Question 1: Choice of Field and Future Expectations

How did you choose your field and what are your primary expectations of your future career? (300 words)

Coming of age as the Human Genome Project reached its fruition, I was obsessed with genetics and learned early on that the genome is a powerful tool with profound application to understanding disease. I was frustrated by undesirable effects of drugs, and began mentally constructing an ideal future where a simple blood test could inform the physician of nearly every aspect of a patient's health status, dictating treatment and future steps. The advent of high-throughput genomics, proteomics, and other -omics technologies brings this vision of personalized medicine closer to reality.

Innovation is best paired with a solid foundation, and I began my journey at MIT by obtaining dual degrees in mathematics and biological engineering. After graduating, I trained as a bioinformatician 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. The Hertz Fellowship can deepen my doctoral studies by allowing me the freedom to forge novel collaborations between computation and biomedicine, as I have witnessed groundbreaking research occur at the intersection of disciplines.

To maximize my impact on revolutionizing medicine through personalized 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 conferences. I cannot wait to co-chair the Student Council Symposium at the International Society for Computational Biology's annual conference and to continue the circulation of innovation in bioinformatics. Moreover, I will strive to apply my discoveries to create something new, whether a new device or a new paradigm; as true power lies in application. 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.

(303/300 words)

Brainstorming choice of field/proposed field of study questions

1. Freedom to explore science in the romantic sense, how newton and Leibniz had worked at their craft because they found it to be interesting
    a. Unabashed/unrestricted ambition
    b. Explore my scientific passions utilizing the resources provided by my institution and transforming them into something great
    c. Romantic scholar of applied sciences
    d. I will be able to propose an idea and instead of a being brushed off as crazy by a potential research mentor, I will have the freedom and the clout behind the Hertz foundation to begin a conversation about **how** this idea could be accomplished rather than being ignored
    e. Forge my own path
        i. Collaborate with both a computational and clinical advisor based on their research interests and not their funding
    f. Joining a community of passionate and brilliant colleagues
        i. Discuss our individual fields, learn about condensed matter physics and distributed computing
2. Career
    a. Currently in a Master's program but aim to finish the normally 2-year program in one year.
        i. Applying to PhD programs now
    b. Want to be Director of national institute
    c. Revolutionize medicine through personal genomics
    d. Genomics is not enough – need epigenomics, proteomics, and a way to integrate it all. I am an integrator
    e. My vision:
        i. A simple blood test will inform the physician of nearly every aspect of the patient's health status, and provide meaningful information for treatment and future steps
        ii. a world where patients get exactly the drug that will help them and not be excessively exposed to other, imperfect drugs
            1. my vision is most pertinent to cancer
        iii. Eg cancer treatment – try a variety of different drugs
        iv. Right now we focus our sequencing efforts on the tumor itself
            1. But recent evidence has shown that tumors release small cells that circulate in the blood stream (circulating tumor cells, or CTCs)
            2. I will harness this phenomenon and draw blood from a patient to sequence their circulating tumor cells, link this to the tumor integrated systems biology, then we could suggest drugs for a patient based exclusively on quantitative evidence
                a. Need a temporal model to understand tumor evolution and response to drugs
            3. I see this mechanism becoming a company that I would lead
                a. Currently, the highest demand for such specific treatment is in cancer, and as a result, that is where the data is
                b. But I am disease agnostic and aim to create a disease-free world
        v. It's one thing to discover something and add that knowledge to the world, but the real power comes in applying this discovery to create something new, whether a new device or a new paradigm.

Question 2: Proposed Field of Study

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

In my future career, I aim to merge the disciplines of mathematics, computer science, and biology to cure human disease. Production of biological data is no longer a bottleneck; rather, computational interpretation of these data is one of the great challenges in biology. I will apply informatics to life sciences by pioneering a future of personalized medicine harnessing advancements in –omics technologies. This vision is currently most pertinent to cancer, where national efforts such as The Cancer Genome Atlas (TCGA) are developing vast data repositories for genome analysis.

The clinical manifestation of cancer is so complex and multimodal, caused by interaction of a variety of inherited and environmental factors and potentially leading to widespread metastasis, that traditional treatment methods of surgery, chemotherapy, and drug therapy are insufficient. However, the combination of cancer biology and informatics provides fresh insight into important problems such as development of drug resistance.

