How did you choose your field and what are your primary expectations of your future career? (331/300 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, we used bioinformatics to help create drugs to target specific disease-causing mutations, but also realized most conditions cannot be described by a single gene. I am electrified by this complexity and am thrilled to use bioinformatics to solve current and future biological problems.

My three goals for my future career are to (1) learn, (2) teach, and (3) collaborate.

(1) In my graduate work, I will continue to study biology by merging the fields of alternative splicing and single-cell analysis to understand how differences in individual cells' RNA regulation affects disease. After graduation, I want to become a professor at a top research institution, where I will be surrounded by incredibly intelligent people who will question, support, and challenge my research. I hope to be constantly humbled by my colleagues.

(2) Training the next generation of scientists is critical for the United States' future success. I taught bioinformatics modules to high school and undergraduate students, and a key component of my teaching style is mirroring authentic research. I want learners to relish the struggle through a problem because it leads to marvelous moments of insight.

(3) The quote, "None of us is as smart as all of us," is increasingly true in integrative fields such as bioinformatics. We cannot do bioinformatics as hermits; we need experimentalists, clinicians, and biotechnologists to drive forward our understanding of biology. As a professor, I will forge connections with experts in other disciplines to create new fields of research. With my colleagues, I want to create a collaborative research institute which partners with biotechnology companies to transform research advancements into realized pharmaceuticals as quickly as possible. This way, the discoveries aren't locked away in an ivory tower, but used to improve public health.

How do your proposed field of study and career constitute an application of the physical sciences or engineering? (270/300 words)

Biology, mathematics, and computer science are applied in bioinformatics to solve biological problems at a rate and accuracy that was unprecedented before technological advances such as high-throughput sequencing and parallel computing. Many of these problems were inaccessibly large before bioinformatics, such as measuring sequence similarity, exploring genome-wide transcription factor binding, or discovering associations between clusters of expressed genes. This has accelerated discoveries and technology development in biomedical research and provided a framework for high-throughput analysis.

At the Broad Institute of Harvard and MIT, I helped develop REVEALER, an algorithm which finds associations between continuous phenotypic output such as gene expression, or response to RNA-interference (RNAi), and discovers additional putative activators or repressors. In particular, we found an additional suppressor of RNAi against the oncogene, KRAS; this suppressor is located on an amplified piece of chromosome 8, 8q24. With REVEALER, we found this association in a matter of hours instead of months, compared to traditional Excel spreadsheet methods. Experiments have verified suppression of KRAS RNAi via presence of this amplicon.

I want to continue solving biological problems with bioinformatics, such as "How do we solve cancer?" While daunting, I believe that framing cancer as a complex engineering problem enables an integrated quantitative approach. Outside of cancer research, a "cure for cancer" sounds tempting, but we've come to understand is that each disease is unique in each individual. Beyond tumor-wide profiling, there are subpopulations of cell types within each tumor and we can tailor treatments to patients based on this tumor composition to prevent relapse. Each cancer is its own disease, requiring tools to understand the disease in the context of the individual. I want to create these tools.

What are the considerations involved in your choice of graduate school? (328/300 words)

I chose to come to the University of California, San Diego (UCSD) because while some schools are strong in pure biological, engineering, mathematics, or computational sciences alone, UCSD has experts in all of these fields actively collaborating on interdisciplinary projects.

In particular, I want to work on both single-cell and alternative splicing research, the crossroads of which are only available here at UCSD with Prof. Gene Yeo. Single-cell research is critical because 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 interactions within the mosaic of individual cells help the larger cancer survive, and Yeo has a prototype of a single-cell sorting device that I will use for my graduate work.

To study cancer survival, I need access to tools for studying alternative splicing, an exciting new field because almost every disease in humans has a known mutation in a non-coding region such as a splice site. Yeo's lab is at the forefront of alternative splicing research, a part of international efforts to catalogue all regulatory features of RNA, and is the best place for me to research alternative splicing in cancer at the single cell level.

Additionally, San Diego is home to many top-notch biological research facilities, such as the Sanford Institute for Regenerative Medicine, where Yeo's lab is, the J. Craig Venter Institute where Dr. Roger Lasken, the world's expert in single-cell analysis is, and the Moores Cancer Center, where I will collaborate with world-renowned cancer researchers such as Dr. Thomas Kipps to obtain patient samples. San Diego is also a major

biotechnology hub, which adds to the infectious excitement for biological innovation present in the air.

Provide a concise resume, in chronological order, with dates, recapitulating significant periods of technical and other creative activity since high school graduation.

------ Undergraduate ------
--- 2006 ---
- September -
Began study at the Massachusetts Institute of Technology (MIT), Cambridge, MA in Mathematics and Biological Engineering.

