**Identified transcription factor binding sites:** In my first undergraduate research experience, I worked in Prof. Martha Bulyk's lab at the Division of Genetics at Brigham & Women's Hospital. The Bulyk lab studies regulation of transcriptional networks by performing protein-binding microarray (PBM) experiments. PBMs are double-stranded DNA microarrays, a completely new technique to assay binding sequences of transcription factors (TFs). I incubated the PBM with purified TF, fluorescently tagged the TF, then scanned and processed with software that detected each spot on the PBM. I performed further analysis with lab software to detect the binding preference of each TF with weighted nucleotide positions in the hypothesized 8-mer binding site. This produced a Logos binding motif, an image of DNA letters where the size of the character corresponded with the strength of binding. I discovered the binding sites of mouse homeodomain transcription factors, known for their critical, but not fully understood role in development. The results of my experiments contributed towards a 2008 publication in the journal *Cell*. Understanding what sequences bind murine TFs can advance medicine by applying this knowledge to corresponding human systems such as cancer, atherosclerosis, and diabetes, diseases commonly found with mutations in TFs and their binding sites. This research experience exposed me to biological computation, and I was fascinated by how a computer could interpret the minutely polka-dotted microarrays and transform them into binding motifs, which inspired me to further explore the field of quantitative analysis of biology.

**Created robust null models in sequence analysis:** After my experience in the Bulyk lab, I was admitted to the competitive Janelia Farm Summer Scholars program to work with Prof. Sean Eddy at Howard Hughes Medical Institute Janelia Farm Research Campus in Ashburn, Virginia. One of the main projects is HMMER, software that performs searches of biological sequences using homology, or evolutionary relatedness, which can help elucidate how genomes can change over time. These homology searches are performed using Hidden Markov Models (HMMs), a sequence of states, where each state encodes a pattern of the protein, such as a domain. The homology of two sequences is evaluated by first aligning each sequence to an HMM, then comparing the distance between the two HMMs to their distance from the null model, an HMM of a generic protein. This generic protein does not exist biologically, but must be created *in silico* to compare homology. I used HMMs to create a more robust null model for HMMER homology searches. I completed this project in two months in the Python programming language and spent the last month rewriting the software in C, the language of HMMER. My project provides a more accurate analysis of sequence homology, improving engineering of nucleic acids and proteins in basic science research, which can have broad impact in therapeutic design in medicine.

**Analyzed synthetic lethality in yeast genome networks:** Wishing to continue genome research, I worked for 6 months in the MIT Computer Science and Artificial Intelligence Laboratory in Prof. David Gifford's lab. I worked on yeast metabolic genome networks found by protein-protein interaction data. While it is known that different yeast strains express different genes, it is not known how various functional combinations of genes are chosen, and especially how deletion of one gene can kill one strain but not confer lethality in another. These genome networks can be modeled where each node is a protein, and each edge is the interaction between the two, and the weight of each edge representing the binding affinity. For this project, I learned graph theory and the Java programming language to implement the network. It is difficult to predict which genes in a network are lethal, and I hypothesized that methods of information content in graphs can inform lethality of a gene. However, my metrics indicated biological networks are so complex, that *in silico* genome networks must expand their model to incorporate epigenetics and metabolites to create a robust computational representation of an organism's

livelihood in response to synthetic lethalities, paving the way for multicellular organisms and tissues, allowing for safe *in silico* testing of pharmaceuticals.

**Detected immune response by T-cell receptor sequence enrichment:** My second project in the Gifford lab was for 6 months, where I worked in collaboration with Prof. Hidde Ploegh's Whitehead Institute lab to parse and analyze T-cell receptor (TCR) sequences. T-cells are a key component of adaptive immunity and bind invading antigens with their receptors. Thus their receptors must be highly varied to provide immunity against a wide range of antigens. The diversity in TCR sequences is determined by three regions, variable (V), diversity (D), and joining (J). Each region is coded into the genome in multiple variations to allow for inherent diversity with over a million possible VDJ combinations. I created random TCR sequences and tested how frequently a particular VDJ sequence must occur for detection by statistical methods. I found that even a slight (2x) increase in frequency was detectable, a promising result for diagnosis as invasion by an antigen increases the frequency of a particular VDJ combination by 10x. This is helpful not only for detecting single sequence enrichment, but also in more real-life situations where many VDJ combinations may be enriched at once. This project was an exciting convergence of bioinformatics research and clinical medicine as its implications allow health professionals to screen an individual's vaccination and antigen exposure history.