My goal is to revolutionize cancer treatment by developing a dynamic model of within-tumor interactions to understand drug sensitivity and cancer evolution through integrated genomics and epigenomics. I am especially interested in sequencing single-cell tumor cells, a promising approach as most cancers are incapable of growing in a monoclonal population on a petri dish. My model will capture the temporal response to therapeutics by encoding the interactions between individual cells and predict the magnitude of the population's response, how quickly the surviving resistant population may relapse, and what combination therapies to pursue to avoid relapse. Currently, the highest demand for tissue-specific modeling is in cancer, but I am interested in modeling other disease tissues through collaborations with leading research hospitals such as Massachusetts General Hospital.

As a professor, I will embody the bimodality of bioinformatics by obtaining appointments in both medicine and applied mathematics, and pursuing collaborations with physicians and biologists. My quantitative skills will filter noisy biological data into actionable information to promote health and combat disease.

(312/300 words)

Question 3: Choice of Graduate School

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

I considered pursuing an MD/PhD dual degree, but realized I would contribute more to society if I focused my efforts on research. Thus, instead of pursuing a dual degree, I want to do my PhD in a medical environment.

Currently I am in a Master's of Bioinformatics program at University of California – Santa Cruz (UCSC), where I intend to finish the two-year program in one year. This is a feasible goal as I have experience in accelerated programs - I received two bachelor's degrees from MIT in four years. Since UCSC does not have a medical school, I am currently applying to PhD programs. It is important for me to pursue my doctoral studies where there is a large effort towards clinical integrated genomics because I know these labs will have the first opportunity to access new data collected by novel methods, and therefore the first opportunity to develop algorithms that can save lives.

For example, attending Harvard Medical School's program in Bioinformatics and Integrative Genomics (BIG) would offer me the opportunity to capitalize on existing collaborations, and forge novel ones with clinicians from the fertile hospital community. Specifically, I am interested in working with the Children's Hospital Informatics Program for their innovations in health care informatics, such as the Informatics for Integrating Biology and the Bedside (i2b2) program. The i2b2 network allows researchers and clinicians from around the world to collaborate on patient data, creating a platform for distributing innovations in clinical bioinformatics. The true mark of applicability is in widespread usage, and I want to be in an environment that has the capability to implement my tumor-analysis algorithms in a clinical setting, and where I can surround myself with people with this collaborative mindset.

Some researchers in Harvard BIG that I am interested in working with include Professors Bonnie Berger, Isaac Kohane, and Eric Lander for their pioneering efforts in personalized medicine.

(296/300 words)

[Children's also has a program I participated in called Sustainable Medical Apps, Reusable Technologies (SMART) apps for Health. I made a breast cancer risk predictor webapp based on the NIH's breast cancer risk predictor (translated C++ to python, learned HTML/javascript in a week and wrapped the code in a webapp), but it was incomplete when I submitted it. The app is no longer featured on the SMART app website, and I want to talk about this experience in my essay somehow but I'm not sure how to integrate it.]

[Another side note: My top choice of school is a four-way tie Harvard Med Bioinformatics and Integrative Genomics (BIG), Stanford Biomedical Informatics, MIT Computational Systems Biology, and Harvard Engineering and Applied Sciences. But I know the most about Harvard Med BIG so I wrote about that.]

Academic honors

2008-2009 Gordon-MIT Engineering Leadership Program

Question 4: Chronological Resume

Provide a concise resume, in chronological order, with dates, recapitulating significant periods of technical and other creative activity since high school graduation. Omit activities only distantly related to your professional development. Include workshops, summer schools, a general description of all courses of study pursued (e.g. "3 quarters of Differential Equations") and degrees expected or awarded (dates, institutions, fields). Separate your undergraduate activities from your graduate activities (if/as applicable) with a single dashed line. (300 words)

Objective: Apply bioinformatics and integrative genomics to predict clinical outcome and infer the most effective disease treatment.