--- 2007 ---
- June to September –
Worked in Prof. Martha Bulyk's laboratory on transcription factor binding sites via protein binding microarrays, contributing to a publication in *Cell* journal.

--- 2008 ---
- January -
Participated in Undergraduate Professional Opportunities Program to hone networking and professional skills

- June -
Work from Prof. Bulyk's laboratory was published in *Cell* journal: "Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences." 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. Cell (2008) PMID: 18585359

- June to August -
Won a summer scholar position at Howard Hughes Medical Institute Janelia Farm Research Campus to work with Sean Eddy on using Hidden Markov Models to create a better null model for protein similarity search.

--- 2009 ---
- January to December -
Worked in Prof. David Gifford's laboratory on comparing synthetic lethality networks of yeast strains, and in silico modeling of T-cell receptor enrichment.

--- 2010 ---
- January to June -
Worked in Prof. Sebastian Seung's laboratory on neuron orientation in rabbit retina

- June -
Earned two S.B. degrees in Mathematics and Biological Engineering at MIT

MIT Course highlights, lower level:
1 semester each calculus, differential equations, linear algebra, computational science and engineering
1 semester each chemistry, organic chemistry, genetics, biochemistry
2 semesters physics with laboratory
1 semester computer programming

MIT Course highlights, upper level:
1 semester each probability, discrete math, number theory, computational molecular biology, physical mathematics, theoretical computer science
1 semester each thermodynamics; cell biology; biological engineering lab; biological instrumentation and measurement; biomechanics; fields, forces, and flows; biological engineering design; computational systems biology

------ Graduate Work ------
--- 2010 ---
- July -
Shadowed pediatrician Dr. Adwoa Kwashi at 37 Military Hospital in Ghana for one week

- November -
Began work in Dr. Jill Mesirov's laboratory at the Broad Institute of Harvard and MIT on REVEALER algorithm which discovers additional putative activators of oncogenic pathways

--- 2011 ---
- July -
Presented REVEALER at Intelligent Systems for Molecular Biology (ISMB), largest international computational biology conference, held in Vienna, Austria

- September -
Began Master's degree in Bioinformatics and Biomolecular Engineering at University of California, Santa Cruz (UCSC).

- September to December -
Worked in Prof. Josh Stuart's laboratory on comparing ovarian cancer to basal breast cancer

--- 2012 ---
- January -
Taught bioinformatics curriculum to Advanced Placement Biology high school students

- January to June -
Worked in Prof. Nader Pourmand's laboratory on a single-cell RNA-sequencing differential expression analysis pipeline

- April to June -
TA'd Computational and Systems Biology, a graduate level course

- May -
Work from Dr. Mesirov's group was published in the journal Science Signaling: 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
And in Journal of Clinical Oncology: 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 Clin Oncol (2012) PMID: 22559819

- June -
Taught stem cell curriculum using bioinformatics to minority undergraduates
Earned MS in Bioinformatics and Biomolecular Engineering at UCSC

- July -
Presented single-cell analysis pipeline developed under Prof. Pourmand, at ISMB in Long Beach, CA
Co-chaired ISMB Student Council Symposium

UCSC Course highlights
1 quarter each: bioinformatics models and algorithms, molecular biology, Bayesian statistics, computational genomics, biotechnology and drug development, applied gene technology, graph algorithms

- August -
Finished RNA-Sequencing Differential Expression package, posted code and documentation to Github and wrote a blog post on usage.

- September -
Began PhD degree in Bioinformatics and Systems Biology at UC-San Diego

- September to present -
Rotating in Prof. Trey Ideker's laboratory, comparing transcription factors in two yeast species: S. pombe and S. cerevisiae.

Academic Honors

------ Undergraduate ------
Gordon-MIT Engineering Leadership Scholar (2008-2009)
One of two (out of a thousand) undergraduates to double major in Mathematics and Biological Engineering

------ Graduate ------
First person to finish M.S. in 9 months at UCSC
First 1st year graduate student to TA a graduate class in UCSC BME program

Fellowships, Scholarships, etc (* is for nationally-competitive fellowships)

------ Undergraduate ------
Howard Hughes Medical Institute Janelia Farm Research Campus Summer Scholar (2008)
Cold Spring Harbor Undergraduate Research Program (2008 and 2009, declined for other opportunities)
------ Graduate ------
UCSC Regent's Fellowship (2011-2012)
*NSF Graduate Research Fellowship Honorable Mention (top 20% of applicants, 2011-2012)

Previous Research

1) "Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences." 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. Cell (2008) PMID: 18585359

2) "Creating a better null model for HMMER protein domain search." PI: Dr. Sean Eddy. Location: Howard Hughes Medical Institute Janelia Farm Research Campus. June-August 2008.

3) "Differences in synthetic lethality networks in Saccharomyces cerevisiae strains S288C and Sigma1287b." PI: Prof. David Gifford. Location: MIT Computer Science and Artificial Intelligence Laboratory. January - June 2009.