**Discovered neuron orientation via computational image analysis:** As I continued through my genetic research, I began to wonder how else biological data could be examined. I joined Prof. Sebastian Seung's lab at MIT's Department of Brain and Cognitive Sciences to pursue computational image analysis. Using three-dimensional electron microscopy images gathered from shaving off slices of brain tissue, neurons are delineated, or segmented, from one another. I analyzed a segmentation of neuronal tissue from rabbit retina, specifically the inner plexiform layer (IPL). The IPL is highly layered and plays a critical role in visual signal processing. My analysis of the IPL segmentation revealed that most neuron fragments in this 3D image are fairly flat along the axis of information transmission, and do not significantly traverse the direction perpendicular to of signal flow. Since in biology, structure and function are interrelated, this showed that flow of visual information is the same as the orientation of neuron fragments.

**Revealed new cancer activators via integrated genomics:** Most recently, I worked at the Broad Institute of Harvard and MIT in Cambridge, MA, where I worked with Prof. Jill Mesirov and in collaboration with Prof. Todd Golub to develop REVEALER, an algorithm that integrates genomic and functional data to infer new associations. For example, a researcher may know that a certain oncogenic gene signature is overexpressed in many cancers, and that this overexpression can be caused by a mutation in the governing oncogene. However, there are many cases where there is no mutation in that oncogene and yet the pathway is highly expressed. REVEALER finds novel candidate activators of this signature by removing samples which already have a mutation in the original oncogene, and searching for the top genomic feature that explains the remaining samples. This ambitious project combines state-of-the-art computational methods with genomic analysis in the same way I hope to do in my future research. My interest in personal genomics is particularly suited towards cancer research, where the role of genomics tends to be much more significant in the expected clinical outcome of cancer patients than for patients of other diseases.

**Publications:** Berger MF *et al.*, *Cell* (2008); Botvinnik OB *et al.*, *Intelligent Systems for Molecular Biology Conference* (2011); Botvinnik OB *et al.*, REVEALER paper (in preparation); Wood KC *et al.*, experimental application of REVEALER (in review).

**Differential gene activity of adjacent cells elucidates intercellular interactions**

**KEYWORDS:** RNA-Seq, single-cell analysis, nanopipette, biofilms, networks, bioinformatics.

**BACKGROUND AND MOTIVATION:** High-throughput RNA sequencing (RNA-Seq)[1] has revolutionized gene expression analysis by producing a high-resolution snapshot of a cell's transcriptome. Recent advances allow RNA-Seq with very small sample sizes[2] paving the way for single-cell transcriptome analysis.

Single-cell RNA-Seq is advantageous as they allow for robust determination alternative splice sites, a key component of transcriptome analysis[3]. By analyzing single-cell transcriptomes cross-referenced with physical cell positions, I will use differential gene expression programs of adjacent cells to infer cell-cell interaction networks.

Intercellular interactions play a key role in biofilms of pathogens. Biofilms are the primary mode of bacterial infection in humans, especially in catheter and prosthetic heart valve infections of hospital inpatients. Biofilms improve resistance to antibiotics by 10-1000 fold[4], and recent advancements allow for high-throughput visualization of biofilm formation and survival of *Vibrio cholerae*[5]. *V. cholerae* therefore is an excellent model system for developing methods of analyzing interactions between adjacent cells. I will study single-cell *V. cholerae* transcriptomes to elucidate drug resistant intercellular interactions.

*HYPOTHESIS: Differential gene expression (via RNA-Seq) of adjacent drug resistant and respondent* V. cholerae single *cells within a biofilm uncovers intercellular interactions.*

**RESEARCH PLAN:** I will develop an *in silico* model of *in vitro* heterogeneity of *V. cholerae* by analyzing single-cell RNA-Seq data of several populations and time points. Heterogeneity of cell populations will be elucidated by observing differential expression of adjacent cells and deducing intercellular interaction networks. I hypothesize the heterogeneity of a cell population can be modeled by extracting network information from many single-cell transcriptomes.

As an example collaboration of how this research plan can be implemented, I will work at University of California, Santa Cruz (UCSC) supervised by Prof. Yildiz, Prof. Nader Pourmand, and Prof. Josh Stuart, harnessing the strengths of Yildiz's extensive work in *V. cholerae* biofilms, Pourmand's nanobioinstrumentation and Stuart's algorithmic systems biology expertise.

