Application of Fast Estimation of Privacy Risk in Functional Genomics Data (FANCY) to Various Genome-Wide Data Types

Abstract

The recent surge in the field of personalized genomics has led to increased support and attention to the implications of such research. One of the pressing issues in the field is the personal privacy associated with a generation of large datasets from patients' biological material. Functional genomics data such as gene expression levels specifically is becoming clinically relevant due to the direct relationship between the genome and the diseases, however determining privacy leakage from such data is difficult due to the different sequencing techniques that are used. FANCY is a tool that rapidly estimates the amount of private information leakage from functional genomics data and identifies whether the data is "safe" or "not safe" enough to be released publicly. In order for the tool to truly be versatile, it must be applied to a variety of datasets and be user-friendly for the researchers that are not familiar with the field of genome privacy. In this proposal, we plan to develop an intuitive interface and a web server that makes it easier for researchers to utilize FANCY. We will then use this new interface to quantify the private information leakage from widely used functional genomics data sets from different sources and create a database associated with the risk of sharing these publicly available datasets. This research will help data producers to assess the risk to the patients' privacy before the release of the data.

Background and Significance

With personalized genomics becoming the future of the biomedical industry, information about ones genomics is rapidly becoming available in a public forum. This has begun to raise concerns about privacy of the patients as in a research and clinical setting more and more genomics assays are performed.. With the advent of the "omics" revolution, the type of the experimental genomic assays that is required to understand the diseases has expanded largely. Functional genomics experiments such as RNA-Seq and ChIP-Seq are some of these many new experimental types that researchers can infer a wealth amount of biology from. However, the efforts to understand the risk of privacy loss to a patient when these new data types are released publicly remain low.

There are two major ways in which these functional genomics data will compromise the privacy of an individual: (1) the comprehensivity of the experiments and the accessibility to the digital information that is controlled. In order to infer important biological information from these experiments,we need to sequence a large amount of individuals with different phenotypes. Therefore the implications of inaccuracies with security would be large-scale and affect individual medical information for a myriad of individuals. (2) Given these circumstances, it has been hypothesized that embedding this genomic information within digital networks and the Internet will create large issues with the management of privacy and and consent in the era of big data. Specific risks include re-identification of an individual and data security fallacies that need to be addressed with both clinical and computational knowledge to ensure that as biotechnology and new experimental assays continue to grow measures are taken to have security systems grow in parallel.

Functional genomics experiments allow for detailed characterization of disease states. Given the large amount of critical information these experiments provide, there is a gap between the utility of such readings and the personal privacy of the individual to which those readings belong.

Functional genomics data allows for a detailed characterization of disease states and susceptibility, which is why scientific funding agencies' are interested in creating databases with functional genomics data from a large number of individuals. This can be problematic as it is important yet difficult to bridge the gap between utility of data and privacy in terms of sharing functional genomics data (3). In comparison to the level of security that exists for DNA sequencing based data, very few tools exist for evaluating and tending to privacy loss in functional genomics data (4).

What is needed to address this problem is a robust method to assess the level of privacy required for function genomic experiments. Before the release of such information to other agencies, it is necessary to quantify the number of possible leakages that could occur. This is possible through genotyping raw sequences and overlapping them with genotypes obtained from whole-genome sequencing of the individual. In order to create a robust way to predict the number of leakages this would create the Gerstein Lab developed FANCY - Fast Estimation of Privacy Risk in Functional Genomics Data. FANCY is a supervised learning method which utilizes a Gaussian Process Regression model and machine learning techniques to infer the number of leaking single nucleotide variants (SNVs) from raw functional genomics data (GAMZE). The model utilizes an array of factors that allow for a 95% confidence level in predicting the number of leakages that could occur given the functional genomic data of an

individual or group of individuals. More specifically, the model utilizes statistical attributes such as standard deviation and skewness to predict the total number of leaking SNVs. The model is intuitive and relatively user friendly - outputting a simple warning message depending on the level of privacy leakages that are predicted.

FANCY utilizes a two-step method, the first being a regression framework that aligns the raw functional genomics to a reference genome. Utilizing previously mentioned statistical attributes true SNVs were determined and low quality SNVs were filtered. The second step was comparing filtered SNVs with the gold standard SNVs that were generated from the whole genome sequence. In doing so, the true number of leaking SNVs could be determined and then sorted based on allele frequency. The dataset utilized to generate FANCY was RNA-Seq data (5) generated individuals in a collaborating project (gEUVADIS) and another data set of 100 individuals generated by the PsychENCODE Consortium (6).

Hypothesis

Given the current limitations in the predictive modeling of possible leakages in the upcoming field of personalized genomics, a supervised learning method that compares SNVs with gold-standard genotyping will be able to conduct a privacy risk assessment and evaluate whether data "can be shared" or "cannot be shared". Such a method will be intuitive to the user and applicable to a wide variety of data types, given the large variation in genotyping that exists currently and will only continue to diversify in the future.

Specific Aim 1: Further Development of FANCY server to Make it User-Friendly

Rationale: While the current FANCY system has been developed in a way to evaluate raw functional genomic data on a scale from safe enough to be shared to cannot be shared, in order

for the tool to be used on multiple platforms, the system must be easy to access and to understand. Therefore, the server website will be improved to have a more functional layout that makes it clear how to utilize FANCY and what the results of the system are. This will be done by editing the source code to create a more functional website in hopes to improve the accessibility of FANCY.