4) "Detecting immune response by T-cell receptor sequence enrichment." PI: Prof. David Gifford. Location: MIT Computer Science and Artificial Intelligence Laboratory. June - December 2009

5) "Determination of neuron orientation via computational image analysis." PI: Prof. Sebastian Seung. Location: MIT Department of Brain and Cognitive Sciences. January - June 2010

6) "Uncovering of new cancer activators via integrated genomics." PI: Dr. Jill Mesirov. Location: Broad Institute of Harvard and MIT. November 2010 - September 2011

I worked at the Broad Institute of Harvard and MIT in Cambridge, MA, with Prof. Jill Mesirov to develop REVEALER, an algorithm that integrates genomic and functional data to infer new associations [A]. 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 using a mutual information (MI) metric to discern between top candidates. This flexible MI metric was also used to determine associations between sensitivity to the melanoma drug PLX-4720 and NFkB gene set expression [B], analysis of response to the myelodysplastic syndrome drug ezatiostat via microRNA and gene set expression [C], and in single-sample gene set enrichment analysis [D]. The REVEALER algorithm found that resistance to the oncogene KRAS RNAi knockdown is related to an amplicon in 8q23-24 in addition to KRAS mutation [E], a finding that had taken another researcher months of looking through Excel files. 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.

Publications:

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

[B] 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

[C] 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 Clin Oncol (2012) PMID: 22559819

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

[E] 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

7) “Comparison of basal breast cancer to ovarian cancer.” PI: Prof. Joshua Stuart. Location: University of California, Santa Cruz. September - December 2011.

8) "Single-cell analysis of paclitaxel resistance in breast cancer." PI: Prof. Nader Pourmand. Location: University of California, Santa Cruz. January - August 2012.

There are some cancers that no matter what you throw at them, they still relapse. In Prof. Nader Pourmand's laboratory at Univ. Calif.-Santa Cruz (UCSC), we were interested in breast cancer drug resistance at the single-cell level, to observe how individual cells escape chemotherapy, specifically paclitaxel which inhibits microtubule elongation, preventing

proper mitotic spindle formation and subsequently cells from dividing [F]. Our collaborator at the Salk Institute extracted individual cells from untreated (6 cells), treated with paclitaxel (6 cells), and survivor (5 cells) MDA-MB-231 triple-negative breast cancer cell line populations. "Triple-negative" breast cancers do not have the mutations in the three major receptors responsible for breast cancer, and cannot be treated with antibody therapy, so patients must be treated with systemic chemotherapies, and resistance to chemotherapies is acutely important in these cases. Our lab performed whole-cell RNA-Seq on these 17 samples using the NuGen Ovation amplification kit and Illumina Hi-Seq sequencer. One of the main technical challenges was consistently analyzing the aligned sequencing reads and producing lists of differentially expressed genes. If two bioinformaticists write their own scripts to perform these tasks, there will be two different results. To increase consistency between sequencing experiment interpretations, I developed an open-source RNA-Seq differential expression (RSDE) pipeline [G] which takes in alignment files and annotation, then outputs genome-wide coverage (visualized via Circos), heatmaps and lists of differentially expressed genes, and RNA-Seq quality control information. Using RSDE to study taxol resistance in breast cancer at the single-cell level, we found the surviving cells had differentially regulated actin cytoskeleton genes, which may indicate that these survivors have some mechanisms to compensate for improper chromosome alignment to the mitotic plate.

Publications:

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

[G] “RNA-Sequencing Differential Expression Pipeline.” Code: bitly.com/VzZ8wR ; blog post on usage: bit.ly/PMMiOc . Botvinnik OB. Posted on Github in August 2012.

9) “Cross-species analysis of transcription factor binding in Schizosaccharomyces pombe and Saccharomyces cerevisiae.” PI: Prof. Trey Ideker. Location: University of California, San Diego. September 2012 - present.

Additional essay: Include here information about your favored extracurricular and leisure time activities since your graduation from high school.

When I tell people that I study bioinformatics, their usual reaction is, "Bio-what?" Despite its rapid growth, bioinformatics is still a new field, and one rarely understood. I first heard of bioinformatics via the Human Genome Project in my eighth grade genetics class, and was captivated by the possibility of completely understanding hereditary disease and personalizing medicine. I want to spread the joy of bioinformatics, and I want high school students to become excited about bioinformatics just as I was.