**AIM 1: Develop RNA-Seq algorithms for analysis of single-cell transcriptomes**

Miten Jain, a fellow graduate student in Prof. Pourmand's laboratory with expertise in nanopipette technology will perform the wet-lab experiments on the eight primary strains of *V. cholerae* used by the Yildiz lab[5]. We will use untreated, and both sick and healthy survivors of antibacterial agents for analysis. Apoptotic cells cannot be used as their intracellular molecules are not localized within a cell. Miten will extract RNA from adjacent diploid cells using nanopipette technology, a novel method in the Pourmand lab using nanometer-sized pipettes to probe single cells and carefully remove RNAs[6]. Miten will use these RNAs to create a cDNA library, and then amplify and sequence the cDNA using an Ion Torrent sequencer.

Current RNA-Seq assemblers do not take into account the sample preparation used to generate the reads, and as we will use a novel nanopipette technique to extract RNA and a specialized in-lab primer library to amplify the cDNA, we require specialized sequencing assembly bioinformatics. I will develop both *de novo* and reference-aligned RNA-Seq assemblers based on previous algorithms[7-8] and collaborate with Miten to design custom methods to account for our unique sample preparation. Specifically, our primers tend to join together, creating "alien" DNA fragments which pollute the read library. To ensure algorithm accuracy, I will test on sequencing reads created from spiked-in known mRNA transcripts.

**AIM 2: Build a network model of intercellular interactions between neighboring cells**

After establishing robust RNA-Seq algorithms I will analyze differential transcripts between pairs of neighboring cells by using the Stuart Lab's PARADIGM[9], a statistical software tool used to infer differential protein activity. Analysis will test differences between two neighboring cells at a time and importantly, must account for differences in cell cycle stage. Using documented checkpoint genes, I will pinpoint the cell cycle phase of each cell, and remove protein isoforms known to be expressed solely as a result of cell cycle[10].

After accounting for cell cycle-specific transcripts, I will analyze differential gene expression programs between neighboring cells in both WT and antimicrobial conditions. I will identify transcripts to differentiate control, drug-respondent, and resistant cells using TopModel (unpublished), a Stuart lab classification program that tests hundreds of machine learning algorithms before settling on the best classifier. I expect to see active transcription in secretory pathways of neighboring respondent and healthy survivors, indicating communication between these cells and identifying specific transcripts potentially useful to survival.

Based on the amount of differential sickness between neighboring cells, I expect to see varying degrees of signaling from the sick cell and responsive secretion from the healthy cell. However, even in comparing cells with similar drug susceptibility, there will be a spectrum of highly expressed transcripts. Using these differentially expressed genes and known *V. cholerae* interaction networks[5,10], I will develop a network of cell-cell interaction between adjacent cells in response to antimicrobial agents. Using intercellular interactions implied by proteins and micro-RNAs found by RNA-Seq, Miten will functionally validate their effects by injecting these biomolecules into unhealthy *V. cholerae* and observe rescue of the survival phenotype.

**BROADER IMPACTS:** Beyond the scope of my work, further extensions of this project could include incorporation of genome sequencing and DNA methylation analysis, time-series experiments to observe changes in cell-cell interactions over time, distinguishing secreted RNAs from endogenous RNAs, and metabolite profiling[11]. In each of these extensions, the main challenge lies in the extremely small sample size of a single cell. The mRNA, DNA, and metabolites would need to be removed simultaneously from a single cell and specifically separated for analysis. Additionally, the replicability of such experiments remains a constant challenge as cells proliferate and accumulate genomic abnormalities affecting nucleotide and metabolite levels.

In collaboration with the Broad Institute and UCSC, I will develop the first-ever *V. cholera* database, which will empower the broader scientific community to design more targeted experiments and avoid redundant studies. Final versions of networks I create from this study will be deposited in this database. Understanding *V. cholerae* population heterogeneity will help develop better antibacterial agents, and reducing inpatient infections.

Importantly, the development of single-cell methods from my work paves the way for analyzing heterogeneity of other organisms, allowing for previously unimagined granularity of intercellular interactions under transcriptional control. Additionally, this work lends itself to analyzing heterogeneity within a tissue, and is especially relevant to tumors as their resistance to drugs may be attributed to survival of individual independent cells, or to groups interacting cells.

