Bioinformatic approaches to understanding gene switching in the model system fruit fly, *Drosophila melanogaster* and the hijacking of the host T-Cell machinery by HIV.

Sanjive Qazi

I. Project Details.

I am requesting funding for Jason Pitt and myself to analyze the genetic data obtained from a fruit fly (*D. melanogaster*) second chromosome screen for genes involved in female sperm storage (collaboration with Margaret C. Bloch Qazi). In addition, genes implicated in affecting sperm storage will be cross referenced with genes involved in human disease progression (HIV infection) from a screen of thousands of genes using the gene chip technology. Model systems such as *D. melanogaster* provide clinical research scientists crucial information to how genes are switched on and off in the cell that have numerous applications in understanding human gene regulation. High-throughput technologies that can measure thousands of proteins and gene products simultaneously now enable intricate molecular description of human disease progression. Furthermore, the technologies have the potential to detect early onset of chronic diseases and prescribe treatment protocols to increase drug potency with reduced toxicity. I will outline and approach to study gene switching and I expect this research to serve as preliminary findings for future grant applications and publications in peer-reviewed journals.

A. A brief description of the proposed project:

1. Introduction

Of the 3 billion letters (DNA sequences) that make up the human genome only about 3% actually code for about 20 000 gene products. The rest is non-genetic, much of it with unknown function. Regions that do not code for protein are termed untranslated regions, some of which serve for binding sites for other proteins that can switch genes on and off (promoter sites). Molecules that bind to these promoter sites are called transcription factors. Each gene contains untranslated regions for promoters and enhancers that constitute binding sites for RNA polymerases; enzymes required for the first step of decoding the gene sequence into proteins, and the general transcription factors. In human cells, these molecules orchestrate remarkably intricate biochemical reactions to regulate gene expression. Binding of regulators to DNA binding sites (enhancers) can regulate the transcription of one or a subset of genes. Furthermore, gene transcription is regulated by DNA sequences that contain multiple regulatory elements that represent binding sites for multiple transcription factors. Through protein-protein interactions, DNA-bound transcriptional regulatory proteins come together to form gene-specific transcriptional regulatory complexes that also require numerous co-activators and co-repressors. Therefore, gene transcription is controlled by specific combination of DNA sequence elements and regulatory proteins that determines whether a gene is active or silent under specific conditions. This suggests that activation of genes is a complex, dynamic process and understanding this will elucidate fundamental genetic responses to changes in cellular environment.

My research focuses on the general question: How are hundreds of genes activated within a cell to affect functioning of a cell or an organ through transcription factor binding to promoter sites? A simple working model to begin to tackle this question is to consider the cell having chemical detectors, which when activated can switch on/off genes within the cell and the resulting manufacture of proteins will change the function of the cell. In this simple working

model, three fundamental processes need to be considered that include cellular recognition, cell-signaling and gene activation. We propose to use two model systems to gain insight into how the regulatory units (promoters and enhancers) activate genes in complex, dynamic networks. The first model utilizes the change in function caused by deleting genes from a chromosome of *D. melanogaster*. After identifying the genes, we will compare the promoter regions to generate new hypotheses of how the gene switching mechanisms operate for sperm storage after mating. The second model examines changes in the expression of transcription factor genes induced by HIV infection of a human cell line. This directly measures the intricate biochemistry that is usurped by the virus to integrate into the host cell and aid its replication.

1.1 Identifying gene switches involved in sperm storage in D. melanogaster.

One approach to better understand the components of these networks is to selectively 'delete' pieces of DNA and then asses function. In the case of *D. melanogaster*, strains of flies that are missing genes from specified regions of their chromosome are readily available to screen for functional defects. Ongoing research performed by Margaret Bloch Qazi and her colleagues have identified hundreds of potential genes that affect female sperm storage in *D. melanogaster*. Female fertility is determined by developmental processes including oogenesis, ovulation, sperm storage, fertilization, and embryonic viability. Identification of genes whose activity in the female effects sperm fate helps elucidate physiological mechanisms of fertility.

The fruit fly, *D. melanogaster*, is a powerful system to identify these types of genes. Genome/proteome screens have identified genes expressed within the lower reproductive tract (Mack et al., 2006; Swanson et al., 2004) where sperm storage occurs, and whose activity corresponds with female responses to male mating stimuli (McGraw et al., 2004).

