**Gerstein lab experience on single cell analysis**

Single-cell RNA-seq

In collaboration with Prof. Nenad Sestan’s and Flora Vaccarino’s group at Yale, together with groups at USC, the Allen Brain Institute and elsewhere, we analyzed large amounts of RNA-seq data to characterize the transcriptome of the human brain during development. The aim of this project was to create a comprehensive map of gene expression and to understand how the human brain changes throughout life. We have already developed RSEQtools, a suite of tools that performs common tasks on RNA-seq data such as calculating gene expression values, generating signal tracks of mapped reads, and segmenting that signal into actively transcribed regions. This project provides a reference atlas of gene expression in different regions of the brain that will provide valuable information to help interpret neurological function and dysfunction. More recently, this collaborative consortium has started work on PsychENCODE, a project aimed at understanding regulatory variants in the context of their functional connections to psychiatric disease. The project’s approach involves a comprehensive examination of the genome, transcriptome, epigenome, and proteome in relation to brain function. We have extensive experience with using single-cell RNA-seq data to deconvolve expression data from bulk tissue. In particular, we have successfully deconvolved the bulk RNA-seq data across 1,866 individuals in PsychENCODE and GTEx using single-cell RNA-seq data via non-negative least squares.RNA sequencing (RNA-seq) of sputum samples can be challenging to interpret due to the complex and heterogeneous mixtures of human cells and exogenous (microbial) material. Therefore, we developed a pipeline that integrates dimensionality reduction and statistical modeling to grapple with the heterogeneity. LDA(Latent Dirichlet allocation)-link connects microbes to genes using reduced-dimensionality LDA topics. We validated our method with single-cell RNA-seq and microscopy and then applied it to the sputum of asthmatic patients to find known and novel relationships between microbes and genes.

Single-cell ATAC-seq

To simulate single-cell ATAC-seq (scATAC-seq) experiments with known cell type labels, we have developed an efficient and scalable scATAC-seq simulation method (SCAN-ATAC-Sim) that down-samples bulk-tissue ATAC-seq (e.g. from representative cell lines) in an organized fashion. Our simulation protocol recreates the homogeneous signal-to-noise ratio in a single scATAC-seq experiment by integrating the different amounts of background noise in separate bulk-tissue experiments and independently samples twice without replacement to account for the diploid genome.

*Reference*

1. Wang, Daifeng, et al. Comprehensive functional genomic resource and integrative model for the human brain. *Science* 362.6420 (2018).
2. Spakowicz, Daniel, et al. Approaches for integrating heterogeneous RNA-seq data reveal cross-talk between microbes and genes in asthmatic patients. *Genome biology* 21.1 (2020): 1-22.
3. Chen, Zhanlin, et al. SCAN-ATAC Sim: a scalable and efficient method to simulate single-cell ATAC-seq from bulk-tissue experiments. *bioRxiv* (2020).

**Gerstein lab experience on pseudogenes**

Pseudogenes are defined as disabled counterparts of functional protein coding genes. Depending on their formation mechanism, pseudogenes can be classified into three groups: processed (created through a retrotransposition event), unprocessed (resulting from a gene duplication event) and unitary (originating from functional genes through a loss-of-function event). Processed and unprocessed pseudogenes have a functional protein-coding counterpart (parent gene) in the same organisms. By contrast, unitary pseudogenes have no parent gene in their host genome, and are characterized by the presence of functional homologs in a syntenic region in a different species or organism. The fourth type of pseudogenes is called polymorphic. These are defined by the presence of a disabling mutation that is not fixed in the population. Specifically, polymorphic pseudogenes are regarded as protein coding genes that possess a loss-of-function variant in the reference genome.

