Field of Interest and the Role of Computational Science. Please describe your chosen research area and what contributing role computational science will play. Computational science involves the innovative and essential use of high-performance computation, and/or the development of high-performance computational technologies, to advance knowledge or capabilities in a scientific or engineering discipline. (294/300 words)

It's important to me to pursue something not just because it's interesting, but because it matters. I first learned of bioinformatics upon the success of the Human Genome Project, and was excited by the potential of personalized medicine. Armed with the genome sequence, scientists used bioinformatics to help create drugs to target specific disease-causing mutations, but didn't always understand when the drugs weren't effective. I am electrified by this complexity and am thrilled to use bioinformatics to solve current and future biological problems.

Current methods to study a complex disease such as cancer use homogeneous cell lines to model tumors, which in reality are a chaotic mix of different, interacting cell types. Thus, while the majority of a tumor may respond to treatment, a subpopulation may be resistant and cause future relapse. We need to study these subpopulations to understand how the variance within the mosaic of individual cells help the larger cancer survive. In my graduate work, I will study individual cells' differences in RNA regulation and their effect on disease via high-performance computation on sequence data with Prof Gene Yeo at UCSD, who has a one of a few single-cell sorting devices.

RNA regulation is critical as mutations in non-coding regions such as splice sites, and the process of RNA editing, modifying the transcribed RNA before translation to protein, are known to be altered in in cancer. Yeo's lab is at the forefront of alternative splicing research and is the best place for me to study cancer survival at the single cell level. To

Research Using High-Performance Computing and/or Large Data Analysis. What new science or engineering would high performance computing or large data analysis and management enable in your area of interest and why do you think this is the case? In particular, what are the challenges that need to be addressed to make this advancement? (328/300 words)

Large data analysis and management would enable comparison of sets of 96 individual cells produced by UCSD's Prof Yeo's single-cell sorter. Currently, sequencing reads from the RNA transcriptome are mapped onto a "reference" sequence. However humans differ from one another by 0.1%, amounting to millions of differences in the 3 billion base pair human genome, and thus every individual has its own "reference" genome. Additionally the DNA replication process is imperfect, so even neighboring cells of the same type aren't identical. Thus, it is impossible to know whether a detected mutation in RNA is from the genome or RNA editing.

To understand single-cell differences between RNA and, I want to perform simultaneous RNA and DNA extraction of individual cells, assemble the DNA sequence as the individual cell's reference genome, then map the RNA reads onto the assembled reference. Using this technology, we can also interrogate the relationship between epigenetics and RNA by measuring DNA methylation, a chemical process inactivating RNA transcription in genomic regions.

Access to DOE CSGF computational resources would expedite single-cell mutation analysis by providing big data storage and high-performance computing. For storage, a sequencing run for a human cell is around 25GB, and performing RNA and DNA-sequencing for a control and test group of 2*96~200 cells creates 10TB of data for one of many experiments. As for computing, 20 samples' RNA sequencing data can be mapped onto DNA in one day, on a cluster of 1000 CPUs each with 2GB memory. Assuming a similar time for DNA-Seq assembly, it would take 20 days to map both DNA- and RNA-sequencing reads of 200 cells. Thus, a fellowship from DOE CSGF would allow me to immediately delve into the biology of genomic versus transcriptomic differences in single cells, an unanswered question.

Program of Study. The fellowship program of study requirement is designed to give you a breadth of competency in fields outside your own that will enhance your ability to perform computational science research. Please describe (in no more than 300 words) how you expect that the courses listed in your planned program of study outside your chosen discipline will contribute to your own research in the future. Describe why you chose these courses and how they will impact your research plans. (300 words)

The electives I chose for my Program of Study are Algorithm Design and Analysis, Quantitative Molecular Biology, Graduate Oncogenes, Numerical Analysis in Multi-Scale Biology, and Statistical Learning.

I chose Algorithm Design and Analysis to learn how to repurpose existing algorithms for new problems, avoid common pitfalls in algorithm design, and take advantage of data structures in algorithms. For our final project, my partner and I wrote about the parallization paradigm MapReduce, and I enjoyed learning a method of improving an existing algorithm by tweaking it to work on parallel processes.