---------- Graduate ----------

University of California - Santa Cruz, Santa Cruz, CA: Sept. 2011 – June 2012 (expected)
M.S. in Biomedical Engineering and Bioinformatics

---------- Post-Baccalaureate ----------

Broad Institute of Harvard and MIT 2010-2011
Prof. Jill P. Mesirov Lab, Bioinformatics Research Technician

- Collaborated with Prof. Todd R. Golub's lab
- Developed single-sample Gene Set Enrichment Analysis (ssGSEA), a fast GSEA for multiple comparison
- Created REVEALER algorithm to unveil new candidate oncogenic activators using a mutual information metric to compare genomic aberrations to cancerous pathway expression
- REVEALER publication in preparation

---------- Undergraduate ----------

----- Education -----

Massachussetts Institute of Technology, Cambridge, MA: Sept. 2006 – June 2010
S.B. in Mathematics
S.B. in Biological Engineering

- GPA 4.2/5.0
- Mathematics Coursework: Probability and Random Variables; Linear Algebra; Computational Science and Engineering (MATLAB); Computational Structural Biology; Principles of Applied Mathematics; Theory of Numbers; Differential Equations; Seminar in Physical Mathematics; Seminar in Theoretical Computer Science
- Biological Engineering Coursework: Genetics; Graduate-level Computational Systems Biology; Biochemistry; Organic Chemistry; Thermodynamics of Biological Systems;

Analysis of Biomolecular and Cellular Systems; Fields, Forces, and Flows in Biological Systems; Introduction to Computer Science and Programming; Principles of Chemical Science, Laboratory Methods in Biological Engineering; Instrumentation and Measurement for Biological Systems; Molecular, Cellular, and Tissue Biomechanics; Biological Engineering Design
- One of two people to graduate with both degrees
- Led a team to win Biological Engineering Design competition for a theoretical cure for atherosclerosis, an inflammatory disease
- [Lightweight Men's Crew (Coxswain), Baker House Executive Committee (Social Chair), Department of Mathematics Tutor, Kappa Alpha Theta Women's Fraternity, DanceTroupe (Dancer, Choreographer, Publicity Chair), Gordon Engineering Leadership Program]

----- Employment -----
MIT Department of Brain and Cognitive Sciences, Cambridge, MA: Jan. -June 2010
Prof. Sebastian Seung Lab, Undergraduate Researcher
- Analyzed neuron segmentations of electron microscopy images of rabbit retina inner plexiform layer slices and elucidated patters in interneuron size, shape, and orientation

MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA: Jan. – Dec. 2009
Prof. David Gifford Lab (Computational Genomics Group), Undergraduate Researcher
- Tested whether measures of information flow can predict gene lethality in different genomic networks
- Built a method to cluster, parse, and interpret T-Cell Receptor sequence data

Howard Hughes Medical Institute Janelia Farm Research Campus, Ashburn, VA: Summer 2008
Dr. Sean Eddy Lab, Undergraduate Researcher
- Improved protein sequence homology search by creating a better null model
- Used Hidden Markov Model (HMM) methods, written in Python, then in C (self-taught)

Harvard Medical School Brigham and Women's Hospital, Division of Genetics: Summer 2007
Prof. Martha Bulyk Lab, Undergraduate Researcher
- Analyzed DNA binding specificities of tissue-specific homeodomain mouse transcription factors using Protein Binding Microarrays
- Resulted in a publication in Cell journal

---------- General ----------
Membership
- Member of Gordon-MIT Engineering Leadership Program
- Student Member of International Society of Computational Biology (co-chairing Intelligent Systems for Molecular Biology conference Student Symposium 2012)
- Student Member of Society of Women Engineers
- Student Member of Association of Women in Science.

Skills
- Programming: Java, C, Python, Perl, R, MATLAB, SVN, LaTeX

- Algorithms: HMMs, Gene Set Enrichment Analysis (GSEA), machine-learning, network analysis, information theory
- Graphic Design: Adobe Illustrator CS 5, OmniGraffle

(539/300 words)

In the space provided below, list, in chronological order, academic honors and distinctions which you have received and the time or time-interval of receipt. Separate your undergraduate from your graduate awards (if/as applicable) with a single dashed line. (Include title, reason for award, and where/when received.) Use no more than one line per award whenever possible (what, where/when received).

2011-2012 University of California - Santa Cruz, Regents' Fellowship

---

2009 Cold Spring Harbor Laboratory Undergraduate Research Program (declined)2008 Howard Hughes Medical Institute Janelia Farm Research Campus Summer Scholar
2008 Cold Spring Harbor Laboratory Undergraduate Research Program (declined)
2009 MIT Math Department Tutor

Please list the most significant research projects that you have pursued, in chronological order. (Include reference information for those that have been formally documented, presented at a conference, or submitted for publication.)

