Africa, the European and the Middle Eastern region contain areas with low to no malaria transmission. Despite this, the existence of *Anopheles* species with the capacity to transmit malaria is often highlighted as providing the potential for the re-introduction of malaria. (Sinka, Bangs, et al., 2010, p. 2)

Related to this constraint of the impact of current pest control strategies is the constraint of insecticide resistance. This is a major ongoing issue in many pest control initiatives,

mosquito and otherwise. These authors conduct some explicit demarcation boundary work to justify their exclusion of insecticide resistance:

> The bionomics summary of each species is included to accompany the predictive maps as the success of interventions and control methods … in reducing malaria transmission is closely related to the behavioural characteristics of the local DVS [dominant vector species]. This review does not, however, include detailed information relating to insecticide resistance. This was a purposeful omission as it would not be possible to do full justice to this highly dynamic and important aspect of the DVS within the space confines of the current work. Moreover, insecticide resistance is being addressed in detail by other groups, including those at the Liverpool School of Tropical Medicine and the Innovative Vector Control Consortium (IVCC). (Sinka, Bangs, et al., 2010, p. 4)

Additionally, the authors point to another constraint involving human development, which they were also unable to consider in their review:

> Moreover, in an increasingly changing environment, deforestation, the implementation of new irrigation programmes and expanding agricultural development can rapidly alter the composition of the local mosquito fauna, and subsequently influence the control methods required. (Sinka et al., 2011, p. 3)

In sum, the authors of this series of articles addressing the global distribution of dominant vector species of malaria employ strategies of boundary work to demarcate a certain set of behavioral and taxonomic characteristics that they see as relevant to the control of malaria, thus creating an operational (if not comprehensive) definition and classification of the

dominant vector species. This project is an attempt to identify patterns in these plastic aspects of the mosquito behavior that are common among all malaria vectors in order to identify an appropriate intervention strategy to disrupt the transmission cycle. These issues are particularly salient in the authors' analysis of vector species in the Asia-Pacific region:

> Simple, universal species-specific statements regarding the biology of these vectors are nearly impossible due to the locational diversity in behaviour and sympatric distributions of sibling species that contributes to a level of complexity not seen amongst the DVS of other regions. Here we have indicated the *behavioural plasticity* and *locational variation* in species behaviour where possible, and also where known and suspected species complexes exist. However, *until the taxonomic situation is resolved*, the behaviour of many of these DVS will remain unclear. (Sinka et al., 2011, p. 32)

**Conclusion**

This chapter explores the rhetorical form "mosquito" takes in the discourse on pest control research for malaria. Considering the mosquito as a boundary object and investigating the boundary work that scientists engage in to define this object points to some of the inventional strategies that are used in malaria research. As defined by Star and Greisemer (1989), boundary objects are objects that are both robust enough to retain a common identity across different boundaries, and plastic enough to be adapted to local needs. Malaria control research is a highly complex, interdisciplinary area of research, involving parasitology, epidemiology, entomology, genetics, and microbiology, just to name a few. Efficient and effective communication is a necessity. The complex taxonomic status of the

malaria-transmitting mosquitoes presents a great challenge to interdisciplinary communication. These scientists are working around some of these constraints by defining the malaria-transmitting *Anopheles* mosquitoes by the specific set of behaviors related to the blood-feeding cycle and, by extension, malaria transmission. Additionally, they seem to use a form of reasoning that pushes against the assumption that "organisms with similar attributes must be related." These scientists develop an alternative "tree" that uses *Anopheles gambiae* as the "anchor species" (or the trunk) for legitimizing research on other malaria-transmitting species. Other species that are deemed important for malaria research are placed in the outer tiers (or branches) according to their relationship to *Anopheles gambiae,* and/or importance to malaria control.

Because not all vectors of malaria are closely related, and not all closely related mosquitoes transmit malaria, these scientists are required to make the assumption that vector capacity evolved more than once, or that it was lost and reappeared more than once. It then becomes critical to understand *how* and *why* this particular trait evolved in order for scientists to develop a strategy that would be applicable to all species that carry the malaria pathogen. Defining these mosquitoes based upon these specific traits ignores the role of human biology, pathogen biology, and environmental variables. This is not to say that the scientists themselves are ignoring these complexities involved in disease transmission; this definition performs some boundary work to demarcate what these scientists believe will be the most productive area of focus for designing molecular interventions in the transmission cycle.

The complexities around the taxonomic status of the dominant vector species (DVS) of malaria pose significant constraints to rhetorical invention. In the case of malaria research,

the inventional framework provided by evolutionary theory falls short because the species that are known to transmit malaria are not closely related. This constraint led researchers to employ a reasoning strategy that uses the special *topos* of "vector capacity" in place of "evolutionary relatedness" to develop a system for organizing and prioritizing species of interest to malaria researchers. This reasoning strategy is much like tree-thinking in that it establishes a central point (the trunk), and places concepts on branches that radiate outward from this point. Replacing evolutionary relatedness with vector capacity enables the researchers to not only formulate a stable method of communication about dominant vector species of malaria, but also leave room to continue to use evolutionary thinking as a method of rhetorical invention. Additionally, this creates a system of justifying research on a specific species in terms of its vector capacity, meaning that now the "outliers" in the *Anopheles* genus can now be researched as rigorously as others that are more closely related to *Anopheles gambiae.*

To put this back into the terms of boundary objects, evolutionary relatedness remains rigid to create a stable method of communication, and the behaviors related to "vector capacity" become the plastic features that adapt to the local needs of the area or species of study. In this way, malaria researchers are using points of tension between their practical needs and goals as researchers and the theoretical framework of evolution as productive sites of rhetorical invention. The complex taxonomic status of malaria vectors created the exigence for developing an alternative *topos* of invention: vector capacity.

# CHAPTER 4: "IT STARTS BY LOOKING AT THE GENE": THE RHETORICAL CONSTRUCTION AND RE-CONSTRUCTION OF MOSQUITOES

So far in this dissertation I've explored genome databases from the point of view of the developers and taxonomists. Chapter one details the exigence for genome databases in genetic engineering and traces the connections among rhetorical invention, genome databases, and taxonomy. Chapter two looks at the point of view of database developers, specifically VectorBase developers. The two analyses in chapter two show how the developers are envisioning their audience, who are the community of users that research vector-borne diseases, and what kinds of arguments are favored in the structure of the database, which is informed by IDODEN and IDOMAL, the disease ontologies for dengue and malaria, respectively. Chapter three looks at the point of view of taxonomists, who research the evolutionary relationships among different species of *Anopheles*. In this chapter, I look closely at one community of genome database users, one target audience of VectorBase, to understand their motivation for using genome databases, the values they attach to them as a research tool, and how they facilitate rhetorical invention in the laboratory.

I conducted a series of interviews with graduate students, technicians, and principal investigators in one laboratory at a research university in the United States. While not all are users of VectorBase specifically (although many of them are), these interviewees use one or more genome databases on a regular basis, whether a little or a lot. In this chapter, I provide a brief description of the laboratory where I conducted interviews, an overview of the participants and the structure of the interviews, and finally a topical analysis of responses to

five significant questions in the interviews. As I will demonstrate, these particular researchers integrate databases throughout the research process, but with emphasis at the beginning and end, to build groundwork and contextualize their research projects into the research community as a whole. While many of these databases they consult are built around one specific set of organisms, many of these researchers, whether graduate student or faculty, consult a portfolio of different databases as a way to fill gaps in information, retrieve different types of data, or to compare data across species. Additionally, while many of these researchers pointed out inconsistencies and gaps in genomic data, many indicated that they rely on this data to give themselves a "starting point" for research.

This analysis will demonstrate how mosquitoes serve as rhetorical boundary objects, being invented and reinvented across different projects and for different purposes. The species name and vector capacity are robust, unchanging features of these mosquitoes, while individual genes function as flexible features, and provide points of scientific and rhetorical exploration. These robust and flexible features of the mosquito enable these researchers to invent in both senses: discover genes that were not previously known, and create new genetically engineered lines of mosquitoes in order to control disease. The robust features of the mosquito provide a provisional stasis point at the species level, enabling the researchers to explore the organism at the genetic level in a more nuanced way, thus inventing the mosquito from the gene up.

**Study Site**

The laboratory in which I conducted these interviews is part of a biotechnology center in a major research university in the United States. This center occupies four buildings in a

central location on the university's main campus. This center is committed to research, education (both graduate and undergraduate), and outreach at the local, state, and global levels. One of its several acknowledged lines of research is vector-borne diseases, including dengue, malaria, West Nile, and yellow fever. This particular research group includes approximately a dozen faculty and their students from the departments of chemistry, biochemistry, and entomology.

**Interview Structure**

With approved exemption from IRB at North Carolina State University (#6199), I conducted thirteen interviews with members of this lab ranging in duration from thirteen to thirty-five minutes. The participants included seven graduate students at varying levels, two lab technicians including the lab manager, one associate-level professor, one assistant-level professor, and two principal investigators (see Table 5). With the exception of one lab technician who only works with *E. coli*, all participants work with *Anopheles* and/or *Aedes aegypti.* Many of them work with multiple species. Nearly all participants identified VectorBase as their primary genome database, but several indicated that they also use FlyBase, NCBI, BLAST (which is a search tool in the NCBI databases), UniProt, and OrthoDB. UniProt is a protein database and OrthoDB is used to identify orthologs, or genes which are the same across different species, typically indicating shared ancestry.

**Table 5 Interview Participants**

| Interview | Duration | Role of interviewee | Primary species of interest | Primary database of choice |
|---|---|---|---|---|
| 1 | 35:17 | PhD Student | Aedes aegypti | VectorBase |
| 2 | 20:06 | PhD Student | Aedes aegypti and several Anopheles species | VectorBase, FlyBase, OrthoDB |
| 3 | 15:35 | Lab Technician | E. coli | NCBI |
| 4 | 23:29 | Associate Professor | Anopheles gambiae | VectorBase, FlyBase |
| 5 | 29:33 | PhD Student | Anopheles stephensi, Anopheles albimanus | VectorBase |
| 6 | 13:01 | PhD Student | Anopheles stephensi | VectorBase |
| 7 | 22:01 | Principal Investigator | Aedes aegypti | VectorBase, FlyBase |
| 8 | 27:47 | PhD Student | Aedes aegypti | VectorBase, UniProt |
| 9 | 18:62 | First Year Graduate Student | Anopheles (unsure of species) | UniProt |
| 10 | 14:61 | Assistant Professor | Aedes aegypti, Anopheles gambiae | VectorBase, NCBI |
| 11 | 23:23 | Lab Manager | Aedes aegypti | VectorBase |
| 12 | 21:44 | PhD Student | Aedes aegypti | VectorBase |
| 13 | 28:27 | Principal Investigator | Anopheles stephensi, Aedes aegypti | VectorBase, NCBI |

The semi-structured interview protocol involved two parts. In the first half of the interview, I asked participants general questions about the nature of their research and the species that they work with. This often included a discussion of their current research projects relevant to *Aedes aegypti* or *Anopheles,* but at times included discussion of research on other invertebrates. Additionally, I asked them how they think their work in one species compares to working with another species, whether they think it is easier, harder, or different in some specific way. For some, they were able to directly compare their own work in multiple species, but for others, they were only able to speculate on what it would be like to

work with another species. In the second half of the interview, I asked more specific questions about the participant's use of genome databases. Often, they volunteered to show me a database that they use and demonstrate some of the features on a laptop. If they did not volunteer to demonstrate this on their own, I asked them to show me some of the features they use. See Appendix B for the complete interview protocol.

In these interviews, I was able to learn how genome databases are integrated into the research process in this particular laboratory. I found that graduate students and faculty generally had similar practices in terms of what databases they used, why they used them, and when they used them. The answers to many of my questions, despite the question, could be summarized very simply: it depends on the research question. The fact that this came up again and again in responses to my questions, suggests that databases play a significant role in these researchers' invention practices. The following sections explore rhetorical invention at three levels, at the level of the species, at the level of genes, and finally at genome databases.

