

**A QUALITATIVE ASSESSMENT OF BIOINFORMATICS TASKS,  
TOOLS AND WORKFLOW WITHIN AN ACADEMIC BIOLOGICAL  
RESEARCH LABORATORY**

**by**

**Nicholas Rolphe Anderson**

**A THESIS**

Presented to the Division of Medical Informatics and Clinical Epidemiology  
and the Oregon Health & Science University School of Medicine in partial  
fulfillment of the requirements for the degree of Masters of Science in  
Biomedical Informatics.

April 2004

School of Medicine  
Oregon Health & Science University

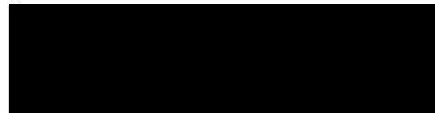
**Certificate of Approval**

This is to certify that the Masters thesis of

**Nicholas R. Anderson**

*“A Qualitative Assessment of Bioinformatics Tasks, Tools  
and Workflow Within an Academic Biological Research  
Laboratory”*

Has been approved



David Robinson, PhD



Christopher Dubay, PhD



Joan Ash, PhD

# TABLE OF CONTENTS

|   |            |
|---|------------|
| <b>ACKNOWLEDGEMENTS .....</b>                               | <b>III</b> |
| <b>ABSTRACT.....</b>  | <b>1</b>   |
| INTRODUCTION .....  | 1          |
| METHODS .....   | 2          |
| RESULTS .....   | 3          |
| CONCLUSION.....   | 3          |
| <b>INTRODUCTION.....</b>                                    | <b>4</b>   |
| <b>BACKGROUND .....</b>                                     | <b>6</b>   |
| <b>LITERATURE REVIEW .....</b>                              | <b>8</b>   |
| BIOINFORMATICS TASKS AND CHALLENGES .....                   | 8          |
| QUALITATIVE WORK IN BIOMEDICAL INFORMATICS .....            | 10         |
| QUALITATIVE WORKFLOW ANALYSIS IN SOFTWARE ENGINEERING ..... | 12         |
| <b>LABORATORY ORGANIZATION.....</b>                         | <b>15</b>  |
| <b>TECHNICAL BACKGROUND.....</b>                            | <b>16</b>  |
| <b>MICROARRAY GENE EXPRESSION ANALYSIS OVERVIEW .....</b>   | <b>17</b>  |
| EXPERIMENTAL VARIATION.....                                 | 17         |
| EXPERIMENTAL DESIGN.....                                    | 18         |
| BIOLOGICAL PREPARATION.....                                 | 19         |
| SAMPLE ACQUISITION.....                                     | 19         |
| SAMPLE PREPARATION .....                                    | 21         |
| RNA ISOLATION AND EXTRACTION.....                           | 21         |
| MICROARRAY SLIDE PREPARATION .....                          | 22         |
| SLIDE SCANNING .....  | 23         |
| FEATURE EXTRACTION .....                                    | 23         |
| <b>METHODS .....</b>  | <b>27</b>  |
| DESIGN .....  | 27         |
| SETTING.....  | 27         |
| SOFTWARE.....   | 29         |
| SUBJECTS .....  | 31         |
| DATA COLLECTION.....  | 31         |
| OBSERVATIONS.....   | 32         |
| INTERVIEWING.....   | 32         |
| OBSERVATIONAL SHADOWING.....                                | 33         |
| DATA ANALYSIS AND INTERPRETATION .....                      | 33         |
| <b>RESULTS/DATA ANALYSIS.....</b>                           | <b>35</b>  |
| ROLE-SPECIFIC APPROACH TO SOFTWARE TOOLS.....               | 35         |

|   |           |
|---|-----------|
| INFORMATION SYSTEMS AS MEASUREMENT TOOLS.....           | 39        |
| EVOLUTION OF SEMI-STRUCTURED INFORMATION SUMMARIES..... | 43        |
| <b>DISCUSSION.....</b>                                  | <b>46</b> |
| TASKS AND TOOLS.....                                    | 47        |
| WORKFLOW.....   | 48        |
| LIMITATIONS.....  | 50        |
| FUTURE RESEARCH.....                                    | 51        |
| <b>CONCLUSIONS.....</b>                                 | <b>53</b> |
| <b>REFERENCES.....</b>                                  | <b>54</b> |
| <b>LIST OF TABLES.....</b>                              | <b>56</b> |
| <b>APPENDIXES.....</b>                                  | <b>60</b> |
| APPENDIX A: INTERVIEW QUESTION GUIDE.....               | 60        |
| <i>Personal History.....</i>                            | 60        |
| <i>Daily/Weekly Workflow.....</i>                       | 60        |
| <i>Research/Communication.....</i>                      | 60        |
| <i>Technology.....</i>                                  | 60        |
| <i>SOP/Protocols.....</i>                               | 60        |
| <i>Bioinformatics data management.....</i>              | 61        |
| <i>Bioinformatics Software and equipment.....</i>       | 61        |
| APPENDIX B: PROJECT INFORMATION SHEET.....              | 62        |
| APPENDIX C: SAMPLE INTERVIEW TRANSCRIPT.....            | 63        |
| <i>Biostatistician 1.....</i>                           | 63        |
| APPENDIX D: PRE-IDENTIFIED THEMES.....                  | 72        |

## **Acknowledgements**

I would like to thank the staff of the Toxicogenomic and Proteomics laboratories of the local university where I conducted my study. I would also like to thank my thesis committee, in particular Christopher Dubay, for being a supporter of my strengths in systems and an advocate of qualitative work in this field, and Bill Smith for helping me convey the ideas in written form. Joan Ash has been invaluable as a qualitative researcher resource, and I greatly appreciate her frankness and encouragement. Thanks also to Ken Guappone, who reviewed a subset of my raw data and provided a much needed second set of eyes.

I would also like to thank Karen Eden. Though she was not on my thesis committee, her support throughout the MS program was invaluable. My fellow students were also of great help in this process, and I hope to work with them again. Finally, I want to thank my fiancé, my family and friends for reassuring me of the relevance of this degree.

## **Abstract**

### *Introduction*

Modern biological research is increasingly dependent on information technology. The Human Genome Project, together with increased computer availability, has catalyzed the development of a range of tools designed to deduce and catalog genetic information. Evolutionary versions of these tools are being incorporated into academic and commercial laboratories around the globe in order to better mine and extend biological genomic data. Institutions both small and large are regularly generating and contributing huge amounts of biological data to the already vast resources available on the Internet. In order to participate in this sort of research, laboratories are faced with a variety of decisions when selecting specialized scientific information tools. In many cases, this first requires an evaluation of roles and goals within the laboratory specifically as they relate to being most scientifically competitive within their respective fields.

Clearly defined information management and data analysis roles are relatively new to traditional biological research, which is experiencing an “information revolution”. Data management strategies must now be increasingly planned in advance, and strenuous efforts must be made to anticipate and standardize experimental data organization. Researchers are faced with coupling the complex problems unique to biology with the domains of computer and information science, statistics and business. Ultimately, the need for tools and data management methods for genetic-related research is an understood necessity, but how to evaluate such tools, and what additional roles and models will be needed to successfully implement them is still undefined. By evaluating

the perceptions related to experimental workflow within newly formed academic toxicogenomic and proteomic research laboratories during the implementation phase of a commercial gene expression analysis system, lessons may be learned which could aid in the design and implementation of future bioinformatics systems in similar environments.

### *Methods*

The purpose of this study was to evaluate the information workflow and roles related to the use of a commercial gene expression analysis system in toxicogenomics and proteomics research. The study used qualitative techniques combining both informal interviewing techniques with in-lab participant observations<sup>1,2</sup> to gather data over a six month period. It used the responsive/illuminative approach to identify researcher and staff perceptions relating to information use of this type of system. Seven laboratory employees attached to the toxicogenomic and proteomics laboratories were asked to participate in this study, which comprises the current inclusive range of information and technical roles necessary for experimental research within the laboratories. Each individual was considered as an individual case study and was analyzed in comparison to several organizational themes defined in advance. The group of individuals was also considered as a functional team involved in collaborative research, and their approaches to specific collaborative requirements of the lab related to experimental information management are described.

## ***Results***

The analysis of the coded interview data together with the observational field notes yielded information as to how the laboratory staff collectively approached the technology and requirements of their research work. Considered as themes, these issues fell into areas relating to role-specific approaches to information tools and evolution of information summaries from the use of such tools. Considered separately, each individual provided a unique viewpoint on the information needs and workflow within the laboratories, even when comparing similar roles. The evolutionary aspect of the communication surrounding experimental results suggested that more mentorship could be provided for the data analysis and presentation roles of the researchers, akin to the mentorship that is customarily given to education on specific laboratory technical task.

## ***Conclusion***

Small research teams are close knit and depend on collaboration and open communication in order to maintain appropriate quality control and insight into their experimental research. The use of the longitudinal qualitative methodology was helpful in identifying how the information needs of researchers within an academic biological laboratory are evolving. Observing the evolving experimentation process from beginning to end within the growing framework of information support tools provided a vantage point for identification of the positive and negative aspects of the information communication process. These findings have the potential to help small laboratories implement similar data analysis systems.



## **Introduction**

The purpose of this study was to evaluate the impact of bioinformatics software tools and roles on the experimental workflow of an academic research laboratory involved in genetic research. Many internal and external forces drive research in a laboratory environment, and any study of a laboratory environment can only be considered a temporal “snapshot” of its functions at a certain point in time. Internally, these forces can include both social and technical factors, such as the experiences and intent of the primary investigator, the progress of a specific set of experiments, or the skills and communication of the research staff relating to particular research material, equipment or techniques. External forces can come in the form of reaction to new peer reviewed published research, the requirements and availability of funding or equipment, or relationships to other academic, industry or government groups involved in similar research.

Laboratories involved in modern genetic research now have the option, and perhaps the requirement, to evaluate and incorporate a wide selection of hardware and software analysis equipment to aid in their work. Much of this equipment is “research-technology” or “multipurpose devices for detection, measurement and control that were conceived and developed by a community connected to both science and industry”.<sup>3</sup> In the past this equipment may have improved a single specific process within a laboratory, such as a new centrifuge or set of reagents. However, the unique information requirements of genetic research increasingly involve systems that both replace entire

systems of equipment while simultaneously requiring new non-traditional roles and expertise to maintain and fully utilize.

## **Background**

The intense public and professional interest in genomics that followed the success of the Human Genome Project spawned an enormous amount of data and technologies that are now readily available to researchers and industry. The ability to explore genetic level biology is now critical to any current biological research. The increased information complexity is driving the need for information management specifically to aid this form of research, requiring academic laboratories to evaluate not only how they do research, but how best to optimize that process in order to continue to do good science and, more practically, to remain competitive for funding.

One such technique, Microarray Gene Expression Analysis (MGEA), is a methodology increasingly in use in genomic laboratories. It is helping scientists seek the genetic basis of disease by allowing detailed comparison of which genes are expressed or repressed in the scrutiny of two or more samples of biological tissue. Because it can be used to explore genetic profiles of disease, it is increasingly being used to help make prognostics of disease course by correlating them with clinical outcome. A recent study of breast cancer has been used to identify a “gene expression signature” which can predict clinical outcome of breast cancer.<sup>4</sup> The article states, “The gene expression profile will outperform all currently used clinical parameters in predicting disease outcome.” It is widely anticipated that similar profiles may be identified that can aid in other disease areas.

The area of focus of MGEA within the laboratory I studied is toxicogenomics, defined as “The study of how genomes respond to environmental stressors or toxicants”.<sup>5</sup> The (neuro) toxicogenomics laboratory has the mission of using MGEA and protein-profiling technologies to investigate how cells of various organisms respond to environmental factors. “Exposures early in life may result in health disorders evident at birth, during development or in later life” says the senior scientist and director of the center. “A greater understanding of the impacts of natural and synthetic chemicals on human development will minimize illness in the population and provide for a healthier workforce.” A fundamental goal of the toxicogenomic laboratory is to attempt to identify biomarkers of exposure to neurotoxins present in industrial solvents.

## **Literature Review**

Inquiry into the implementation effects of software systems on the workflow of biological laboratory research work is underrepresented in the bioinformatics literature. The few specific works that involve qualitative inquiry within a biological laboratory focus more on philosophy of science or process modeling rather than impact of information tools.<sup>6,7</sup> In order to design and evaluate the study, I researched four areas: organizational behavior theory<sup>8</sup>, qualitative analysis in clinical healthcare<sup>1,2</sup>, qualitative analysis and project workflow in software engineering<sup>9</sup>, and bioinformatics specific literature appearing in both biology and commercial journals.<sup>10-17</sup> The first three of these areas focused on research that used qualitative methodology to study small groups of highly technical individuals, and focused on communication, roles and workflow. The bioinformatics literature specific to MGEA was frequently consulted as to be able to put the rapidly changing dependencies of MGEA research and bioinformatics tools into context.