Examples:
1) "Controlled Fusion in the Shadow of the French Alps: My Summer Internship at Grenoble"; June-August 2002.
2) "Peculiarities in Gene Transcription Regulation in wingless Drosophila Mutants", prepared/presented as part of the Student Research Seminar Series; U of Calif. Report, UCRL Report 1029-04, Sept 01-May 02.
3) "Deficiencies in Stereospecific Iodination of Thyroxine Precursors Isolated From Chernobyl-Region Voles", J. Exotic Endocrine Chem. 239, 3365 (June 2003), K. Early, M. Y. Name, L. Late, and M. Middle.
Finally, choose one or two projects that best exemplify your own creativity and discuss in more detail what you personally contributed to them.

1. "Variation in Homeodomain DNA Binding Revealed by High-Resolution Analysis of Sequence Preferences." Cell. June 2008, M.F. Berger, G. Badis, A.R. Gehrke, S. Talukder, A.A. Philippakis, L. Pena-Castillo, T.M. Alleyne, S. Mniamneh, O.B. Botvinnik, E.T. Chan, F. Khalid, W. Zhang, D. Newburger, S.A. Jaeger, Q.D Morris, M.L. Bulyk, T.R. Hughes. (from work performed June-August 2007)
2. "Building a more informed null model for HMMER"; presented as part of Howard Hughes Medical Institute Janelia Farm Research Campus Summer Scholars program, June-August 2008.
3. "Information flow through the yeast metabolome"; January-June 2009.
4. "Enrichment of T-Cell receptor sequences"; June-December 2009.
5. "Orientation of rabbit retina inner plexiform layer neurons", January-June 2010.
6. "Discovery of alternative cancer pathway activators with REVEALER"; presented at Intelligent Systems for Molecular Biology 2011 conference, 1 paper in review, 3 papers in preparation (1 as first author), November 2010 – August 2011.
7. "Analysis of pan-cancer biomarkers"; September 2011-present.

Each of my research experiences was a step towards understanding how numbers, letters, and pixel values describing biological data can be used to uncover the intricacies of biology and improve healthcare. Starting with bench molecular biology at Martha Bulyk's lab, where I was first exposed to how a computer could interpret the minutely polka-dotted microarrays and turn them into a lettered rainbow of a DNA binding motif. The discovery of computational biology led to analyzing biological sequence data in Sean Eddy's lab, and I continued my interest in informatics by working on computational systems biology in David Gifford's lab. I became interested in image analysis and started working on computational neuroscience in Sebastian Seung's lab. Realizing my true passion lay in computational genomics, I worked on cancer genomics at the Broad Institute.

I was admitted to the competitive Janelia Farm Summer Scholars program to work with Sean Eddy at Howard Hughes Medical Institute (HHMI) Janelia Farm Research Campus in Ashburn, Virginia. In Dr. Eddy's lab, I was fascinated by the power of mathematical models of biology, but at the same time I came to appreciate their weaknesses. We discussed the popular sequence alignment algorithm BLAST at length, analyzing its pairwise method that ignored highly repetitive patterns. For example, a protein sequence that has dozens of lycine residues in a row may be aligned to a sequence that is merely highly enriched for lycine but does not carry the striking continuation of repeats. But HMMER, software developed by the Eddy lab, used the machine learning technique of Hidden Markov Models (HMMs) to 1) infer patterns within an amino acid or DNA, 2) use these patterns to align sequences, and 3) infer evolutionary relatedness, or homology. However, while HMMER is a more robust method of analyzing protein and nucleotide sequences given a pattern such as groups of hydrophobic residues or secondary structure, it is imperfect in assigning residues to a position-specific active site or in predicting exact lengths of domains and sequences. 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. If the sequences are closer to the generic protein than they are to each other, then they are dissimilar. As a summer researcher, I created 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 major impact in therapeutic design in medicine.

Most recently, I worked with Prof Jill Mesirov at the Broad Institute of MIT and Harvard, where I developed an algorithm called REVEALER, which can help discover new candidate activators of cancer, aiding in the understanding of the bigger picture of cancer genesis as well as identifying new drug targets. REVEALER accomplishes this by analyzing a panel of samples, removing samples that already have a known activator, and searching for activators of the remaining samples using a mutual information metric. This metric acutely extracts relevance from noisy, entropic data. For example, if the entropy in dataset *t,* H(*t*), matches up with the entropy in dataset *s*, then there is some non-entropic ordering of these data, and that is the mutual information of *t* and *s.* Besides the strength of metric, one of the limitations of REVEALER is it only considers one activator (or repressor) at a time, so joint effects of mutations or other genomic events in combination are ignored.