**REFERENCES:** 1. Wang ET *et al.*, *Nature* (2008). 2. Tang F *et al.*, *Nat. Meth.* (2009). 3. Modrek B *et al.*, *Nuc. Acid Res.* (2001). 4. Rasmussen TB and Givskov M, *Int. J. Med. Microbiol.* (2006). 5. Peach KC *et al.*, *Mol. Biosyst.* (2011). 6. Karhanek M *et al.*, *Nano. Lett*. (2005) 7. Grabherr MG *et al.*, *Nat. Biotech.* (2011). 8. Trapnell C *et al.*, *Nat. Biotech.* (2010). 9. Vaske CJ *et al.*, *Bioinformatics* (2010).
10. broadinstitute.org/annotation/genome/vibrio_cholerae/ 11. Rubakhin SS *et al.*, *Nat. Meth.* (2011).

Fascinated by my middle school genetics class as the Human Genome Project reached its fruition, I learned early on that the genome is a powerful tool with profound application to understanding biology and became obsessed with genetics. I worked in a nematode lab at the nearby University of Oregon, where I realized genetics in an organism is far more complex than Punnett squares. I attended Massachusetts Institute of Technology (MIT) where I was exposed to high-throughput genomics, proteomics, and other -omics technologies which inspired me to pursue computational biology research. To maximize my impact on revolutionizing biological research through high-throughput genomics, I want to be a professor at a major research institution, where I will foster an open-source atmosphere by publishing groundbreaking papers, mentoring students, and organizing international conferences.

At MIT, I was exposed to a variety of aspects of the scientific process—research, creativity and teaching. I had strong computational biology experiences throughout college, which led to graduating with dual degrees in mathematics and biological engineering, one of two people in class of one thousand. I was also heavily involved in MIT Dancetroupe (DT) and while I never danced before college, I quickly picked it up and became a leader in DT. As a choreographer, my specialties were in hip-hop and "tutting," sharp, angular hand and arm movements named after King Tutankhamen for their resemblance to Egyptian silhouette paintings. Outside of creating imaginative choreography, I enjoyed teaching and was especially proud when I watched formerly shy students perform boldly onstage. To balance my craze for dance, I also find solace in playing the cello, which I have done on and off since fourth grade. I began playing again after college, when I was able to rediscover my passion for the instrument and even put on a small recital amongst my close friends.

After graduating, I trained as a bioinformatician in cancer genomics at the Broad Institute of Harvard and MIT. The interdisciplinary environment was inspiring because the algorithm design process was not only an iterative process, but also a convergent symbiotic evolution of biologists and mathematicians. I twice presented my work from the Broad Institute this summer at Intelligent Systems for Molecular Biology (ISMB), the largest bioinformatics conference and at the ISMB Student Council Symposium (SCS), and was invited to co-chair ISMB SCS 2012. Planning this event has been an exciting international experience, collaborating with peers from India, Brazil, Belgium, and Nigeria to continue the circulation of innovations in bioinformatics. I am looking forward to cross-cultural collaboration and providing the opportunity to attend ISMB to students from developing countries.

As an immigrant from the former Soviet Union, I am well aware of the difficulties faced by immigrants and subtle discrimination present in both professional and private life. I sought to ensure that new immigrants had equal opportunity to healthcare through equal access to healthcare and obtained a certificate in Russian-English medical interpretation. Through interpreter training, I was exposed to a variety of ambiguous ethical situations, such as maintaining confidentiality in interpreting for the same patient but for different practitioners, ameliorating reaction to hearing a cancer diagnosis, and experiencing disrespect to female interpreters--interpreters in Russia are primarily male. I am well aware of the Russian discrimination against women, which my mother was exposed to when she was attending the prestigious Moscow Institute of Physics and Technology, known as the "Russian MIT." There, one of her male peers asked her, "What are you doing here? Why aren't you at home having babies?" To advocate for women in Russia, this summer I will travel to Moscow to help develop the biomedical research programs at the MIT/Skolkovo Institute of Technology (SkTech), where

I will organize women's leadership workshops in addition to interviewing potential researchers for SkTech.

Women's inequality in the USA is subtler than in Russia. Instead of overt sexism, there are subtle barriers to women achieving success in both technical and nontechnical fields, and I seek to eradicate them. As a volunteer for Science Club for Girls (SCFG), a Boston-based afterschool program for girls K-6 to get excited about science and math, I taught an anatomy curriculum to 2nd graders and loved seeing their excitement about science. These were easily the most rewarding hours of my week. But when I judged a science fair at an inner-city middle school, I was disappointed. This school didn't have SCFG and I saw that regardless of race, the greatest divide between the fantastic and the mediocre science projects was gender. The girls did not attempt the hard, high-risk projects. Instead they did the tried and true: making crystals out of Borax at different temperatures, investigating invisible ink, and growing plants with varying amounts of sunlight. The boys made a bike that could charge a cell phone, a drum set out of terra cotta plant pots and a piezoelectric, and a static electricity generator. I was dumbfounded that at 7th grade, the gender gap in scientific ability was so marked. I realized while I could teach 2nd graders about exciting science, we as a nation need to shift our messages of women in science.