Specific Aim 2: Application of Server to New Data Types and Creation of a Database

Rationale: One of the overall goals to answer the growing challenge of protecting patient privacy while allowing for the growth of scientific progress is to ensure that FANCY can be applied to a variety of data types. I plan to do this by applying the tool to two data types taken from a various consortium that the Gerstein Lab has worked with.

The first data type that I will work with is PsychENCODE Consortium (6). This is an online resource that was generated across 1866 individuals. It has created single-cell expression profiles for many cell types and also allows for the building of a gene regulatory network that links genome-wide association study variants to genes. As it is a relatively large study, this is a good example of a large data type that is susceptible to leakage. I will apply FANCY to this database by utilizing Gaussian Process Regression, calculate the depth per base pair by using samtools (7) and identify statistical attributes such as standard deviation and skewness. I will then filter for low quality SNVs, and overlap the remaining variants with the gold standard SNVs provided by the dataset. Finally I will ensure that FANCY develops an estimation of rare vs. common variants and sort the categories based on allele frequency. In order to evaluate how effective the tool was on this new data type, I will calculate the confidence level, with the goal of being in the 95%. The second data type that I will work with is ENCODE. ENCODE is a

public data set that stores a wide range of functional genomics arrays from participating mapping centers. The same process that was conducted with PsychENCODE will be utilized here and once again confidence level will be evaluated and improved upon.

I will then create a database with the accession numbers of the experiments from ENCODE and PsychENCODE with the associated risk to the privacy that users can easily query.

Potential Pitfalls and Alternate Strategies

One of the possible challenges with this project is the complete evaluation of FANCY on different data types. Given that all data types are provided with some variation, it is important that the tool is able to adapt to these changes and still determine the privacy risk that is there for the individual. Furthermore, if a functional genomics data set leaks less than a certain amount of variants (more than 1000) then the risk of re-identifying an individual from that data set needs to be extremely accurate. For this instance, FANCY is accurate. However, if the number of leaking variants is relatively low (less than 1,000), then the evaluation of privacy may not always be accurate, especially for larger data types. For that, I will further evaluate the features used in machine learning and improve the prediction by exploring techniques other than gaussian process learning. An alternative strategy for this scenario is to test with $FANCY_{Low}$ which is designed with more precision in instances where the variations are at a minimum. The same steps will be taken to determine the success of the tool on these different data types and where improvements can be made.

Personal Statement

I have always been interested in the field of personalized genomics. In high school, I was particularly interested in staying on the frontline of the field, utilizing technologies such as

CRISPR/Cas9 in my work. During my sophomore year of high school I had the opportunity to meet Dr. Jennifer Doudna, and watch her speak about the exciting applications of CRISPR/Cas9. Excited to see where the technology would go, I joined a lab at Oregon Health and Science University's Knight Cancer Institute focused on utilizing CRISPR/Cas9 to identify genes that were causing resistance to an otherwise highly effective drug in Acute Myeloid Leukemia. Using the technology I was able to narrow the search for correlation to resistance from a candidacy of over 2000 genes to 11 possible genes that had correlation. My fascination with the technology led me to continue to work in the sector of biomedical engineering for 2 more years, trying to find ways to improve the was CRISPR/Cas9 plasmids were delivered in a non-invasive fashion.

My senior year of high school I was given the opportunity to attend and showcase my work at the International Nanomedicine and Drug Delivery Symposium (NanoDDS). The conference further ignited my passion for research in personalized genomics, but also brought into question the limitations of the field. During many of the panels and discussions a major issue was approval from organizations to allow for testing on a personal level due to privacy issues. Realizing that this would be a significant blockade in the development of the field, I started to grow an interest in bioinformatics and the importance of privacy in genomics.

Now that I have had the opportunity to take classes in Computer Science at Yale, I know that I want to combine my knowledge of the field with my passion for genomics. I am looking forward to having the opportunity to work with Gerstein Lab, and to gain more skills in machine learning techniques, coding practices and how personalized genomics and privacy can intersect in a cohesive fashion. I have had over 4 years of research experience from high school and understand how to read primary literature, interpret data and conduct important background

research. By having this exposure I know that I will learn how to think deeply about problems with a Computer Science lens, and in doing so, improve my skills in programming while understanding how to apply those skills in a way that will help me in my future endeavors. Therefore, if I can obtain funding through the First Year Summer Fellowship, I know that I will be taking a step towards developing the problem-solving skills that will help me in my future career.

References

1. Vaszar, L. T., Cho, M. K., Raffin, T. A., Squassina, A., Nielsen, L. F., Møldrup, C., … Stanford University Center for Biomedical Ethics. (2004, November 4). Privacy issues in personalized medicine. Retrieved from https://www.futuremedicine.com/doi/abs/10.1517/phgs.4.2.107.22625

2. Getting Started. (n.d.). Retrieved from https://www.encodeproject.org/help/getting-started/

3. Gamze Gürsoy, Charlotte Brannon, Fabio Navarro, Gerstein MB. FANCY: Fast Estimation of Privacy Risk in Functional Genomics Data *Bioinformatics*

4. National Institute of Health data sharing policy. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-17-110.html

5. Lappalainen T et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature , 2013;501:506-511

6. Wang D et al. Comprehensive functional genomic resource and integrative model for the human brain. Science , 2018;362:6420

7. Li H et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics, 2009;25(16):2078-2079.