### ***Bioinformatics Tasks and Challenges***

Of the large and rapidly growing body of literature specifically devoted to MGEA issues, most is concerned with evaluation and extension of statistical analysis techniques<sup>14,18</sup>, reviews of available software systems<sup>11-13</sup>, or applied research in which MGEA technologies are mentioned only as they relate to the quantitative methods. The most relevant available information relating to MGEA use within a generalized laboratory facility comes in the form of technical protocols that are based on specific products or software packages<sup>10</sup>. These protocols are typically time sensitive, must be updated

regularly to remain useful, and tend to focus on specific portions of the MGEA process – mostly due to the mix of tools, techniques and individuals that are unique to every laboratory. As a result, they are a critical component of the research, but typically exist within a relatively narrow portion of the overall research workflow.

The management of genomic data is a common concern within MGEA. One author recently stated that scientists often express some confusion, a tendency to focus on what is already known, and a sense of weariness when confronted with the full range of tools and data available for any specific experiment.<sup>19</sup> To this end, a range of commercial tools have evolved to help manage the data, as well as increasing standardization within the experimental processes themselves.<sup>20</sup> These tools and standards will undoubtedly have an impact on current and future labs attempting to extend research in this field, though most of the technologies are targeted towards biologists with strong computer and statistical skills. The academic research into Information Overload (IO) is attempting to define methods in order to be able to measure the degree by which individuals are experiencing IO, so to be able to provide organizational models and tools for minimizing it within overwhelmingly complex information environments – of which genomic research is an obvious candidate.<sup>15</sup> Some researchers within the biological fields do not agree that there is an issue however; states Nicholas Cozzarelli, editor of the Proceedings of the National Academy of Sciences (PNAS) “I think it’s great ... I don’t feel there is information overload. The number of researchers has increased, and accuracy has increased, and the number of things to attack has increased”.<sup>21</sup> This perception undoubtedly varies from institution to institution, based on a range of factors.

### *Qualitative Work in Biomedical Informatics*

A growing niche of biomedical informatics involves the qualitative study surrounding the implementation and consequences (both intended and unintended) of computerized systems that support information management and workflow in clinical environments. Systems range in scale and potential impact from analysis of e-mail use between clinicians and patients to the implementation evaluation of full-scale hospital management systems.

A recent study explored unanswered questions regarding the use of email as a physician-patient communications tool, such as “How does this new technology impact the patient-physician relationship”, and “How do physicians integrate this technology into their practice?”. Using qualitative interviewing to gather data from 52 responding physicians (out of an initial sample of 88 who expressed interest), two researchers independently coded the audio taped and transcribed interviews, yielding 642 comments. Several issues emerged from the 4 broadly coded domains, such as improved continuity of care and increased flexibility in responding to non-urgent issues, as well as on-going concern of integration of email into daily workflow. The study concluded that identification of situations where this technology was most effective, coupled with careful integration of e-mail into office flow need to be addressed before the provider-clinician email communication diffuses into most practices.<sup>22</sup>

Computerized Physician Order Entry (CPOE, or alternatively POE) systems implementation has had varying success due to several technological and social factors. Despite obvious benefits, CPOE implementation has often been disruptive, often due to coordination issues between physician/nurse relationships, issues with the actual computer interface, or physical environment issues such as availability of the systems. A recent observational case study was conducted to evaluate information roles within an ICU which were clearly defined prior to the implementation of the system, with the desire being evaluation of post-implementation.<sup>23</sup> In this particular study there were a significant number of individuals to observe, in this case 50, which included physicians, nursing staff, pharmacists and a respiratory therapist. The researcher identified three conceptual themes that were specific to the ICU workflow implementation: *coordination redundancy*, *computational interface* and *work location*. Adaptation of this study to my research area was difficult due to the lack of advance knowledge of clear roles, the changing state of the research workflow dependent on these tools, the “formative nature” of the laboratory itself, and the small number of subjects I observed.

A broader cross-site study of CPOE implementation evaluated multiple perspectives of users from the vantage point of an external, multidisciplinary team unaffiliated with any of the sites under observation. The use of the multiple perspectives approach allowed the researchers to view a variety of overlapping aspects relating to technical, organizational, personal and environmental issues. The team met repeatedly to develop the patterns and themes encoded from the study, data gathered from observational shadowing, interviewing and focus groups. They also enlisted an objective outsider to independently



validate the resulting themes taxonomy. They proposed ten themes and 24 sub themes surrounding perception of a successful implementation of CPOE, most of which reflected aspects the multiple perspectives organizational framework initially identified. Their conclusions are specific to the CPOE situation, but they suggest two items relevant to my work, that users should be included in setting requirements in designing and prototyping such systems, and that in order for a system to be meaningful to a user, the organization of data must be local and individualized.<sup>24</sup>

Similar informative studies have been accomplished with smaller groups by a single investigator. In her 2002 Masters thesis, Erica Danielson assessed “Changes in Nurses’s Workflow in Response to the Implementation of an Electronic Charting System” through a study of five nurses at OHSU.<sup>25</sup> Danielson’s approach used a combination of observational shadowing and interviews to evaluate the workflow impact of the implementation of this system, and yielded three categories: *communication issues*, *technical issues*, and *true changes in workflow*.

In summary, qualitative research within biomedical informatics domains has evolved a strong foundation of techniques that can be applied to the study of information workflow in similar professional work environments of biological research laboratories.

### ***Qualitative Workflow Analysis in Software Engineering***

Methods from the academic study of software engineering work teams can be applied to this biological research work environment for the following reasons: both research

laboratory staff and software engineering teams are often small, highly specialized and closely interconnected, and are often coping with the construction of a particular work through rigorous quality control mechanisms. Their technical environment is also changing frequently, and yet their internal workflow and communications processes are highly dependent on human behavior. As much software engineering takes place inside a person's head, gathering descriptive data can be difficult. Seaman, together with the Software Engineering Laboratory (SEL), which is sponsored jointly by NASA/Goddard Space Flight Center, Computer Sciences Corporation and the University of Maryland, suggests three qualitative methods for evaluating software engineering teams.<sup>26,27</sup>

- *Meeting Observation:*

Since much of the communication in software engineering occurs within meetings, Seaman suggests having two observers conduct strict non-intrusive observation within the meetings, then “cross-code” their results. Much the same, within laboratories the weekly meetings serve the crucial point at which experiment design and progress is discussed.

- *Participant Observation:*

“Think aloud protocols” require the subject to verbalize his or her thought process such that the observer can understand the process going on. However, “such protocols are limited by the comfort level of the subject and their ability to articulate their thoughts”. Within a laboratory this may be different, as much of biological science is learned in a collaborative educational environment, and virtually all

researchers have “rotated through” various laboratories in order to learn specific techniques in one-on-one encounters.

- *Participant Interviews:*

Semi-structured interviews designed with the interview having foreknowledge of some portions of the projects were most useful for generating descriptive data.

Seaman strongly suggests that due to the technical nature of the data acquired from such interviews (and this also applies to the other two methods mentioned above), the qualitative coder should be fairly experienced in the vocabulary and technical domain of the subject being studied.

Seaman also speaks of the need to study “non-technical issues and the intersection between the technical and non-technical in software engineering.” Though the published scientific literature is relatively small in regards to peer-reviewed case study analysis of these groups, the industry press itself has generated a number of books relating to what factors contribute to software projects success or failure.<sup>9,28</sup> These case studies are often clear models of similar environments, but in many cases suffer from the lack of an observer enmeshed within the projects themselves, and are usually described in the traditional (to software engineering) “post-mortem” style.

## **Laboratory Organization**

Making sense of genomic data on a per-experiment level requires the combination of highly specialized skills. Typically overseen by a Principle Investigator (P.I.) with a PhD or MD (or both), an academic biological research laboratory employs a diverse group of individuals. The roles deemed necessary to do genomic research include biological researchers skilled in a specific domain area (such as toxicogenomics or neurobiology), laboratory managers, laboratory technicians, biostatisticians, information technologists, and bioinformaticians. In academics these roles are typically filled by a combination of faculty and staff, as well as undergraduate, graduate and post-doctoral students. All of these individuals contribute to the tending of a collective set of data, the sole goal of which is to accurately describe one of several specific experimental scenarios. In order for this data to be considered scientifically valid, it and all the steps leading to it must be reproducible. This requirement of reproducibility has become critical in the complex experiments undertaken in genomics or proteomics research.

One difference between commercial companies doing this form of research and academics is that in academia, laboratories typically involve persons “in transit” such as pre and post-doctoral students or other levels of student, who may or may not be with the laboratory for long periods of time. This is usually a rich cultural environment, with a range of backgrounds, ethnicities and communication styles involved in day-to-day collaboration. Thus an underlying foundation of academic research is the constant training and education that is central to the processes of experimental methodology. Mentorship of research techniques and information management processes is an expected

portion of academic research work, which is both aided by new approaches to research that evolve within an academic environment, and hindered somewhat by the need to constantly educate so-called “short-timers”. In a science as information rich as genetic research, tools and methods that can reduce some of the variability induced by constant change are thus highly desirable.

The laboratories studied within the center are in their late inception stages, having been formed approximately 18 months prior to this writing. Both laboratories are for practical purposes independent, but their small staff and close proximity, coupled with their research into different aspects of toxic effects on biological processes warrants a close collaboration. The formalized reporting responsibility structure of the laboratories is detailed in the following diagram by lines, with the overlapping contributions of several members shown as the intersection of the two labs (*Figure 1*)

## **Technical Background**

For the purposes of this thesis, it is helpful to review the complex processes that constitute the physical methodology of generating slides for MGEA. The various roles within the laboratory are involved in various cycles along the iterative experimentation design and analysis process.(Figure 2) The next section is a basic overview of this methodology, and focuses on the toxicogenomic laboratory.

## Microarray Gene Expression Analysis Overview

Broadly speaking, MGEA concerns the comparison of two or more biological tissue samples to whether the genes expressed in a treatment sample differ significantly from the genes expressed in a control sample. This is accomplished in two distinct phases, the laboratory experimental procedures, and the data interpretation procedures. The former is heavily dependent on modern biological procedures and equipment, coupled with highly controlled experimental protocols, in order to generate experimental results that serve as data sources for the latter phase. An entire process from biological sampling to gene expression profiling, if run continuously during normal working laboratory hours, could be accomplished in about three full days. A typical experiment involved in such a process would generate anywhere from two to twenty or more gigabytes of data. It is likely that in the future this process will be further optimized, generating ever more enormous amounts of data in shorter periods of time. The laboratory proceeds in three generalized stages, within each of which there is a great deal of opportunity for variation, and each of which is highly sensitive to steps leading up to it.

### *Experimental Variation*

A defining characteristic of MGEA is the constant diligence required in order to minimize variation in each of the steps leading to extracted data. A workflow diagram in the list of tables illustrates some of these dependencies, which shows the major steps in a typical experiment as well as the dependencies that need to be considered in order to maximize the resulting data quality (*Figure 3*). Each of the is now briefly described in more detail.

### *Experimental Design*

This, the first stage of MGEA and indeed most laboratory experimentation, has evolved such that it is now the subject of graduate level statistics courses. Since the amount of data generated from such experiments is vast, the roadmap of the experiment must be exhaustively planned so that a directed focus can be maintained throughout the experiment. This process has to provide clear answers for the (non-exhaustive) list of questions:

- What is the experimental hypothesis?
- What conditions are to be compared?
- How many subjects will be needed?
- How and when will the tissues be acquired from these subjects?
- Will there be adequate tissue for each comparison to provide for good statistical power and sample size when processed?
- What type of microarray slides will be needed, and how many?
- Does the experiment design provide for redundant slides in case of damaged or faulty slides?
- What reagents and equipment will be used in the preparation and analysis of the slides?
- What quality assurance steps will be undertaken during the process, when, and by what criteria will they be evaluated?

This stage requires input from virtually all of the roles within the laboratory and is a familiar to researchers across the domains of science. The unique requirements of MGEA, however, have elevated the roles of the biostatistician in such work, as careful attention is required to minimize the identification of false patterns in the amassed data at later stages. Since this is largely one of design, the variation incurred at this stage, if approached correctly, is minimal. This stage is unique as it has no set time guidelines, and can thus take from minutes to months to clearly define.

### ***Biological Preparation***

This is also known as the “wet-lab” stage. In the case of laboratories working with animals, this concerns the physical management and subsequent harvest of tissue from the biological subject, followed by the RNA extraction and the preparation of the microarray slide. If working with human beings, this might involve biopsy samples from specific organs, but the processing steps would be similar.

### ***Sample Acquisition***

Biological material processing is a traditional laboratory bench work role, but is made more rigorous in MGEA by the need to follow strict protocols in order to minimize random human, equipment or process induced variation. The end result of this procedure is the generation of the primary physical MGEA experimental product - the microarray slide. These slides, coupled with the documentation of steps leading up to their production, are the source of most subsequent raw data in this form of research. As a result they exist at the conceptual juncture of the physical laboratory and the information world. The process surrounding preparation of these slides and the related protocols



undergo regular improvements as more insight and technical improvements are disseminated within the research community.