We have much experience in annotating null & LoF alleles. Our work on LoF is related to our work on pgenes, which are basically fixed version of LoF alleles in the human population. Much of this is in the context of 1000G & CMG (rare disease) consortium. Genome sequencing studies indicate that all humans carry many genetic variants predicted to cause loss of function (LoF) of protein-coding genes, suggesting unexpected redundancy in the human genome. We applied stringent filters to 2,951 putative LoF variants obtained from 185 human genomes to determine their true prevalence and properties. We estimated that human genomes typically contain ~100 genuine LoF variants with ~20 genes completely inactivated. We also found a number of previously uncharacterized LoF variants likely to have phenotypic effects. For instance, we identified three separate LoF variants in *PKD1L3* and one in *PKD2L1*; the protein products of these two genes form a putative sour taste receptor complex, so these variants may underlie variation in sour taste sensitivity between humans. We identified 26 known recessive disease-causing mutations, including mutations associated with the severe early-onset conditions Leber congenital amaurosis, harlequin ichthyosis, osteogenesis imperfecta and Tay-Sachs disease.. We also identified 21 strong candidates for novel disease-causing mutations.

We developed a pipeline called ALoFT (annotation of loss-of-function transcripts), a method to annotate and predict the disease-causing potential of loss-of-function variants. The main features of ALoFT include (1) functional domain annotations; (2) evolutionary conservation; and (3) biological networks. Using data from Mendelian disease-gene discovery projects, we showed that ALoFT can distinguish between loss-of-function variants that are deleterious as heterozygotes and those causing disease only in the homozygous state. When applied to de novo putative loss-of-function variants in autism-affected families, ALoFT distinguishes between deleterious variants in patients and benign variants in unaffected siblings.

*Reference*

1. SM Baxter, JE Posey, NJ Lake, N Sobreira, JX Chong, S Buyske, EE Blue, LH Chadwick, ZH Coban-Akdemir, KF Doheny, CP Davis, M Lek, C Wellington, SN Jhangiani, M Gerstein, RA Gibbs, RP Lifton, DG MacArthur, TC Matise, JR Lupski, D Valle, MJ Bamshad, A Hamosh, S Mane, DA Nickerson, Centers for Mendelian Genomics Consortium, HL Rehm, A O'Donnell-Luria (2022). *Genet Med* 24: 784-797.
2. DG MacArthur, TA Manolio, DP Dimmock, HL Rehm, J Shendure, GR Abecasis, DR Adams, RB Altman, SE Antonarakis, EA Ashley, JC Barrett, LG Biesecker, DF Conrad, GM Cooper, NJ Cox, MJ Daly, MB Gerstein, DB Goldstein, JN Hirschhorn, SM Leal, LA Pennacchio, JA Stamatoyannopoulos, SR Sunyaev, D Valle, BF Voight, W Winckler, C Gunter (2014). *Nature* 508: 469-76.
3. DG MacArthur, S Balasubramanian, A Frankish, N Huang, J Morris, K Walter, L Jostins, L Habegger, JK Pickrell, SB Montgomery, CA Albers, ZD Zhang, DF Conrad, G Lunter, H Zheng, Q Ayub, MA DePristo, E Banks, M Hu, RE Handsaker, JA Rosenfeld, M Fromer, M Jin, XJ Mu, E Khurana, K Ye, M Kay, GI Saunders, MM Suner, T Hunt, IH Barnes, C Amid, DR Carvalho-Silva, AH Bignell, C Snow, B Yngvadottir, S Bumpstead, DN Cooper, Y Xue, IG Romero, 1000 Genomes Project Consortium, J Wang, Y Li, RA Gibbs, SA McCarroll, ET Dermitzakis, JK Pritchard, JC Barrett, J Harrow, ME Hurles, MB Gerstein, C Tyler-Smith (2012). *Science* 335: 823-8.
4. Balasubramanian, S., Fu, Y., Pawashe, M. *et al.* Using ALoFT to determine the impact of putative loss-of-function variants in protein-coding genes. *Nat Commun* **8,**382 (2017). https://doi.org/10.1038/s41467-017-00443-5