The mechanism of sperm storage has broader implications for how cells move to specific locations, how they are recognized and how cells communicate with each other. Movement of sperm into the reproductive tract activates hundreds of genes in the female reproductive tract and can give us insight into the mechanisms of gene activation. Since many biochemical mechanisms are conserved from insects to human cells, understanding how genes are activated in *D. melanogaster* has applications in functioning of human cells.

I am interested in how genes are activated because many human disease states result from the dysfunction in cell recognition, cell signaling and gene activation pathways.

1.2 Animal viruses as model systems to study activation of gene networks.

Viruses that infect mammalian cells have served as a good model system to study the regulation of gene expression. This is because the virus uses the host cell machinery to carry out its own processes therefore providing us with an opportunity to observe regulatory mechanisms operating in animal cells.

Human immunodeficiency virus type 1(HIV-1) is the etiological agent of AIDS. According to the most recent estimates, 36.1 million persons worldwide are infected with human immunodeficiency virus (HIV) and more than 16,000 new infections occur daily (Sepkowitz, 2001). HIV-I is a member of the lentivirus subfamily of retrovirus. All retroviruses contain minimally three genes, gag, pol and env encoding the structural proteins and enzymes required for virus replication. Lentiviruses have a more complex genome than other retroviruses with HIV-1 containing six addition regulatory and accessory proteins (tat,rev, vif,vpu,vpr and nef) that are required for the integration of the virus into the human cell and its replication (Brass et al., 2008)

The remarkable feature of these types of viruses is that just 6 viral accessory proteins usurp hundreds of proteins in the host cell machinery for its own integration into the cell chromosome and replication of viral particles. A knock down screen using small RNA molecules have identified more than 250 human cell HIV-dependency factors are required for the virus life cycle (Brass et al., 2008).

In our studies (in collaboration with Dr. Uckun at the Parker Hughes Cancer Center), we interrogated more than 12,600 gene products using the Affymetrix gene-chip technology; these are microarrays that have small portions of genes bonded to an inert substrate to which a fluorescently labeled sample is added and visualized, to identify genes are switched on/off in the presence of the virus. Microarrays allow the determination of the temporal sequence of host gene induction and suppression during the course of HIV infection. Previously this technology has been used to study host gene expression changes caused by several viruses including cytomegalovirus (Zhu, 1998) and HIV (Corbeil, 2001; Geiss, 2000). Early changes in gene expression were characterized from studies of HIV-1 infection in cultured human immune response cells,CD4+ cells (Corbeil, 2001; Geiss, 2000) confirming the activation of biochemical pathways required for the incorporation of viral DNA into the human cell DNA, cell defense and cell death. Microarrays allow for the temporal sequence of gene induction and suppression to be followed and hence determining the key regulatory events leading to the cell response and then its destruction.

Recently we have reported the characterization of a novel highly potent antiviral drug, Stampidine (STAMP)/HI-113. STAMP, stavudine-5'-[p-bromophenyl methoxyalaninyl phosphate], is a novel aryl phosphate derivative of stavudine (Uckun et al., 2002; Vig et al., 1998). Three strains of HIV virus were used in this current study to determine the common biochemical pathways modified by viral infection over three time points (24 hour, 48 hour and 7 days). These experiments identified both previously defined as well as previously undefined host genes whose expression was altered as the result of HIV infection. In addition, we measured the effects of the anti-viral reverse transcriptase inhibitor, STAMP on these gene expression profiles.

2. Research Goals:

We will obtain two large data sets from the *D. melanogaster* screen and the HIV infection studies. The overall goal will be to characterize how networks genes are activated using these two model systems: genes involved in sperm storage in *D. melanogaster* and networks of genes that are affected by HIV infection of human cells (note that the traditional approach such as pursued by Dr. Bloch Qazi's lab is to identify specific proteins involved in sperm storage).

2.1 D. melanogaster screen

1. Identify all the genes missing from the deleted portions of the second chromosomes that have sperm storage effects.

2. Compare the gene sets from the published gene expression profiles.

3. Identify biochemical pathways represented by the genes from the published expression studies and the deletion mutants.

4. Identify promoter regions for the *D. melanogaster* genes.

2.2 HIV infection studies.

1. Identify genes that are affected by HIV infection of human cells.

2. Determine sets of co-regulated genes using hierarchical cluster analysis. These are genes that show similar levels of expression across all time-points with HIV treatment.

