NIH Person's comments

Specific focus on points 3, 4 and 6 would be great. I would also not count publication as a penalization but accept that this was a misstep.  In point 1, I would elaborate more on what is being proposed to be done, and which portions are different from what has been published within the article, and why these need to be funded.

Further things MG wants to include in blue

Very rough text from MG in gray

We'd like to provide a rebuttal to the criticisms in the form of a few key bullet points.

**\* 1 - By effort, grant is half theory & half software**

The grant has three aims. The first two are developing a theory about privacy and the last one is developing practical software to instantiate this. This last aim, of course, is a key point of the RFA we are responding to. We would also like to point out that while there's only a single aim devoted to software development, the work is not really commensurate with this, so we would estimate that at least half of the work, if not more, in the grant would be devoted to practical software development.

**\* 2 – Regarding the 2016 Nature Methods Publication**

We would like to make it clear that we did not anticipate at all when we were writing the grant, we would so quickly finish the work and have it immediately come out in publication. . We just got one set of referees reports and then it was in. This is very lucky for a paper in a highly ranked journal and we believe is a mark of the quality and innovation in our approach. However, this caught us a bit off guard with respect to the grant. . We think that our Nature Methods is a solid proof of how well our approach to developing the formalism is perceived. We believe this justifies better our need to be funded for continuing our work on genomic privacy. We also would like to emphasize that there are several extensions to the work that were not explored in our paper, which we are proposing to accomplish in the context of the grant. These are listed below.

**\* 3 - $145 K/yr to cover the software & some theory extensions**

As we've pointed out above, the published paper is the formalism but there's no practical software for comprehensive analysis and protection of genomic datasets from breaches on privacy. Given this, we believe we could accomplish the remaining work in the grant, the software development, with a budget of a little more than half what we asked for, spread over 3 years -- ie $145 K/yr direct cost. We will also extend some of the theoretical considerations further to incorporate the points that are raised in reviewers’ comments.

**\* 4 Need to Extend Formalism to Precisely Specify How Phenotypic Data can be Anonymized to Mitigate Linking Attacks**

|  |
| --- |
| One of the criticisms we got in the review was that while we described how to quantify the predictability in the information content from the eQTLs, and showed how they could be used in linking attacks, we did not describe a practical procedure for removing the characterizing information from a gene expression set, and hence rendering it more private, or anonymizing it. We agree, while we did alude to how this could be done, we do agree that this could be fleshed out more and we anticipate additional theoretical work needed in the grant period to actually carry this out.**\* 4.1 Approaches for Phenotypic Data Anonymization** |
| Initially, we will study two basic approaches for anonymizing a phenotype dataset. First, we will study how the phenotypic measurements associated with some of the eQTL's that cause highest leakage could be simply removed. We will prioritize the eQTLs to select the ones that have highest predictability and information content. We will develop threshold models that are functions of these two metrics and remove a small number of data points associated with the selected eQTL's. One challenge in the selection of eQTLs to be removed is that the number of combinations of eQTLs increases exponentially. We will evaluate greedy and iterative approaches for selection of the data points that will be removed. The second approach is to introduce noise into the phenotype data to secure it. The noise addition decreases the predictability of the data. We will evaluate several functions and levels of noise functions. After we secure the phenotype dataset we will show that one cannot perform a linking attack on the resulting anonymized data sets and that the anonymized data sets are not biased in terms of various considerations, for example, the noise addition (and data point removal) must be done carefully to not bias the dataset. An attacker should not be able to easily identify which gene expression levels have been altered by noise addition. As yet another approach, we will also study the hybrid anonymization techniques where data point removal and noise addition methods will be utilized for anonymization. This will allow us to have a more flexibility on the anonymization procedure.**\* 4.2 Quantification of Biological Utility of Phenotypic Datasets**The utility of the biological dataset must be considered while the data points are removed or noise is added to the dataset. The main factor for the anonymization is to assess the tradeoff between utility and anonymity of the datasets. In particular, the increase in the anonymity of the data (decrease in predictability and information content) versus decrease in biological utility must be tracked in the course of anonymization. First, we will extend the information theoretic framework to account for measuring the utility of the dataset. We will explore the possibilities of comparing the distributions of phenotype measurements, in information theoretic terms, before and after anonymization steps. For example, for a gene expression dataset, if the distributions of gene expressions for the anonymized genes are too distant from each other, the anonymization can be stopped. We will also formulate the biological measures and incorporate them in the estimation of the data utility. For example, we will evaluate the possibility of conserving the set of differentially expressed genes while the an expression dataset is being anonymized. As a first step in building the anonymization, we will perform computational experiments for simulating the QTL datasets and perform simulated anonymizations of these datasets to estimate how well each anonymization strategy works in practice. |
| **\* 4.3. Large Scale Testing of the Anonymization Formalism**After we show how anonymization can be performed efficiently and effectively, we then have to extrapolate these conclusions from the data sets at hand which cover eQTLs in a certain number of tissues to larger and larger data sets. Here, we will greatly benefit from the new GTEX and newly genetared large scale pancancer data sets, for example ICGC’s pancancer analysis of whole genomes datasets (PCAWG). GTEX data sets that have recently come out where we can take the data set and show how with increasing the number of tissues and progressively higher coverage in more individuals, the number of eQTLs increases, but there is some form of saturation in terms of the eQTLs with high predictability. We hope to show that by still removing only a fairly small number of eQTL's and gene expression levels one could still get a private data set. |
| **\* 4.3. Integration of the Anonymization Formalism with Existing File Formats**Related theoretical work has to also be done in relation to the number of variants removed through the MRF file format. This file format removes a substantial amount of sensitive variant information from an RNA-seq experiment. This is accomplished by censoring the read sequence information in the sequenced data file, e.g., fastq file. However, it is still possible to quantify the expression levels of genes and perform linking attacks. Another complication is that MRF file formatting does not remove structural variants, particularly those in non-coding regions. We will use simulations to study how the structural variant information can be inferred from RNA-seq data and it can be used as a source of sensitive information leakage. We will analyze gene expression sets that have progressively higher and higher depth and estimate the predictability of structural variant genotypes. We believe that there can be substantial sensitive information leakage through structural variant genotypes. |
| Overall, we hope to show that the amount of information leakage after removing the obvious variants with MRF, and the strong EQTL's, is fairly large, and one has mostly a private data set. However, even after strong anonymization, there will still be information leakage. We will quantify the remaining information leakage, and aim at bounding it at some user defined number. The extended information theoretic framework will be used to perform this quantification. We anticipate this will require a substantial amount of theoretical work, and then obviously a large amount of software development and practical simulation adapting it to specific data sets. |