**Gerstein lab experience on allelic analysis**

We have developed a computational tool, AlleleSeq, for the construction of personal genomes. The tool integrates an individual’s genomic variation data (SNVs, indels, and SVs) into the reference genome. Phase information of heterozygous variants is also incorporated, producing maternal and paternal haplotypes. Chain files generated by the program can be used to account for coordinate offsets between the individual’s parental haplotypes and the original reference genomic sequence. We have previously constructed the personal diploid genome, splice-junction libraries and personalized gene annotations for NA12878 (available as a resource at alleleseq.gersteinlab.org). Furthermore, we have implemented this on a larger scale in a recent publication where we built 382 personal genomes using the variant call sets from the 1000 Genomes Project.

We have extensive experience with analyses of allele-specific expression and binding and developed a pipeline, AlleleSeq, to measure and detect allele-specific events. We have spearheaded allele-specific analyses in several major consortia publications, including ENCODE and the 1000 Genomes Project. We have annotated variants associated with allele-specific expression and binding in a large pool of individuals from the 1000 Genomes Project. These results were made available as an online resource, AlleleDB (alleledb.gersteinlab.org). Most recently, we constructed a high-resolution map of allelic imbalances in DNA methylation, histone marks, and transcription in 71 epigenomes from 36 distinct cell and tissue types from 13 donors.

*Reference*

1. Rozowsky, Joel et al. “AlleleSeq: analysis of allele-specific expression and binding in a network framework.” *Molecular systems biology* vol. 7 522. 2 Aug. 2011, doi:10.1038/msb.2011.54
2. Chen, J., Rozowsky, J., Galeev, T. *et al.* A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals. *Nat Commun* **7,**11101 (2016). https://doi.org/10.1038/ncomms11101

**Gerstein lab experience on variant prioritization**

Variant effects at multiple levels

We were among the earliest to use statistical models to perform systematic network reconstruction of PPI networks by integrating high-throughput data. The task of identifying interactions between biological elements is known as network reconstruction. It requires integrating both large-scale yet noisy data from high-throughput experiments and reliable yet incomplete data from small-scale experiments. Reconstructed networks enable us to infer the direct and indirect effects of variants, and we have worked closely for ~15 years to use networks as a means of prioritizing and investigating variants. We have leveraged network properties (such as centrality) to evaluate the functional significance of genomic variants, revealing that the correlations between them depend on the network type. We devised methods to identify generalized hierarchies (with loop structures). We have also developed HirNet (a method for determining network hierarchies). We also investigated the roles of network bottlenecks (i.e., nodes through which many "shortest paths'' pass) in the context of protein networks.

Genomic variants can also lead to disruptions of network connections. As a result, some recurrent patterns may be perturbed, such as TFs that frequently co-regulate target genes. We have developed the DiNeR method for identifying such changes and analyzing their consequences to downstream gene expression programs. On a larger scale, some network perturbations may propagate to cause major network rewiring. We have developed the TopicNet method to measure such rewiring in transcriptional regulatory networks. We have also applied this idea to study network rewiring in cancer cells, as part of our efforts in producing a general resource for cancer research based on ENCODE data.

We also have experience in estimating the effects of noncoding variants using data integration based on inter- and intra-species conservation, loss- and gain-of-function events for transcription-factor binding, enhancer-gene linkages and network centrality, and per-element recurrence across samples. Using data from multiple sources, we have developed a weighted scoring system to prioritize variants based on their predicted phenotypic effect. We also developed RADAR which extends FunSeq2 by combining RNA regulome and tissue-specific information. We have also built the GRAM pipeline, which is a generalized framework to predict the cellular expression modulating effect of a noncoding variant by incorporating TF disruptive information, histone modification, and cell-specific expression and regulatory network information. GRAM can be extensively applied to fine map causal variants within an LD associated region.