3. Identify biochemical pathways represented by the genes affected by virus treatment at the three time points. Also identify which genes are required for viral replication by using an anti-viral drug, Stampidine, to reverse the viral effects.

4. Identify the promoter regions of the co-regulated gene products.

Both projects will require use of sophisticated statistical techniques applied to large data sets. These include screening of significant effects using multivariate data analysis tools such as hierarchical clustering techniques and analysis of variance. To understand potential functions of genes we propose to data mine using gene sequence data bases (NCBI BLAST for sequence comparisons, Genebank/OMIM for gene descriptions in human cells, Flybase for *D. melanogaster* gene information) and biochemical databases (GenMAPP, TRANSFAC).

3. Project Description:

3.1.1 Screening deletion mutants of D. melanogaster for sperm storage effects.

Potential genes that affect female sperm storage were identified using a *deficiency screen*. A *deficiency* is a type of genetic rearrangement resulting in the deletion of a small number of genes. Normally animals have two copies of every gene. Flies possessing a deficiency have only one copy of a small number of genes and two copies of all other genes. Deficiency screens are used to identify genes that, when present in only a single copy, are insufficient to direct normal characteristic expression (St. Johnston, 2002). Animals showing this condition are called *haploinsufficient*. For example, if normal sperm storage organ development requires two copies of gene *SSO*, then a female from the deficiency strain missing one copy of gene *SSO* is predicted to have malformed sperm storage organs and store fewer sperm. Each deficiency stock contains a deficiency in a distinct location within their genome. By screening all available deficiency stocks, one is screening most of the genome. Since deficiency screens conducted in other laboratories have detected haploinsufficient effects on a variety of *D. melanogaster* traits such as heart rate and headless behavior (Ashton, 2001) it is reasonable to try this type of screen for a complex process like female sperm storage. Changes in progeny production over time were consistent with sperm storage defects in 23 deficiency lines. Of 83 assayed lines spanning the 2^{nd} chromosome, 27.7% had phenotypes consistent with defects in sperm storage: decreased progeny production over time relative to controls. This is consistent with the physical or functional loss of stored sperm.

3.1.2 Proposed data analysis of the D. melanogaster screen.

Gene-ontology and biochemical pathways.

In trying to understand function of genes, databases are now being designed that have standardized terminology for biological processes. Research into genes implicated in cell function can be hampered by the wide variations in terminology that makes it difficult to describe gene function. For example, in searching for new targets for drug action, the scientist may be interested in the gene products that are involved in protein synthesis. Different data bases may describe this process as 'translation', whereas another uses the phrase 'protein synthesis' thereby hindering investigations into what is known about these proteins. We will study these biological processes using a software tool GenMAPP, which allows for the mapping of discovered genes from an experimental manipulation onto these Gene Ontology classifications to create a list of pathways most affected in the cell (Gene MicroArray Pathway Profiler) (Doniger, 2003). The Gene Ontology (GO) Consortium is creating a defined vocabulary of terms for scientists describing the biological processes, cellular components and molecular functions of all genes. GenMAPP organizes the biological processes in a tree-like structure for and links gene-expression data to the GO hierarchy. A gene product may be located in multiple cellular components or active in one or more biological processes to perform many molecular functions. This information can be comprehensively captured using the defined gene ontology terms giving a more complete understanding of gene function. In addition, genes can be organized into to associated pathways so that any cellular process can have many gene components. We were interested in describing the genes implicated in sperm storage effects at the level of known biological pathways using gene ontology classifications (eg. DNA repair, RNA breakdown, Protein degradation, Cell cycle). These pathways will be compared to other published studies

examining sperm storage in *D. melanogaster* that have measured gene expression changes in thousands of genes.