Quantitative Molecular Biology will prepare me to manage the stochasticity in quantifying RNA and DNA molecules in single cells. There is stochasticity of the act measuring, the production and degradation of RNA, and coupled with oscilliatory cell-cycle specific gene expression, we need to model this stochasticity to compare cells in the algorithm development.

I chose Graduate Oncogenes to become intimately familiar with the transformation process of a normal tissue to a tumor. This course covers which genes are involved in cancer, including their roles, so I can understand how individual cells act in cohort to create cancer. (not offered in 2012-2013, taking it 2013-2014)

I chose Numerical Analysis in Multi-Scale Biology to learn the state-of-the-art in modeling biological systems from molecules, to cells, to tissues, to organisms. This course will focus on metabolic modeling, but I will use concepts from this class to model gene expression in single cell systems.

I chose Statistical Learning, a seminar class in which students present papers on the topic, to get experience presenting heavily technical statistical learning papers, understand what techniques are employed throughout the field as a whole, and think about the problems in bioinformatics that could benefit from tweaked versions of existing algorithms.

List of publications: Please include a list of publications authored or co-authored by the applicant. (3200 chars)

Botvinnik OB*, Kim W*, Abzeed M, Tamayo P, Mesirov JP. Exploring the Landscape of Genomic Abnormalities using Functional Profiles of Oncogenic Pathway Activation and Dependency. (in preparation) *These authors contributed equally to this work.

Birger C, Botvinnik OB, Tamayo P, Mesirov JP. Single-sample GSEA compares gene set enrichment of multi-sample experiments. (in preparation)

Botvinnik OB, Lopez-Diaz F, Lee W, Pourmand N. Single-cell analysis of taxol resistance in breast cancer. Poster at Intelligent Systems for Molecular Biology Conference. Long Beach, California, United States (2012)

Botvinnik OB. RNA-Sequencing Differential Expression Pipeline. Code: bitly.com/VzZ8wR ; blog post on usage: bit.ly/PMMiOc. Posted to Github on August 2012.

Goncearenco A, Grynberg P, Botvinnik OB, Macintyre G, Abeel T. Highlights from the 8th ISCB Student Council Symposium, Long Beach, California, USA, July 13 2012. BMC Bioinformatics. (2012)

Galili N, Tamayo P, Botvinnik OB, Mesirov JP, Brown G, Raza A. Prediction of response to therapy with ezatiostat in lower risk myodysplastic syndrome. J Hematol Oncol (2012) PMID: 22559819

Wood KC, Konieczkowski DJ, Johannessen CM, Boehm JS, Tamayo P, Botvinnik OB, Mesirov JP, Hahn WC, Root D.E, Garraway LA, Sabatini DM. Miniaturized functional screening reveals genetic modifiers of therapeutic response in melanoma. Sci Signaling (2012) PMID: 22589389

Botvinnik OB, Tamayo P, Mesirov JP. Discovery of novel candidate oncogenic activators with REVEALER. Poster at Intelligent Systems for Molecular Biology Conference. Vienna, Austria (2011)

Berger MF, Badis G, Gehrke AR, Talukder S, Philippakis AA, Peña-Castillo L, Alleyne TM, Mnaimneh S, Botvinnik OB, Chan ET, Khalid F, Zhang W, Newburger D, Jaeger SA, Morris QD, Bulyk ML, Hughes TR. Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. Cell (2008) PMID: 18585359

Laboratory and Research Experience/Other Employment. Begin with current or most recent employment. Please include employer, dates employment started and ended, position, and nature of work. (4000 chars). Examples:

Laboratory Experience example:

Los Alamos National Laboratory, Summer 2012, Lab assistant, Research

Naval Research Laboratory, 5/2011–8/2011, Research Assistant, Immunosensor Research

Student Employment example:

Michigan State University, Engineering Department, 12/2011 - 5/2012, Lab Technician, Research.
University of Pennsylvania, 5/2011 - 8/2011, Research Assistant, Developed stochastic models of neurons.
University of California - San Diego, Prof. Gene Yeo's lab, 01/2013-present, Rotation Student. Developing experimental methods for simultaneous RNA and DNA extraction from a single cell.

University of California - San Diego, Prof. Trey Ideker's lab, 09/2012-12/2012, Rotation Student. Compared transcriptional interaction networks of two distantly related yeast species, S. cerevisiae and S. pombe.