Supervised machine learning framework to assign pathogenicity to SVs

Recently, we have developed a supervised machine learning based framework to prioritize somatic and germline SVs. The underlying hypothesis of our approach is that various genomic and epigenomic features of disease SVs are very distinct from benign SVs observed among healthy populations. Furthermore, such differences can be sensitively identified in appropriate tissue specific contexts. Thus, we have developed tissue specific models that quantify pathogenicity score by comparing genomic and tissue-specific epigenomic features of a given SV to known benign SVs. To build our somatic and germline machine learning models, we leveraged SVs from the Pan-Cancer Analysis of the Whole Genomes (PCAWG) Project, Genome Sequencing Program (GSP) and the 1000 Genomes (1KG) Project. We employed tissue-specific epigenomic data from Epigenome Roadmap, various genomic element annotations, and cross-species conservation metrics to generate our feature sets. Overall, our approach achieved high accuracy in identifying pathogenic somatic deletions (mean AUC of 0.865) and duplications (mean AUC 0.835) across multiple cancer types. Similarly, our germline models showed high accuracy in identifying pathogenic germline SVs in cancer (mean AUC 0.8) as well as cardiovascular disease (mean AUC 0.79).

Moreover, we also evaluated the contribution of various features toward performance of our somatic and germline models. As expected, we found SV length and overlap with conservation related features had the most significant contributors to the predictive performance for these models. Additionally, various noncoding and epigenomic features, including overlap with 3' and 5' UTR, sensitive regions and H3K4me3 signals had significant contribution toward the predictability of our somatic deletion models suggesting an essential influence of SVs on cis-regulatory elements. Similarly, the predictive performance of somatic duplication models primarily depends on overlap with known heterochromatin annotation, UTRs, and sensitive regions. In our germline model, along with various regulatory region features, we also observed strong contribution of 3D-genome annotation toward predictability of our models.