3.2.1 HIV infection of Human T-cells.

We measured levels of RNA transcripts using an U95 GeneChip microarrays from Affymetrix, which interrogates the expression level of 12,625 genes. The chips are widely used for a variety of experimental applications and spotted on each chip are a robust series of controls to minimize chip to chip variation. On this type of an array each human gene is represented by at least one probe set composed of multiple probe pairs (16-20 pairs). Each probe pair consists of two sets of 25mer sequences. One set is a perfect match (PM) and the other set has a 13th base mismatch (MM) to serve as an internal control for the signal produced by the perfect match probe. A quantitative Signal metric was used to measure the level of each transcript on the chip (developed by Affymetrix). The algorithm calculates the signal using a one-step Tukey's Biweight Estimate, which determines a weighted mean that is relatively insensitive to outliers. The estimated real signal is calculated by subtracting the log of the Perfect Match intensity from the stray signal estimate. The mismatch signal is used to estimate the stray signal where appropriate. The probe pair is weighted more strongly to calculate the Signal if the signal is closer to the median value.

All virus treated samples were compared to the no-virus control at each time point (eg. RTMDR treated at 24 hour was compared with 24 hour control without virus). Two metrics were used to select for genes that changed the level of expression. The first metric was obtained by taking the difference in the Signal of the virus treated relative to the control (virus – control). The second metric used the Signal Log Ratios (SLR) calculated by the Affymetrix software. The SLR compares each probe pair on the virus treated sample to the corresponding control array and takes the mean of the log ratios of the two arrays (base 2 is used so that a 1 unit change corresponds to a two fold change). As with the Signal metric the SLR is a weighted mean (Tukey's Biweight). This method minimizes differences due to probe binding co-efficients and cancels out differences due to individual probe pair intensities.

Affymetrix	No drug			1nM STAMP		
No. replicates	24hr	48hr	7day	24hr	48hr	7day
No virus	4	4	1	3	3	3
RTMDR	2	2	1	3	3	3
IIIB	1	1	1	2	1	1
BR92019	1	1	1	2	1	1

Table shows the experimental treatments and the number of replicates for each treatment.

Viral Strains: RT-MDR is an NRTI-resistant and nonnucleoside analog RT inhibitor-resistant laboratory strain of HIV. HIV RTMDR-1/MT-2 (catalog no 252) HTLVIIIB and

92BR019 (catalog no 1778; envelope subtype B) were obtained from the NIH AIDS Research and Reference Reagent Program.

3.2.2 Proposed data analysis of the HIV infection data.

Analysis of Variance.

To extract useful biological significance of gene expression changes, sources of biological and experimental variation have to be adequately accounted for and considered in the statistical design. Analysis of variance (ANOVA) techniques provide an efficient and powerful way to analyze designs in which there are multiple experimental treatments and complex factorial structures. We propose used a one-way ANOVA technique to identify differentially expressed genes using a time or drug factor (JMP Software, SAS, Cary, N.C.) to screen for the most significantly affected genes. Figure 1 shows the genes that were affected by virus treatment alone. We propose to extend this analysis to genes that were reversed by anti-viral drug treatment (STAMP).

Clustering of virus affected transcripts.

Genes that were consistently affected by virus treatment at 24hr and 48hr and 7 days were organized using a hierarchical clustering algorithm (639 genes shown to be affected out of the 12625 measured). The genes are arranged on the vertical axis such that similar expression patterns are placed adjacent to each other ordered by the cluster algorithm. Each square represents the log ratio of the expression value relative to the mean expression of the control group for each gene at each time point. The clustering is represented graphically as a colored image in which the expression indicated by red represents up-regulation (increase in expression relative to mean control signal value) and green represents down-regulation of genes (decrease in expression relative to mean control value). Therefore each gene has an expression profile across all the treatments shown by the pattern of red and green across all treatments. The clustering algorithm finds similar expression pattern of a pair of genes and joins them together, then clusters of genes are joined together to form larger and larger clusters until all of the genes are joined into one giant cluster. The joining of the cluster is shown by the branch structure that connects individual expression patterns with larger groups. Two closely related gene expression patterns show short branch lengths at the point of joining. More distant related expression patterns join at increasing branch lengths.

Figure 1 shows that genes are co-regulated at each of the time points. We propose to map these changes in gene expression using the GenMAPP software to identify the biochemical pathways represented by these gene products such as shown in Figure 2 (MAPK kinase pathway).

Analysis of the gene regulatory regions using TRANSFAC database.