**Analytical Method**

After conducting these interviews, I noticed clear patterns emerging in all responses. As described above, much of the interview protocol focused on the decisions made by the lab members on the experimental design scale, e.g. comparing work with *Aedes* to *Anopheles,* choosing laboratory strains, choosing genetic components, and choosing genome databases to consult. That said, I noticed the responses increased in complexity and nuance with the level of the researcher. Principal Investigators gave the most nuanced and complex responses, understandably so, as they are the ones more often making these decisions, then training

graduate students and laboratory technicians accordingly. For that reason, I have chosen to focus on these two interviews for an in-depth analysis. Evidence from other interviews are used to corroborate the patterns I found in the PI transcripts, or to note potential exceptions.

Interviews were conducted and voice-recorded by me and transcribed by a professional transcription service. The interview transcripts were first segmented by speaker turn, then longer responses were segmented further by breaks in speech, or the completion of a single thought. Next, I selected segments to be coded; this included any responses that were direct responses to interview questions, and excluded any segments that were spoken by me or were tangential to the interview protocol.

Selected segments were first coded based on four themes addressed in the interview protocol to which they responded: 1) comparing *Aedes* and *Anopheles*, 2) choosing laboratory strains, 3) choosing genetic elements, and 4) choosing and consulting databases. Typically, responses followed this order, as this is the overarching structure of the interview protocol (see Appendix B), but given the semi-structured, conversational nature of the interviews, the responses occasionally presented in a mixed order.

After coding responses based on theme, I then coded the selected segments for one, two, or all three reasoning families used by Walsh (2010): causal, dimensional, and comparative. These are the same reasoning families I used in Chapter 2 to categorize the *topoi* I indentified in the IDOMAL and IDODEN. Given the small amount of data used in this analysis, this simple coding system enabled me to identify clear patterns in each of the two interviews, and to directly compare these patterns to what I discuss in Chapter 2.

I defined each of these reasoning families and provided examples for a second coder. After a second coder and I independently coded one interview, I refined the definitions to clarify some disagreements. I have reproduced the final definitions and examples in Table 6. After this second pass on coding, I generated two reliability scores. Treating each possible code combination as unique (e.g. "Causal and Dimensional" is not a match to "Causal, Dimensional, and Comparative," "Causal and Comparative," "Causal," or "Dimensional"), we reached a simple reliability score of 86%. Treating each of the three codes as a match to any combination that included that code (e.g. "Causal" is a match to "Causal and Comparative," "Causal and Dimensional," and "Causal, Comparative, and Dimensional") our simple reliability reached 96%.

**Table 6 Coding Definitions for Reasoning Families (based on Walsh 2010)**

| Family | Definition | Example |
|---|---|---|
| Causal | Any reasoning that connects one process, concept, or event to another. | "What is the biological **effect** of all those genes that are missing? It's going to **change** its behavior in some way." (Interview 7) |
| Dimensional | Any reasoning that depends upon a continuum of potentials, a procedure, or time. | "**Slightly more** difficult, but it's not rearing, it's just you have to keep it going **all the time**, **constantly**, especially when you have transgenic lines, multiple lines. Can't stop, have to **keep on going**." (Interview 13) |
| Comparative | Any reasoning that depends upon polarities or discrete entities. | "That's just in **the lab strain**, so it's missing a lot of genetic components that presumably **the wild strain** would still have" (Interview 7) |

The following sections present the results of this analysis in three domains: the species, the genes, and the genome databases. This organization follows the basic structure of

the interview protocol (see Appendix B), and allows me to construct a narrative of invention in this lab, taking a top-down approach by tracing boundary objects at the level of the species, to the laboratory strain, to the gene.

Before breaking down the results at each level of invention, it is helpful to see the total frequency of each reasoning family in both interviews with Principal Investigators. Table 7 shows the total occurrences of each reasoning family, whether occurring alone or combined in a single coded segment. The relative frequencies presented in this table show the proportion of occurrences of that family in each discussion area. For example, the causal family appeared ten times in discussions of laboratory strains, or 56% of the total coded segments in discussions of laboratory strains. Overall, the causal family appeared most often across the entirety of both interviews, coded 71 times, or 56% of all coded segments. Dimensional and comparative appeared relatively the same amount, 21% and 23% of coded segments, respectively.

The fact that these researchers rely primarily on causal thinking is not surprising, considering that the main goal of this community, beyond just this lab, is to explore how disease is transmitted by mosquitoes to humans, and how we might intervene to stop this transmission. What is interesting, then, is to explore the intersections of causal thinking with dimensional and comparative thinking. This shows us when causal thinking is simply not enough for these researchers. In the following sections, I will focus primarily on these intersections in each discussion area, with the exception of discussion of comparing species, as this area had no overlaps among reasoning families, likely because this area of conversation was brief.

**Table 7 Relative Frequencies of Coded Segments Showing Reasoning Families in Each Discussion Area**

| | Species | Lab Strain | Genes | Databases | Total coded segments in each reasoning family |
|---|---|---|---|---|---|
| Causal | 2 (67%) | 10 (56%) | 18 (56%) | 41 (56%) | 71 (56%) |
| Dimensional | 1 (33%) | 1 (6%) | 7 (22%) | 17 (23%) | 26 (21%) |
| Comparative | 0 (0%) | 7 (39%) | 7 (22%) | 15 (21%) | 29 (23%) |
| Total coded segments in each discussion area | 3 (100%) | 18 (100%) | 32 (100%) | 73 (100%) | 126 (100%) |

## Inventing the Species

### *Comparing Aedes aegypti and Anopheles in the Lab*

In Chapter 3, I discussed the complex nature of the *Anopheles* taxonomy and the constraints that this imposes on malaria research and vector control. *Aedes aegypti,* the primary vector for dengue fever, has a much simpler taxonomy than *Anopheles*. It is a relatively well-understood species with a fully sequenced genome and well defined breeding habits. Unlike *Anopheles,* it is not thought of as currently undergoing any speciation events, nor are there any known cryptic species or disease-transmitting sibling species. Given these stark differences between the two, I asked interviewees if they considered one to be easier to work with than the other. Interviewees who were able to answer based their choice either on the ease of rearing and physically working with the mosquito, embryos, or larvae, or on the complexity of the genome of that particular species. The taxonomic status of any species did not arise as a concern for these researchers. Their answers were generally based on what each

individual researcher needed to accomplish in the lab. For instance, one graduate student discussed the relative ease of injecting embryos in different species:

> Graduate Student: I have always heard, in maybe a couple of weeks that I have working with [*Anopheles*] *stephensi*, that *Aedes* [*aegypti* embryos] are much hardier.
>
> Me: Hardier?
>
> Graduate Student: Hardier as in, so I do a lot of embryonic injections, now this might just be an indifference. As a lab we work with *Aedes* a lot. I was fortunate enough to be trained by someone who was very good at *Aedes* injections, embryonic injections. I've heard that [*Aedes*] are much more inclined to be tampered with as a species. I'm not sure exactly why that is. (Interview 1)

One Principal Investigator, thinking more globally in terms of the laboratory, also argued that *Aedes aegypti* is easier to work with on the basis of maintaining a constant supply of living mosquitoes and embryos:

> Principal Investigator: *Aedes aegypti* is probably slightly easier because you don't have to keep the colony going all the time because you can let egg dry, and store it for a few months. That's really a big advantage there. (Interview 13)

Those researchers who worked directly with the genome, i.e. computational work, generally stated that *Anopheles* were easier to work with than *Aedes aegypti.* They indicated that the *Aedes aegypti* genome was several times larger than any *Anopheles* genome and had many repeating sequences, making it more difficult, requiring more time to do any computational work:

> Graduate Student: The *Aedes aegypti* is five times larger because it has so many repeats. In genomics, it's much more computationally intensive. (Interview 2)

These researchers seem to value the mosquito for both its material constraints (the ability to sustain a colony, inject embryos, etc) and its informational constraints (the genome sequence) simultaneously. Whether they thought of one as easier to work with than the other depended on the nature of their work, whether it was based primarily on wet bench work like embryonic injections, or on computational work like building phylogenetic trees.

It is significant to note here what is missing from these comparisons of the *Aedes aegypti* mosquito and *Anopheles* mosquitoes: any discussion of taxonomic status, relatedness to other species of mosquitoes, or vector capacity. These issues would require a much broader view of mosquitoes and the diseases they transmit. Many of these interviewees indicated that this was the ultimate purpose of their research, but when asked further about working with these mosquitoes, they were more attuned to the nitty-gritty of day-to-day work. The notable absence of discussion about vector capacity, taxonomic status of these mosquitoes, or relatedness to other species of mosquitoes (which would employ dimensional and comparative reasoning in addition to causal) suggests that these issues are assumed by these researchers. If we are considering mosquitoes as boundary objects, these features are the robust characteristics. The flexible characteristics of the mosquito can be better understood looking at the choices that are made regarding a laboratory strain.

*Choosing laboratory strains*

As I discuss in Chapter 1, organisms used for laboratory research are quite different from their wild counterparts. Researchers have a choice of what they choose to use in their lab. For

nearly every species used in laboratory research, there are a number of different sub-types, or "strains," that are developed and distributed by laboratory supply companies. Each strain has different advantages and disadvantages.

The significance of this can be further understood with some historical context. One of the first model organisms, and perhaps the most widely used model organism today, *Drosophila melanogaster,* the common fruit fly, was originally brought into the lab because its habits and seasonal cycle matched the needs of academic life—it could quickly be baited, trapped, bred, and easily sustained through a semester. Thomas Hunt Morgan, who first used the fly for experimental research, chose the species because it was *not* domesticated; he felt research should be done with wild animals. However, through Morgan's breeding practices and his identification of several mutations, *Drosophila* eventually became a "biological breeder reactor" where "[t]he more mutants turned up, the more crosses had to be done to work them up. The more crosses were done, the more mutants turned up. The process was autocatalytic, a chain reaction" (Kohler, 1994, p. 47). It was through the Morgan Laboratory and the "Fly Boys'" (and some of their wives') research on the fruit fly that the genetic map was invented, and the laboratory-bred mutant *Drosophila* became a cornerstone of genetic research (Kohler, 1994).

One of the advantages of using a laboratory-bred strain of a specific organism is its standardization. As one interviewee, a principal investigator, explained to me, "To be able to do the type of work [we do] you need to have a consensus genome that everybody has access to, so that people are talking about the same things" (Interview 7). *Drosophila*, through hundreds (if not thousands) of rounds of breeding and inbreeding, was made into a

specialized organism that suited the needs of the Morgan laboratory. The same has been done for many other species, including mosquitoes. For *Aedes aegypti* and several *Anopheles* species there are several strains that are bred and sold for laboratory research.

Given the cost and infrastructure that is required to rear and maintain a colony of mosquitoes, the Principal Investigators make the ultimate decision about what strains they will keep alive in their lab. Graduate students have some choice based on the nature of their thesis projects, but many early career students as well as technicians were not able to speak much about the advantages and disadvantages of certain strains. Some could not even recall the type of strains the lab keeps.

Despite whether the interviewee had much of a choice in the lab strains they use, the response was similar across the board: the strain you use depends on what your research question is, or what you need to do. In short, it is a rhetorical choice, in the classic Aristotelian sense of finding the best available means. For example, one Associate Professor explained:

> One consideration is, what research questions we can answer with this strain. For example, if you want to study hybrid sterility in the *Anopheles gambiae* complex, [then] we want to choose [a] species that [is] possible to breed, to grow, and to get the next generation. Then we start a next generation of these mosquitoes to see what characters are affecting, for example, male sterility. We can look at that type. Another consideration is how easy to keep the colony. For example, if you have a choice of several colonies from the same species, we would prefer that colony that we can

> easily keep in our lab because some strains are much more difficult to keep. (Interview 4)

As indicated by this researcher, theoretical questions like "What kind of strain do I need to answer my research question?" are folded together with questions of accessibility of a strain, and practical concerns like rearing and maintaining a colony of mosquitoes. The graduate students I interviewed seemed more concerned with accessibility, but their responses still generally addressed suitability to their research question as well. For example, one graduate student stated,

> It's more like what we can get access to the fastest. If we have it in the lab, that's what we use. (Interview 2)

Of course, what is available in the lab is a decision that is made primarily by the Principal Investigators. The following section presents the results of a close topical analysis of the two interviews with Principal Investigators. Table 8 shows the co-occurrences of reasoning families in discussions of choosing laboratory strains. The relative frequencies presented are relative to the total number of coded segments in discussions of laboratory strains. For example, causal and comparative thinking co-occurred 5 times, or 42% of all coded segments in discussions of laboratory strains. Figure 5 shows the relative proportions of occurrences and co-occurrences of reasoning families in the discussions of laboratory strains with the two Principal Investigators.