*Reference*

1. Rozowsky, J., Euskirchen, G., *et al*. (2009) PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat Biotechnol* 27, 66-75.2. Harmanci, A., Rozowsky, J., & Gerstein, M. (2014) MUSIC: identification of enriched regions in ChIP-Seq experiments using a mappability-corrected multiscale signal processing framework. *Genome Biol* 15, 474.3. Habegger, L., Sboner, A., Gianoulis, T. A., Rozowsky, J., Agarwal, A., Snyder, M., & Gerstein, M. (2011) RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries. *Bioinformatics* 27, 281-3.4. Du, J., Leng, J., Habegger, L., Sboner, A., McDermott, D., & Gerstein, M. (2012) IQSeq: integrated isoform quantification analysis based on next-generation sequencing. *PLoS One* 7, e29175.5. Lu, Z. J., *et al.* (2011) Prediction and characterization of noncoding RNAs in C. elegans by integrating conservation, secondary structure, and high-throughput sequencing and array data. *Genome Res* 21, 276-85.6. Boyle, A. P., Araya, *et al.* (2014) Comparative analysis of regulatory information and circuits across distant species. *Nature* 512, 453-6.7. Yip, K. Y., Cheng, C., Bhardwaj, *et al.* (2012) Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biol* 13, R48.8. Fu, Y., Liu, Z., Lou, S., Bedford, J., Mu, X. J., Yip, K. Y., Khurana, E., & Gerstein, M. (2014) FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol* 15, 480.9. Yan, K.-K., Gürkan Yardimci, G., Yan, C., Noble, W. S., & Gerstein, M. (2017) HiC-Spector: A matrix library for spectral and reproducibility analysis of Hi-C contact maps. *Bioinformatics* , .10. Yan, K.-K., Lou, S., & Gerstein, M. (2017) MrTADFinder: A network modularity based approach to identify topologically associating domains in multiple resolutions. *bioRxiv* , 097345.11. Cheng, C., Alexander, *et al.* (2012) Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Res* 22, 1658-67.12. Cheng, C. & Gerstein, M. (2012) Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells. *Nucleic Acids Res* 40, 553-68.13. Cheng, C., Yan, K.-K., Yip, K. Y., Rozowsky, J., Alexander, R., Shou, C., & Gerstein, M. (2011) A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. *Genome Biol* 12, R15.14. Gerstein, M. B., Kundaje, A *et al.* (2012) Architecture of the human regulatory network derived from ENCODE data. *Nature* 489, 91-100.15. Yan, K.-K., Wang, D., Rozowsky, J., Zheng, H., Cheng, C., & Gerstein, M. (2014) OrthoClust: an orthology-based network framework for clustering data across multiple species. *Genome Biol* 15, R100.16. Wang, D., He, F., Maslov, S., & Gerstein, M. (2016) DREISS: Using State-Space Models to Infer the Dynamics of Gene Expression Driven by External and Internal Regulatory Networks. *PLoS Comput Biol* 12, e1005146.17. Wang, D., Yan, K.-K., Sisu, C., Cheng, C., Rozowsky, J., Meyerson, W., & Gerstein, M. B. (2015) Loregic: a method to characterize the cooperative logic of regulatory factors. *PLoS Comput Biol* 11, e1004132.18. Mu, X. J., Lu, Z. J., Kong, Y., Lam, H. Y. K., & Gerstein, M. B. (2011) Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 Genomes Project. *Nucleic Acids Res* 39, 7058-76.19. Habegger, L., Balasubramanian, S., Chen, D. Z., Khurana, E., Sboner, A., Harmanci, A., Rozowsky, J., Clarke, D., Snyder, M., & Gerstein, M. (2012) VAT: a computational framework to functionally annotate variants in personal genomes within a cloud-computing environment. *Bioinformatics* 28, 2267-9.20. Balasubramanian, S., Fu, Y., Pawashe, M., McGillivray, P., Jin, M., Liu, J., Karczewski, K., MacArthur, D. G., & Gerstein, M. (2017) Using ALoFT to determine the impact of putative loss-of-function variants in protein-coding genes. *Nat Comm* in press, .21. Khurana, E., Fu, Y., Chen, J., & Gerstein, M. (2013) Interpretation of genomic variants using a unified biological network approach. *PLoS Comput Biol* 9, e1002886.22. Clarke, D., Sethi, A., Li, S., Kumar, S., Chang, R. W. F., Chen, J., & Gerstein, M. (2016) Identifying Allosteric Hotspots with Dynamics: Application to Inter- and Intra-species Conservation. *Structure* 24, 826-37.23. Kumar, S., Clarke, D., & Gerstein, M. (2016) Localized structural frustration for evaluating the impact of sequence variants. *Nucleic Acids Res* 44, 10062-10073.24. Chen, J., Wang, B., Regan, L., & Gerstein, M. (2017) Intensification: A Resource for Amplifying Population-Genetic Signals with Protein Repeats. *J Mol Biol* 429, 435-445.25. Lochovsky, L., Zhang, J., Fu, Y., Khurana, E., & Gerstein, M. (2015) LARVA: an integrative framework for large-scale analysis of recurrent variants in noncoding annotations. *Nucleic Acids Res* 43, 8123-34.26. Rozowsky, J., Abyzov, A., *et al*. (2011) AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol Syst Biol* 7, 522.28. Chen, J., Rozowsky, J., Galeev, T. R., Harmanci, A., Kitchen, R., Bedford, J., Abyzov, A., Kong, Y., Regan, L., & Gerstein, M. (2016) A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals. *Nat Commun* 7, 11101.29. Djebali, S., Davis, *et al.* (2012) Landscape of transcription in human cells. *Nature* 489, 101-8.31. Gerstein, M. B., Rozowsky, J., *et al.* (2014) Comparative analysis of the transcriptome across distant species. *Nature* 512, 445-8.32. Gerstein, M. B., Lu, *et al.* (2010) Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science* 330, 1775-87.33. Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Hsi-Yang Fritz, M., Konkel, M. K., Malhotra, A., Stütz, A. M., Shi, X., Paolo Casale, F., Chen, J., Hormozdiari, F., Dayama, G., Chen, K., Malig, M., Chaisson, M. J. P., Walter, K., Meiers, S., Kashin, S., Garrison, E., Auton, A., Lam, H. Y. K., Jasmine Mu, X., Alkan, C., Antaki, D., Bae, T., Cerveira, E., Chines, P., Chong, Z., Clarke, L., Dal, E., Ding, L., Emery, S., Fan, X., Gujral, M., Kahveci, F., Kidd, J. M., Kong, Y., Lameijer, E.-W., McCarthy, S., Flicek, P., Gibbs, R. A., Marth, G., Mason, C. E., Menelaou, A., Muzny, D. M., Nelson, B. J., Noor, A., Parrish, N. F., Pendleton, M., Quitadamo, A., Raeder, B., Schadt, E. E., Romanovitch, M., Schlattl, A., Sebra, R., Shabalin, A. A., Untergasser, A., Walker, J. A., Wang, M., Yu, F., Zhang, C., Zhang, J., Zheng-Bradley, X., Zhou, W., Zichner, T., Sebat, J., Batzer, M. A., McCarroll, S. A., 1000 Genomes Project Consortium, Mills, R. E., Gerstein, M. B., Bashir, A., Stegle, O., Devine, S. E., Lee, C., Eichler, E. E., & Korbel, J. O. (2015) An integrated map of structural variation in 2,504 human genomes. *Nature* 526, 75-81.34. PsychENCODE Consortium, Akbarian, S., Liu, C., Knowles, J. A., Vaccarino, F. M., Farnham, *et al.* (2015) The PsychENCODE project. *Nat Neurosci* 18, 1707-12.35. Cancer Genome Atlas Research Network (2015) The Molecular Taxonomy of Primary Prostate Cancer. *Cell* 163, 1011-25.36. Li, S., Shuch, B. M., & Gerstein, M. B. (2017) Whole-genome analysis of papillary kidney cancer finds significant noncoding alterations. *PLoS Genet* 13, e1006685.38. Fang, L. T., Afshar, P. T., Chhibber, A., Mohiyuddin, M., Fan, Y., Mu, J. C., Gibeling, G., Barr, S., Asadi, N. B., Gerstein, M. B., Koboldt, D. C., Wang, W., Wong, W. H., & Lam, H. Y. K. (2015) An ensemble approach to accurately detect somatic mutations using SomaticSeq. *Genome Biol* 16, 197.