The care and handling of the biological sources, which in the case of the toxicogenomic and proteomic laboratories involves mice and rats respectively, is the first major source of variation to which MGEA is sensitive. For this reason the protocols for animal handling are extremely rigorous as to how the animals are handled, the conditions to which they are exposed, the methods by which samples are extracted, and the process by which the samples are stored. Since human understanding of genetic effects is still largely inferential, steps are taken to ensure that the ability to compare the sample of interest at resulting stages is maintained. In addition to control animals, for example, there are “sham” control animals, which do not receive the control treatment, but merely undergo the trauma of a sharp needle. Samples from these “sham” animals will ultimately allow their comparison against the “true” controls in order to isolate any hypothetical gene expression purely related to any metabolic reaction of being repeatedly injected.

The toxicogenomic animal harvesting laboratory is depicted in the list of tables (*Figure 4*). The typical equipment is very basic, a humane guillotine, scalpels, sterilization supplies and a range of test vials for material storage. The primary protocol adaptation of this process to accommodate MGEA concerns the speed in which it is done (approximately 6 minutes per animal by my watch), and the care to avoid any animal-animal contamination between extractions. A MGEA specific solution known as

RNAlater® is used to store the individual samples by inactivating RNases and stabilizing the RNA within the tissues or cells. This helps eliminate variation that might occur in tissue degradation that could become an issue later in the process. This process is generally performed as quickly as possible for a given set of experimental samples in order to isolate changes that might continue to occur in treated biological samples.

### *Sample Preparation*

The sample preparation steps are all performed in the toxicogenomic laboratory, a self-contained area the size of a small house containing approximately a million dollars worth of modern gene expression analysis equipment. It is depicted in the list of tables for clarity (*Figure 5.*)

### *RNA Isolation and Extraction*

RNA isolation from biological samples is now typically performed with commercial kits and recommended protocols for their use. This helps ensure that the material is prepared in a consistent manner, which again helps isolate variation later in the process.

The RNA isolation procedure follows several steps that vary slightly depending on the kit and sample in question (for example, liver or brain).

- **Homogenize** tissue or cells
- **Centrifuge** material
- **Extract RNA**
- **Store RNA** for later slide preparation.

Quality checking is done at regular stages by equipment such as the NanoDrop®<sup>29</sup> Spectrophotometer and the Agilent Bioanalyzer<sup>30</sup>, coupled with analysis software designed specifically for each device. Equipment used in these steps is state-of-the-art, and provides relative ease for in-process tests that identify variance early in the process. This stage of processing a given set of samples lasts approximately 12 hours.

### ***Microarray Slide Preparation***

Preparing the microarray slides from the extracted RNA is the last physical step prone to variation from environmental or human factors. The choice of what style of slide and what subsequent scanner is used determines the protocol by which the slides are created. In the case of the toxicogenomics lab, the primary supplier is Agilent, which provides 60 base pair oligonucleotide mouse chips with 22,000 (22K) spots per slide. The slide preparation procedure follows protocols specific to each chip and is highly sensitive to variation. The specific protocol at the time of this writing is the Agilent Low RNA Input Fluorescent Linear Amplification Kit, and the slides are Mouse G4121A.

The processing steps are:

- **Prepare pairs of samples RNA** (determine concentration)
- **Amplification:** cDNA synthesis from RNA for each sample
- **Fluorescent labeling** of samples (different color for each sample)
- **Purify Amplified RNA**

- **Combine** pairs of samples together
- **Hybridize** samples or “probes” to microarray chip
- **Wash and dry** the microarray slides

The end result of this is a microarray slide on which each dot represents a specific gene and its activity as a reaction to the probes. This process takes approximately 24 hours to complete, of which the majority of time is devoted to the 17 hour microarray hybridization procedure, which is does not require monitoring.

### *Slide Scanning*

This process is concerned with identification and measurement of the spots, or genes, which have been expressed on the slides in question. This is also the point at which traditional biological methodologies are completed - from this stage onwards the experiment is pure data and subject to interpretation only. Slides are placed within a microarray scanner and shuttled over a stationary laser which records the intensity of each labeled fluorescent probes. This generates a visual image comprised of a combination of red or green dots, which is accompanied by a detailed manifest of physical characteristics identified, such as intensity, size, location, etc.

### *Feature Extraction*

The basic goal of this step is to identify the genes that are highly expressed from the raw images, and then extract the gene ID’s that those spots represent from the layout file provided by the manufacturer, resulting in a list of genes that were hypothetically affected by the treatment under scrutiny. This process can be performed many times in

order to optimize the statistical methods used to evaluate the data to the particular slides and extract the most statistically significant data from the images. There is an emerging discipline of statistics devoted to this stage, as well as a range of commercial tools offering integrated solutions for the data extraction and management.

The processes within this stage are roughly stated below, but may occur in different orders based on whether they are incorporated within a specific software tool, package, or if they are performed manually.

- **Visual Inspection**

A cursory step taken to see whether any extreme physical artifacts have degraded the quality of the slide, such as scratches, washing or drying issues.

- **Normalization**

Undertaken to determine whether any portion of the preparation and scanning of the slide yielded unusable data, and how to optimize the image quality for that which is usable.

- **Feature Extraction**

The measurement of relative spot intensities are to determine which genes were expressed or repressed, and to what degree.

- **Expression Analysis and Filtering**

A high level review of the extracted features to identify thresholds by which subsets of “interestingly significant genes” can be isolated.

- **Annotation**

Linkage of contextual information with the identified genes, such as roles played in genetic pathways, protein products, and literature regarding the genes in question.

- **Interpretation**

The fundamental human science portion of the process where true knowledge is extracted from the datasets.

After normalization, the spot intensities are processed to determine which genes were expressed or repressed, and to what degree. This will generate a large list of data with a probe ID that refers to a specific gene. The list of genes is then filtered to isolate only those genes that were expressed or repressed to a significant threshold, a measurement which is commonly used is whether the gene is 1.5 fold changes over or under it's unexpressed state, and to what degree of confidence this is expressed. This yields a list of highly expressed genes.

This gene list is then annotated based on the known function of each gene, which can be used to gain more information about the genes of interest as needed, such as the protein products, participation in metabolic pathways, or inclusion in published papers on the topic. This final stage is often one of collaboration, with the researchers specialized in various biological domains contributing to the teasing of information from the set of genes and associated information. This is aided by the increasing sophistication of tools that can integrate various genetic information sources together into comprehensive data sets.

These last three steps can be accomplished in a variety of ways and can consequentially generate a range of information: via a collection of statistical analysis techniques within a common spreadsheet such as Microsoft Excel together with freely available Internet genomic database searches, via a collection of specialized statistical packages such as the freely available BioConductor package<sup>31</sup>, or via a comprehensive analysis system such as Rosetta Resolver. Each option has advantages and disadvantages depending on the scope of the work within a specific laboratory, as well as other more subjective factors.

## **Methods**

### *Design*

This was a longitudinal qualitative study of a group of individuals involved in quantitative research within the emerging research domains of toxicogenomics and proteomics. The data was gathered using the illuminative/responsive methodology of informal semi-structured interviewing techniques, participant observation in laboratory meetings and direct observation during shadowing of laboratory researchers engaged in specific tasks. This research strategy is appropriate when the possible hypotheses are both difficult to measure quantitatively, and the goal is a descriptive understanding of a complex environment in context.<sup>32</sup>

### *Setting*

The study was conducted with staff working within the toxicogenomic and proteomics laboratories of a research center within a local university. The center has more than 100 scientists and research staff and is involved in basic and applied research into preventing illness and disability related to hazardous chemicals and their effects.<sup>33</sup> The two laboratories are intended to be self-contained, but have close collaboration with each other on experiments. Members of the proteomics staff participate in all toxicogenomics meetings, experiments and forward planning.

The toxicogenomic laboratory is one of five research centers selected and funded in 2001 by the National Institutes of Health (NIH) and National Institute of Environmental Health Sciences (NIEHS) to pioneer development of the NIH/NIEHS Toxicogenomics Research



Consortium (TRC). This is an initial five-year research project involving the development of the toxicogenomic research field, which at this writing is approximately three years old in its present form. The laboratory is tasked with collaborating with the other institutes to improve the understanding of how environmental and occupational agents adversely impact health.

The other academic partners of the TRC are Duke University, Fred Hutchinson Cancer Research Center (University of Washington), Massachusetts Institute of Technology and the University of North Carolina – Chapel Hill (*Figure 6*). The commercial partners are Paradigm Genetics<sup>34</sup> and Science Applications International (SAIC)<sup>35</sup> which are responsible for support ranging from specialized equipment and protocols to maintenance and development of the Chemical Effects on Biological Systems (CEBS).<sup>36</sup>

The toxicogenomic laboratory has been operating since late 2002, and considered itself at full operational status in July of 2003. Until recently it has been performing its microarray analysis work with a combination of its own internal scientific resources coupled with, as needed, external expert collaboration. It has spent the formative period of its existence establishing SOP's and experimental protocols for itself while initiating early experimentation. Since the study of toxicogenomics involves the chemical agents effect on biological systems, the highest priority of the researchers is to ensure that the effects being studied are truly due to the toxins themselves as opposed to some confounding influence.

## *Software*

The toxicogenomics laboratory uses several primary commercial software packages. Of these, several are closely tied to laboratory equipment and would not function without them. During the wet laboratory stage, two pieces of software are used to drive the two primary pieces of measurement equipment that allows the laboratory technicians to ensure RNA. Respectively, these pieces of equipment are known as the NanoDrop and the Agilent Biolanalyzer. The Agilent microarray scanner is also driven by software manufactured by Agilent. All of these software packages have a Microsoft Windows application interface.

The centerpiece of the information analysis is Rosetta Resolver version 3.0, developed by Rosetta Biosoftware, the software arm of Rosetta Inpharmatics. This software is in use, according to Rosetta's statements, by "over half of the major pharmaceutical companies in the US". It is additionally in use by approximately half of TRC members.

Incidentally, Rosetta Inpharmatics is now a wholly owned division of Merck Pharmaceuticals as of 2001, which is a major player in the commercial pharmaceutical industry. During the gene expression analysis phase, other pieces of software are being used in analysis, most notably BioConductor, based on the R statistical processing language, and SpotFire, a competitor to Rosetta Resolver.

The Resolver software runs on a Solaris system, but provides a Microsoft Windows application for system use. Rosetta Resolver is intended for use in the management of MGEA experiments within a laboratory. It is not intended to be a Laboratory

Information Management System (LIMS). The system provides the following broad functionality:

- Large-scale data storage for gene expression data.
- Comprehensive statistical analysis functionality for gene profile extraction
- Centralized repository of genetic information from a wide variety of sources that can be easily linked to experimental data as needed.
- Definition and storage of experimental protocols
- Wide range of visualization possibilities.

The Rosetta system is implemented in a client/server model with the server running on a Sun Solaris 9 high performance computer, the data storage being through Oracle 9i. Microsoft Windows 2000 workstations access the server via Rosetta client software, which is controlled by a strict simultaneous licensing model. Access to experimental data through the client software is managed by membership in groups defined by the Rosetta administration software, which was suggested by Rosetta but defined by the P.I. and coordinator of the toxicogenomic laboratory. Management of the hardware and Solaris software is performed under an on-going service agreement with Information Technology Group (ITG) of the university; Center staff performs management of the Rosetta Server, with assistance from Rosetta Inpharmatics.

The software and hardware for the Rosetta system each cost approximately \$150,000, with the burden of payment being split mostly evenly between the center for the software and

the ITG for the hardware. Access to the system initially is limited to two “floating licenses”, meaning that up to a maximum of two clients could be connected to a system at any time, but that the number of users within the system was not limited. Pricing for each additional license is approximately \$25,000. Combined with the other analysis software and equipment, the toxicogenomic laboratory has approximately one million dollars in streamlined gene expression analysis equipment.

### ***Subjects***

The subjects were all paid employees of the center, engaged within either the Toxicogenomic or Proteomics laboratories, and will be referred to as “researchers” for the remainder of this document. The Institutional Review Board allowed me to interview a total of seven researchers without requiring informed consent. These researchers were individuals that I had grown to know over a period of four months prior to the start of the study through interactions within the weekly laboratory meetings. Each of the researchers had one or more of the following roles: Principal Investigator, Post-Doctoral Researcher, Laboratory Manager, Coordinator, Laboratory Technician, Biostatistician and Bioinformatician.

### ***Data Collection***

Each researcher was interviewed in November and December of 2003. Each interview was approximately 40 minutes, was audio taped, and followed a semi-structured interview guide focusing on seven main areas. (*See Appendix A: Interview Question*

*Guide*) I transcribed each interview, and stripped the identity of each interviewee to maintain subject confidentiality.