Looking closely at these intersections of causal and comparative thinking, these co-occurrences focus primarily on comparing laboratory strains to wild mosquitoes, and understanding the nature and effect of the differences between the two. Before comparing the

laboratory strain to the wild strain, these PIs determine what is available, widely used, and, by extension, deemed credible by the scientific community as a whole. As one PI explained,

> To be able to do the type of work [we do] you need to have a consensus genome that *everybody has access to* [causal reasoning], so that *people are talking about the same things*. [comparative reasoning] (Interview 7)

For this PI, while accessibility is his primary justification for choosing a strain, this is not necessarily his ideal justification. Later, he moves to the comparative reasoning family, describing his ideal justification for choosing a laboratory strain:

> PI: If there was a better quality genome or a wild strain we would much rather use the wild strain. But we're stuck because *we need to have access to the genome* [causal reasoning]. It'll speak the same language as everyone else who's working on it
>
> …
>
> Me: What makes you say you would rather use [a wild strain]?
>
> PI: … [The laboratory strain is] *missing a lot of genetic components that presumably the wild strain would still have* [comparative reasoning]. *What is the biological effect of all those genes that are missing? It's going to change its behavior in some way*. [causal reasoning] Obviously, those mutations are not detrimental to the survivor in a survival of the organism in a laboratory, but they probably do change its physiology to some respect. So I would rather work with a strain that does not have so many deleterious mutations in it. Those are things that have been fixed over time. There's a lot of other mutations that are present at lower levels that cause a lot of problems in

> these laboratory strains. The wild strains would be much more fit, *much more representative of the biology of this organism* [comparative reasoning]. (Interview 7)

Why does this PI resort to using what is accessible (causal reasoning), if his ideal preference would be one that is more representative of the organism (comparative reasoning)? This suggests that the need to communicate and exchange data across boundaries trumps the desire to narrow the gap between the wild organism and the laboratory strains, or communication outweighs scientific accuracy. This point alone demonstrates not only the value of communication, but also the rhetorical awareness these scientists bring to their work. I would not say that communication is valued *at the expense of* scientific accuracy, but rather that scientific accuracy is improved by this nuanced rhetorical awareness. In addition, this shows how mosquitoes are functioning as rhetorically constructed boundary objects. They are constructed and reconstructed across different laboratories and across different research projects, adapted for specific purposes. It is not only new scientific knowledge that is being invented in this laboratory, but the mosquito itself is being invented and reinvented as an informational resource in digital form to help these scientists respond to specific research exigencies. This process helps these researchers discover new information about the disease-transmitting mosquitoes, and also helps them to develop new genetic constructs that would potentially be used to develop a genetically modified mosquito. The following section turns to the latter type of invention.

**Table 8 Relative Frequencies of Reasoning Family Co-Occurrences in Lab Strain Discussion**

| | Dimensional | Comparative |
|---|---|---|
| Causal | 1 (8%) | 5 (42%) |
| Dimensional | | 0 (0%) |

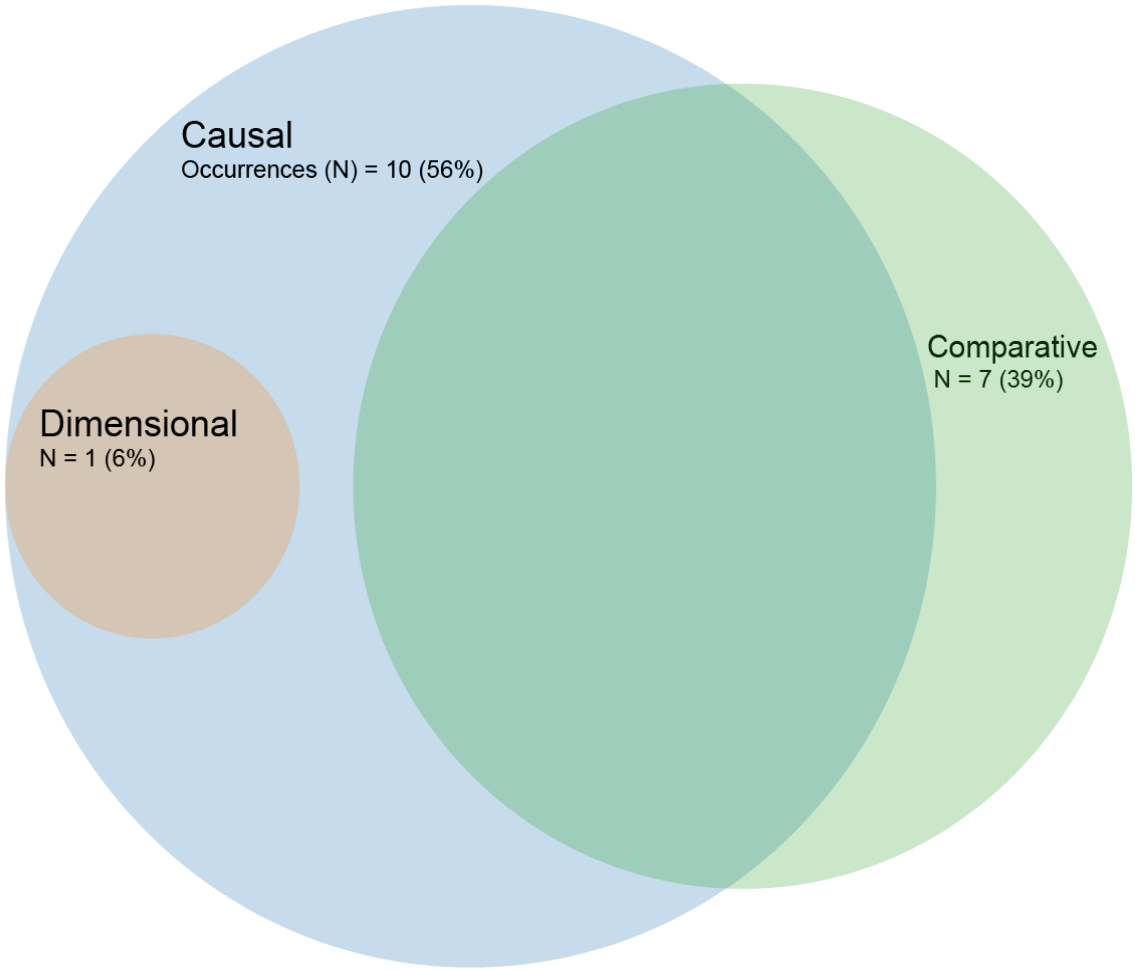



**Figure 5 Relative Proportion of Reasoning Families in Discussion of Laboratory Strains with Principal Investigators.** Frequencies indicate totals for each reasoning family (both single occurrences and co-occurrences).

**Inventing Genes: Choosing Genetic Components**

Building genetic constructs is both a means and an end for researchers in biotechnology-related fields. These can be used as a tool to better understand some aspect of the natural genetic makeup of an organism, or developed and patented into a new technology that can be introduced to the market. Constructs built for either purpose are patentable; constructs that are initially built as a laboratory tool may eventually be adapted for a marketable application. In the case of this laboratory, researchers build constructs to help them look for another gene, and then that construct may be used again to help build a genetically modified mosquito that would be somehow incapable of transmitting disease. At risk of oversimplification, these researchers are looking for genes that express specific attributes they would like to be amplified or expressed throughout an entire population of organisms. Once they have identified their "gene of interest" they couple this gene with a "promoter" that would ensure that it is attached and transcribed into the target organism's genome:

As with the discussion of laboratory strains, some of these constructs, are familiar ones that are well understood by the community as a whole, and used over and over again by researchers in many different laboratories:

> Graduate Student: For doing experiments, we're choosing stuff with the right expression profile that we have characterized. There's actually not a big library of components to make constructs in mosquitoes. The same ones get used over and over again. (Interview 2)

There is not much of an interest or motivation to “reinvent the wheel” in this laboratory. Additionally, part of the motivation for using the same components again and again may be for the sake of using “tried-and-true” methods, and being able to borrow information from others that are known to be credible sources. One principal investigator discussed this explicitly:

> Principal Investigator: We rely on things that have been published and things that have been preferentially characterized already. (Interview 7)

When choosing which constructs to use, albeit from a small and well-defined portfolio, these researchers primarily use the metaphors of time and space when choosing which construct to use. This was described explicitly by the first interviewee:

> Graduate Student: I guess you got to think about time and space. Where do you want it to be? What exactly do you want to be promoted? At what time do you want it to be promoted? Where do you want it to be promoted? If you want it to be promoted in the embryos, then you choose a promoter which the transcription factors tend to be upregulated in embryonic development. If, instead of the embryo, you want it to be in the ovaries, then you'd chose an ovarian promoter. If you wanted it to be in the ovaries after a blood meal, you chose a promoter which is only activated after a blood meal. You have to think about time and space. (Interview 1)

A close look at the interviews with PIs reveals more detail about how these metaphors of time and space are used in choosing genetic components.

In this analysis for choosing genetic components, we see a distribution of reasoning families similar to that used in discussions of choosing laboratory species (Figure 5) as for

choosing genetic components (Figure 6). Once again, it is not surprising that the majority of this discussion drew from the causal reasoning family, as these researchers are ultimately interested in identifying the genetic basis for how a species transmits a disease. Therefore, it is more interesting for the purposes of this dissertation to focus on where causal reasoning is not enough, where it overlaps with comparative and dimensional reasoning (see Table 9 for relative frequencies of co-occurrences of reasoning families).

In the conversations about choosing genetic constructs, dimensional thinking was coupled with causal thinking to explore the evolutionary history of a specific gene (i.e. phylogeny). For example, one Principal Investigator discussed how he pulls together information about genes in different species ("orthologs") to see how sex-determining genes may have evolved over time:

> "Then you can take your y gene and make a phylogenetic genetic tree together with all these orthologs from the autosome. By looking at a tree, you can make a hypothesis to see whether this y copy came from the autosome. If so, when did that happen?" (Interview 13)

Dimensional thinking, when co-occuring with causal thinking, nearly always drew from the *topos* of time, but in several different ways. As in the example above, these PIs considered specific genes in terms of their evolutionary time. This is thinking of time on a large scale, but the PIs also considered time on a smaller, local scale, in terms of the development of one generation of organisms:

> "For promoters we take into account the expression profile of the gene. Is it temporally limited? Is it spatially limited?" (Interview 7)

By spatial and temporal limitation, this PI means they are looking for a gene that is expressed in a certain area of the mosquito (spatial) and expressed at a certain point of development, be it embryonic, larval, or adult (temporal).

As with the discussions of choosing laboratory strains, one PI used comparative thinking coupled with causal thinking to compare genes and their purposes across different areas of the genome, different species, or different sexes:

> Principal Investigator: Ideally you want to use the native promoter of that gene.
>
> Me: What do you mean by native?
>
> PI: Say, for example, when we look at GUY1, which is a gene that we found in *Anopheles*. We think it is important in sex determination, and to try to make a tran but it's on the Y chromosome. We're trying to make a trans-gene. So not only the male, the female would also express this gene. In that case, ideally you just want to use its own promoter, GUY1's own promoter, so it would be mimicking its own transcription pattern.
>
> Me: So using a promoter in *Anopheles* that is from *Anopheles*?
>
> PI: Right, for that particular purpose. (Interview 13)

Here, the PI is explaining how he would ideally make a "transgene" that works across both sexes by inserting the sex-determining gene (GUY1) using the promoter that naturally occurs with that gene, but is transcribed in both sexes. This PI seems to value keeping genetic components as true to the naturally occurring genome as possible. This is a similar line of thinking to the discussion of choosing laboratory strains, where the PIs seem to ideally want

a strain that is as close as possible to its wild counterpart. In short, they want in interfere with nature as minimally as possible but still achieve the desired results.

