New Milestones

\* Year 1 – Development of mathematical formalism for quantifying information leakage & privacy attacks and anonymization of the data In particular, focus will be on developing new metrics for measuring the biological utility of phenotypic datasets and the trade-off between utility and sensitive information leakage. We will build solid formalisms using information theory for modeling the sources of sensitive information leakage. For example in DNA-sequencing: single nucleotide polymorphisms, structural variants, and indels; in RNA sequencing: gene expression levels, splicing patterns and alternative splicing, measures of unannotated transcription, transcript level expression quantifications, and signal tracks; in ChIP sequencing: the signal tracks and peaks calls. We will mathematically unify these diverse set of data types and data summarizations and propose simple metrics of information leakage and biological utility. We will also develop anonymization strategies for protecting data at different risk levels that the users deem acceptable for their scenario. We will finally propose risk management protocols for different genomic data types that are easy to deploy. We will also develop different types of attacks scenarios and test the risk management protocols under different types of attacks.

\* Years 1 & 2 – Development of practical software instantiating the above formalisms. Software development will be aimed at making the formalisms developed in year 1 practical for everyone’s use. We will distribute the software from github where it will be open to public for download and comments. The software will be tested vigorously on large scale computes. The code will be updated to encompass new types of attacks on different data types.

\* Year 2 - Large-Scale Deployment of the Risk Management and Anonymization Formalism

We will deploy the implemented software on real data types. Some of the large scale sequencing projects like TCGA, ENCODE, GTex, GSP, etc will be the most suitable applications of the developed software tools. We will build datasets that are anonymized at certain risk levels and test those using different attack scenarios.

\* Years 2 & 3 - Develop file formats systematically removing private information (Integration of the Anonymization Formalism with Existing File Formats) We will develop new data file formats that enable efficient distribution of the anonymized datasets. These data formats will be supported by our software packages for easy accession and processing. Our lab proposed MRF format previously that was implemented with RSeqTools software package. We will extend MRF’s capabilities for ensuring more inclusive protection of privacy.

\* Year 3 - Instantiate other sources of extremity into practical software

We have shown that extremity is a simple yet very effective concept that can be used to breach privacy in linking attacks. The simplicity of these attacks are rather concerning with respect to protection of privacy and they must be studied in greater depth. We will study different extension of extremity in different data types and scenarios and implement these attacks into the software package. We will also develop anonymization techniques against the extremity attacks.