Integration of a Personal Genome Analysis Tool into Integrated Genome Browser

The cost of sequencing a human genome has fallen from $3 billion dollars in 2001 to approximately $1,000 today, making it practical for the first time in human history for individuals to have their own personal genomes sequenced. To date, around 228,000 human genomes have been sequenced and current estimates project that this number will double each year[1]. This flood of new data requires new approaches to analyze, visualize, and share this data. Making data accessible to more people is becoming even more important because companies like 23andMe are selling genotyping services directly to consumers, bypassing the clinic.

A pioneer in the direct-to-consumer genotyping space, 23andMe offers an inexpensive consumer whole genome genetic testing kit. For $99 an individual can determine their genomic sequence at 1 million positions throughout their genome. For cost reasons, 23andMe does not report the DNA base letters for every position in the human genome. Instead, they report single nucleotide polymorphisms (SNPs), DNA bases that often differ between individuals. Many of these SNPs are utterly harmless; they don't affect health but instead simply reflect our natural human diversity and ancestry. Other SNPs, however, can cause problems. This occurs when these SNP bases alter the encoding of proteins or other functional elements that need to work correctly to maintain human health.

The overall goal of our proposal is to help people with little training in biology to understand these differences and how they can affect genome function. Ultimately, we want everyone who gets their genome sequenced (or their SNPs identified) to understand the molecular basis of how differences in their DNA affect their genes, and ultimately their health. We hypothesize that by making these data available in a zoomable, flexible, fun-to-use graphical format, we can achieve this goal. Unfortunately, there is currently no way for consumers to interactively view his or her own genomic data or to investigate their data in relation to current research. Therefore, we will write a program that allows consumers to easily convert and import their 23andMe data into the Integrated Genome Browser and add new features to IGB for personal genome exploration and analysis.. In addition, we will provide training videos to educate consumers about genes, genetic markers, and how differences between genomes can affect health.

The specific aims of the proposed project are:

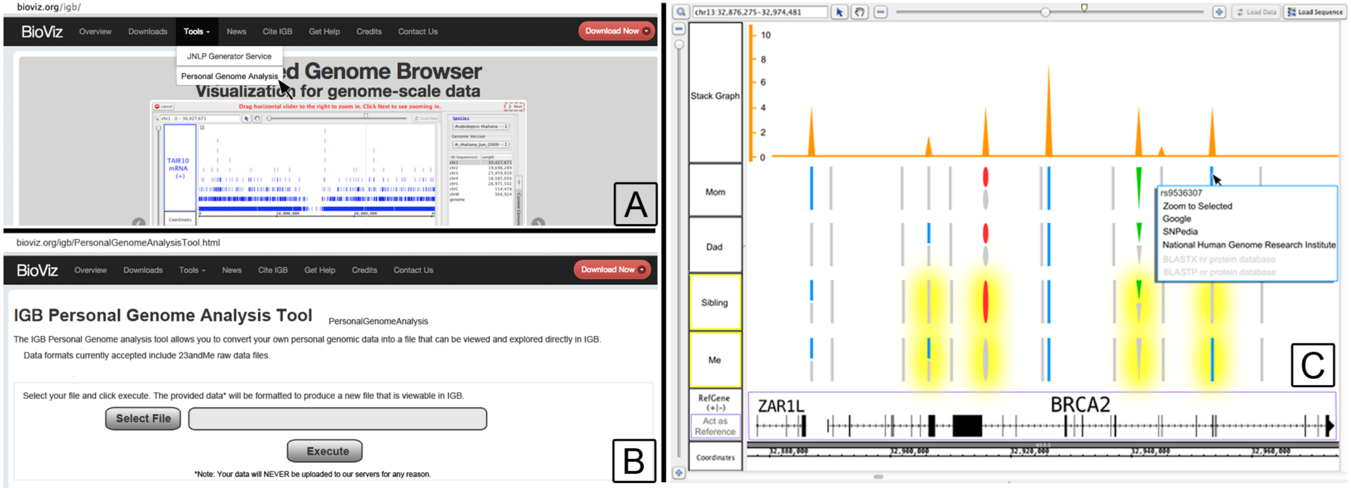
1. Write a program to parse 23andMe data into an IGB readable file. We'll deploy this on the IGB web site (BioViz.org) under the "Tools" menu.
2. Develop an IGB plug-in that will enable users to "link out" to Web resources that help them explore genetic differences between their personal genome and other genomes.
3. Develop educational materials to help new users understand genetics and genomics. To test our materials, we'll consult with computer science students whose last biology coursework was pre-college.
4. Work with experts in personalized genomics to design better user interfaces. For this, we will work with Dr. Rachel Karchin of Johns Hopkins University and Dr. Steve Chervitz, Research Scientist at Personalis.