The algorithm is highly extendable and is also capable of finding deactivators or features driving drug sensitivity. It can be used to advance basic biological research in addition to clinical findings as researchers can search through a panel of cell lines or yeast strains to uncover new genomic programs based on expression, survival, or protein abundance data. REVEALER discovers putative causal relationships, saves researchers time in performing otherwise tedious analyses, and produces consistent, reproducible results. I led the collaborative effort of REVEALER, working with computational biologists to analyze the algorithm's effectiveness, wet-lab researchers and physicians to determine biological and clinical relevance, and software developers to flesh out examples and ensure usability.

I presented my work on REVEALER in Vienna at the 2011 Intelligent Systems for Molecular Biology, the largest annual computational biology conference. I was impressed by the truly international nature of the conference; people from Spain, Brazil, Nigeria, India, and North America gathered to share their love of computational biology. There, I participated in both the main conference and the Student Council Symposium, where I was struck by my peers' talks about finding the relevant mutations in autism, refining algorithms for protein folding, and developing web-based environments to share computational methods and tools with bench biologists. Inspired by the welcoming and scientifically rigorous environment, I volunteered to co-Chair the next symposium and I am looking forward to working with the students from around the world to create another intellectually stimulating event.

Personal Essay: Include here information about your favored extracurricular and leisure time activities since your graduation from high school. (word limit unspecified)

As a child, I heard my parents and their friends complain about immigrant children losing their native language, and I was determined to not be one of them. I attended Russian parties with my family and sat at the "adults" table to practice Russian. In undergrad at MIT, I went to Russian-speaking events and took Russian language classes at Harvard. After college, I sought to ensure that new immigrants had 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 and having knowledge of the patient's medical condition that could affect a procedure, 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 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?" I have experience similar prejudice from Russian men in the US. They are so surprised to hear that a young woman such as myself is attended MIT and is not just looking for a rich husband.

Women's inequality in the US is subtler than in Russia. Instead of overt sexism, there are remaining 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. It was easily the most rewarding few hours of my week. But when I judged a science fair at different middle school, I was disappointed. This school didn't have SCFG and I saw that beyond race, the greatest divide between the fantastic and the mediocre science projects was gender. The girls weren't going for the hard, high-risk projects. Instead they did the tried and true: making crystals out of Borax at different temperatures but not recognizing the difference between crystalizing in liquid or in air, investigating invisible ink, and growing plants. 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 difference in scientific ability between the genders was so great. I realized that while I could teach 2nd graders about science and get them excited about my own subjects, there was a psychological shift that we as a nation need to make happen.

Women in the media are not portrayed as strong, capable, and smart. Instead they are sexy, manipulative, and vapid. I am involved in "Nerd Girls," a national movement to dispel myths about women in science and engineering. Women and girls today are pressured into looks and not smarts as a result of a lack of media coverage of smart women. I know that I was pressured by this media. Through Nerd Girls, I am creating stories for young women to look up to and are inspired by, so they know it's OK to be smart. Additionally, I am organizing a screening on UCSC campus on November 1st 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 also want to use this film as a forum for students of color, as well. While I am Caucasian, I recognize that Black, Latino/Latina, East Asian, South Asian, and Middle Eastern people are also highly underrepresented in the media's depictions of high-ranking officials in both the public and private sector, and this must change. This discussion should occur in groups of all genders and races for the first step to eliminate discrimination is to realize that it occurs.

My passion for encouraging strong and confident personalities is not limited to speaking and teaching, but is extended through dance and music as I enjoy the influence of many methods of expression. While I had never danced before college, I quickly picked it up and became a leader in MIT DanceTroupe as a choreographer and publicity chair. As a choreographer, my specialties were in hip-hop and "tutting," sharp, angular finger, hand, and arm movements named after King Tutankhamen for their resemblance to Egyptian silhouette paintings. More than inventing creative choreography, I enjoyed teaching dance, especially to students new to dance as I once was. I loved watching their progress over the course of the semester and was especially proud when I watched formerly shy students perform their hearts out onstage. As publicity chair, I rebranded DanceTroupe by designing a new, consistent logo that is still in use today, three years later. Previous logos had only lasted a year at most, and I wanted to ensure that a single emblem conveyed the diversity in dance styles that DanceTroupe had to offer: hip hop, contemporary, street jazz, tap, and ballet. 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 picked it back up right after college, when I was able to rediscover my passion for the instrument and even put on a small recital amongst my close friends.