We will compile a list of genes from the 2nd chromosome of *D. melanogaster* and the clustered genes from the HIV infection studies to analyze the untranslated regions of these genes. This will give us deeper insight into the gene regulatory process that lead to activation of hundreds of genes. The complex interplay between the transcription factors in the switching of genes will result in better understanding of how the cell responds to changes in its environment. We propose to use TRANSFAC® and TRANSCompel® databases, which have compiled

published data on factors that regulate gene expression allowing for the generation of new hypotheses for mechanisms of this type of regulation. The primary data in the two databases (e.g. DNA-binding sites in TRANSFAC®, composite elements in TRANSCompel®) are based on experimental evidence and extracted by curators from peer-reviewed papers. Suitable data are identified and entered via an input client, making use of controlled vocabulary and various automated functions, into a relational database. The data can then serve for (sequence-based) predictions by certain programs, e.g. Match™ (5) (for matrix-based transcription factor binding site searches), Patch™ (for pattern-based transcription factor binding site searches) and P-Match™ (6) (for a mixture of matrix- and pattern-based binding site searches). (Matys et al 2006).

The two data sets will yield lists of genes that re affected by sperm entry for the *D. melanogaster* screen and genes switched on/off by the HIV screen. These lists will be used to generate potential gene-regulatory elements from known interactions of proteins and DNA.

*** Data in Figures omitted for confidentiality. Contact S. Qazi (sqazi@gustavus.edu) for more information. –Margaret Bloch Qazi

References

Ashton, K, Wagoner, A. P., Carrillo, R., and Gibson, G. (2001). Quantitative trait loci for the Monoamine-related traits heart rate and headless behavior in *Drosophila melanogaster*. *Genetics* 157, 283-294.

Brass AL, Dykxhoorn DM, Benita Y, Yan N, Engelman A, Xavier RJ, Lieberman J, Elledge SJ. (2008). Identification of Host Proteins Required for HIV Infection Through a Functional Genomic Screen. *Science* Jan 10 [Epub ahead of print].

Corbeil, J., Sheeter, D., Genini, D., Rought, S., Leoni, L., Du, P., Ferguson, M., Masys, D. R., Welsh, J. B., Fink, J. L., Sasik, R., Huang, D., Drenkow, J., Richman, D. D., and Gingeras, T. (2001). Temporal gene regulation during HIV-1 infection of human CD4+ T cells. *Genome Res* 11, 1198-204.

Doniger SW, Salomonis N, Dahlquist KD, Vranizan K, Lawlor SC, Conklin BR. (2003). MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol.;*4(1):R7.

Geiss GK, Bumgarner RE, An MC, Agy MB, van 't Wout AB, Hammersmark E, Carter VS, Upchurch D, Mullins JI, Katze MG. (2000). Large-scale monitoring of host cell gene expression during HIV-1 infection using cDNA microarrays. *Virology*. Jan 5;266(1):8-16.

Mack PD, Kapelnikov A, Heifetz Y, Bender M. (2006). Mating-responsive genes in reproductive tissues of female *Drosophila melanogaster*. *Proc Natl Acad Sci U S A.* Jul 5;103(27):10358-63.

Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE, Wingender E. (2006). TRANSFAC® and its module TRANSCompel®: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res*.; 34(Database issue): D108–D110.

McGraw LA, Gibson G, Clark AG, Wolfner MF. (2004). Genes regulated by mating, sperm, or seminal proteins in mated female *Drosophila melanogaster*. *Curr Biol.* Aug 24;14(16):1509-14.

Sepkowitz KA. (2001). AIDS--the first 20 years. *N Engl J Med*.;344(23):1764-72.

St. Johnston, D. (2002). The art and design of genetic screens: *Drosophila melanogaster*. Nature Reviews 3, 176-188.

Swanson WJ, Wong A, Wolfner MF, Aquadro CF. (2004). Evolutionary expressed sequence tag analysis of Drosophila female reproductive tracts identifies genes subjected to positive selection. *Genetics*;168(3):1457-65.

Uckun FM, Qazi S, Pendergrass S, Lisowski E, Waurzyniak B, Chen CL, Venkatachalam TK. (2002). In vivo toxicity, pharmacokinetics, and anti-human immunodeficiency virus activity of stavudine-5'-(p-bromophenyl methoxyalaninyl phosphate) (stampidine) in mice. *Antimicrob Agents Chemother*.;46(11):3428-36.

Vig R, Venkatachalam TK, Uckun FM. (1998). D4T-5'-[p-bromophenyl methoxyalaninyl phosphate] as a potent and non-toxic anti-human immunodeficiency virus agent. *Antivir Chem Chemother.;*9(5):445-8.