I participated in the weekly laboratory meetings for a period of seven months, occurring both prior to and during the study, which were held for at least two hours every Wednesday. These meetings were not audio recorded but field notes were taken, and a timeline of issues was maintained. (*Figure 7*)

### ***Observations***

I was the sole observer in this study and functioned in two separate roles. During the weekly meetings, I functioned in a participant/observer role in order to keep track of the needs and requirements of the experiment and level work. During the shadowing of laboratory bench work or analysis, I functioned in a passive role and tried my best to keep out of the way of the researchers I was shadowing.

### ***Interviewing***

All potential interview subjects were initially contacted by email or in person and asked if they would be interested in participating. Upon scheduling an interview, I provided a basic outline of topic areas (*See Appendix B: Project Information Sheet*). Each interview was held in a comfortable environment without outside distractions. Each interviewee was tape-recorded; additional interview field notes were also taken. The set of interviews lasted approximately 40 minutes each. A sample transcript of an interview is included in (*See Appendix C: Sample Interview Transcript*)

### ***Observational Shadowing***

Due to somewhat unpredictable day-to-day nature of this form of research, observational shadowing occurred with prior approval of the P.I. and laboratory manager such that it could be coincided with specific events in the experimental process. Unplanned shadowing of each subject was not practical, as observing researchers writing papers was not illuminative, nor desired by the researchers. Instead, the shadowing focused on three physical locations: the laboratory itself, the laboratory meeting room and the animal-harvesting laboratory. Each shadowing lasted between one and four hours.

### ***Data Analysis and Interpretation***

Transcriptions of the interviews were entered into the qualitative analysis software (QSR NUD\*IST version 6), and coded based on three systems, in the following order:

- My perceptions relating to the structure of the questions in the interview guide.  
*(Appendix A)*
- My perceptions relating to list of six themes identified prior to the interview portion of the study. *(See Appendix D: Pre-identified Themes)*
- Emergent themes either not identified, at intersections, overlap or union to the prior described coding systems.

A random sampling of interview transcripts was provided to a pre-doctoral student in medical informatics who had some training in the methods of ethnographic research,

along with the list of pre-identified themes (*Appendix D*). The student independently coded the transcripts based on his perceptions, which were then included into the Nudist software node structure. A weakness of this study is that its interpretative nature would have been greatly aided by a full-time co-investigator able to concur or not concur with my observations, but this was not possible within the scope of this study.

## Results/Data Analysis

Several re-occurring themes relating to information use, bioinformatics software systems and workflow emerged from the analysis procedures.

- Role-specific approach to software tools
- The treatment of information systems as measurement tools
- Evolution of semi-structured information summaries

### *Role-specific Approach to Software Tools*

Approaches and relationships to the information technology software were closely related to the roles and responsibilities of each individual within the lab. While the more specialized interfaces of measurement tools such as the NanoDrop and Biolalyzer were seen by all the researchers as essentially stand-alone instruments, the Rosetta Resolver system was considered variably as an analysis system, a data storage device, or a somewhat nebulous entity requiring extensive training and dedicated time to use.

Both P.I.'s saw the Rosetta system as a necessary piece of equipment and were (at least initially) excited about the advertised capability being able to measure and compare large numbers of experimental results. However, neither P.I. expected to use the system directly.

*P.I. 1:* "Well, I want to find out what it can do, but I never imagined that I could be using it all by myself and make sense of the data. It is probably too much to handle" (#254)



The biostatistician also saw the system as powerful and interesting, but was reserving judgment as to whether he would depend on it as much as was expected, as he was more familiar with other analysis tools. While he was expecting to use it, was required by the P.I. to use it, and was awaiting formal training on it, he was not convinced that its underlying statistical methods could be encapsulated into “one-click” functionality. He indicated that he would continue to use his preferred validated statistical methods for analysis. When interviewed he had not yet been trained on the system, though he had been asked to validate some data output from the system against his own methods. He was somewhat professionally skeptical but was reserving judgment until he had a better opportunity to compare future experimental data.

*Biostatistician:* “There are things that seem to be a lot of, well - we use this because it seems to work, not because there is a theoretical basis as to how error propagates from one source to the next. In statistics we know how error propagates and we can build it in there in a statistical way, but that is not what they seem to be doing in all the stages.” (#471)

The laboratory manager and toxicogenomics technician were most directly related to the system and all the laboratory steps leading up to it, and it was upon them that the initial setup and use fell most heavily. They were increasingly enthused as data output from the system sparked collaborative discussion within the meetings. Of the laboratory workers, and apart from the biostatistician, they had the strongest computer skills, as well as the

most direct relationship with the experimental process, and were perhaps the in the best position to see benefits of the system on their workflow.

*Lab Manager:* “It's fairly straight forwards, as long as you understand what you want to use it for.” (#487)

The post-docs and the proteomic lab staff expressed optimism and enthusiasm for the technology, but were more concerned with getting results of their own experiments than exploring the system. They presumed that they would be trained on the system at some point, but did not see the immediate relevance to their own work. Additionally, the data of their own experiments was not imported into the system, which gave them little initial reason to use it. Due to the relatively long duration of time between the system being installed and their “formal” training, most lost interest or were busy with other work only peripherally related to the capabilities of the system. To this point, most of the statistical work had been handled via a relationship with either the biostatistician within the laboratory or via another statistician located elsewhere on campus, and they were comfortable with this relationship.

Each of the researchers considered the Rosetta system as being interesting and useful, but few (with the exception of the laboratory manager and the toxicogenomics technician) were prepared to claim ownership. Most were not encouraged to use the system on their own and were told they needed to wait until the next version (4.0) was released, when additional formal training would be provided - which was scheduled for several months in

the future. This tended to have the Rosetta Resolver system adopted by those who had the most time to use it, and who were closest to the day-to-day experimental process.

### *Information Systems as Measurement Tools*

The researchers tended to consider software systems primarily as tools or equipment by which a biological characteristic could be measured. They were extremely comfortable and proficient with the highly specialized laboratory equipment that could be calibrated and evaluated, but less so with the extensive capabilities contained within the Rosetta system, or increasingly required by bioinformatics and biostatistical analysis. This appeared to be at least partly due to the requirement of reproducibility inherent in MGEA experimentation, and partly because the newness of the science itself tended to encourage focus on specific portions of the experimental process and less on the broader data abstraction possibilities of system such as Rosetta Resolver. The researchers were generally more content with interfacing with the human biostatisticians, who manually helped them focus and validate their experimental design and generate a small subset of “distilled” data.

*Biostatistician:* “...we would get talking and we would discover that there were quite a few more issues that need to be addressed, that there were more levels to their experiment than what they thought. They had different animals, they have experiments stretching over different days or something so there are things that need to be addressed, then they'd come and we'd sit down and talk about it...”

(#127)

The laboratory had been constructed, in part, by the implementation of somewhat traditional yet specialized laboratory instrumentation adapted for specific purposes

appropriate to MGEA workflow. When approaching the acquisition of the Rosetta Resolver system, the primary concerns discussed in meetings by the majority of the staff related to how it would be supported within the laboratory, with the implicit expectation that once installed and tested, it would be utilized as a sort of data microscope (or “macroscope”), and applied when and where necessary. The researchers, and perhaps the vendor itself, were not as concerned with the workflow integration issues resulting from this type of software system, assuming (perhaps correctly) that the process by which the system would be used would be refined several times and that discussing cultural integration effects was at best premature, and at worse naïve. The solution to most concerns brought up in the meetings was that they would be handled “in the training”. The exception to this view came from the statistician, who expressed some concern regarding the general trend towards “one-click” analysis and felt that more emphasis should be put on defining expectations of data analysis by the system.

After the system had been installed, and when asked individually what was the greatest specific fear of management for this sort of data, other members of the staff echoed the opinions of biostatistician, particularly after the system was itself installed and early work performed on it.

*Lab Manager:* “I don't think there is a fear. I think it would be very wise for us to just sit down and discuss what exactly we are going to be doing with the data, and how it is best used. There is no fear about it; it is just trying to understand what we do with all of it.” (#320)

days, not just this lab, all labs are not only dealing with variations in biology, they are also dealing with technical variations. Even though my background is genetics, I know that most differences between two species are even from one batch. Each day we get a new technique, causing new variation. So that's why I think we should not only look at biological variation but also technical variation.” (#189)

*Lab Manager:* “Quality control, that would be important to have those in place prior to doing anything with the data. And then after that would be to have a clear understanding of what exactly what tools would be needed to do the task, and then it would be very nice for parts of the data to be looked at by different individuals such that they could take that data and see if it can answer the questions that need to be answered for that particular experiment.” (#268)

*Lab Manager:* “Yes, to a certain degree, we have one experiment which is almost done, and we are checking to make sure that everything is, as they say, ‘i's dotted and t's crossed’ (laughter) and all the data is believable. We don't want to put things up there that are not believable, cannot be reproduced. The most important thing for us is to have data that is reproducible.” (#326)

The biostatistician had a slightly different approach, emphasizing overall management of the data generated by the process:

*Biostatistician:* "...centralizing the data, tracking the changes from start to finish, trying to minimize the human involvement for changes. I see scientists who sometimes go through their data and I tell them 'I need to have you use a uniform type of coding for something', and they go through it and do it by hand, so I still get errors because of type, so using things." (#300)

Overall, when faced with a system that had such potential, the researchers desired a method by which the data used within the system as well as the statistics used for evaluation could be insured of reproducibility. Since the system itself could not guarantee this, and it had never been a core role within the laboratory, several researchers indicated that they would be comfortable with the role of the biostatistician remaining much as it were – as to provide professional expertise to help the on-going education of the system.

*Biostatistician:* "The expertise though, at least here in the center, seems to be one where people identify outside expertise, in terms of a piece of software or a person who can deliver it, and then go and try and secure that, try to make it internal." (#108)

### ***Evolution of Semi-structured Information Summaries***

Communication within the laboratory appeared mostly semi-structured and was focused primarily around the experimental process. It occurred via email, in person or by phone - with no evident preference or pattern apart from the biostatistician who is now located in

Japan and mostly dependent on email. It was, however, widely agreed that the primary information communication tool within the laboratory was the regular Wednesday meeting. Within the meetings the primary information dissemination methods were paper photocopies or printouts of data and articles, accompanied by interpersonal discussion and planning. Computers or other interactive information analysis tools were rarely if ever used. The importance of the regular meeting was often referred to during the interviews:

*Lab Manager:* “The most important part is the discussions in the laboratory meeting. Those are the most important parts because we have to discuss and understand what the problems are and that is the most important part of keeping track and keeping contact.” (#81)

*P.I. 1:* “Typical communication in lab is by regular weekly meeting, which consists of individuals within the lab as well as those interested in the topic on a rotating basis. The lab meeting is the central point to ensure good experimental design.” (#17)

*Lab Tech 2:* “And also we have a weekly meeting. If something unusual is happening we get information from the others.” (#121)

During the meetings themselves, I observed a gradual evolution and refinement of experimental data presentation formats to suit the changing information needs of the



meeting attendees. Initially much of the laboratory work analysis, with the caveat that in the early stages of the experiment there were few results to analyze, was discussed based on domain knowledge of the researchers and topical concerns of the P.I.'s. As the Rosetta system became operational and the workflow kinks were ironed out, discussion became more commonly driven by group interpretation of experimental data - usually lists of genes or results from relevant peer reviewed papers. This experimental data then underwent a further refinement as the laboratory technicians utilized the software tools to provide better methods to visualize the data distilled from the progressing experiments.

The P.I.'s and post-docs became most involved in the discussion when they had a set of clear comparisons to discuss, or when the data was progressively refined to support a direction of inquiry from a previous meeting. When presented with the raw data for a given experiment, the researchers were entirely capable of analyzing it, but it was a somewhat slow process while the basics context was verbally hashed out. It is possible, however, that the change in participation that I noted during the increasing use of output from the Rosetta system may be based on other factors, such as workload, or may be biased by my lack of neurological biological knowledge.

It is again to be noted that despite the fact that this information was increasingly being generated from the Rosetta system or other software analysis tools, computers were never used in the meetings to examine the data, the preferred information artifact was a printout. Knowledge extraction was accomplished almost entirely through discussion and debate.

## **Discussion**

When considering the effect of sophisticated analysis tools on the workflow of the laboratories, it must be stated that core to the character of the toxicogenomic and proteomic laboratories is the academic relationship the larger university. Most, if not all, of the strategic decisions within the laboratories are made with this in mind, from the experimental methodology to the acquisition of research technology, as well as the day-to-day work involving the coming and goings of other persons. The culture of the laboratory is one of academic research, and to this end it is assumed by the director of the laboratories that there is ample future opportunity for involving complementary disciplines in the core research areas that the laboratories have expertise in. It is due to this belief that I was even able to conduct my research, as I represented the first of what the laboratories hope to be a steady stream of graduate students in biomedical informatics.

With this in mind, it also needs to be stated that the observation of a research facility within an academic setting will generate insight significantly different from the observation of a laboratory with a similar purpose within a commercial biotechnology firm. Despite the structural similarities of the qualitative studies of CPOE implementation and software engineering organizations described in the literature review, this “snapshot” of the laboratories may be outdated by the time this thesis is published. Despite these limitations, an evaluation of the tasks, tools and workflow within these laboratories can provide planning criteria for other academic laboratories interested in implementing similar systems.