These researchers, whether professor or student, seem to have more freedom over what genetic components they choose to use than over what laboratory strains they use. For laboratory strains the PIs are bound by what genome is available and widely accepted, and the graduate students are bound by what the PI chooses to house in the lab. Of course, this laboratory only provides so much space for storing and maintaining colonies of mosquitoes. However, researchers are incredibly less constrained by the "space" for storing genetic components—genome databases. The fact that any number of genes can be explored and studied just within the physical space of a personal computer enables these researchers to think of more nuanced justifications for choosing certain genetic components than their justifications for choosing laboratory strains.

The previous section discussed how these researchers are, by necessity, choosing laboratory strains based on what is available to them. The discussion of choosing genetic constructs gets more nuanced, but there is still little discussion about how different constructs might be expressed differently in a different laboratory strains or in different species. It seem then, that these researchers are accepting provisional definitions of species (even if only temporarily) as a way of creating a point of stasis at the level of the species, and enable inquiry at the genetic level. This makes defining species and even defining laboratory strains not an arguable point (even if only for the time being), enabling the researchers to productively and effectively communicate information at the genetic level, saving time on contextualizing their arguments at the species level. The following section turns to the

medium that seems to enable this stasis with the species and enable nuanced inquiry at the level of the genes—genome databases.

**Table 9 Relative Frequencies of Reasoning Family Co-Occurrences in Choosing Genetic Components**

| | Dimensional | Comparative |
|---|---|---|
| Causal | 5 (23%) | 3 (14%) |
| Dimensional | | 0 (0%) |

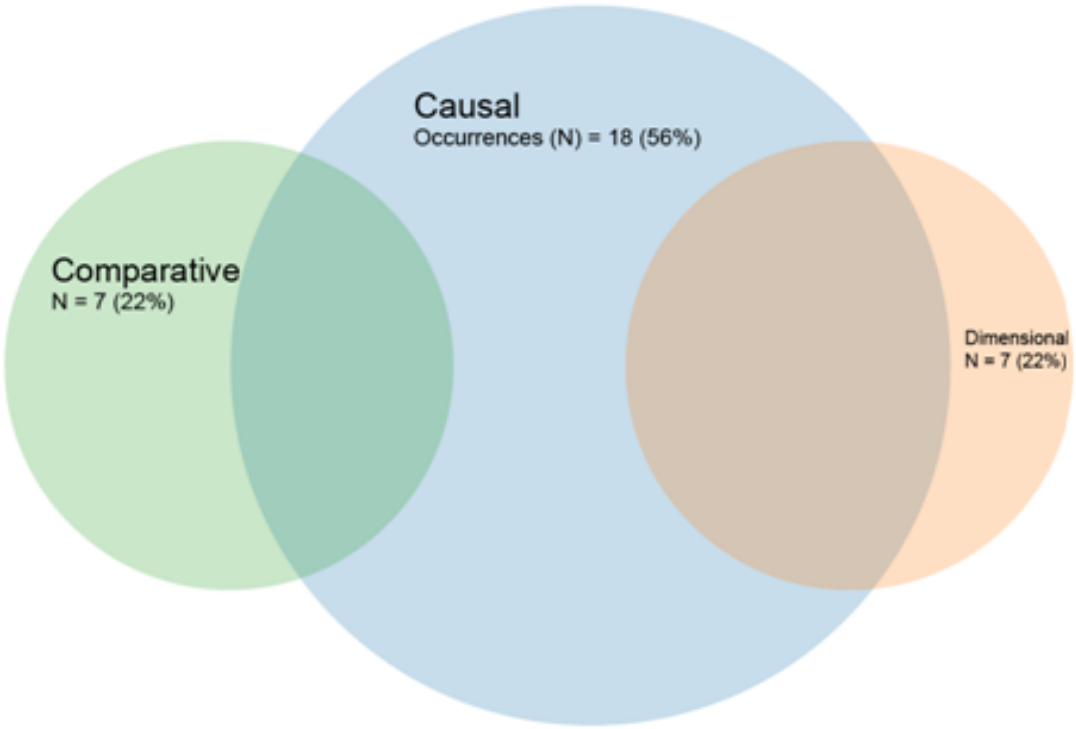



**Figure 6 Relative proportions of reasoning families in choosing genetic components.** Frequencies indicate totals for each reasoning family (both single occurrences and co-occurrences).

**Choosing genome databases**

The genome databases are where much rhetorical invention happens with these researchers. The primary rationale for considering what database to consult is based on its defined scope. Unsurprisingly, the majority of the researchers I interviewed identified VectorBase as the primary genome database they consult, since the mosquito genomes housed in this database were the primary object of their research. However, nearly all interviewees consulted others as well, often including FlyBase and the databases hosted by the National Center for Biotechnology Information (NCBI). The initial rationale for choosing

which database to consult simply depended on what organism you were researching. As one Associate Professor explained,

> The consideration is simple: FlyBase is only information about flies, not about mosquitoes. Although you can find orthologous genes for mosquitoes using FlyBase but that's about it, you cannot do anything else if you want to study mosquitoes. (Interview 4)

Even with this simple consideration of content, each database has different value to these researchers, depending on the nature their research. The interviewees indicated that they move back and forth among several databases for different purposes, or to fill gaps:

> Graduate Student: [I use] VectorBase probably the majority of the time. Sometimes, if I find what's maybe lacking in VectorBase, the next one I'd probably go to is FlyBase, which is the free fly database. Besides that, I use the NCBI if you're going back ever further, maybe if you want to make some phylogenic trees or something.
>
> Me: When you say, "Going back even further," you mean?
>
> Graduate Student: You have a wealth of information in Drosophila, which is like the gold standard, not, I guess, the function of certain genes aren't there 100 percent. You can search however many other organisms are in much larger databases, if you want to see some kind of hint for what it is. Most of the time, FlyBase has it about right. Not only that, they have all their orthologs and all that down pat. Most of the time VectorBase does the trick. If not, then I'll go FlyBase. (Interview 1)

This graduate student demonstrates some additional considerations of choosing a database, drawing on the *topos* of comparison. This researcher is using *Drosophila* as a point of

comparison for understanding the gene of interest. He is explaining here how he uses FlyBase to identify orthologs, which are the same gene in a different species, presumably indicating shared ancestry between two or more organisms.

In addition to the scope of the database, these researchers consult a portfolio of databases where they identify gaps, faulty annotations, or any other errors in the genomic data. Some of the inconsistencies stem from computer-generated gene models. One graduate student explained the relationship between the model and the actual sequence:

> Graduate Student: I'm in no form or fashion a computer science person, but they use predictive algorithms, using orthologs, to come up with a predicted gene model for something like CU80. Then they'll go through and say, "This certain percentage matches to this in Drosophila, so this is where we think this intron, this exon, and all these things are." It'll go through and it'll generate a gene model, which can be right, can be wrong. Some of the genes, CU80 was actually one of them, I went in and … designed primers, and sequenced the whole mRNA.
>
> There was a nine base-pair discrepancy between the gene model and the actual sequence that we were getting out of the genome. Effectively, all there was, was a nine base-pair difference. It predicted that those nine base pairs, those three codons were going to be there. They weren't actually there, at least in our wild-type strain.
>
> Me: What's then the advantage of comparing to a model, if it might be right, might be wrong?
>
> Graduate Student: The advantage with the models is it gives you a fantastic starting point. (Interview 1)

These researchers are using databases to “do homework” on their gene of interest; they are both acknowledging the messiness and inaccuracies of the data and also working with that to help develop new tools and ideas to suit their purposes at hand. It is not surprising then, that most of these researchers indicated that they primarily use databases at the beginning and end of a project, to gather information needed to develop their project, and then to refine their results and contribute information back to the community at the conclusion of a project.

The close topical analysis of the interviews with the PIs, once again showed a pattern similar to discussions of laboratory strains and genetic components—they primarily drew from causal reasoning, with some overlap in dimensional and comparative thinking. Table 10 shows relative frequencies of co-occurring reasoning families, and

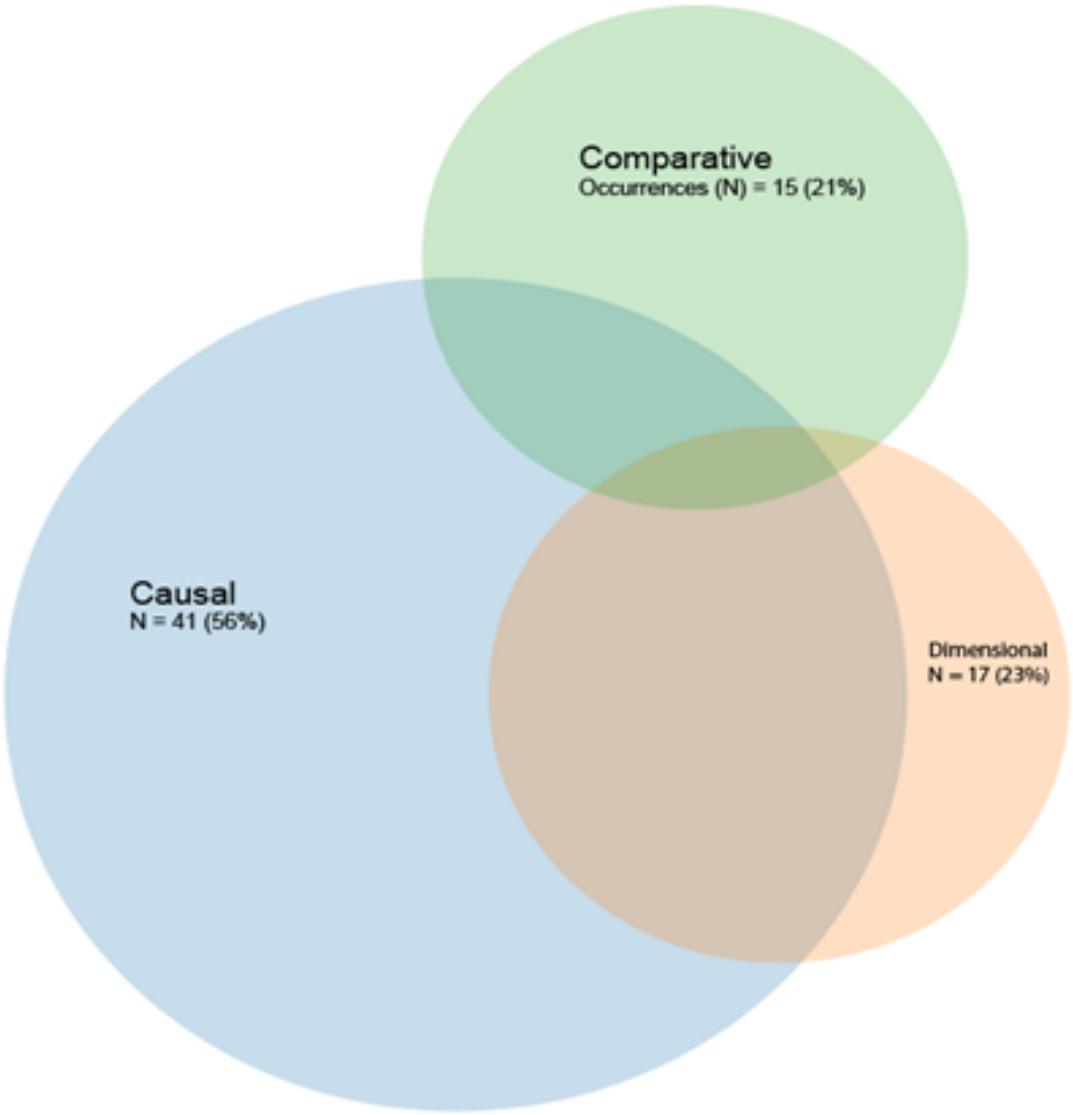



**Figure 7** shows proportional occurrences of each family and overlap.

Looking at the co-occurrences with causal reasoning, these PIs draw from the dimensional reasoning family to discuss the speed of databases (indicating they value doing work quickly), and reasoning based on how genes have changed in the databases (in the sense of having been updated, or changed evolutionarily). When drawing from both causal and comparative thinking, these PIs are thinking about comparing data:

> Principal investigator: What we do is actually go to [a database] and see *if people have sequenced a male genome* [causal], either the RNA or the genomic DNA. We look at the raw data just to see if, in those raw databases, "*Are there any sequences that look like X?*" [comparative] That's one example. (Interview 13)

To get a full sense of how genome databases influence rhetorical invention, it is helpful to understand how these databases are integrated into the research process from beginning to end. One Principal Investigator provided a hypothetical narrative to demonstrate when he consults databases during a research project. This aptly illustrates many of the *topoi* I've discussed so far. First, he describes his starting point as a point of comparison between *Drosophila* and mosquitoes:

> Principal investigator: I use them more often in the beginning and the end. Typically, a project will start with either focusing on a gene to begin with. For example, let's say somebody has published a paper on *Drosophila* and they've identified that *gene X does something Y* [causal], and I think, *"Oh, I wonder if the mosquito ortholog of that gene does something similar."* [comparative] (Interview 7)

Again, it is not that surprising that this PI uses primarily causal thinking in his research, given that the main focus of the lab (and the community as a whole) is to identify and describe genetic causes for disease transmission. What is interesting is when the researcher moves from causal reasoning in *Drosophila* to comparative reasoning with the mosquito.

