Specific Aims [needs to 1 page, ~750 words]

The genomic characterization of thousands if not millions of individuals promises to be very useful for medical research, allowing us to find the subtle correlations between genetic variants and various disease states necessary for future advancements in medicine. However, the very nature of next generation sequencing (NGS) – i.e., the rapid technology that can decode DNA's unique and complete description of each and everyone one of us-- that permits the understanding of the genetic underpinnings of disease, also results in a dataset that is extremely private and exceptionally revealing. More troubling, large swaths of these datasets are shared amongst non-sequenced family members, making consent unlikely for many who have a lot to lose from the disclosure and/or misuse of the data.

Much of the attention in genomic privacy is focused on variant datasets, however other, -omic datasets could not only be potentially very useful, but might be less problematic with regard to privacy concerns. These other datasets are drawn from, amongst others, functional genomics experiments such as RNA-Seq, Methyl-Seq and ChIP-Seq. These experiments probe quantitative endophenotype traits such as gene expression levels or levels of methylation. Furthermore, while these quantitative traits can be correlated with particular variants in the framework of eQTL and other associations, these correlations are often weak and are not particularly revealing. However we propose to show that with a large enough number of available QTL correlations, a substantial amount of privacy information can be gleaned from functional genomics datasets.

AIM 1

[JC2MG150915] As our first goal, we aim to develop a mathematical formalism to calculate the amount of information leakage from a gene expression set that is publicly available. Specifically, we will examine the predictability of particular eQTLs, especially extreme ones, and how one trades off between privacy and public access to these gene expression sets.

[MGold] As our first goal, we aim to develop a mathematical formalism to characterize the information leakage from a gene expression set that could allow that set to be potentially linked to a variety of genotypes. This formalism will calculate the amount of characterizing information in a genotype dataset in a set of predicted genotypes and also the predictability of particular eQTLs and how one trades off between these. [D2]

AIM 2

[JC2MG150915] In the second aim, we will demonstrate how one can instantiate this mathematical formalism from Aim 1 to create a practical 'linking attack' (explain briefly?). We will use particular outlier gene expression levels and show that it is possible to associate a set of particular genotypes to an individual, thereby potentially revealing sensitive individual-specific information. We will show how straightforward this can be with the use of publicly available gene expression datasets.

[MGold] In the second aim we will demonstrate how one can instantiate this particular mathematical formalism to create a practical linking attack. We will use particular outlier gene expression levels and show that it is possible using these outliers to often avert a gene expression dataset coming up with a set of particular genotypes to link it to an individual, perhaps revealing sensitive information. We will show how easy this is instantiate practical gene expression datasets.

AIM 3

[JC2MG150915] For the third aim, we will develop software to simulate a privacy attack, specifically on the outliers of the data, and implement our mathematical formalism to characterize the amount of information leakage in gene expression and other quantitative datasets. Then, we intend to demonstrate how information leakage can be reduced by a variety of simple file format manipulations and abstracting extreme gene expression levels that are the most identifiable. Here, we build on our previous work in developing a simple file format for gene expression analyses, which removes a lot of privacy-problematic variants. By taking into account the way genomic structural variants and particular splicing events can be reflected in gene expression and subsequently selectively removing and/or noising over them in gene expression datasets, we will be able to produce a dataset that allows public access with minimal information leakage. It will be an accomplishment and of value to make even a portion of a dataset publicly shareable. Thus, we intend to develop an easy-to-use software that through a number of mathematical manipulations, will result in a fraction of a gene expression dataset that can be shared publicly without compromising the individual's privacy. Furthermore, by characterizing the amount of information leakage, we can convey more explicitly to an individual what they are consenting to, in terms of the release of their genomic data. This will involve trying to simplify the idea of information leakage, which is more formally expressed in information-theoretic terms, so that it can be understood and transcribed in the framework of a consent form.

[MGold] For the third aim, we will develop software for implementing our mathematical formalism to simulate an outlier attack and characterizing the information leakage in gene expression and other quantitative datasets. Then we intend to demonstrate how information regarding the leakages can be reduced by a variety of simple file format manipulations and removing much of the extreme gene expression levels that are characterizing for an individual. Here we build on our previous work developing a simple file format for gene expression analysis, which removes a lot of privacy-problematic variants. In taking into account the way structural variants and particular splicing events can be reflected in gene expression and selectively removing and/or noising over them in gene expression datasets, we will be able to produce a dataset that while still having some information leakage, could nevertheless still be used in a public context.

Even having some parts of a dataset publicly sharable is a valuable accomplishment, thus we intend to develop easy to use software that, through a number of mathematical manipulations, will result in a fraction-of-gene-expression dataset that can be shared

publicly without compromising the individual's privacy. Furthermore in characterizing the amount of information leakage an individual consenting to their release has a clear idea of what they are consenting to. This will involve trying to boil down the information leakage, which is expressed in information theoretic terms but in simpler ways that can be understood in the framework of a consent form.[DG2MG: not sure what this means]