Zhu H, Cong JP, Shenk T. (1997). Use of differential display analysis to assess the effect of human cytomegalovirus infection on the accumulation of cellular RNAs: induction of interferon-responsive RNAs. *Proc Natl Acad Sci U S A*.;94(25):13985-90.

4. Nature of the Collaboration:

Jason was a Mayo Scholar and researched the potential for pharmacogenomic discoveries to improve disease diagnosis and prediction of drug toxicity. He has a good appreciation of the limits and promise of the new genomic technologies that will serve us well during the summer research.

I intend to support this collaboration several ways. First, I have already developed the investigations and general methodology upon which Jason will elaborate. He will learn how to explore high level research grade databases for scientific investigations (GenMapp, TRANSFAC, BLAST, FlyBase). I will teach him how to analyze high-dimensional datasets that have hundreds of treatments and thousands of variables to build new models and generate new hypotheses for future experiments. I will mentor Jason giving him room to develop his own ideas and interpretations, but will also monitor and discuss his progress with him. Third, I will encourage Jason's participation in the undergraduate summer programs in Nobel Hall (Sigma Xi and Merck). These programs will provide Jason with opportunities to interact with other summer research students and a forum in which to present ongoing research informally during the summer and again formally in a research symposium at the end of the summer. While I do not 'offer' these programs to Jason, they are available to him and will contribute to her training and our collaboration.

Jason is inquisitive, dedicated, intelligent, and very eager to conduct this type of informatic investigation. I expect that his level of sophistication in experimental design and analysis will increase greatly from this experience. I am confident that Jason will contribute novel perspectives to the interpretation of the screen results by analyzing and interpreting the vast amount of data I will make available for him. Finally, Jason has strong writing skills. I anticipate that he will contribute to the writing of both manuscripts and his perspective will be a unique and invaluable resource as we develop theory.

I expect that through our discussions of both primary literature and our experimental results Jason and I will develop a better understanding of fly genetics and applying bioinformatics to clinical disease states.

Jason is also planning to write a senior honors thesis to research if scientific data is enough to enable genomic technologies to be accepted into mainstream healthcare. His knowledge gained from the summer research experience will serve well to develop the thesis from a deeper understanding of these types of genomic screens.

B. A clear statement of anticipated outcomes

My main goal at Gustavus is to establish an initiative for Bioinformatics in Clinical Research. I expect to continue four major collaborations from this initiative: Dr. Fadil Santosa at the IMA, University of Minnesota to participate in multiple workshops and research programs examining high dimensional data sets; Dr. Nilima Nigam at McGill University to develop dynamic models for gene regulation; Dr. Fatih Uckun for providing data from clinical trials and treatment of patients at the Parker Hughes Cancer Clinic; and Dr. Margaret C. Bloch Qazi analyzing the *D. melanogaster* screen data.

The proposed analysis is expected to generate 4 papers within a year. Two papers for the *D. melanogaster* project: one will report the main biochemical pathways affected by sperm transfer; the other paper will report the findings of the binding sites that may be involved in the gene switching. The two papers for the HIV infection study will report: the clusters of genes and biochemical pathways induced by virus invasion; and the other will report the transcription factors that may be recruited by virus proteins and used for integration. Any similarities in the gene regulation or the genes between the *D. melanogaster* and HIV screens will enable the development *D. melanogaster* as a model system to study gene regulation applicable to human disease. The papers will serve as foundations for grant applications to NIH and NSF.

C. Likely Placement for Publication

BMC Bioinformatics for two *D. melanogaster* papers.
Journal of Virology for the HIV infection papers.

D. Anticipated research completion date

Phase I: Compile lists of genes by the end of Spring Semester.
Phase II: Analysis of regulatory units and writing of manuscripts by the end of summer

II. Participant details

A. Names and brief biographies of participants

1. Sanjive Qazi

I received my Ph.D in the department of Biochemistry at Bath University (U.K.) characterizing muscarinic acetylcholine cellular receptors in the locust brain (1992). I gained fundamental training in basic biology using insect model systems during my academic fellowships At Tufts University (1992-2000). Here I used enzyme/binding kinetics, computational biology, biophysics to understand information processing in caterpillar locomotion. This experience enabled me to apply this knowledge in cancer therapy and anti-HIV drug development programs at Parker Hughes Clinics and Paradigm Pharmaceuticals (2000-2007). This is because many drugs used for cancer therapy target cellular receptors. My statistical understanding of noise in data proved to be useful in cancer patient survival outcome analysis; assisting the drug development teams in IND preparations for a promising anti-HIV compound (Stampidine); Internal Review Board evaluations of toxicity and efficacy, studies. In the research realm I tackled the statistical issues that arose from measuring thousands of variables from high-throughput technologies using relatively small sample sizes.