This is where the genome database comes in. He continues by searching a genome database for components in the mosquito that are similar to the gene in *Drosophila*:

> I'll start going to the database, finding that gene, finding any paralogs, orthologs, *figuring out the phylogeny of that gene, how it's evolved* [dimensional]. Then, we may *clone pieces of that gene, express it in a mosquito, knock it out,* [dimensional] things like that. During those experiments, then we're just doing work in the lab, but it starts by looking at the gene.

Here this PI is drawing on dimensional thinking in terms of looking at how that particular gene has changed over time, and the experimental process of cloning, expressing, and then knocking out (or "deactivating") a gene.

Finally, his narrative leads to an explicit exercise of definition. He began discussing the definition of a gene in *Drosophila*, then moved to comparison to mosquitoes, to evolutionary time, and then finally back to questions of definition of the gene of interest so that he can interpret his results:

> *What are those genes? Can we get any information about them? What's their ontology and what families do they fit into?* What physiological functions do they have, things like that. We need to get all that from BioMart on VectorBase because that has whatever descriptions are available for those genes. We know what they are and what they're related to and what not, so we can interpret our experiments. (Interview 7)

This researchers moves from causal reasoning (gene X does function Y in *Drosophila*), to comparative reasoning (is the same gene in the mosquito?), then to dimensional thinking (considering its evolution and the procedure of cloning, expressing, then knocking out). This

movement from causal to comparative to dimensional reasoning sets him up to produce a nuanced definition of his gene of interest. Genome databases play a key part in this process.

Genome databases, then, carry much of the weight of rhetorical invention in this lab. Researchers consult a portfolio of different databases depending on their specific needs as a researcher. They serve as places where researchers can focus on a specific gene of interest, look at how that gene relates to other genes, whether it is present in other species, when it may have developed and for what purpose. If the main type of question being asked by these researchers is one of definition (what is this gene?), then they approach this stasis of by moving through causal, to comparative, and then to dimensional thinking.

**Table 10 Relative Frequencies of Reasoning Family Co-Occurrences in Database Discussion**

| | Dimensional | Comparative |
|---|---|---|
| Causal | 12 (22%) | 4 (7%) |
| Dimensional | | 1 (2%) |

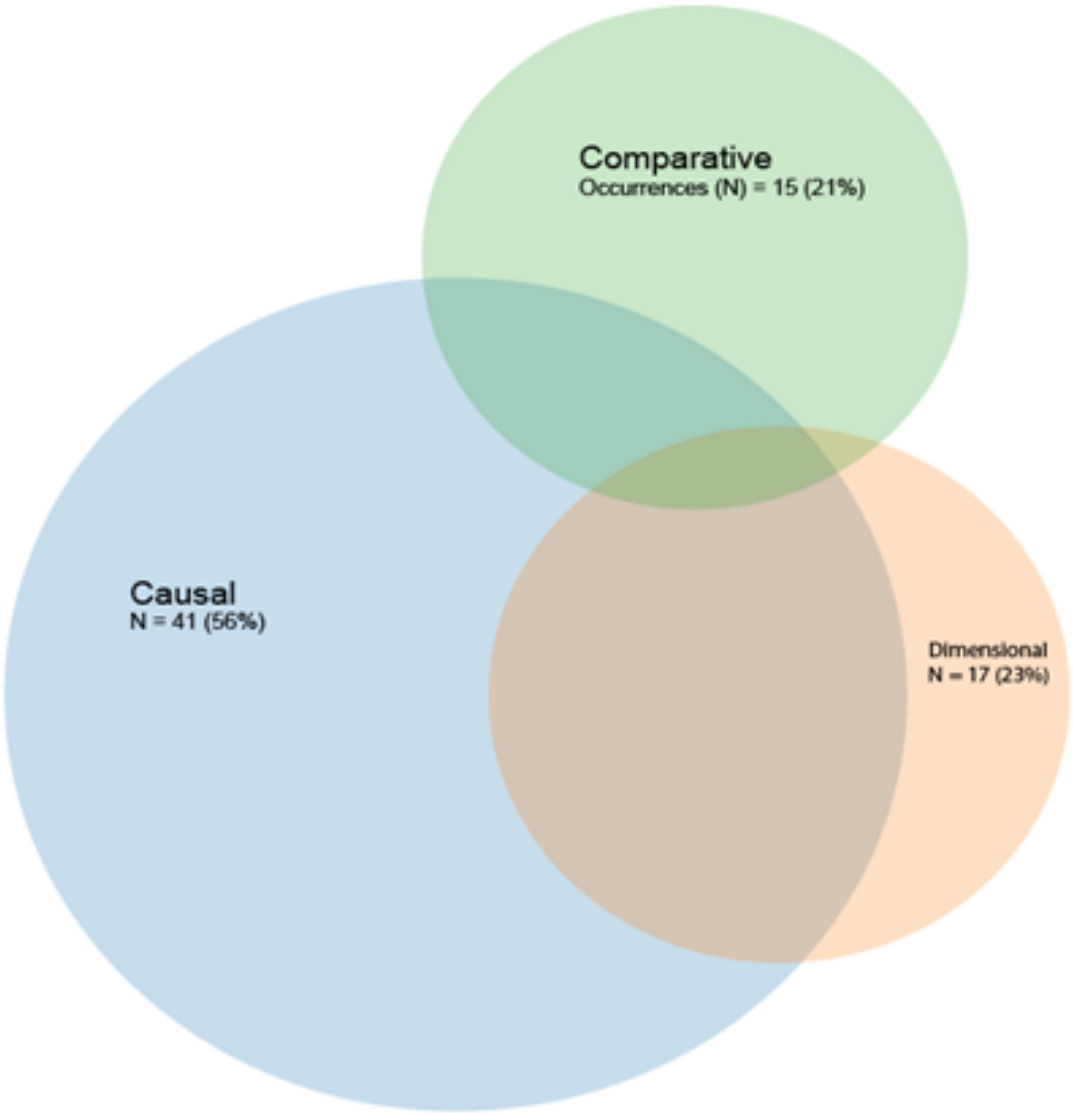



**Figure 7 Relative proportion of reasoning families in choosing databases.** Frequencies indicate totals for each reasoning family (both single occurrences and co-occurrences).

## Conclusion

Genome databases provide the medium for inventing and reinventing mosquitoes for different purposes. When asked why they make certain decisions they do regarding laboratory work, the scientists I interviewed said, again and again, that it depends on the

research question. Or, you could say, it depends on the rhetorical situation. In these interviews, the scientists were providing warrants for their decisions in the lab based on what they understood to be the best available means to accomplish their task at hand.

The notable absence of discussion about vector capacity, taxonomic status of these mosquitoes, or relatedness to other species of mosquitoes suggests that these issues are assumed by these researchers. As a result, the discussion of comparing *Aedes* to *Anopheles* species was relatively brief and did not vary much in substance with each interviewee; they gave similar responses across the board: *Aedes* has a bigger and more complex genome, but it is much easier to store in the lab.

We see a bit more nuance in definition when these scientists discuss specific laboratory strains. Causal reasoning still dominated these conversations, but comparative reasoning increased slightly. Even though responses were lengthier and somewhat more complex, these researchers did not have much freedom in their choice of laboratory strains. Graduate students and research technicians were bound by what was available to them in the lab, and PIs were bound by what genome was available to the community and what was considered to be the most widely used strain. At the same time, the PIs were able to discuss their ideal considerations for a laboratory strain. Drawing from the comparative reasoning family, they primarily discussed the differences between the laboratory strains and natural types. At the genetic level, the researchers coupled causal reasoning (e.g. “gene X performs function Y”) with both dimensional (e.g. “change over time”) and comparative (e.g. “male and female”) reasoning, demonstrating even more nuance than what was demonstrated in discussions about species and laboratory strains.

Moving to the level of the gene, causal reasoning, again, dominated conversations, and comparative and dimensional reasoning appeared roughly the same. These results, coupled with the explicit process of definition in this area of discussion, indicates that these researchers have even more flexibility in this area. These researchers use metaphors of time (When did a gene evolve? When is this gene expressed?) and space (Where is this gene expressed?) to define specific genes of interest.

There are similar results in the discussion of databases. Again, causal reasoning dominated and comparative and dimensional reasoning appeared nearly the same amount. These researchers have a similar level of freedom in choosing which databases to consult as with choosing which genetic components to use. Across all interviews, they indicated that they consulted a portfolio of different databases based on their particular need at the moment.

As I discussed in Chapters 1–3, clearly defining a "species" can sometimes be an untenable process, especially if we want to hold the assumption that species change over time. So why does this laboratory seem to take definitions of *Aedes* and *Anopheles* unquestioningly? This can be answered looking at the scope of this lab. This laboratory looks at only a few specific species of mosquitoes and their genetics. They are interested in learning more about the function of specific genes. The species of mosquito (e.g. *Anopheles gambiae*) provides a rhetorically stablized object around which researchers are able to ask questions. In other words, the species is the robust feature of the mosquito, and the genes are the more plastic features. These researchers are taking a bottom-up approach, starting with the genes to understand how the species transmits disease, and can be manipulated into a tool to help minimize the transmission of disease.

# CHAPTER 5: CONCLUSION

This dissertation explored the use of genome sequence databases in genetic engineering for disease control. What I hope these analyses demonstrate is that *databases are rhetorical*. They respond to an exigence and they are adapted to specific audiences and specific needs. They are then used to formulate new arguments, respond to new exigencies, and address specialized audiences. As a result, they persuade users towards certain beliefs about the world in the way data are organized. While *data* certainly never speak for themselves, as soon as database designers sort, classify, or define groups of data, they impose a symbolic system onto data that create certain rhetorical effects on the user. So far, I have explored the rhetorical nature of these databases from the perspective of the developers and their understanding of intended use, from the perspective of the organizing principles of the database, and from the perspective of one group of users. In this final chapter, I intend to further clarify the story that these analyses tell for the impact of genome databases on rhetorical invention.