### *Tasks and Tools*

The traditional laboratory tasks within the MGEA experimental process were heavily dependent on carefully defined protocols relating to specific instrumentation and biochemical preparation. These tasks were considered core technologies in the laboratory work and were enthusiastically mentored to other laboratory technicians and pre-doctoral students as part of a tradition of collaborative practicums. With respect to the bioinformatics tools, however, there were no task protocols related to the data analysis. Where the primarily laboratory equipment had a long history of use, and indeed could be used in other laboratories to different ends, the software systems were still new to most of the laboratory staff. The closest things to a protocol for the use of the systems were the software manuals relating to the features and management of the system.

In respect to the Rosetta system, issues of training limitations led to it being utilized by only a limited number of individuals. Approximately five months after the system installation, only the toxicogenomic lab technician, the laboratory manager and an undergraduate intern had used the system, though resulting data in the form of printouts had been shared in the later weekly meetings. A contributing factor to this was that none of the other researchers had data within the system to investigate, primarily due to a lack of expertise and time by those who were familiar enough with the system to help with such a thing. There was example “training” data within the system which researchers could explore functionality, but researchers were not encouraged to do this as it was discovered that the new version was due out in the relatively near future, and little effort

should be made to master the existing system until formal training had been provided on the newer version. This emphasis on formal structured training appeared to be one factor that limited enthusiasm of the researchers, and continued dependence on the more traditional methods of consulting the biostatistician.

### *Workflow*

The overall workflow of the laboratories is based on iteration and refinement of process, with an end goal of consistent reproducibility for the experimental data at all stages.

During the initial stages of the toxicogenomic laboratory setup, the commercial suppliers provided protocols for the microarray preparation procedures, which provided outlines for the wet-lab workflow. These were sometimes modified slightly based on input from other TRC members such as Paradigm systems. Other outside triggering events that might have an effect on experimental protocols and workflow were brought up occasionally - the most significant being the discovery of ozone being a possible source of microarray slide degradation, which needed to be tightly controlled.

From my inquiries, this appeared to be the primary means by which protocol and workflow were altered – by input from outside authorities. In respect to the resulting data from the experiments, there appeared to be very little organized workflow aside from the weekly meetings. This evolved somewhat as the laboratory technician and laboratory manager became more familiar with the system, but was largely conducted in a group exploration format, and depended on biological interpretation based on current roles present at the meetings. Overall, despite the implementation of the Rosetta system, the laboratory still depended on the presence of the biostatistician to aid in interpretation.

Part of this appeared to be due to the desire for independent validation, and part was that the statistician could provide verbal summaries of the information, where the tools could not.

As it stands, the workflow has bottlenecks that may be dramatically improved with additional specialized outside expertise. Focused attention on data import and export among the various systems would help integration of the vast array of genetic information sources available on the Internet. In particular, this is crucial for planning for the management and support of the evolving toxicogenomic specific standards and databases such as MIAME/Tox, as well supporting TRC responsibilities to the CEBS. Though the laboratories have good relationships with ITG and biostatisticians, this domain of information is highly evolving and requires full-time specialized attention. Though these skills could be initially brought in from outside, as the laboratories tend to do with biostatisticians, they would not be fully effective without integration at all steps of the experimental process, and would best be aided by being part of a mutual educational collaboration with the researchers. At present, the experimental design phase concentrates on the physical and statistical requirements appropriate to each individual experiment. It would be wise to include overall data design and expectations within experimental design stage.

Overall, the laboratory itself seemed to be off to a good start. However, if its greater goals include more of a service model of laboratory work it will require additional interwoven roles of bioinformatics support within the core laboratory team, as well as a

leadership directive to support these types of roles within the laboratory culture. With the workflow and ever increasing requirements of consistency being dependent on so many individual processes, it is clear that evolution of how information workflow is best managed will become an even greater necessity.

### *Limitations*

There are three limitations that need to be considered with respect to this study. The first is that this study was conducted with only one observer. Since this study was one that depended on observation and interviews for all data, there was opportunity for “observer bias” to affect the evaluation of the data. To this end I attempted to identify bias that related to my “pre-identified themes” prior to the interviewing portion of the study. Additionally I involved a fellow student in this process, which helped to verify the emergent themes independently of my own observations. In future work this could be greatly aided by a co-researcher being involved at the design and evaluation portions of the process.

Secondarily was the possibility that I lost some of my objectivity during the period of time I studied the labs. As the nature of the laboratory was collegial and intellectually stimulating within a university educational setting, it was difficult to remain completely detached from the process. However, since my background does not include biology of any form, I believe that my objectivity was kept at a distance due to my lack of domain knowledge. This allowed me to concentrate on the interactions and information seeking behaviors in regards to the software tools within the laboratory. Conversely, my lack of

domain knowledge may also have been a limitation in understanding the more informative questions to ask.

The third limitation was the single round of interviews conducted. Ideally, the study would have been aided by repeating interviews with key informants in an evolution of grounded qualitative theory. However, the rapidly changing nature of the laboratory as it “got up to speed”, coupled with the difficulty in scheduling additional interviews prevented this.

### *Future Research*

In order to better analyze role-specific information requirements, designing a future study based on a multiple perspectives model would aid in identifying local information seeking tasks that are relevant to each type of user. This model could be formed based on the primary roles identified in this study, or generated as a result of surveys provided to other laboratories engaged in similar implementation processes. It would be useful to compare information workflow perspectives of laboratories that lack core bioinformatics tools or roles with laboratories that do not.

Analysis of what methods are best suited for implementing a more integrated use of software analytical tools in academic biological laboratories would be of benefit to both the laboratories as well as to progressing academic education. Students of varying experience and goals are regularly cycling through these laboratories, though most have overwhelmingly biological ambitions. Evaluation of whether implementing a mentored

tutorial of data analysis techniques that co-exists with the technical biological practicums work would be interesting. This might be studied by following a cohort of students involved in laboratory rotations, with the sample drawn from a range of interests, and their experiences analyzed during the process.



## **Conclusions**

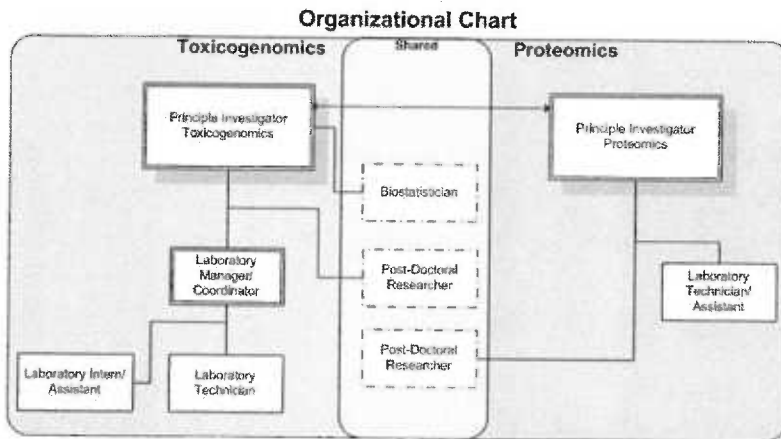
To my knowledge, this study was the first of its kind to attempt to apply qualitative analysis techniques to identifying information roles, tasks and workflow themes within a biological research laboratory. With this in mind, it should be considered an early pilot study of this domain area. It has identified issues that may need to be addressed for successful implementation of this type of software analysis systems in other similar academic laboratories, where the dual requirements of operation are research and education. Hopefully, the information revealed in this study will be useful to the laboratory staff, as well as bioinformatics students and researchers interested in improving education, system functionality and workflow integration in this area.

## References

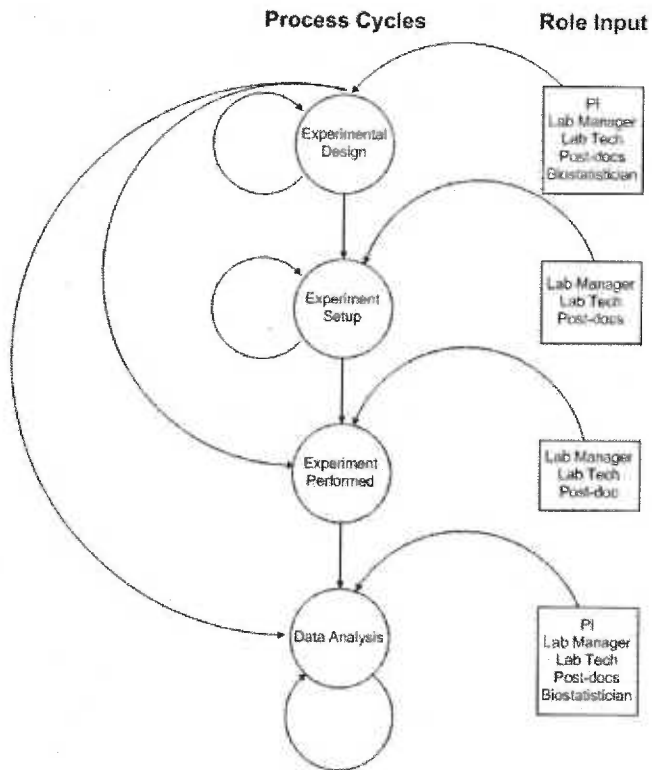
1. Crabtree, B., Miller, W. *Doing Qualitative Research* (Sage Publications, 1992).
2. Padgett, D. K. *Qualitative Methods in Social Work Research: Challenges and Rewards* (Sage Publications, 1998).
3. Joerges, B., Shinn, T. *Instrumentation Between Science, State and Industry* (ed. Joerges, B., Shinn, T.) (Kluwer, 2000).
4. van 't Veer, L. J. et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415, 530-6 (2002).
5. *Department of Energy- Human Genome Project Information*  
[http://www.ornl.gov/sci/techresources/Human\\_Genome/glossary/glossary.shtml](http://www.ornl.gov/sci/techresources/Human_Genome/glossary/glossary.shtml)  
Accessed: 1/30/04
6. Latour, B., Salk, J. *Laboratory Life: The Construction of Scientific Facts* (Princeton University Press, Princeton, NJ, 1979).
7. Chin, G., Schuchardt, K., Myers, J., Gracio, D. *Participatory Workflow Analysis: Unveiling Scientific Research Processes with Scientists* in Participatory Design Conference, Graduate Center, City University of New York, NY 2000
8. Robbins, S. *Organizational Behavior* (Prentice Hal, Upper Saddle River, NJ, 2003).
9. Jones, C. *Patterns of Software System Failure and Success* (Patterns of Software System Failure and Success, 1996).
10. *Bioprotocol Homepage* <http://bioprotocol.bio.com/protocolstools/browsdesc.jhtml>  
Accessed: 1/24/04
11. Matthiessen, M. W. Affordable biocomputing for everyone: using the Internet, freeware and open-source software. *Trends Biochem Sci* 27, 586-8 (2002).
12. Matthiessen, M. W. BioWareDB: the biomedical software and database search engine. *Bioinformatics* 19, 2319-20 (2003).
13. Grewal, A., Stockton, J., Bolger, C. Tools for discovery: gene expression enterprise solutions. *Curr Opin Drug Discov Devel.* 6, 333-8. (2003).
14. Bassett, D., Eisen, M., Boguski, M. S. Gene Expression Informatics - It's all in your mine. *Nature* 21, 51-55 (1999).
15. Dennis, C. Biology Databases: Information Overload. *Nature* 417 (2002).
16. Robbins, R. J. Bioinformatics: Essential Infrastructure for global biology. *Journal of Computational Biology* 3, 465-478 (1996).
17. Gittelman, M., B. Does good science lead to valuable knowledge? Biotechnology firms and the evolutionary logic of citation patterns. *MANAGEMENT SCIENCE* 49, 366-391 (2003).
18. Rajagopalan, D. A comparison of statistical methods for analysis of high density oligonucleotide array data. *Bioinformatics.* 9, 1469-76. (2003).
19. Kaminsk, i. N., Friedman, N. Practical approaches to analyzing results of microarray experiments. *Am J Respir Cell Mol Biol.* 27, 125-132 (2002).
20. Hubbard, T. Biological information: making it accessible and integrated (and trying to make sense of it). *Bioinformatics.* Suppl2 (2002).
21. Fogarty, M., Bahls, C. in *The Scientist* (2002).
22. Patt, M. R., Houston, T.K., Jenckes, M.W., Sands, D.Z., Ford, D.E. Doctors who are using e-mail with their patients: a qualitative exploration. *Journal of Medical Internet Research* 5, online (2003).