My work with Parker Hughes Cancer Clinics required me to innovate statistical techniques focused on delivering patient-tailored care at the facility. In this scenario, each patient was screened with imaging technologies, cancer markers and drug sensitivity trials. It was my goal to extract statistically significant factors in the patient objective response given the multitude of different variables. My work was performed in collaboration with clinicians, nurses, scientists and students. Success in my data analysis resulted in a peer-reviewed clinical journal publication highlighting prolonged survival of patients treated at our facility. It would be a pleasure to offer my perspective to students wanting to pursue health related professions.

I have mediated work in multi-disciplinary teams to gain novel insights into biological function and therapeutics: last two jobs resulted in co-authorships in more than 40 peer reviewed articles with 24 Scientists, 5 Graduate students, 5 Undergraduate students, 3 Oncologists and a Nurse.

My research uses both theoretical and experiential knowledge in diverse collaborative projects to tackle a wide range of problems in the preclinical development of 6 drugs (anti-cancer, anti-HIV and anti-allergy); published 6 papers on receptor characterization, 5 on signal transduction pathways, 10 papers screening for drug resistant HIV strains, 5 on structure activity relationships, 3 on *in-vivo* toxicity studies, 3 on drug formulation, 8 on pro-drug activation and 3 on clinical efficacy of anti-cancer drugs. I implemented data standards for hypothesis testing and federal regulatory requirements in the award of grants to fund drug discovery efforts (anti-HIV, anti-cancer compounds).

I gained extensive experience using analytical tools to identify key gene expression changes induced in cancer cells or virus infections using high-throughput screens (>12000 genes interrogated) requiring in-depth knowledge of data structure: probability of finding false negative, true negative, false positive and true positive genes, number of outliers, data inconsistencies, correlation patterns and trends.

2. Jason Pitt

I was born and raised in Faribault, Minnesota. There I attended Faribault Senior High School where I graduated with high honors and was captain of the baseball team. Though I no longer play baseball, I was fortunate enough to coach the Faribault VFW baseball team both as an assistant and head coach.

Currently, I am in my junior year at Gustavus and am a biology and philosophy double major. My academic accomplishments include dean's list honors during all semesters of attendance, as well as being named to the President's list three consecutive years. I was a keynote speaker on philosophy, science, and religion at the Association of Congregations Conference spring of 2007. In fall of 2007, I was a teaching assistant for general chemistry and introductory biology laboratories. I have conducted research on gender communication with Dr. Richard Martin, and have research patented technologies pertaining to pharmacogenomics as a part of the Mayo Clinic's Mayo Scholar's Program. I am also currently organizing a senior biology honors thesis that with emphasize biomedical ethics relating to genetic predisposition to both breast cancer and leukemia using recently acquired clinical research data. Furthermore, I will be writing a philosophy honors thesis most likely focused on human ethical development as related to evolution.

My extracurricular involvement at Gustavus has included Peer Assistants, and the Servant Leadership Program. Within the St. Peter community I have been trained as a Sexual Assault Advocate by Crime Victim Services of Nicollet County, and have been working as an emergency medical technician for St. Peter Community Hospital and the auxiliary St. Peter Area Ambulance since fall 2006. Outside of St. Peter, I have been a now and again intern for the emergency department at District One hospital in Faribault, and I am also a recent member of The Coalition for Genetic Fairness.

Once my time at Gustavus has concluded I plan to attend medical school, while also considering the possibility of a joint degree in public health or law. By participating in bioinformatics research with Dr. Qazi I hope to not only attain significant experience in a dynamic field, but also receive subsequent direction as to further career opportunities and interests.