Chapter 2 took the perspective of the developers of one specific database, VectorBase. In this chapter, I showed how the developers shape the specific exigence to which this database responds, and how they understand the community of users (their audience). In addition, this chapter looked at the organizational structure of this database to identify what types of arguments are favored by the structure of the database itself. We see in this analysis that the developers emphasize the capacity of VectorBase to integrate data, and an emphasis on the community as consumers rather than producers of data. This is perhaps a result of the overwhelming amount of data that is undoubtedly housed in VectorBase,

shifting the focus from production and collection of additional data to the organization and usability of existing data. This second analysis considered the Infectious Disease Ontology for Malaria (IDOMAL) and Dengue (IDODEN) that structure part of VectorBase. This analysis reveals the types of arguments that are favored by the database by outlining the lines (or "places," to use Aristotle's metaphor) of reasoning that the ontology designers considered to be acceptable by the community of users. In this way, these ontologies reflect a tension between the familiar and the unfamiliar, perhaps discouraging the unfamiliar. Walsh (2010) argues that STEM research has shifted to emphasize collaboration and consensus rather than dispute and difference. The *topoi* that I identified in the disease ontologies support this argument. Given the goal of databases to facilitate collaboration and exchange of data this is unsurprising. However, this potentially leads to some loss in the inventive capacity of the database by deemphasizing metaphorical (Leff, 1983), or transpositional (Prelli, 1989), thinking, where one thing is considered in terms of something that seems to be entirely unrelated.

Chapter 3 explores the rhetorical form the mosquito takes in the discourse on pest control research for malaria. In the world of molecular control of malaria, it seems that scientists are defining the malaria-transmitting *Anopheles* mosquitoes by a specific set of behaviors related to the blood-feeding cycle and, by extension, malaria transmission. These scientists are using a form of tree-thinking that pushes against the assumption that "organisms with similar attributes must be related," by using *Anopheles gambiae* as the "anchor species" for legitimizing research on other malaria-transmitting species. Because not all vectors of malaria are closely related, and not all closely-related mosquitoes transmit

malaria, these scientists are required to make the assumption that vector capacity evolved more than once. It then becomes critical to understand *how* and *why* this particular trait evolved in order for scientists to develop a strategy that would be applicable to all species that carry the malaria pathogen.

The complexities around the taxonomic status of the dominant vector species (DVS) of malaria pose significant constraints to rhetorical invention. In the case of malaria research, the inventional framework provided by evolutionary theory falls short because the species that are known to transmit malaria are not closely related. This constraint led researchers to employ a reasoning strategy that uses the *topos* "vector capacity" in place of "evolutionary relatedness" to develop a system for organizing and prioritizing species of interest to malaria researchers. Employing a *topos* that is relevant and useful to this particular community enables the researchers to not only formulate a stable method of communication about dominant vector species of malaria, but also continue to use evolutionary thinking, or "tree thinking," as a method of rhetorical invention. This involves developing an alternative "tree" that uses *Anopheles gambiae* as the "anchor species" (or the trunk) for legitimizing research on other malaria-transmitting species. Other species that are deemed important for malaria research are placed in the outer tiers (or branches) according to their relationship to *Anopheles gambiae,* and/or importance to malaria control. Evolutionary relatedness remains a robust feature of the boundary object to create a stable method of communication, and the behaviors related to "vector capacity" are the plastic features that adapt to the local needs of the area or species of study.

Chapter 4 turns to the audience of this genome database, one group of targeted users. In this chapter, I provide an analysis of interviews I conducted with practicing scientists in the area of genetic engineering for dengue and malaria control. Genome databases provide the medium for inventing and reinventing mosquitoes for different purposes. The *topoi* of vector capacity and taxonomic status, or relatedness to other species of mosquitoes, are notably absent from these interviews, suggesting that the researchers assume the vector capacity of the mosquitoes they work with. In this lab, this assumption had a physical manifestation, as many doors in and out of the lab were marked for "authorized personnel" only and required key access because they worked with live pathogens.

The interview responses became increasingly nuanced and detailed as the conversation moved from comparing *Aedes* and *Anopheles* mosquitoes, to discussing laboratory strains, and finally to discussing genetic components. Looking at each of these areas of discussion in detail, I found that researchers primarily use causal reasoning, but as the nuance increases, the researchers depart into comparative and dimensional thinking. What this tells me is that these researchers are focusing primarily at the genetic level of the mosquito, and in order to do so, they are accepting provisional definitions of the species, and in some instances the laboratory strains as well, in order to open up the genetic level of the organism for scientific inquiry. Genome databases are an essential tool in this stabilization-destabilization process, stabilizing the organism at the species level in order to destabilize at the genetic level.

It is helpful to compare the results from Chapter 4, regarding actual users, to the results from Chapter 2, regarding designers and intended use. In chapter 2, I argued that the

designers of VectorBase have shifted their focus from collecting data to organizing and integrating that data. The database and ontology developers seem to be focused on integrating data in order to facilitate causal and dimensional thinking in a community of interdisciplinary researchers. In the analysis of IDODEN and IDOMAL, I showed how the organization of the database emphasizes causal reasoning. This emphasis on causal reasoning is also shown in the analysis of interviews with Principal Investigators in Chapter 4, further supporting the argument that the community emphasizes collaboration and consensus building. The researchers I interviewed recognized, again and again, how they rely on the work of others to move forward on their own work. They use laboratory strains that have been sequenced and widely used by others. They use genetic components that have, likewise, been sequenced and widely tested by others. And, of course, they rely on the information that is contributed by others to genome sequence databases to help build their constructs and interpret their results. Taking the results of these two chapters together indicates that this community is perhaps bound together by the value of consensus building, and the value of having consistent and reliable modes of communication of their research. Comparative and dimensional thinking, which would ordinarily represent classification and categorization, are notably rare in the analysis of IDODEN and IDOMAL, and are likewise rare in discussions with PIs on the *Anopheles* and *Aedes* mosquitoes. This suggests that the community takes definitions of species as a point of stasis, if only temporarily, in order to enable the more nuanced work at the genetic level.

**Implications for rhetorical theory**

If genome databases are built in response to an exigence (Chapter 2), help to formulate new arguments (Chapter 3), and address a specialized audience (Chapter 4), then what do databases do for theory in rhetorical invention? I hope this dissertation demonstrates that boundary objects, like the mosquito, can serve as provisional stasis points to allow for exploration at a different stasis. Star and Griesemer (1989) developed the idea of boundary objects as a mechanism to explain how cooperation occurs despite the heterogeneous nature of scientific work. What this project adds is that boundary objects provide points of stasis among different social worlds, enabling members of these different social worlds to cooperate and collaborate. Mosquitoes facilitate collaboration by providing "standardized" material (e.g. the Liverpool strain and its genome) for laboratory research that provides stable points of communication, and also enables researchers to ask questions at a lower level: the gene. By providing stasis points for more specialized research at a lower level, boundary objects enable a family of *topoi* to emerge around the object, providing places to search for meaningful utterances. In the case of genetic engineering, this allows for quite literal places to search in genome databases.

Looking at databases as literal places to search for meaningful utterances follows most closely to Wallace's definition of *topoi* as "an orderly way of searching for meaningful utterances" (Wallace, 1972, p. 395). More recent work in the rhetoric of science has defined the *topoi* as beliefs, norms, and values that function as warrants in an argument (L. Walsh, 2010) and as resolved stases (Graham & Herndl, 2011). Being a standardized way of organizing all data on dengue, malaria, and their respective mosquito vectors, databases

provide an explicit way of searching for meaningful data and arguments. Wallace's definition follows the traditional sense of "invention" in rhetorical studies, that of "coming upon what already exists," despite the common usage of "invention" in English as "contriving something that never existed before" (Miller, 2000, p. 130). While databases do, quite literally, provide places to search for stock arguments, they also reflect beliefs, norms, and values of a given community in the way they are structured. VectorBase promotes consensus-building as a norm, and provides stable definitions (the second stasis) of the dominant vector species (DVS) that transmit malaria and dengue. This more complex definition of *topoi* that suits databases helps to bring genome databases into the realm of invention as innovation as well as discovery. In using and re-using laboratory strains and genetic components that have been heavily used and annotated by others in the field, they are using "stock arguments" to build something new, to build a genetically modified mosquito that behaves in a way not found in nature, in order to control the spread of malaria or dengue.

This dissertation began as an exploration of the practices of classification, categorization, and defining mosquito species. In effect, these are stases themselves. For instance, the stasis of definition is, by default, also a practice of classification. By defining what something is, what it is not, what is relevant to the object, one is participating in practices of classification. Boundary objects provide stases, then a class/category of *topoi* emerge around that object/stasis. To put this in Star and Griesemer's terms, the stasis point is the rigid feature of the object, and the *topoi* are the flexible characteristics. Looking at how specific boundary objects provide provisional stases, then evoke a class of *topoi* would be a fruitful direction for future research in rhetorical invention in science. For instance, in the

case of vector-borne diseases, one could explore how the mosquito is invented in different domains of discourse, such as scientific and technical literature, prevention and control guidelines from the World Health Organization, and popular news media in different areas of the world.

**Limitations and Future Work**

To a certain extent, my arguments are tautological. I went into this work with the assumption that genome databases facilitate rhetorical invention, and much of my argument reiterates that point. Additionally, I began this work with the plan to consider the mosquito as a rhetorical boundary object in order to understand how it is defined by a specific research community. This approach is not unlike the approach of my research participants. I began with these grand assumptions in order to observe the rhetoric surrounding the mosquito and the rhetoric surrounding genome databases on a more granular level.

Some specific limitations include the quantity of data I was able to collect and analyze. In Chapter 4, the major analysis focused on two interviews that totaled less than one hour in length. While this analysis was able to tell me a great deal about the inner workings of this lab, more interviews with other PIs, graduate students, postdocs, lab technicians, and faculty would potentially tell a very different story. I intend to expand the range of data collection with future research. Additionally, other data sets I use are limited simply due to what is available at the present time. Specifically, IDOMAL, IDODEN, and the developers' reports which were the focus for Chapter 2 provided only thirteen codable "relations" in the ontologies, and 21 pages of codable reports. In addition, the fact that the lead author on the reports varied, the differences I find may be an artifact of individual writing styles. In the

future, I will add analyses of other disease ontologies related to mosquito-borne diseases, and other developer reports as they are published.

To achieve the level of technical understanding I sought for the present study, I focused on just one highly specialized research community, one small group of organisms, and one type of database. This type of study study could be modified and applied to many other fields of scientific research, many other types of species, or many other types of scientific databases, and come to different conclusions of rhetorical invention and boundary objects. Additionally, while this is beyond the scope of the present study, this study would be very well complimented by a usability study of genome sequence databases. The present study focuses on the rhetorical constraints, be them enabling or restrictive, of genome databases, but it could be the case that researchers are well trained to look for these very constraints and work with or around them in some way. In other words, it should not be assumed that any recommendations I offer in light of this work are not already being addressed by users, even if subconsciously.

**Ethical Implications: The Case of Zika Virus and CRISPR/Cas9**

Since I began writing this dissertation, there have been two relevant developments in the world of genetic engineering and disease control: the invention of CRISPR/Cas9 and the Zika virus outbreak in Latin America. Zika has been compared in many respects to dengue, as it is transmitted by the same mosquito and causes similar flu-like symptoms. What is different about Zika is its potential connection to microcephaly, a condition in which infants exhibit smaller than average skull sizes. Some Latin American countries have urged women to delay pregnancy as a result of the outbreak. These outbreaks in Latin America, following a

similar outbreak in French Polynesia in 2014, led the World Health Organization to declare Zika an international public health emergency in February 2016 (Roberts, 2016). Genetically modified mosquitoes have been a part of this conversation as both a potential solution (Allen, 2016) and a possible cause (Tickell, 2016) of the outbreak. The specific technology being discussed as both a possible cause and potential solution is the Oxitec mosquito, which is intended to decrease the population of the *Aedes aegypti* mosquito by spreading a lethal gene into the population that causes mosquito larvae to die before reaching adulthood. The Oxitec mosquito was being released in Brazil prior to the Zika outbreak.

A new gene-editing technology, CRISPR/Cas9, is also being explored as a possible solution to the Zika crisis. In a nutshell, this technology enables scientists to edit the genome in a way that a gene is "driven" through an entire population at a 100% inheritance rate, even if that genetic change leads to the destruction of the entire species (Esvelt, 2016). This powerful new technology could be enlisted to eradicate diseases like malaria, dengue, and Zika by attacking the vector that transmits them. It could also be used in controlling invasive species that pose a threat to biodiversity, such as rats on islands. Of course, in the wrong hands, it could even be employed to cause detrimental effects to species that are beneficial to us and the environment, or even our own species. While this may seem far-fetched to some, it is still right to give us pause in employing such a powerful technology.