One of my first big successes towards developing secondary school bioinformatics education was at the University of California-Santa Cruz (UCSC), where I was part of a team of students and a professor that pioneered the development of bioinformatics teaching modules to use in AP Biology (AP Bio) classes. My topic was "Genes and

Disease," in which students chose a gene related to a familiar disease or condition and explored its properties - protein size, biological function, chromosomal location, and so on. One creative group explored "maple syrup urine syndrome," and found that the gene associated with this condition was associated with sugar metabolism. This bridge from genetic defect to disease causation was an anomaly, as most of the students found the links between disease and genes to be very tangled. There are many genes associated with complex diseases such as cancer and diabetes, and I was happy to hear the students realize how difficult it is to pinpoint an individual cause of a condition. Pleased with the success of the pilot project, we released the “Bioinformatics in AP Bio” curriculum on the web for any teacher to use, and it has since been viewed by hundreds people from around the world. I am continuing to refine this unit and develop others, working with biology teachers at underserved high schools in San Diego. It’s amazing how much of bioinformatics can be done from any computer on the web, even at just a library, and I hope that in teaching underserved students these skills, I can help empower them to pursue science.

This summer, I built on my early experiences and developed a curriculum based around the fundamentals of stem cell biology and taught it to underprivileged and minority undergraduates at UCSC embarking on a summer research project. I worked with the university’s Professional Development Program, which emphasizes teaching in a diverse environment. My favorite aspect of the course, though it was also the most difficult, was planning the "inquiry" part of the curriculum, where learners developed their own unique method of investigation. In many ways, the method of inquiry models authentic research techniques, and thus is both fun and useful for the students. Our process goal was for learners to realize that cell identity is a continuous spectrum and to become more comfortable with conflicting information–an important reality of research. We structured our curriculum around how neural stem cells differentiate via an intermediate to pre-neurons. We first showed students an analogous stem cell system–embryonic stem cells differentiating into cardiomyocytes without an intermediate. We wanted the students to come up with the idea of an intermediate on their own, and it was hard to give enough hints without giving the answer away. Each of the three facilitators taught one of the methods of exploring the particular biological technique: microscopy, biochemistry, and bioinformatics (me). For me, it was difficult to create a module that was challenging enough that the students would be engaged the whole time, but not watered down. I debated whether to clean up the data or not, but decided that would diminish the experience of mirroring authentic research, as research data is not clear-cut whatsoever. In the end, leaving the data in its realistically messy form was the right decision. Given the complexity of the biological system, the amount of data was overwhelming and my group missed some of the details of neurological differentiation, but they certainly attained the process goal of realizing that defining cell identity is non-trivial, and information can be quite conflicting.

In addition to teaching, I also enjoy giving back to the research community through organizing conferences. This summer I co-chaired the Student Council Symposium (SCS) at Intelligent Systems for Molecular Biology (ISMB), the largest international computational biology conference, held in Long Beach, CA. We received over 100 abstract submissions and accepted 50, with the top 10 as oral presentations and

the remaining as posters. All told, we had 60 attendees from all six habitable continents, and a day full of three fantastic keynote speakers and ten excellent student talks. Planning for this event began in the previous September, and it was thrilling to watch the event unfold, meet authors and watch student talks, after almost a year of planning. I especially enjoyed meeting the recipients of the travel fellowships, who flew in from Europe, South America, and Asia and were very grateful to experience ISMB. The SCS is a fantastic venue for students to give a research talk to a broad audience outside of their lab and home countries, and some for the first time in English. I was honored to be able to facilitate such a life-changing event for so many of my peers.

On a more personal level, I like to explore how successful people spend their time through my blog, and my most popular post (http://bit.ly/PMMkWz) was about a role model of mine, Prof. Eric Lander, a mathematician and geneticist who pioneered the Human Genome Project and was pivotal in creating the interdisciplinary research facility, the Broad Institute. In a news article about him, and what I was most struck by was not the boasts of his genius, but his quiet determination. In a video interview, he talked about being "stubborn through the struggle" of a problem, that it takes persistence to create good science. His words continue to inspire me through my scientific struggles.

Outside of science, 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 delight in playing the cello, which I played seriously from fourth grade to high school. 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.

By engaging in teaching, mentoring, and the scientific community, I have realized the importance of each unique contribution. This has strengthened my research as it has pushed me to pursue even more interdisciplinary science and will help me in my future career. I intend to become a professor and together with my colleagues, build a research center that integrates genomics, proteomics, metabolomics, and any incredible future inventions to solve biological problems.

*I know they’re looking for “dedication” here and I’m afraid that the activities I **can** mention aren’t consistent enough to count.*

Dance? but I only did it in college and wasn’t that good, ie didn’t join any competitive teams. Though I did choreograph several dances.

Cello? But I only just re-started playing (since late September). Played 8 years through HS, stopped in college, now I’m picking it up again because I missed playing.

Blogging? I only started in January

Teaching? Can talk about Bioinformatics in AP Biology

Sports/Exercise could be another thing.

Cooking.

Scientific programming

Organizing the NSF review

Supplemental information: Publications, Research proposal as a single document (<2mb)