I found a synergistic relationship between my research, my social and mentoring activities, and my passion for dance, as each fuels me to continue to strive harder and delver deeper to bring the best out of myself and others. While many people might see my activities as separate, I believe engaging both sides of the brain keeps me healthy and positive. My fluid transitions in activities have enabled me to look at my research and activities in new ways, pushing the cutting edge of innovation, becoming a role model in science, and positioning me to teach the next generation of engineers.

Brainstorming Personal Essay ideas

1. Dance
    a. Never danced before college
    b. Became a leader in MIT dancetroupe as a choreographer and publicity chair
    c. Hip hop, 'tutting', and contemporary dance
2. Women
    a. Nerd Girls
    b. WiSE at UCSC
    c. Women's group at Broad
        i. Led a discussion centered around a TED talk by Sheryl Sandberg, the Chief Operating Officer of Facebook
    d. Science Club for Girls
3. Russian
    a. Interpreter classes
4. Immigrant
    a. Thankful for my family's move to the US from Soviet Russia
5. Ghana?
    a. Shadowed doctor
        i. But really just went to see Ghana with Kwasi not a service trip…
    b. Recognized need for bioinformatics in developing countries
        i. Hertz fellowship would give me the freedom to pursue development of cheap devices for bioinformatics
6. Co-chairing ISMB Student Council Symposium 2012
    a. Aim to improve feedback to presenters
    b. Advocate for female students in ISMB (what does this even mean?)
    c. Want to increase publicity and interest in student symposium from top tier US universities
7. Role models
    a. Eric Lander (Human genome project)
    b. Robert Langer (innovator in biomedicine, member of National Academy of Science, Engineering and Medicine!)
    c. Bonnie Berger (first female faculty hired by MIT Math department)
    d. Jill Mesirov
    e. Ben Zeskind (MIT PhD in 3 years, founded Immuneering, a clinical bioinformatics consulting firm to pharmaceutical companies)
8. Baker House
    a. Social
        i. Doubled attendance of social events by improving publicity and innovating on traditional entertainment (hired a band)

Discovery of novel candidate oncogenic pathway activators with **REVEALER**

Olga B. Botvinnik*, Pablo Tamayo, Jill P. Mesirov

* presenting author: botvinn@broadinstitute.org

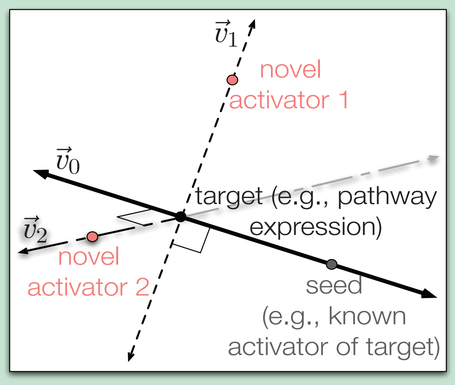
BROAD INSTITUTE 7 Cambridge Center Cambridge, MA 02139 USA

1. Why REVEALER?

- Mutations in oncogenes commonly activate downstream signaling pathways
- However, oncogenic pathway expression can be high in the absence of oncogenic mutation
- REVEALER can discover novel pathway activators

2. What REVEALER does

Input: gene expression, mutation, and copy number data, and a specific pathway's gene set and activating features.



Discovery: Right, space of all genomic features and pathway *target*. ***REVEALER finds orthogonal candidate pathway activators*** to $\vec{v}_0$, the line connecting the pathway and initial features *seed*. It does this by excluding samples which already have an activating feature and uses a mutual information metric to "fish out" the best feature that explains activation in the remaining samples. This step is repeated until an activating feature is found for all samples, or feature quality is below a cutoff.

Output: Plot of original and new features that explain activation.

Note: This is a general method that has many applications. The target can be RNAi data, drug response, or expression as shown here. Also, features can match inactivation.