Given the incredible power of this technology, and given that the Zika virus mostly affects disenfranchised women in impoverished areas of Latin America, some experts have opened up discussions on how to proceed in an ethical way that gives voice to those who are

most affected by the diseases we are seeking to control (Zielinski, 2016). Some experts have been quite vocal about proceeding with caution and humility (see Esvelt, 2016).

These conversations reside primarily in the stases of quality and judgment, focusing on the question "How do we implement these technologies in an ethical way?" While I don't intend to entirely dismiss this line of thought, this question assumes a determinist approach to the technology, that the technology *will* be used, it's merely a question of when and how. This determinist approach is partly due to positioning the gene drive technology as the primary boundary object for collaboration. This boundary object, then, is explored in primarily the stases of quality and judgment, taking for granted many technical details and risk factors, which would be issues in the stases of fact and definition. This is forcing technical issues related to gene drives (which would ordinarily reside in the first and second stases) directly into the stases of quality and judgment. These technical issues are then prioritized over other concepts that would reside in the stases of quality and judgement, and perhaps over-simplifying a very complex social and cultural issue.

If we changed the question to one that would naturally reside in the stasis of quality, such as, "How can we decrease human suffering?" the conversation would change dramatically. This would displace the technology and put suffering, including disease, as the primary boundary object for collaboration. Shifting the stasis point in this way would, by necessity, include questions of how to decrease oppression, disenfranchisement, *and* disease. This would displace the technology from the center of the conversation, eliminating the determinist approach, and move those affected by disease to the center. Boundary objects are needed to provide provisional points of agreement to enable collaboration, but collaborators

should continually interrogate what is being used as a provisional point of agreement, exercising caution against determinist lines of thinking.

This dissertation demonstrates how one community of researchers creates points of stasis in an indefinitely complex scientific situation in order to enable rhetorical invention at the genetic level, ultimately leading to the invention of technologies like CRISPR/Cas9. I propose that this rhetorical strategy be applied to governance and ethics as well. The desire to decrease human suffering—from disease, oppression, or otherwise—could provide this community with a *topos* where conversations about how to employ various technologies in an ethical way would become more productive, providing an agreed-upon normative goal that can serve as a warrant to strengthen arguments regarding medical intervention. If we can agree on disease as a stasis point, then the mosquito would become a productive boundary object for discussions of how it can be best controlled to reach this goal, through genetic engineering or other techniques.

## REFERENCES

Allen, G. (2016). Genetically Modified Mosquitoes Join The Fight To Stop Zika Virus. Retrieved February 29, 2016, from http://www.npr.org/sections/goatsandsoda/2016/01/26/464464459/genetically-modified-mosquitoes-join-the-fight-to-stop-zika-virus

Ankeny, R. a., & Leonelli, S. (2011). What's so special about model organisms? *Studies in History and Philosophy of Science*, *42*, 313–323. doi:10.1016/j.shpsa.2010.11.039

Aultman, K. S., Gottlieb, M., Giovanni, M. Y., & Fauci, A. S. (2002). Anopheles gambiae genome: Completing the malaria triad. *Science*, *298*(5591), 13.

Beatty, J. (1992). Speaking of Species: Darwin's Strategy. In M. Ereshefsky (Ed.), *The Units of Evolution* (pp. 227–246). Cambridge: MIT Press.

Benson, D. a., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2013). GenBank. *Nucleic Acids Research*, *41*(D1), 36–42. doi:10.1093/nar/gks1195

Benton, M. J. (2000). Stems, nodes, crown clades, and rank-free lists: is Linnaeus dead? *Biological Reviews of the Cambridge Philosophical Society*, *75*(4), 633–648.

Besansky, N. J., Ashburner, M., Carlton, J. M., Coetzee, M., Collins, F. H., della Torre, A., … Wirtz, R. (2008). *Genome Analysis Of Vectorial Capacity In Major Anopheles Vectors Of Malaria Parasites*. Retrieved from https://www.vectorbase.org/projects/genome-analysis-vectorial-capacity-major-anopheles-vectors-malaria-parasites

Bowker, G., & Star, S. L. (1999). *Sorting Things Out: Classification and its Consequences*.

Cambridge, Mass: MIT Press.

Burke, K. (1966). Terministic Screens. In *Language as Symbolic Action* (pp. 44–62). Berkeley: University of California Press.

Ceccarelli, L. (2014). Rhetoric of Science and Technology. In *Ethics, Science, Technology and Engineering: A Global Resource* (2nd ed., pp. 621–625). Macmillan Reference USA.

Creager, A. N. H. (2001). *The Life of a Virus: Tobacco Mosaic Virus as an Experimental Model, 1930-1965*. Chicago: University of Chicago Press.

Crosswhite, J. (2008). Awakening the Topoi: Sources of Invention in The New Rhetoric's Argument Model. *Argumentation and Advocacy*, *44*(Spring), 169–184.

de Queiroz, K., & Gauthier, J. (1994). Toward a phylogenetic system of biological nomenclature. *Trends in Ecology & Evolution*, *9*(1), 27–31. doi:10.1016/0169-5347(94)90231-3

Endersby, J. (2007). *A Guinea Pig's History of Biology*. Cambridge, Mass: Harvard University Press.

Esvelt, K. (2016). Strategies for Responsible Gene Editing. Retrieved February 29, 2016, from https://www.project-syndicate.org/commentary/crispr-gene-drive-editing-rules-by-kevin-m--esvelt-2016-01

Fahnestock, J., & Secor, M. (1988). The Stases in Scientific and Literary Argument. *Written Communication*, *5*(4), 427–443. doi:10.1177/0741088388005004002

Farrell, T. B. (1991). Practicing the Arts of Rhetoric : Tradition and Invention. *Philosophy & Rhetoric*, *24*(3), 183–212.

Galperin, M. Y., Rigden, D. J., & Fernandez-Suarez, X. M. (2015). The 2015 Nucleic Acids Research Database Issue and Molecular Biology Database Collection. *Nucleic Acids Research*, *43*(Database Issue), D1–D5. doi:10.1093/nar/gku1241

Ghiselin, M. (1969). *The Triumph of the Darwinian Method*. Berkeley: University of California Press.

Gieryn, T. F. (1983). Boundary-Work and the Demarcation of Science from Non-Science: Strains and Interests in Professional Ideologies of Scientists. *American Sociological Review*, *48*(6), 781–795.

Giraldo-Calderon, G. I., Emrich, S. J., MacCallum, R. M., Maslen, G., Dialynas, E., Topalis, P., … Lawson, D. (2015). VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases. *Nucleic Acids Research*, *43*(Database issue), D707–D713. doi:10.1093/nar/gku1117

Graham, S. S., & Herndl, C. G. (2011). Talking Off-Label: The Role of Stasis in Transforming the Discursive Formation of Pain Science. *Rhetoric Society Quarterly*, *41*(2), 145–167. doi:10.1080/02773945.2011.553764

Graves, H. (2005). *Rhetoric in(to) Science: Style as Invention in Inquiry*. Cresskill, NJ: Hampton Press.

Hotez, P. J. (2010). A plan to defeat neglected tropical diseases. *Scientific American*, *302*(1), 90–94, 96. doi:10.1038/scientificamerican0110-90

Hull, D. (1988). *Science as a process: An evolutionary account of the social and conceptual development of science*. Chicago: University of Chicago Press.

Kohler, R. E. (1994). *Lords of the Fly: Drosophila Genetics and the Experimental Life*.

Chicago: University of Chicago Press.

Krzywinski, J., & Besansky, N. J. (2003). Molecular systematics of Anopheles: from subgenera to subpopulations. *Annual Review of Entomology*, *48*, 111–39. doi:10.1146/annurev.ento.48.091801.112647

Lawniczak, M. K. N., Emrich, S. J., Holloway, A. K., Regier, A. P., & Et Al. (2010). Widespread Divergence Between Incipient Anopheles gambiae Species revealed by Whole Genome Sequences. *Science*, *330*(October), 512–515.

Lawson, D., Arensburger, P., Atkinson, P., Besansky, N. J., Bruggner, R. V., Butler, R., … Collins, F. H. (2007). VectorBase: A home for invertebrate vectors of human pathogens. *Nucleic Acids Research*, *35*(Database issue), D503–D505. doi:10.1093/nar/gkl960

Lawson, D., Arensburger, P., Atkinson, P., Besansky, N. J., Bruggner, R. V., Butler, R., … Collins, F. H. (2009). VectorBase: A data resource for invertebrate vector genomics. *Nucleic Acids Research*, *37*(Database issue), 583–587. doi:10.1093/nar/gkn857

LeFevre, K. B. (1987). *Invention as a Social Act*. Carbondale, Il: Southern Illinois University Press.

Leff, M. (1983). Topical Invention and Metaphoric Interaction. *The Southern Speech Communication Journal*, *48*(3), 214–229.

Leonelli, S. (2012). Introduction: Making sense of data-driven research in the biological and biomedical sciences. *Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*(1), 1–3. doi:10.1016/j.shpsc.2011.10.001

Leonelli, S., & Ankeny, R. (2012). Re-thinking organisms: The impact of databases on model organism biology. *Studies in History and Philosophy of Biological and*

*Biomedical Sciences*, *43*(1), 29–36. doi:10.1016/j.shpsc.2011.10.003

Mayr, E. (1982). *The Growth of Biological Thought: Diversity, Evolution, and Inheritance*. Cambridge: Belknap Press.

Megy, K., Emrich, S. J., Lawson, D., Campbell, D., Dialynas, E., Hughes, D. S. T., … Collins, F. H. (2012). VectorBase: Improvements to a bioinformatics resource for invertebrate vector genomics. *Nucleic Acids Research*, *40*(Database issue), D729–D734. doi:10.1093/nar/gkr1089

Miller, C. R. (2000). The Aristotelian Topos: Hunting for Novelty. In A. Gross & W. Keith (Eds.), *Rereading Aristotle's Rhetoric*.

Miller, C. R. (2005). Novelty and heresy in the debate on nonthermal effects of electromagnetic fields. *Rhetoric and Incommensurability*, 464–505. Retrieved from http://www4.ncsu.edu/~crmiller/Publications/MillerParlorPress05.pdf

Miller, C. R., & Selzer, J. (1985). Special Topics of Argument in Engineering Reports. In L. Odell & D. Goswami (Eds.), *Writing in Non-Academic Settings* (pp. 309–341). New York: The Guilfors Press.

Mitraka, E., Topalis, P., Dritsou, V., Dialynas, E., & Louis, C. (2015). Describing the Breakbone Fever: IDODEN, an Ontology for Dengue Fever. *PLOS Neglected Tropical Diseases*, *9*(2), 1–19. doi:10.1371/journal.pntd.0003479

Nene, V., Wortman, J. R., Lawson, D., Haas, B., Tu, Z. J., Loftus, B., … Birren, B. (2007). Genome Sequence of Aedes aegypti, a Major Arbovirus Vector. *Science*, *316*(22 June), 1718–1723.

Normile, D. (2013). Surprising New Dengue Virus Throws A Spanner in Disease Control

Efforts. *Science*, *342*(October), 2013. doi:10.1126/science.342.6157.415

Nothstine, W. L. (1988). "Topics" as Ontological Metaphor in Contemporary Rhetorical Theory and Criticism. *Quarterly Journal of Speech*, *74*, 151–163.

Pietsch, T. W. (2012). *Trees of Life: A Visual History of Evolution*. Baltimore: The Johns Hopkins University Press.

Prelli, L. J. (1989). *A Rhetoric of Science: Inventing Scientific Discourse*. Columbia, SC: University of South Carolina Press.

Rader, K. A. (2004). *Making Mice: Standardizing Animals for American Biomedical Research, 1900-1955*. Princeton: Princeton University Press.