B. Explanation of how this project fits into the career of the faculty

I am establishing an independent research program, which will form the foundation for studying Bioinformatics in Clinical Research. I plan to include undergraduates in this research to the extent that I can fund their activities. Once complete, the results are sufficient for submission to peer-reviewed scientific journals. Finally, preliminary results from these studies will be used to apply for future funding from internal and external sources. I plan to further develop collaborations with University of Minnesota (Dr. Fadil Santosa, Mathematics), McGill University (Dr. Nilima Nigam, Mathematics) and Parker Hughes Cancer Center (Dr. Fatih Uckun, Oncology).

Including undergraduates in my research will provide students with training in the the new and emerging fields of Bioinformatics, Pharmacogenomics and Systems Biology. In addition to working collaboratively on experimental design and execution, students and I will read and discuss primary literature. These conversations are important for developing understanding, generating hypotheses, and supporting arguments.

I thoroughly enjoy working with undergraduates and in my previous positions I always included them in my research plans that led to peer-reviewed publications. I find great creativity in young minds from the kind of mentoring relationship that can develop between teacher-scientists and student-scientists. More practically, conducting research with students greatly aids me in understanding how they grasp concepts; in itself an enlightening process. This understanding, in turn, informs how I can develop my own instruction in the lab classes I run at Gustavus.

C. Explanation of how this project fits into the educational trajectory of the student.

Like many other students, I was first introduced to genetic analysis and bioinformatics in a Cell and Molecular Biology course. In our laboratory, we inserted a pSK plasmid into bacterial cells via a virus. If the plasmid was integrated into the genome of the bacterial cells we expected to see variance in growth responses between experimental bacteria and their controls when plotted onto MacConkey plates containing various antibiotic compounds. We also analyzed DNA sequences of the experimental bacteria, entered the subsequent data into the BLAST search engine, and determined the identity of the particular gene the plasmid inserted into the bacterial genome. Though challenging at the time, I was intrigued how easy it was to manipulate the genetic makeup of an organism.

I was then privileged enough to be selected as a part of the Mayo Scholars Program to conduct research on patents concerning pharmacogenomics. Pharmacogenomics is the study of how different therapeutic drug metabolization responses can occur in an individual due to variances found in his or her genome. Our patents contained variants in genetic sequences that occurred on both coding and non-coding regions of genes. Some of these patents were potential indicators of predisposition to various cancers, joint-disease, mental disorders, and asthma. During this project, I was thoroughly exposed to bioinformatics due to massive amounts of variables, which are both genotypic and phenotypic, contained in each individual's genome that could influence drug response. Furthermore, I was responsible for any ethical considerations that may arise during the pharmacogenomic testing and implementation process. Consequently, I was not only exposed to the high detailed research of bioinformatics, but also to the delicate ethical implications included in any research with genetic foundations.

Analyzing the genetic response to HIV infection and treatment has many similarities to that of pharmacogenomics. This is apparent when analyzing the microarray data sets provided by the Affymetrix GeneChip where we see similar trends in patient response to treatment, but yet certain individuals had greater response to drug therapy than others. By undergoing genetic research on genes involved with HIV infection pathways will bring a wide variety of benefits. Primarily, I will be exposed to the field of virology, which is one I have not had the opportunity to pursue. Also, I will be able to expand knowledge of pharmacogenomics and gene expression obtained during the Mayo Scholars program. Since much emphasis will be place on the human gene expression related to HIV, I will also learn how model systems can be applied to clinically derived data by Dr. Qazi's intentions of performing gene knockout studies on D. melanogaster. These knockouts will allow me to expand my knowledge in a field that first grasped my interested in genetics. The bioinformatics, pharmacogenomic and well as model system aspects of this project will be tremendously helpful while conducting my senior biology honors thesis, which will give recommendations and guidelines as how to ethically conduct testing for pharmacogenomic responses in a clinical research setting using gene profiling data. Along with my honors thesis, I intend of continuing genetic HIV research with Dr. Qazi during the 2008-2009 school year.

Working under the guidance of Dr. Qazi, I expect this experience to be an amazing learning opportunity, as well as a fun and exciting way to spend my summer. He is an excellent instructor and has given me knowledge about many scientific and non-scientific topics as well as career paths. Currently, I plan to apply for medical school in fall 2008. I have considered pursuing a joint degree in medicine and law or public health. However, I hope to gain further direction as to my career goals by participating in this research project.