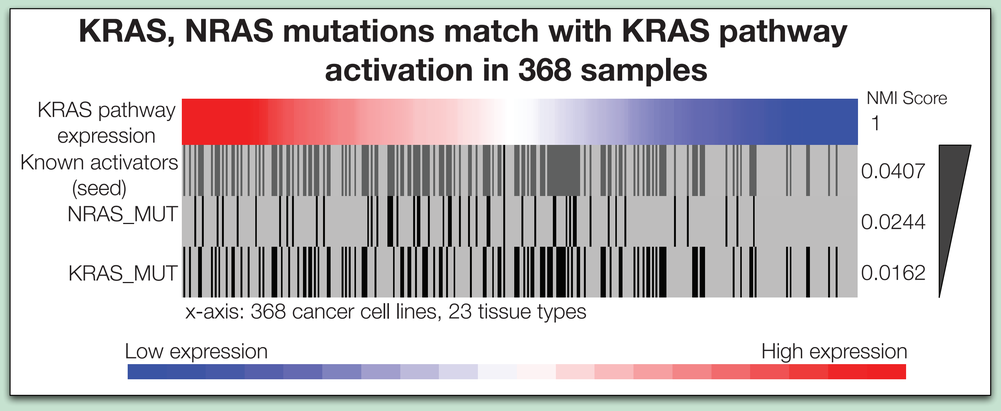
3. REVEALER relies on a sensitive mutual information metric for feature selection

Mutual information (MI) measures the gain in information from two data sets, each having entropy H, where MI is the union of the entropies. MI is more sensitive than correlation measures such as Pearson.

The normalized MI metric (NMI, right) is the $MI(X,Y)$ of two data sets normalized by their joint entropy $H(X,Y)$, which takes data sparseness into account.

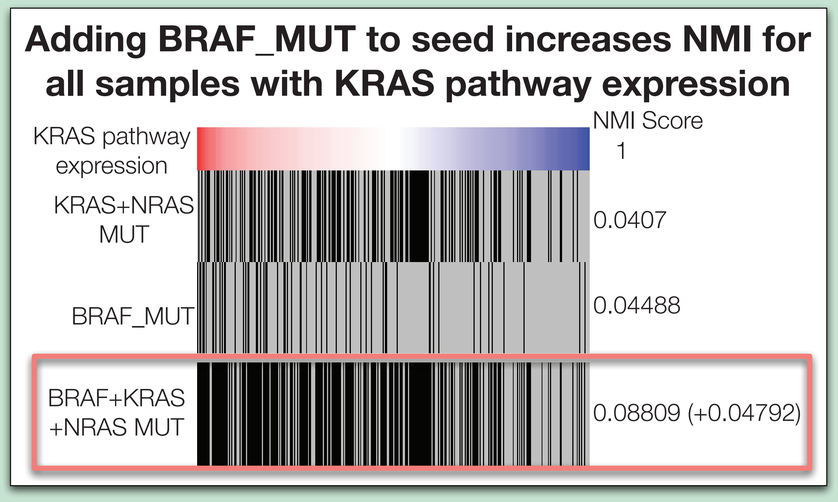
$$NMI(X,Y) = \frac{MI(X,Y)}{H(X,Y)} = \frac{H(X)+H(Y)}{H(X,Y)} - 1$$

4. REVEALER iteratively discovers oncogenic activator candidates



Above, KRAS pathway expression and RAS family mutation data* from 368 cancer cell lines of 23 tissue types is shown. Together, NRAS and KRAS mutations create a *seed* of known activators. An NMI score of each activator and the pathway is calculated to indicate the relative contribution of each mutation.