Ribes, D., & Bowker, G. C. (2009). Between meaning and machine: Learning to represent the knowledge of communities. *Information and Organization*, *19*(4), 199–217. doi:10.1016/j.infoandorg.2009.04.001

Roberts, M. (2016). Zika-linked condition: WHO declares global emergency. *BBC News*. Retrieved from http://www.bbc.com/news/health-35459797

Scott, R. L., Andrews, J. R., Martin, H. H., McNally, J. R., Nelson, W. F., Osborn, M. M., … Zyskind, H. (1971). Report of the Committee on the Nature of Rhetorical Invention. In L. F. Bitzer & E. Black (Eds.), *The Prospect of Rhetoric: Report of the National Developmental Project* (pp. 228–236). Cliffs, NJ: Prentice-Hall.

Sinka, M. E., Bangs, M. J., Manguin, S., Chareonviriyaphap, T., Patil, A. P., Temperley, W. H., … Hay, S. I. (2011). The dominant Anopheles vectors of human malaria in the Asia-Pacific region: occurrence data , distribution maps and bionomic précis. *Parasites & Vectors*, *4*(89). doi:10.1186/1756-3305-4-89

Sinka, M. E., Bangs, M. J., Manguin, S., Coetzee, M., Mbogo, C. M., Hemingway, J., … Hay, S. I. (2010). The dominant Anopheles vectors of human malaria in Africa , Europe and the Middle East : occurrence data , distribution maps and bionomic précis. *Parasites & Vectors*, *3*(1), 117. doi:10.1186/1756-3305-3-117

Sinka, M. E., Rubio-palis, Y., Manguin, S., Patil, A. P., Temperley, W. H., Gething, P. W., … Hay, S. I. (2010). The dominant Anopheles vectors of human malaria in the Americas: occurrence data, distribution maps and bionomic précis. *Parasites & Vectors*, *3*(72), 1–26.

Star, S. L., & Griesemer, J. R. (1989). Institutional Ecology , ' Translations ' and Boundary Objects : Amateurs and Professionals in Berkeley ' s Museum of Vertebrate Zoology , 1907-39. *Social Studies of Science*, *19*(3), 387–420.

Stevens, H. (2013). *Life out of Sequence: A Data-Driven History of Bioinformatics*. Chicago: University of Chicago Press.

Strasser, B. J. (2008). GenBank--Natural History in the 21st Century ? *Science*, *322*(October), 537–538.

Tickell, O. (2016). Pandora's box: how GM mosquitos could have caused Brazil's microcephaly disaster. Retrieved February 29, 2016, from http://www.theecologist.org/News/news_analysis/2987024/pandoras_box_how_gm_mosquitos_could_have_caused_brazils_microcephaly_diasaster.html

Topalis, P., Mitraka, E., Bujila, I., Deligianni, E., Dialynas, E., Siden-Kiamos, I., … Louis, C. (2010). IDOMAL: an ontology for malaria. *Malaria Journal*, *9*(230), 1–11. doi:10.1186/1475-2875-9-230

Wallace, K. R. (1972). Topoi and the Problem of Invention. *The Quarterly Journal of Speech*, *58*, 387–395.

Walsh, L. (2010). The Common Topoi of STEM Discourse: An Apologia and Methodological Proposal, With Pilot Survey. *Written Communication*, *27*(1), 120–156. doi:10.1177/0741088309353501

Walsh, L. (2013). Resistance and Common Ground as Functions of Mis/aligned Attitudes: A Filter-Theory Analysis of Ranchers' Writings About the Mexican Wolf Blue Range Reintroduction Project. *Written Communication*, *30*(4), 458–487. doi:10.1177/0741088313498362

Wilson, G., & Herndl, C. G. (2007). Boundary Objects as Rhetorical Exigence: Knowledge Mapping and Interdisciplinary Cooperation at the Los Alamos National Laboratory. *Journal of Business and Technical Communication*, *21*(2), 129–154.

Winsor, M. P. (2013). Darwin and Taxonomy. In M. Ruse (Ed.), *The Cambridge Encyclopedia of Darwin and Evolutionary Thought* (pp. 72–79). New York: Cambridge University Press.

Zielinski, A. (2016). The Ethical Risks of Engineering Mosquitoes Into Extinction to Stop Zika. Retrieved February 29, 2016, from http://thinkprogress.org/health/2016/02/24/3752711/gene-drives-mosquitoes-zika/

# APPENDICES

## APPENDIX A

**VectorBase Report Topoi with Definitions and Examples**

1. **Application:** Any mention of a potential marketable or patentable application resulting from VectorBase-related projects

    a. **"**The aim of these projects was to better understand the biology of the pathogen through its genome, with the goal of identifying *new therapeutics* and thus shorten the time from therapeutic lead to *marketable product*, a notoriously slow process."

2. **Genome as method:** Any explicit mention of experimental methods, especially the application of VectorBase data to experiment design. Additionally, any mention of genome data being used as a vehicle towards learning more about an organism or biological process.

    a. **"**The availability of the 'Culex' *genome annotation facilitates comparison* of the three main families of mosquitoes (Anopheline, Aedine and Culicine) with the model dip- teran Drosophila melanogaster."

3. **Efficiency:** Any mention of the simplification or ease of workflow within the VectorBase interface, or as a result of using VectorBase, including increased speed or timeliness of work.

    a. **"**The *simplicity of the submission process* in conjunction with community representative involvement in data quality consistency checks (e.g. does the submitted sequence translate correctly) ensures that any required discussion and error correction *happens in a timely manner*."

4. **Identifying gap:** Any mention of an area where more knowledge, resources, or data are needed or are lacking

**a.** **“**A more holistic approach to improving our understanding of these pathogens *needs to include* intermediary vectors where they exist.”

**5. Cost:** Any explicit mention of increased or decreased cost of laboratory tools or processes

**a.** **“**Over the past few years the *cost* of genome sequencing has fallen dramatically making it feasible to sequence the genomes of vectors and complete our knowledge of the triumvirate of species involved in many parasitic diseases.

**6. Funding:** Any explicit mention of funding sources for VectorBase or other entities

**a.** **“**VectorBase is *funded* by the National Institute of Allergy and Infectious disease (NIAID) as part of a group of Bioinformatics Resource Centres (BRCs) (http://www.brc-central.org/) aiming to provide web-based resources to the scientific community for organisms considered to be causing or trans- mitting emerging or re-emerging infectious disease.”

**7. Collaboration:** Any mention of two entities collaborating or partnering together to improve VectorBase or benefiting from the collaborative affordances of Vectorbase; often coincides with “community”

**a.** **“**VectorBase is involved in all the stages of genome analysis: first-pass annotation of new genome sequences *in collaboration with* the sequencers, re-annotation of existing genome sequences and submission of these data sets to the public nucleotide databanks.”

**8. Submission:** Any mention of any portion of the process of submitting data to VectorBase

**a.** **"***Data can be submitted* to the VectorBase Population Biology Resource via spreadsheet forms using open source tools to assist with formatting and ontology term selection (ISA-Tab (27) and Phenote, http://www.phenote.org). Genotypes are submitted to the variation resource in standard VCF format (5).

**9. Review:** Any mention of the data review process in VectorBase, including annotation and re-annotation processes

**a.** "Once an annotation is finalized, *additional analyses are performed* such as our standard orthology/paralogy relationship predictions (6) and *cross-referencing to other resources*. This system was *trialled* for the R. prolixus and G. morsitans genomes."

**10. Accuracy:** Any mention of the accuracy or precision of data or technical process, or corrections to those data or techniques

**a.** **"**These data include *corrections* of gene structures and relevant metadata such as gene symbols and citations."

**11. Consistency:** Any mention of efforts towards organizing, storing, ordering data in a consistent manner, including any mention of "ontologies"

**a.** **"**The collection of experimental and sample-related metadata has been aided through our development of *ontologies and controlled vocabularies* for vector-specific data, such as field-associated samples, pathogen transmission and in-secticide resistance"

**12. Integration:** Any mention of data being linked, cross-referenced, compared or connected in some way

**a.** **"***Integration of these data with existing gene sets* has greatly improved reference gene sets (e.g. An. gambiae) and has led to a new 'patch' build system that uses heuristics to merge manual and automated gene predictions to allow more frequent gene set updates"

**13. Search/retrieval:** Any explicit mention of the process of data mining or searching or retrieving data from VectorBase or other entity

**a.** **"**We have also implemented *data mining* tools, such as the HMMER package (http://hmmer.janelia.org/) to build profile hidden Markov models from multiple sequence alignments which can then be used for sensitive database *searching* using statistical descriptions of a sequence families consensus"

**14. Manual/automatic:** Any explicit mention of a process being manual or automated (ie performed by human vs computer)

**a.** **"**The annotation of the An. gambiae genome is being *manually* appraised using the GMOD annotation tool Apollo (4). Currently, over 50% of the genome has been com- pleted including the entirety of the chromosome arms 2L, 2R and X."

**15. Future work:** Any mention of work to be done (construed in the future tense), or mention of future directions or goals

**a.** **"**A number of Anopheles species *will be targeted* for genome sequencing (http://www.vectorbase.org/Docs/ShowDoc/?doc=White Papers) and the reduction in cost means that individual labs can produce significant amounts of sequence data from species or isolates. The integration and management of these data will be a major challenge for *the coming years*"

**16. Community:** Any mention of users, a community of scientists, work being outsourced to a community, etc. Often co-occurs with "collaboration"

**a.** **"**Other material of interest to the vector community is being incorporated, including the newly developed controlled vocabu- lary of mosquito anatomy (http://obo.sourceforge.net/detail. cgi?mosquito_anatomy) and other vector-related ontologies."

**17. Customization:** Any mention of the database interface being tailored for/by a specific user

**a.** **"**The standard display methods provide a wide variety of options that can be customized by a submitter to best suit their data."

**18. Reference to past report:** Any explicit mention to a previously published VectorBase report; often co-occurs with "past growth"

**a.** **"**As anticipated in *our previous update* (2), analyses of populations and variations at the genomic level have increased significantly"

**19. Past growth:** Any mention of improvement to the database construed in the past tense

**a.** [From 2015] "A major overhaul of the PopBio infrastructure *was undertaken in 2012–13* and a few of the more user-visible changes are described here. URLrobustness and data maintainability *were improved* by the allocation of stable IDs for projects, samples and assays. The submission procedure *was streamlined* to a single ISA-Tab (9) spreadsheet submission.

**20. Breadth/scope of data:** Any mention or *display* of the breadth, scope, or variety of data included in VectorBase

**a.** **“**VectorBase currently hosts *nine genomes* of which the majority are mosquitoes, reflecting their importance in disease agent transmission. The seven corresponding species are: *Anopheles gambiae (three genomes, for the PEST, Mali-NIH and Pimperena colonies), Aedes aegypti, Culex quinquefasciatus, Glossina morsitans, Ixodes scapularis, Pediculus humanus and Rhodnius prolixus*”

## APPENDIX B

**Semi-structured interview protocol**

Thank you for your willingness to chat with me. Do you mind if I video record and voice record our conversation?

I'm conducting these interviews to learn more about how you use genome databases in your research. First, I'd like to learn some more about your research specifically.

Can you tell me, in general terms, about the nature of your research?

What specific laboratory strain(s) of [AEDES OR ANOPHELES] do you use?

What considerations do you make when choosing which [ANOPHELES OR AEDES AGYPTI] strain to use in your laboratory?

How would you say your work compares to research in [OPPOSITE, AEDES OR ANOPHELES]? Would you say it's an easier or harder species to work with?

What considerations do you make when choosing which promoters, genes and other components to put into a construct to use in designing an experiment?

I'd like to turn now to more specifics about what databases you use in your research.

Do you have a specific database you like to use?

What considerations do you make when choosing which database to use?

At what point in your research projects do you use genome databases?

Walk me through [DATABASE] and show me what features you use specifically.

What features are particularly helpful for you in this database?

Could you give an example of the type of questions you have used this database to help you answer?

Do you ever develop new ideas after browsing through this database? What would be an example?

What do you see as specific limitations of [DATABASE]? Do you have tricks to work around those limitations?

Is there anything else you think I should know?