Women in the media are not portrayed as strong, capable, and smart. Instead they are over-sexualized, manipulative, and vapid. I am a role model through *Nerd Girls*, a national movement to dispel myths about women in science and engineering. Girls today are pressured into looks and not brains as a result of a lack of media coverage of smart women. Through Nerd Girls, I am creating stories for young women to look up to and be inspired. I am also involved in developing modules for AP Biology to help teach biological concepts though bioinformatics at local Santa Cruz high schools. Additionally, I am organizing a screening on UCSC campus of *Miss Representation*, a new documentary exploring the dearth of strong and capable women in media, and will lead a discussion of the film afterwards. There is an accompanying K-12 curriculum to the film and after this initial screening, I will teach this curriculum to local Santa Cruz schools. I want to use this film as a forum for students of color, as well. While I am Caucasian, I recognize that minorities are highly underrepresented in the media's depictions of scientists and high-ranking officials in both the public and private sector, and this needs to change. A discussion of how media portrayal of women and minorities colors perception of their abilities should occur in groups of all genders and races, for only when all parties realize that discrimination occurs can it stop.

My fluid transitions between research, dance, music, service, mentoring, have enabled me to look at bioinformatics in new ways and push the cutting edge of innovation, positioning me to become a professor and teach the next generation of scholars. As a professor, I will have a broad impact by working at the interface of biology and computation by obtaining appointments in both biology and applied mathematics, where my quantitative skills will filter noisy biological data into novel discoveries. I will be a "collaboration junkie," (source: Prof. Aviv Regev) and seize opportunities to work with people of diverse experiences and expertise because our strengths together can solve problems in completely unexpected ways. I had the opportunity to witness groundbreaking research occur at the intersection of disciplines because I had the freedom to forge novel collaborations between computation and biology, I am now doing a Masters in one year in Bioinformatics at the University of California, Santa Cruz (UCSC), where I am deepening my understanding of computational biology. The NSF Graduate Research Fellowship Program would give me the understanding to pursue more substantial research projects that involve both computation and biology at UCSC.

# Score for Botvinnik, Olga

## Intellectual Merit Criterion

### Overall Assessment of Intellectual Merit

Excellent

### Explanation to Applicant

Very well written proposal that outlines some very interesting and relatively sophisticated research plans. The student clearly has the necessary requisites to succeed in her chosen career and I found her personal statement and philosophy compelling.

## Broader Impacts Criterion

### Overall Assessment of Broader Impacts

Very Good

### Explanation to Applicant

The candidate provides a fairly good documentation of previous activities that meet the broader impacts criteria laid out by the National Science Foundation. A more detailed plan for her future activities in this realm would strengthen the proposal.

# Score for Botvinnik, Olga

# Intellectual Merit Criterion

## Overall Assessment of Intellectual Merit

Very Good

## Explanation to Applicant

The applicant has a fine scholastic record and extensive undergraduate research experience that was well-described. Although the proposed research plan is well-written, with clearly-articulated hypotheses and aims, it is not clear how much of it was independently-conceived. The research is on the cutting edge of the biological sciences.

# Broader Impacts Criterion

## Overall Assessment of Broader Impacts

Excellent

## Explanation to Applicant

The applicant has disseminated her results in presentations and seminal publications, and she is passionate about the importance of conducting research that is collaborative and international in scope. The applicant is an eloquent spokesperson on the role of women in science, and she discusses ways to enhance the participation of women in science, She has also been involved in outreach to K- 12 students. The overall importance of his research for scientific understanding was well-described. The applicant could better articulate how the research will benefit society.

# Score for Botvinnik, Olga

## Intellectual Merit Criterion

### Overall Assessment of Intellectual Merit

Excellent

### Explanation to Applicant

The applicant has intellectual capabilities and demonstrated research experiences; she has designed and conducted experiments, generated and analyzed results, and published some of them. She has demonstrated that she can work as part of team or as an independent researcher.

## Broader Impacts Criterion

### Overall Assessment of Broader Impacts

Very Good

### Explanation to Applicant

The proposed research is an important one and once completed successfully, it is expected to generate new knowledge that will benefit society. The applicant has a history of successful mentoring and empowering others including women. The current research plan does not include plans for outreach/training/dissemination.