Experience on Genomic Privacy Analysis

# Previous experience in tool development and file formats for anonymizing sequence information

We have developed a number of tools and data formats to handle the increasingly large quantities for data generated by RNA-Seq experiments. For example, we have developed the Mapped Read Format (MRF), a compact data summary format for short, long and paired-end read alignments that enables the anonymization of confidential sequence information.

# Previous experience in RNA-seq and ChIP-seq computational technology development and data analyses

We have extensive experience in RNA-seq tool development and analysis1. For RNA-seq, we have developed RSEQtools, a suite of tools that enables anonymization of sequence information and quantification of annotated RNAs and identification of splice sites and gene models 2.

These tools consist of a set of modules that perform common tasks such as calculating gene and exon expression values, generating signal tracks of mapped reads and segmenting that signal into actively transcribed regions. RSEQtools is implemented in C and the source code is available at<http://rseqtools.gersteinlab.org/>. In addition, we have developed IQseq, a computationally efficient method to quantify isoforms for alternatively spliced transcripts 3. Our Database of Annotated Regions with Tools (DART) package contains tools for identifying unannotated genomic regions enriched for transcription, as well as a framework for storing and querying this information4. We developed incRNA, a method that uses known ncRNAs of various classes as a gold-standard training set to predict and analyze novel ncRNAs 5. We have also worked on characterizing the sources of transcription in the genome. A large fraction of the transcription comes from genomic regions not associated with standard annotations, representing ‘non-canonical transcription’. We also have extensive experience in characterizing the functions of genes and non-coding elements via expression data through clustering and network analyses.A group of genes in a co-expression cluster have often been demonstrated to be responsible for a common function. While there are well-known algorithms for expression clustering such as hierarchical clustering, spectral clustering and K-means, we have developed several novel methods. We developed a spectral biclustering method for co-clustering genes and conditions. More recently, we developed a new clustering framework, OrthoClust, for simultaneously clustering network data across different contexts6. OrthoClust is able to identify conserved and specific components across different networks. We also have extensive experience in tool development for ChIP-Seq datasets 7–10.

We also serve as the data integration hub in the exRNA consortium (<http://exrna.org/>), which is generating hundreds of RNA-seq and small RNA-seq samples. Other notable consortia for which we have been involved in pipeline construction and big data processing and analyses include the BrainSpan project ([http://www.brainspan.org](http://www.brainspan.org/)/), which collected RNA-seq data for 8-16 brain structures in each of 13 brain developmental stages11, as well as the PsychENCODE project (<http://psychencode.org/>).

### Previous work on various social and practical aspects of privacy

We also have been active in raising privacy concerns with regard to large-scale genomic datasets. Genomic information axiomatically uniquely identifies its owner. Moreover, and perhaps more problematic, individuals represented in genomic datasets implicitly add the genomic information of their close relatives (who share much of the genomic data by relatedness) who likely have not consented to having their data included in the dataset. We have suggested in a number of publications that a combination of technological, regulatory and policy changes might best serve to protect individuals described by this arguably unanonymizable data12. The policy and regulatory changes would reflect changing norms where we, as a society, no longer dogmatically desire anonymity in many areas of our lives, or at the minimum, have come to peace with the lack of privacy in the modern age13. In acknowledging the changing realities, instead of regulating how to seek out data, we suggest that regulations focus on limiting instances wherein DNA related data can be used to harm; for example, by limiting employment or insurance opportunities. We expect that such regulatory changes might further reduce the need for anonymity of heretofore sensitive data. Corresponding technological changes might include considering both how data is stored as well as where that data should be stored14. We have suggested, for example, using cloud based storage options to control and monitor access to data sets and limiting the ability and need to download data to inherently more insecure personal and/or portable computers15. We have also proposed creating "stub-datasets" that have the look and feel of the typical online data sets, but that would be freely available to all researchers. These data sets, while sharing many of the same statistical characteristics, with their larger cousins, would not present privacy concerns as large, and consequently, could be used to develop and profile code before deployment on real datasets.

# REFERENCES

1. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10,** 57–63 (2009).

2. Habegger, L. *et al.* RSEQtools: A modular framework to analyze RNA-Seq data using compact, anonymized data summaries. *Bioinformatics* **27,** 281–283 (2011).

3. Du, J. *et al.* IQSeq: Integrated isoform quantification analysis based on next-generation sequencing. *PLoS One* **7,** (2012).

4. Rozowsky, J. S. *et al.* The DART classification of unannotated transcription within the ENCODE regions: Associating transcription with known and novel loci. *Genome Res.* **17,** 732–745 (2007).

5. Lu, Z. J. *et al.* Prediction and characterization of noncoding RNAs in C. elegans by integrating conservation, secondary structure, and high-throughput sequencing and array data. *Genome Res.* **21,** 276–285 (2011).

6. Yan, K.-K. *et al.* OrthoClust: an orthology-based network framework for clustering data across multiple species. *Genome Biol.* **15,** R100 (2014).

7. Rozowsky, J. *et al.* PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat. Biotechnol.* **27,** 66–75 (2009).

8. Harmanci, A., Rozowsky, J. & Gerstein, M. MUSIC: identification of enriched regions in ChIP-Seq experiments using a mappability-corrected multiscale signal processing framework. *Genome Biol.* **15,** 474 (2014).

9. Bernstein, B. E. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).

10. Gerstein, M. B. *et al.* Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science* **330,** 1775–1787 (2010).

11. Miller, J. a *et al.* Transcriptional landscape of the prenatal human brain. *Nature* **508,** 199–206 (2014).

12. Greenbaum, D., Harmanci, A. & Gerstein, M. Proposed social and technological solutions to issues of data privacy in personal genomics. in *2014 IEEE Int. Symp. Ethics Sci. Technol. Eng.* 1–4 (IEEE, 2014). doi:10.1109/ETHICS.2014.6893418

13. Greenbaum, D. Genomic Data Disclosure: Time to Reassess the Realities. *Am. J. Bioeth.* **13,** 47–50 (2013).

14. Greenbaum, D., Sboner, A., Mu, X. J. & Gerstein, M. Genomics and privacy: Implications of the new reality of closed data for the field. *PLoS Comput. Biol.* **7,** (2011).

15. Greenbaum, D. & Gerstein, M. The Role of Cloud Computing in Managing the Deluge of Potentially Private Genetic Data. *Am. J. Bioeth.* **11,** 39–41 (2011).