**\* 5 - Obvious extensions to work on (ChIP-Seq)**

There are some obvious areas that we can provide further clarification in a formal rebuttal and which would occupy significant effort in the grant. One area that the proposal was consistently, and we feel unfairly, criticized in relation to, was its focus on expression QTLs, or eQTLs, related to RNA-seq, as opposed to, which was seen as a very narrow but important focus. We picked this narrow focus because of the large amount of RNA-seq data available, as opposed to other functional genomics data, such as ChIP-seq and Methyl-seq. We felt this would give us a concrete case to focus on. However, we believe that our formalism that we've developed, and the resulting software, will be easily applicable to use in other functional genomics data sets for which QTL datasets exist. For instance, there is a growing compendium of QTL datasets for ChIP-Seq datasets, these can be easily incorporated into our analysis framework.

|  |
| --- |
|  |
| In the course of the grant, we actually plan to do this, showing how the formalism can easily be applied to ChIP-seq data sets. We also expect that many more functional genomics data sets will be generated, most possibly during the grant period, that will help us greatly. Some of these datasets comprise many different tissues and some of them include even population level information. We see these becoming available from the GTEx Consortium, and also from another consortium, such as PsychENCODE, which is planning to generate ChIP-seq data sets on almost a thousand individuals. We plan to show how the formalism can easily be applied to these data sets, reducing the privacy. These datasets will make it much easier to study other data types. |
| Although we do not anticipate that the formalism will need to be changed much to adapt to these new dataset, it will certainly need to be extended. Unlike RNA-seq data, the ChIP-Seq data processing pipelines (and generally other functional genomics pipelines) generate different types of data abstractions. For example, the “peak calls” from ChIP-Seq data will most likely be treated differently from the ChIP-seq signal profiles. Although one could think that the peak calls do not convey much sensitive information, we believe there may be certain portions of the genome where peak calls may reveal underlying sensitive genetic information. There are also other considerations in ChIP-Seq. For example, some of the targets of the ChIP-Seq experiments exhibit broad mark patterns (like histone modifications) while others exhibit much more punctate patterns (like transcription factors) on the genome. It is necessary to consider how new data types will be incorporated to estimate predictability and information content.  |

**\* 6 –Collaborative work between PIs already developed**

We were criticized for not having a collaborative work plan. However, we wish to point out that both the Yale and UCB teams have taken important roles in large consortia and collaborated together with each other. We therefore believe this is a non-issue.

**\* 7 - Keen to get one of the few grants for privacy**

Finally, we are very keen to be funded by this RFA. We'd like to point out that it is very important for the work in the laboratory to get dedicated funding for privacy. Privacy is a particularly important and recurring problem for many areas of genomics, and we have worked on various privacy issues in this area for more than a decade. We've worked on privacy for many years, patching together funding for this from a variety of sources, but not really have any proper funding . We have been scrapping together resources to do the work that we've done. Even without formal funding we've managed to do a number of conceptual, ethical, and theoretical pieces on it. The central topics of these were about the social implications, measuring the information content, and thinking about what type of formats and computer set ups would be necessary to enable secure genetic computing. However, to get into practical solutions and software development we'd need funding. Thus, we'd like to get even a small amount of official NIH support for this effort.