University of California - Santa Cruz, Prof. Nader Pourmand's lab, 01/2012-06/2012, Master's Student. Developed RNA-Sequencing differential expression pipeline for single cell analysis.

University of California - Santa Cruz, Prof. Joshua Stuart's lab, 09/2011-12/2011, Rotation Student. Found similarities between ovarian and subtypes of breast cancer, specifically basal and luminal B.

Broad Institute of Harvard and MIT, Dr. Jill Mesirov's lab, 11/2010 -- 8/2011, Associate Computational Biologist. Developed REVEALER, an algorithm that integrates genomic and functional data to infer new associations between phenotypic and genotypic data via a novel mutual information metric.

MIT Department of Brain and Cognitive Sciences, Prof. Sebastian Seung's lab,1/2010 -- 6/2010, Research Assistant. Analyzed neuron orientation in the rabbit retina inner plexiform layer (IPL) and found that neuron fragments in this 3D image lie along the axis of information transmission, and do not significantly traverse the direction perpendicular to of signal flow.

MIT Computer Science and Artificial Intelligence Laboratory, Prof. David Gifford's lab, 1/2009 -- 12/2009, Research Assistant. Compared metabolic networks in two strains of the same yeast species, and investigated in silico enrichment of T-cell receptor sequences.

Howard Hughes Medical Institute (HHMI) Janelia Farm (JF) Research Campus (RC), Dr. Sean Eddy's lab, 6/2008 -- 8/2008, Summer Scholar. Used Hidden Markov Models to create a robust null model for HMMER, sequence homology software developed in the Eddy lab.

Harvard Medical School, Prof. Martha Bulyk's lab, 6/2007 -- 9/2007, Research Assistant. Analyzed homeodomain transcription factor binding sites by performing protein-binding microarray experiments, resulting in a 2008 publication in the journal Cell.

Academic Awards and Honors. Include undergraduate and graduate honors (if applicable). (4000 chars)
Academic Awards and Honors example
Highest Honors, U of Michigan, 2009 - 2012
National Merit Scholar, 2009

Dean's List, U of Michigan, 2009, 2010, 2011, 2012
------ Graduate ------
First person to finish M.S. in 9 months at UCSC (Expected time: 2 years)
First 1st year graduate student to TA a graduate class in UCSC BME program
UCSC Regent's Fellowship (2011-2012)
NSF Graduate Research Fellowship Honorable Mention (top 20% of applicants, 2011-2012)
------ Undergraduate ------
Gordon-MIT Engineering Leadership Scholar (2008-2009)
One of two (out of a thousand) undergraduates to double major in Mathematics and Biological Engineering
Howard Hughes Medical Institute Janelia Farm Research Campus Summer Scholar (2008)
Cold Spring Harbor Undergraduate Research Program (2008 and 2009, declined for other opportunities)

Extracurricular Activities. Include technical societies and service organizations. (4000 chars)
Extracurricular Activities example:
Men's Fencing Captain, 2011
Volunteer for Habitat for Humanity
American Physical Society
------ Graduate ------
- Organized UCSC *Miss Representation* documentary screening (2011)
- Organized UCSC and UCSD fellowship critiquing circle (2011 and 2012)
- Organized departmental holiday gift exchange (2011)
- Volunteer tutor for 8th grade Algebra (2011-2012)
- Professional Development Program (inquiry-based teaching)
- Taught bioinformatics "Genes and Disease" module to high school biology class
- International Society for Computational Biology Student Council: Co-Chair for 8th annual Student Council Symposium at Intelligent Systems for Molecular Biology, the largest computational biology conference
------ Undergraduate ------
- Lightweight Men's Crew (Coxswain)
- Everett Moore Baker House Executive Committee (Social Chair)
- Kappa Alpha Theta Women's Fraternity (Zeta Mu chapter)
- Gordon-MIT Engineering Leadership Program
- DanceTroupe (Dancer, Choreographer, Publicity Chair)

Additional comments. If necessary, include any additional comments regarding your application. (2400 chars)
Asdf

Odds and Ends - Research using high performance computing
It is possible to assemble RNA and DNA reads simultaneously, but the RNA assembly algorithm uses only sequences of 25 nucleotides or more, missing out on micro-RNA of

20 nucleotides. Thus even after RNA assembly it is necessary to try to map the unassembled reads onto the assembled DNA.