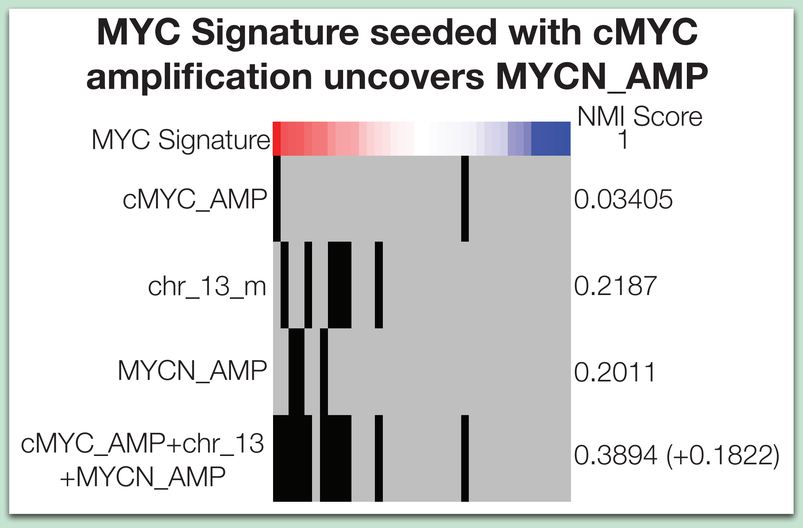
Above, REVEALER removed samples with KRAS or NRAS mutation on the pathway-activated side, as it explains their high expression. Then, high expression of KRAS pathway is used to uncover potential new causes of activation in the genomic feature dataset. Here, BRAF, which shares downstream targets with the RAS family, is found as an alternative activator by REVEALER.



Above, addition of BRAF increases NMI by 0.04792, indicating that BRAF is not only a good match on its own, but is orthogonally descriptive of KRAS pathway activation.

*Expression and copy number data from the Cancer Cell Line Encyclopedia (CCLE), a collaboration between the Broad, Novartis Institute of Biomedical Research (NIBR), and Genomic Institute of Novartis Foundation. Mutation data from The Cancer Genome Atlas (TCGA).

5. REVEALER finds known activators



Analysis of MYC signatures in meduloblastoma patient data from Tamayo, *et al*, *J Clin Oncol* (2011) with cMYC amplification uncovered amplification of another member of the MYC family. NMI increases tenfold with the addition of REVEALER discoveries.

Works in Progress

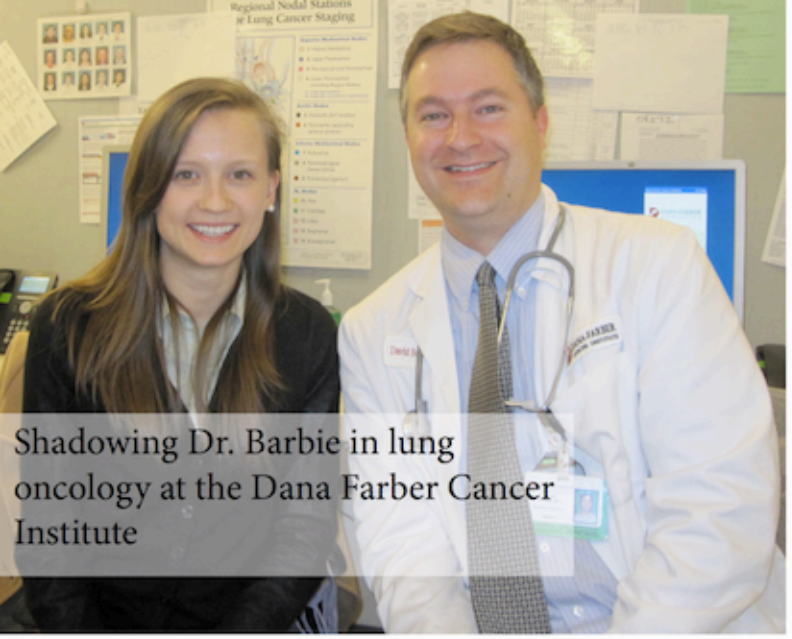
- False discovery rates, p-values. NMI calculation is fast, but thousands of permutations is unrealistic for an average user.
- Discovery of continuous data sets with other continuous data sets. NMI is capable of handling any data type, continuous or discrete, and can elucidate non-linear reactions.
- Integration into GenePattern and release as an R package. A manuscript is in preparation and REVEALER will be released and available online with the full description of the algorithm.

Summary

REVEALER is a novel algorithm that identifies candidate activators by searching orthogonal genomic aberration spaces using a mutual information metric.

Acknowledgements

OBB thanks PT and JPM for their invaluable mentorship and members of Todd Golub's and Levi Garraway's labs for their constant feedback.



Shadowing Dr. Barbie in lung oncology at the Dana Farber Cancer Institute

Final day at 2nd grade "Body Maps" (Anatomy) Science Club for Girls

Presenting REVEALER at ISMB

MIT Dancetroupe Fall 2010 Performance

NICHOLAS CHORNAY—THE TECH