23. Cheng, C. H., Goldstein, M. K., Geller, E. & Levitt, R. E. The Effects of CPOE on ICU Workflow: An Observational Study. *Proc AMIA Symp*, 150-4 (2003).
24. Ash, J. S., Gorman, P.N., Lavelle, M., Stavri, P. Z., Lyman, J., Forunier, L., Carpenter, J. Perceptions of Physician Order Entry: Results of a Cross-Site Qualitative Study. *Methods of Information in Medicine* 42, 313-323 (2003).
25. Danielson, E. A *Qualitative Assessment of Changes in Nurses Workflow in Response to the Implementation of an Electronic Charting Information System* Department of Medical Informatics and Clinical Epidemiology OHSU 2002
26. Seaman, C. B. Qualitative Methods in Empirical Studies of Software Engineering. *IEEE Transactions on Software Engineering* 25, 557-572 (1999).
27. Seaman, C. B. Communication and Organization in Software Development: An Empirical Study. *IBM Systems Journal* 36, 550-563 (1997).
28. Flowers, S. *Software Failure: Management Failure : Amazing Stories and Cautionary Tales* (John Wiley & Sons, 1996).
29. NanoDrop Technologies Homepage <http://www.nanodrop.com> Accessed: 1/24/04
30. Agilent Technologies Homepage <http://www.agilent.com> Accessed: 1/24/04
31. Bioconductor Homepage <http://www.bioconductor.org/> Accessed: 2/1/04
32. Wolcott, H. F. *Writing Up Qualitative Research* (Sage Publications, 2001).
33. Muckter, H. What is toxicology and how does toxicity occur? *Best Pract Res Clin Anaesthesiol* 17, 5-27 (2003).
34. Paradigm Genetics, Inc Homepage <http://www.paradigmgenetics.com> Accessed: 1/24/04
35. Science Applications International Corporation (SAIC) Homepage <http://www.saic.com/> Accessed: 1/28/04
36. Chemical Effects on Biological Systems <http://www.niehs.nih.gov/nct/cebs.htm> Accessed: 1/24/04

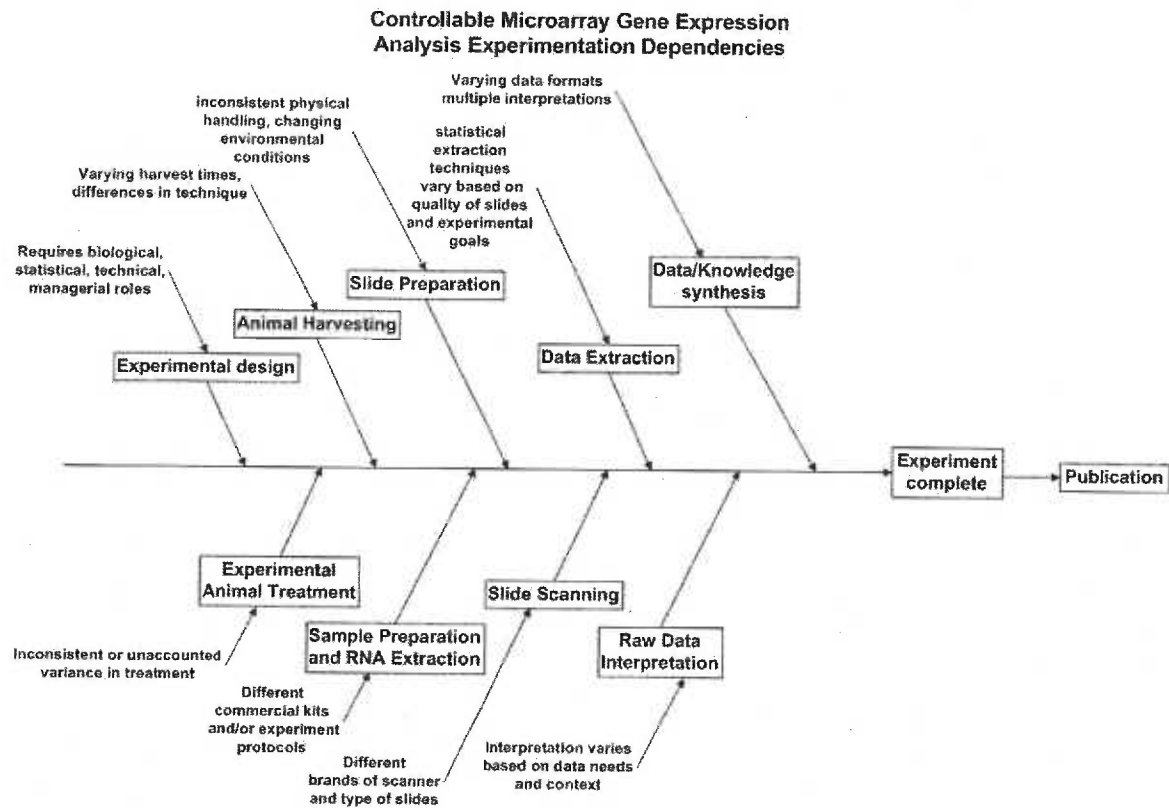
# List Of Tables



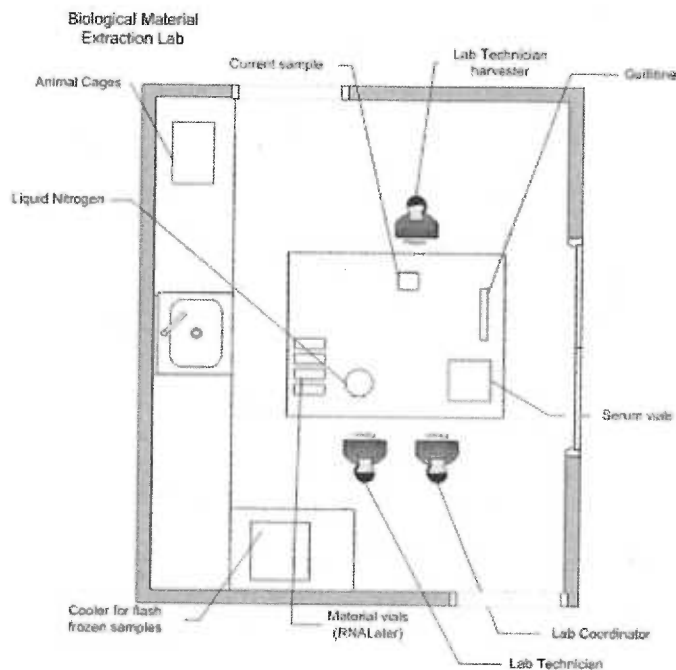
**Figure 1. Toxicogenomic/Proteomic role relationships**



**Figure 2. Experimentation Process and Roles**



**Figure 3. Microarray Gene Expression Analysis Experimentation Dependencies**



**Figure 4. Animal Harvest Lab**

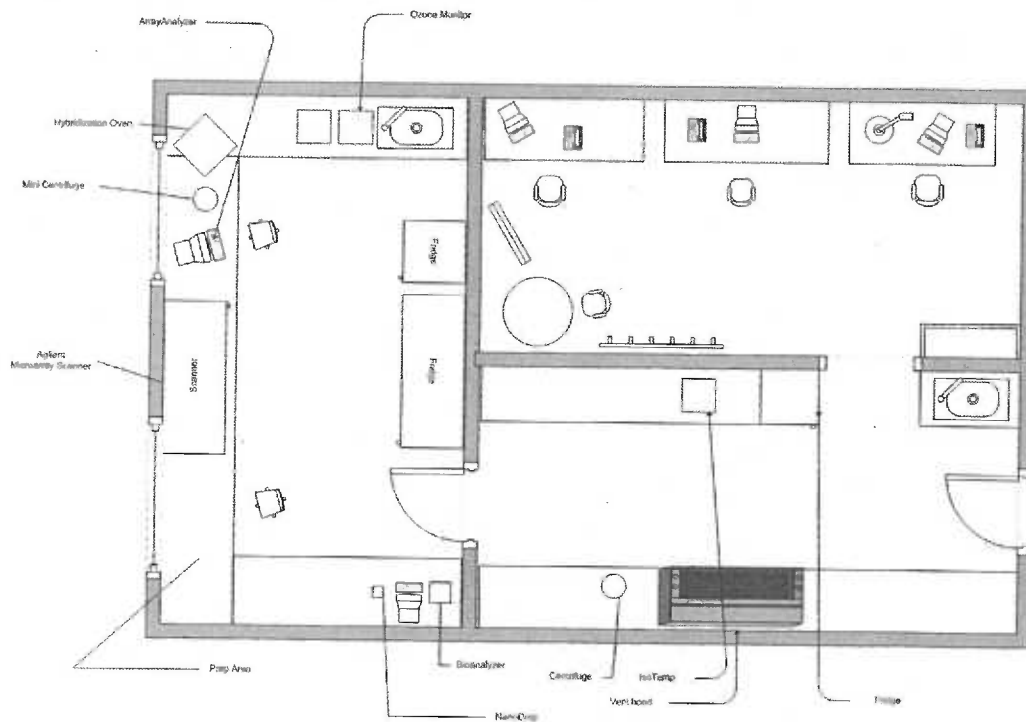


Figure 5. Toxicogenomic Laboratory Floorplan

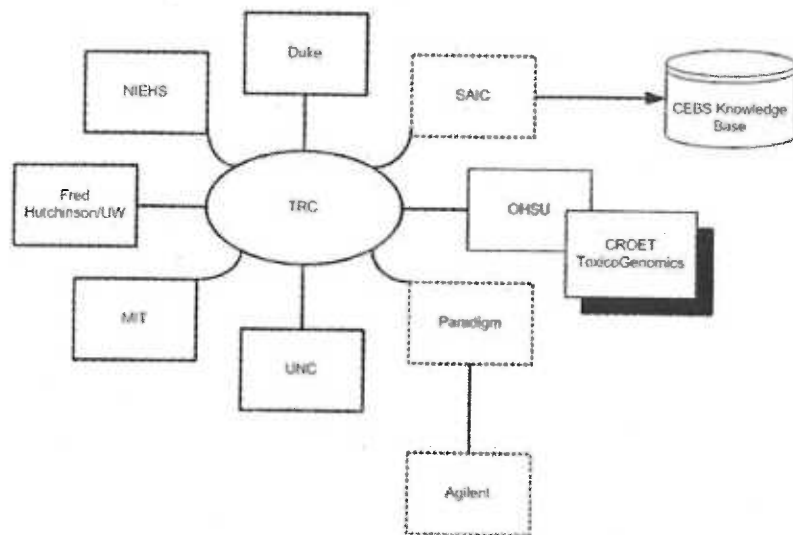


Figure 6. Toxicogenomic Research Consortium Relationships

Study Timeline  
7/7/03 - 2/16/04

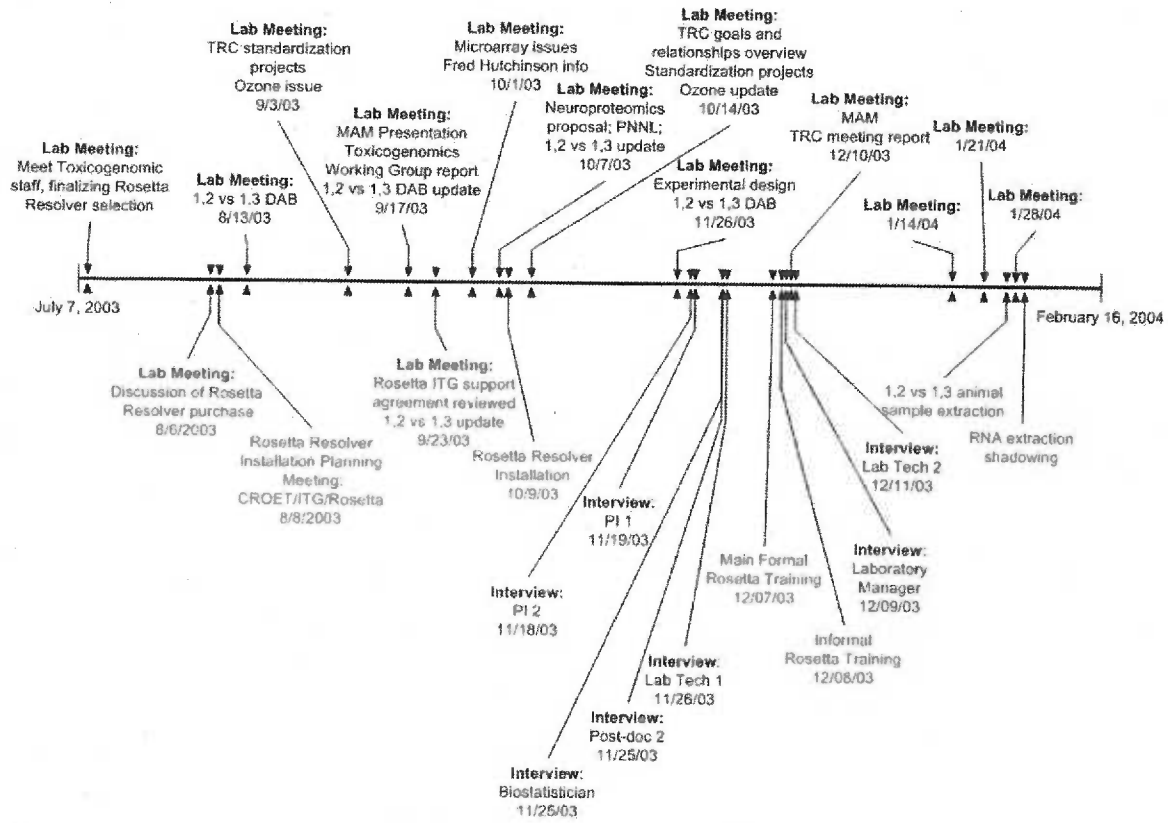


Figure 7. Study Timeline

## **Appendixes**

### ***Appendix A: Interview Question Guide***

#### **Personal History**

- How would you describe your personal role in the laboratory?
- How would you characterize your training to get to this point of your work the lab, was it traditional or non-traditional?
- Is there any skill area you wish that you had more experience in now to do your work in the lab?