The Integrated Genome Browser (IGB) was developed first at Affymetrix and now at UNC Charlotte as a way for researchers to easily view and explore various forms of genomic data[2]. It is a Java-based, rich client genome browser used by thousands of scientists, to serve the needs of consumers seeking to understand their personal genome. As IGB already has thousands of users, and ease of use was heavily focused on during development, the addition of a consumer focused program would be a natural fit.

To develop and test our proposed user interface, we will use the 23andMe data from Dr. Nowlan Freese, who is leading this project. We believe this is important, as it will simulate an end user's experience a consumer using IGB. In our first prototype for how data will be displayed (see Figure 1), a combination of colors and shapes will be used to signify key features and genomic differences. SNPs will appear as vertical bars, and if the SNP is heterozygous the bar will appear broken into two halves, one inherited from the mother and one from the father. When the SNP matches the reference, it will appear greyed out, while those that do not match the reference will appear blue. Insertions and deletions in the genome will appear as triangles and ovals, respectively, and will also change color based on whether they match the reference genome.

As 23andMe data is often shared between family members, we also want to provide an intuitive means of comparing multiple individuals. Users will be able to load multiple tracks and view them simultaneously. If a user wishes to compare a subset of tracks directly they will be able to select those tracks for comparison. Any differences between the selected tracks will appear highlighted in yellow.

Since the vast majority of SNPs are shared, users will also need a way to quickly identify regions where there are differences. The Stack Graph functions to quantify the number of differences between individuals' SNPs, insertions, and deletions, displaying it above the tracks. Once a user identifies a feature of interest, they will be able to right click on it to bring up a menu, which will allow the user to search external sites for more information.



IGB Personal Genome Analysis Tool

The IGB Personal Genome analysis tool allows you to convert your own personal genomic data into a file that can be viewed and explored directly in IGB.

Data formats currently accepted include 23andMe raw data files.

Select your file and click execute. The provided data* will be formatted to produce a new file that is viewable in IGB.

*Note: Your data will NEVER be uploaded to our servers for any reason.

Figure 1. Simple prototype illustrating a personal genomes visualization workflow. (A) Users will visit the IGB home website (Bioviz.org). Under the **Tools** dropdown menu at the top of the page, they will select the Personal Genome Analysis link. (**B**) This will take them to the **Personal Genome Analysis Tool** page, where they'll select a file and run the tool. Note that the code will run in the user's browser and will not need to be uploaded to our server, thus protecting user privacy. **(C)** Users will either stream data directly from the tool into IGB, or first save it as a local file and open it using the **File > Open File** menu in IGB. Because IGB is a local application, and not a web site, users can explore their own data in privacy. Shapes differentiate variation such as SNPs, insertions, and deletions, while colors differentiate between the users data and that of a reference genome.

PROJECT PERSONNEL (in addition to PI Ann Loraine)

Dr. Nowlan Freese, Ph.D. – Dr. Freese will serve as project manager. Dr. Freese leads outreach and training efforts for the IGB project. Has a background in biology, a broad understanding of genome visualization programs, and access to 23andMe data. He will help design and test the user interface, write documentation, help mentor the students, and produce screencasts.

Daniel Narmi – Mr. Narmi is pursuing a Professional Science Masters Degree in Bioinformatics at UNCC. He has a background in both computer science and bioinformatics, and is experienced with Java and javascript. He will help with data wrangling, developing code, designing the UI, and writing documentation.

Computer Science Masters student – to be named. The student will work with Mr. Narmi on programming and will help review documentation. He or she will help demo the tool to other CS students and identify ways the tool can be improved to make it more accessible to our target users, 23andMe customers who are interested in genetics, but may not have formal training in genetics. We have several candidates for this position thanks to interviews we recently conducted to fill a position vacated by Tarun Kanaparthi, who graduates in December.

Steve Chervitz, Ph.D. Dr. Chervitz is a Bioinformatics Scientist at genetic testing company Personalis. An IGB power user, Dr. Chervitz will provide consulting on using IGB to visualize personal genomics data.

BUDGET

We request $6000 stipend for spring 2015 for Daniel Narmi and the Graduate Student in Computer Science, totaling $12,000 for both students. We also request $6000 to support Dr. Freese's effort and $4,000 to purchase a computer workstation and monitor for both students. Lastly, we request $500 to provide an honorarium for Dr. Chervitz and also to provide Amazon gift cards for students and others who help evaluate the interface and documentation. Altogether, our request totals $18,000 (stipends) + $4,000 (equipment) + $500 (honorarium, gift cards).

[1] http://www.technologyreview.com/news/531091/emtech-illumina-says-228000-human-genomes-will-be-sequenced-this-year/

[2] The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets. 2009, Bioinformatics.