#### **Daily/Weekly Workflow**

- What is a typical day or week like in the laboratory?
- Do you feel work in the laboratory follows of a collaborative process, a role-driver process, or a leadership driven process, or some other type of process?
- In the course of an average week, how do you interact with the other members of the lab? In person, by phone, email?

#### **Research/Communication**

- Do feel that there is a new dependence on non-traditional biological research roles in the lab? If so, what roles?
- Do you think the laboratory is dependent on outside expertise, be it vendors or collaborative researchers?
- Is there anything you would prefer to be more constrained to do work in the lab?
- Based on your experience in this work, what set of skills or role familiarity would you like to see in new researchers?

#### **Technology**

- How would you characterize your approach to evaluating new computerized tools for information analysis in the laboratory?
- How would you contrast this to your approach for evaluating new techniques in bench-lab experimentation?
- How do you evaluate new software and tools for the lab?
- Do you feel that you approach new computer tools skeptically?
- In specialized scientific applications such as bioinformatics tools, what do you see as the biggest risk?

#### **SOP/Protocols**



- What are the most time consuming internal processes in experimental research in your laboratory?
- What are the most time consuming external processes that are critical to your experimental research?
- In developing a SOP for laboratory bench work, do you think that there is any room for flexibility?
- How are changes to SOP's brought about?
- Do you think that most biological bench work research methods can be standardized into SOP's such that trained technicians can use them?
- When bench work SOP's/protocols were implemented, were there any unintended consequences?
- Do you think that experiment analysis methods provided by bioinformatics tools can be standardized such that trained technicians can use them?
- The improvements in technology relating to biological research have allowed for certain experimental techniques to be mostly automated, do you feel that there will be diminishing roles in traditional biological research, and if so, what?

#### **Bioinformatics data management**

- Gene expression analysis generates a large amount of data. What do you think are the most important steps in order to ensure accuracy and integrity of this data?
- What is your greatest fear in management of experiment data?

#### **Bioinformatics Software and equipment**

- Do you think that there is any difference in how researchers approach commercial vs. non-commercial bioinformatics tools?
- Do you think that Rosetta Resolver would be used differently if access was free?
- What is the biggest expense of using the system?
- When the system was first installed, were there any unintended consequences?
- At what point will you consider the system to be a valuable investment?
- What are your current concerns with the implementation of the system?
- Do you expect to personally use the system, and what for?

## ***Appendix B: Project Information Sheet***

I am a graduate student in the Department of Medical Informatics and Clinical Epidemiology, and am presently doing my Masters Thesis. The title of my thesis is "*A qualitative assessment of impact on the bioinformatics workflow and protocol evolution in response to the implementation of a commercial gene expression analysis system.*"

The core thesis work involves the workflow analysis of the center Toxicogenomic Laboratory environment and employees, with attention on the flow of and management of information. I will be observing and interviewing employees of the lab before, during and after implementation of the Rosetta Resolver Gene Expression Analysis software. I hope to complete my observations by March of 2004.

When interviewing you, I will not use your name, but I will identify your responses by a classification such as "Researcher A", or "Technician B". You may decline to be interviewed at any time. If you have any concerns regarding this process, you may contact my thesis advisor, Dr. Christopher Dubay, or me at any time.

Nicholas Anderson, BS  
(503) 267 1099  
[andersni@ohsu.edu](mailto:andersni@ohsu.edu)

Christopher Dubay, PhD  
(503) 494 2116  
[dubayc@ohsu.edu](mailto:dubayc@ohsu.edu)

## *Appendix C: Sample Interview Transcript*

### *Biostatistician 1*

Date: 11/25/2003  
Start Time: 4:00 PM  
End Time: 4:44 PM

#### Pre-Interview Impressions

This subject is a statistician employed by the center for use in experimental research. The subject recently moved to Japan and has been continuing work remotely from emailed data. The subject was visiting the center for the first time since leaving for Japan about 2 months previously.

#### Interview Transcript

\*P:

How would you describe your role in the work of the center to date?

\*I:

Mostly as support for the other scientists, so I am somebody who hopefully can catalyze some of the investigations that come along, specific things that they suspect they would like to see, or that they think that they would like to see, they give some input in the questions they would like to have answered, then they set me free to try and coax those answers out of the data whatever way I can. In terms of how that works with the current projects involving microarray and Toxicogenomics, it's much more difficult because I don't think anybody has a clear handle ahead of time in what they want to see in their data, they don't want to necessarily prejudice things by telling me go find all the beta tubulin things to be upregulated, they don't know what it should be so it's more of a complication, because you don't have.. It's a case where you are trying to use statistics for really an investigative testing, and lot of people want to give the rigor behind you know formal hypothesis testing, pre-specified hypothesis, this is what I look for, this is what I found, isn't that great, and a lot of times what we are looking at you know, this is just an investigation, I don't know if they will hold up under a repeated applications, you can run the experiment again and get something totally different, but this is what I saw, these are the patterns, and what these are becoming more of is more of a graphical kind of problem, can you describe the patterns in a way that is more communicable to other people, so that they can appreciate and understand it.

\*P:

SO how would you describe your training to get to this point, do you have a biological background?

\*I:

I have really no biology background, I think the last biology class I had was in high school, my training has been in mathematics as a bachelors degree and statistics as a masters degree. So people around here sometimes refer to me as a biostatistician, that is not by training, that is not by design, I just happen to be a statistician who is working in a biological setting.,

\*P:

Ok

\*I:

Um, I picked up, not really the biology, I picked up the things are really unique to statistics by working with people like X, epidemiology and seeing the different kinds of statistics they use. The Masters degree kind of gave me an overall training in the fundamental core pieces, but you can assemble those pieces together whatever way you need to meet the demands of the field that you are working in. Epidemiology required a certain assemblage, Toxicogenomics requires a totally different assemblage, so the fundamental

building blocks are the same no matter really what your training is, it's really what area you are practicing in seems to give people the name biostatistician or epidemiologist or whatever.

\*P:

Is there any specific skill area as a result of that in your current work that you wish you had more experience at this point?

\*I:

I wish I had a little more experience for Toxicogenomics related to you know, the genomics, the biological part. When the scientists are talking about, you know, 5-prime end and 3-prime end cross linking, this is you know, incorporating with a different link, it is confusion, but luckily for me they can usually distill it down to wanting to know if this is different from this. And so they can specify the very clear question they are directing this to, their insights too.

\*P:

Excellent, um, so your work is largely experiment based.

\*I:

For certain scientists here it is experiment based, um there are people here who have clear designs, clear protocols for how their data are generated. There are other scientists here who are doing, more surveys, more data collection that doesn't really have as clear a design, it's limited to you, know, people who are volunteers who are willing to participate, these are the people we could coral on a certain day to take part in these tests, these are the ones who showed up for the first session but didn't show up for the second so you have to drop them, in those situations this is not a design based kind of things, that's an observation kind of study, so my work has been split here at center, 50/50 some scientists who really have designs, and can guarantee me repeat that they will always have this # of animals, this # of units, treated this exact way, other people can only tell me that they started with this many, or they hoped to get this many in, due to drop out, attrition, here's what we finally found.

\*P:

um, do you feel that work in the laboratory, some of these may not apply to you, do you feel that work follows a collaborative, or role driven, or some other kind of process?

\*I:

Um, yes, it seems of be that each lab is of course driven by a different person. In X lab, (epidemiology), she had a procedure whereby she kind of gathered the data, then starting writing the manuscript leaving big gaping holes, saying here's what this section needs to describe, here's what this section needs to describe then farming it out to the other people to fill in the gaps, so it was very much led by her as to how she wanted to run the process, other people it seems to be driven, but it might be driven more, for instance, in Toxicogenomics I think X is sort of the leader of it, but he sort of micromanages and farms it out to other people who are leading, so naturally when you have different levels of leadership, you get things that could be miscommunicated, misunderstood or people adding the own flavors to the mix at various stages walk away.

\*P:

Ok, um, ok. Do you think that this kind of research is becoming increasingly dependent on outside expertise, biostatisticians, commercial vendors, or anything like that?

\*I:

I think its something that is definitely becoming reliant on outside expertise. The expertise though, at least here in the center seems to be one where people identify outside expertise, in terms of a piece of software or a person who can deliver it, and then try and secure that to try to make it internal. I know that in terms of X (Toxicogenomics P.I.), he wanted to make sure that there was a biostatistician or statistician here in the center, he didn't really like the idea of borrowing some kind of communal statistician on another part of campus, he wanted to have someone who was dedicated, who could work on his projects on his time schedule.

\*P:

Um, I know know that you are based in Japan, speaking perhaps of your prior life here, how did you interact with people in the lab, was it mostly in person, by phone, email?

\*I:

It would be a combination of things, it would usually be a call up by phone, the common thing would be it was a quick thing, would only take 2 minutes

\*\*P:

Laughter

\*I: then we would get talking and we would discover that there were quite a few more issues that need to be addressed, that there were more levels to their experiment that what they thought, they had different animals, they have experiments stretching over different days or something so there are things that need to be addressed, then they'd come and we'd sit down and talk about it, so it was more of a face to face in person kind of thing. In Japan I think I am limited now to questions that they know that they can get answers to in email, because there is not the same back and forth that they could get when I was here in person.

\*P:

Is that, does that lead to less open-ended questions?

\*I:

No, I think that that is leading to more volleys of email; I have conversations now where it seems like I am talking in real time. I send a letter off, and as soon as I send it, 2 minutes later I update it and threes the result from the person so I read it, I send back one of them, 2 minutes later I have another one in my box. So we are communicating, so, depending on how people write it can lead to other kinds of miscommunication.

\*P:

Based on your experience in this type of work, history, being sort of an outside expert, what sort of skills would you like to see within this laboratory environment?

\*I: Skills, as relating to whom?

\*P:

well, do you see any consistent, or as you said at least in the Toxicogenomics lab, they have made an effort to bring in new skills as they need them, how successful do you see that, do you think they need more people like yourself, less like yourself? (Laughter)

\*I:

I think it could help to have an intermediary, somebody who's in charge of managing the data. A lot of times I say, send me your data, and I get this spreadsheet that is organized in a totally bizarre way, I spend more time organizing the data in such a way that I can import it into my stat package, than actually analyzing the data. SO if there is somebody that can filter the data, screen things out, in terms of missing values, getting it in a coherent way that would be a big help.

\*\*P:

So kind of a consistent interface

\*I:

A consistent interface is something that is lacking with a lot of the scientists, not just epidemiology, but the basic scientists, some of them have different kind of computer formats than what I was using so I was getting things in a bizarre file formats that used line feed instead of carriage return line field, so I was getting these files that are just all messed up.

\*P:

Um, so obviously since your domain is very technically based, how do you evaluate new software, or other researchers tools? What steps do you take to evaluate new tools for your work?

\*I:

mmm, there are only a few packages that I have really evaluated. I usually like to restrict myself to packages I can get for free. I have no interest personally in evaluating a package if it is going to cost me several hundred dollars to get it on my desktop. So I look at the ones that are there for free and then I try them out. I try to get a feel for how they are to program, in terms of is it easier for me to program, is it harder for me to program, um, I look in terms of speed, a lot of the applications that I do require looping, and there are certain programs that are very rich in terms of statistics, but are very slow in terms of loops, so some programs win big points in certain points, and lose points in other areas. So consequently I might use some programs to generate a data set then use a different program because it is faster. I look for things in terms of what kind of support is out there for the program. R for instance, has a very wide base of modules that I can plug in there, other programs that I use there is practically nothing out there, so you are borrowing or reverse engineering from other packages into the one you want to use/

\*P:

Do you in the use of specialized scientific applications, and there are so many commercial vs. non commercial, what do you see is the biggest risk for scientific use of these type of tools

\*I:

The biggest risk is that because some of these commercial tools are becoming easier to use in the way of point and click, that they will be placed in the hands of somebody who only knows how to point and click. Um, I've seen things happen with statistical software where what used to be a command line syntax thing where you have to specifically type, ask for, know the syntax, know what to get, it required a certain level of expertise to use, but at least you were guaranteed that when you got a certain result, that was the specific result you asked for and the person who typed in that command knew that was the result they asked for. When that kind of program has been to drag down and drop with, yes I want to do a regression, I need these three variables, go, the out put some people don't really know how to interpret it and so it can lead of people looking in the wrong area for a conclusion, they might cite a certain p value that has nothing to do with the test they are looking for, so my concern would be that although you could get the best tool with the best intention, they person who is using it may not interpret the results correctly, or, it could even be the case with the property software, that it is written but it is not subject to any outside peer review, so one wrong line of code, or something that causes the machine to underflow or overflow, and you can do as many analysis as you want by the results would be garbage

\*P:

Exactly, um. I'll get back to that. When you do you your analysis, what is the most time consuming process?

\*I:

The most time consuming process is getting the data cut in a format that is usable by my stats packages. So formatting it, reformatting it, that sort of thing is very time consuming. The second time consuming part is, a lot of times the people try to have well-developed questions, but they want to look at the same question in many different ways, and even if you can look at the results, and say, well, I know from looking at that so the first approach didn't work, then these other approaches will also lead to no real findings, I would like to move on, but the person who is asking these questions says no, I want to look at these things. So I still have to wind up generating sets of tables of things that are going to be duplicating results that I know are similar to what I had before or that won't differ appreciably from the first set?

\*P:

Does that knowledge come from the process that you can fairly intuitively tell that no amount of message will differ too much from it?

\*I:

It's not smooch massaging, you might be looking at, say, 11 different possible outcome variables, and none of these show any particular pattern, but then they want to start looking at subsets. You know, form a summary of these 6, form a summary of these 8, maybe if you combine this one and this one, take the average of time one and time two, just look at time one vs. time two. Things that you know, if I didn't see a difference between the two most extremes, then there is no way I am going to see a difference between things that are internal. IF the max and min don't differ, the things in between don't differ either. And you can tell them that and they say I want to look at it anyway.

\*P:

Do you; see a point in biostatistics where the process can be standardized such that a generally experienced biological research will be able to

\*I:

Push a button?

\*P:

push a button.

\*I:

Yes, I think that something like that could be done. But for things in genomics think it is too new, and I don't think there is any optimal solution yet. For other areas, I think that yes, you can always get something to push a button, but its very very difficult to see all the ways that the data is collected. Many scientists collect information from 3 rats then somehow combine or pool that information, the # of mutants from rat 1, the # of mutants rat 2 # mutants from rat 3, and then they combine some total mutant, certain analysis behave differently when you combine results than if you look at the 3 separate measures. So even if you might streamline it and come up with a procedure where this does the right kind of analysis for the right kind of data, that fundamental step as to how the data was collected, could make enough of a difference where you don't get the right kind of results.

\*P:

That leads us back to SOP within the lab, have you had any involvement in the specific steps, such as sample preparation, up to the point of the data?

Um, I haven't been around for the sample preparation of the data, but they ask me about the experimental design, as to what they should be using for the animals, as to how the animals should be housed, should they be put in separate cages, should the be in the same cage, so I get certain, so I get certain, should they have a separate water supply, or individual water, I get certain questions that deal with some aspects of operating procedures, but those are really the aspects that deal with statistical issues, when it comes to situations of how the material was harvested, I have no basis to whether you need to switch scalpel blades from one animal to the next.

\*I:

Ok, so obviously gene expression analysis generates a large amount of data, what do you think is the most important steps in order to ensure the accuracy and integrity of this data?

\*P:

A standard way of tracking from start to finish, and also a way to documenting the changes along the way, and hopefully trying to get rid of the older versions so that somebody who is coming along hopefully doesn't use one of the outdated things. What I have seen not just in genomics, when the data is generated, it goes to different people, it goes to different sources, because it is not centralized in terms of where it is being kept, and then different people look at it different ways, so you know, each person changes it a little bit, then might share it with somebody else, so finally when a person gets it, if two people try to compare different results, they are using different data bases.

\*I:

## Hundred monkeys problem

\*P:

Laughter - exactly

\*I:

So centralizing the data, tracking the changes from start to finish, trying to minimize the human involvement for changes, when I see scientists who sometimes go through their data and I tell them I need to have you use a uniform type of coding for something, they go through it and do it by hand, so I still get errors because of type, so using things like search and replace, which is fully automated and catches everything, as opposed to somebody who mistypes or misspells something, these are all things that lead to integrate of the data that is important to the analysis.

\*P:

Standardization where possible?

\*I:

Standardization where possible.

\*P:

What is your concern to the actual management of the data long term?

\*I:

I guess long term management of this data, is that what will happen is as these results, or as advances are made, people will continue to go back and reanalyze data that is already being done using the newest best approach, and then start scratching their head and wondering well, now my results don't agree with what I had before, I get a conflicting signal, and so the ability to have all this data and have it archived lets people look backwards retrospectively and reanalyze it using a new way and say things like well I didn't get the same thing the first time, or having two people in the room and say well I looked at the data but I didn't get the same thing that you did. So Storing the data to allow me to look at it repeatedly, and the more you look at things, the more you are going to find differences

\*P:

So Rosetta is a commercial system, and it has the ability to define roles in regards to access to the data. Do you have a preference on the difference between commercial and non-commercial packages and the limitations to the data based on cost?

\*I:

Based on cost....

\*P:

Well, this is a difficult question for me to parse, Rosetta gives a lot of power for the money, however it provides quite a bit of limitation, in terms of access, do you think that this will change the experimental approach to using these tools if there is more of a trend towards cost for access?

\*I:

I think that there is already that sort of situation going on. When Rosetta first came out and people were uncertain how to analyze this data, it was a very expensive piece of software, it's coming down in price because I think they realize that they have competition from free pieces of software, and that even if you do have Rosetta, you need to have somebody to interpret it, so if you are going to pay for a statistician to interpret it, or for somebody who has the ability to interpret that data, you might as well pay for somebody who has the ability to program and then develop their own tools for specific sets, so I think that marketing something or selling a piece of commercial software is going to become a losing kind of business as time goes on, I think people are getting smarter about how to use these, people are getting better in terms of how to analyze the data, which way to go, so I think that different tools are going to merge as the best ones, and



I don't think that commercial software necessarily has the best way to do it, or if they do, it's going to take a lot of reverse engineering to put the best thing in there.

\*P:

One thing from that, is that Rosetta has a role based tracking method, do you think that the needs to be more of that type of process in this type of management of data?

\*I:

Yes, one of their strengths is that they allow for this kind of tracking, that's not to say that Rosetta is the only place where you can get this, you can hire people who's job is specifically to archive, manage, warehouse the data, and make sure that the people who are using it are using the most up to date version, and that any changes are made are made in a centralized way. I don't think it has to be done by Rosetta

\*P:

How do you think these types of tools would be used differently if these tools were free? IF there is no limitation to the access such that you could spend day staring at the same screen, is there any advantage to arranging to look at it?

\*I:

I always like to think; my feeling is that it is better to look at things with a directed purpose, because the more you stare at it, the chance that you might see a spurious pattern?

\*P:

What do you thing would be the biggest expense in using such a system, people, trying, and money?

\*I:

Such as the Resolver system? The biggest expense...? I don't know

\*I:

Ok, Speaking of terms of your own abilities, what do you think that would be the biggest benefit of using a system when contrasted with your previous methods?

\*P:

Just that it is going to be consistent, I might look at the data in a particular data set on a certain day and say, this data requires me to take the log transform, and then another different set of data comes along, trying to measure the same thing, but because of some peculiarity I look at it and it doesn't quite need the log but needs the square root transform. So each time I am trying to look at the optimal situation or the best situation for this particular set of data, I am not necessarily when I analyze a current set I am not thinking retrospectively what did I think about what did I do before and how did I exactly do it before, I am really treating it as you know, this is a new set of data, which leads to problems, because you know, I pass the results to X, X looks at it, pass a different set off to Y, Y looks at his, X had log transform, Y had square root transform , now they can't talk to one another.

\*I:

So, maintaining the ability to have consistent reproducibility.

\*P:

Absolutely, but consistent reproducibility does not necessarily mean that it is the best way to analyze things, so you sacrifice things, it be that you have something that is reproducible to such a way that no experiment goes through there gets optimized for the analytical approach.

\*I:

At what point would you personally consider the system to be valuable, what criteria would judge it as being a success?

\*P:

Um, it is really hard to judge something as being a success if it is a commercial system, as things that I have seen coming from Rosetta say confidential and propriety, it doesn't come with a list of references indicating where their algorithms came from, it doesn't come with a list of people who are programming, so you don't have a lot of built in faith as to how these algorithms were developed, you just know that this is what the system uses and they say it works because, well, it works. So these empirically derived systems don't always have a lot of basis in the theory, so how to evaluate it professional, so how to evaluate it professionally, I would really evaluate it based on how well it matches the applied with the theory, but you can't really see that because proprietary information is not really completely transparent to how they have assembled their software.

\*I:

Have you had any need to try to validate or evaluate these techniques yet?

\*P:

Yes, I have tried that, and for Rosetta there are certain things that you try to calculate a log ratio for the red and the green, and for certain ones it matches up exactly with what they call their computed log ration for certain values, and for other values it doesn't match up quite exactly, and so you try to backtrack and figure out why. Their background corrected signal is not just what they describe in the manual as just the mean - the background, they have some other fudge factors in there that sort of jitter the data a little bit sometimes, so it is difficult to reconstruct it, and so it is not completely transparent, and so as the researcher you don't want to start sending off emails and say, this this this this, all these things don't match, because you don't want to get back the answer that he forgot to look at this, this simple thing. So I know I am being an alarmist when I see that things don't match, but some of them don't match because of big problems, some of them don't match because of smaller problems, and it is hard to know the big ones from the little ones.

\*I:

This system has only been installed for a while. Where there any unintended consequences from your first interactions with it?

\*P:

I haven't used Resolver yet, when I speak of matching up, I speak of things that people have run for me. And then they send me back the data and I am just looking at columns of data, and I say that this column is supposed to be derived from columns A and B, but if I just use column A & B, it does not give me what this third column is supposed to be. Again, you know, since I am not the one that actually pushed the buttons I don't know if I should be alarmed at internal in the software or the person who has pushed the buttons.

\*I:

Do you expect to personally use the system?

\*P:

I expect that I will one I have gotten the training on it. Once I have the training, I think I would be better suited to commenting on the stability of the algorithms, and whether this is or is not a major problem

\*I:

At this point, apart from the your issues with possible interformulary issues, are you fairly confident that it is advertised?

\*P:

I am confident that it is fast, I am confident that it is complete in what it offers, I am less confident in the appropriateness of all of the methods, because from what I have seen there are certain things that are done that are not the way that you would do it if you were starting from scratch if you were doing it as a homework problem for a statistics class. There are things that seem to be a lot of, well, we use this because it seems to work, not because there is a theoretical basis as to how error propagates from one source to the next, in statistics we know how error propagates and we can build it in there in a statistical way, but that is not what they seem to be doing in all the stages.

\*I:

Do you think that the next version will resolve any of that, is there any knowledge of whether they are in agreement with this?

\*P:

I don't know if it will, at one time we raised some concerns about Aglient feature extraction, I didn't see any kind of information in the documentation at that time, how Aglient made some of their computation, in the most recent version, as I have gone through the data again, I downloaded a new version of the manual and their seems to be more answers to the questions, my suspicions are that maybe they saw the questions that we raised and they tried to answer them in their newest version of the manual. So perhaps these things do get updated, but I don't think they are going to take these things proactively, until somebody says, excuse me, how do you get this result, and how do you use this?

\*I:

One more question - So most of the tools that you use, are generally freely available and have public support, and presumably you use this because you get fairly good response from the community in order to evaluate them to your satisfaction, and if you come up against a brick wall, you have a simple format by which you can share with other people and say what am I doing wrong here

\*P:

With my software I can send a person a snippet of code and say, here's what I wrote, it is giving me this result, I know it should be giving me this result, could you run it and tell me what the improvement is or how it's different?

\*I:

Does Rosetta have statisticians on staff to answer this?

\*P:

I hope there are, when we interfaced with Aglient, our questions went to a technical support rep who interfaced for us with the person who write some of the code, it was sort of ok but it wasn't the best way to interface.

\*I:

Right.

\*P:

<cough>

\*I:

Thanks

\*P:

Thank you.

## ***Appendix D: Pre-identified Themes***

- **Impact on internal collaboration**  
*In what ways does the laboratory communicate, and how does a such a system affect this communication?*
- **Impact on external (consortium) collaboration.**  
*In what ways do the staff communicate with outside agencies (considered either academics, professionals, companies, or other), and how does having such a system affect this?*
- **Impact on existing experiment definition**  
*How does the lab define experiments, and how does such a system affect this?*
- **Impact on experiment results evaluation and visualization.**  
*How does the lab discuss experiments, and how does such a system affect this?*
- **Impact on data management**  
*How does the lab manage the data related to the experimental process, and how does the system affect this?*
- **Overall expectations evaluation.**  
*What other issues led the researchers to the desire to acquire this system?*