**Gerstein lab experience on genomic data privacy**

Functional genomics experiments have a somewhat more nuanced position with respect to privacy than traditional DNA sequencing. Functional genomics experiments are often performed to understand the biology behind phenotypes, such as different diseases, as opposed to DNA sequencing which is performed solely for genotyping. In fact, many of the conclusions we infer from these experiments are not tied to the identity of individuals but represent universal statements about biology (e.g., genes up-regulated in cancer). However, by virtue of the experimental procedure, the raw sequences contain the genetics variants of the research participants. On the other hand, the knowledge of these genetic variants are often not necessary to calculate key quantities from the functional genomics data, such as gene expression values or transcription factors binding peaks. Privacy concerns around functional genomics data due to its dependence on next-generation sequencing are typically addressed by increasingly restrictive access provisions limited to a few authorized parties. These restrictions are in tension with the need for large-scale multi-modal analyses that integrate public and private data. In particular, large-scale consortia such as IGVF that aim to generate and disseminate data to a wider community outside the membership of the consortium itself is in immediate need of different modes of raw data sharing beyond controlled-access. We have previously developed an alternate approach: pTools, a theoretical framework and open-source implementation for privacy-preserving transformations that allow standard functional genomics analyses to be performed openly on private data without compromising security. We have shown that the utility loss from privacy-preserving transformation is less than the differences between two biological replicates of a functional genomics sample. We also developed pTools for the use of 10x single-cell RNA-Seq data and showed that downstream analysis such as read counts and cell type clustering are not affected by this transformation.

*Reference*

1. Prashant S. Emani, Gamze Gürsoy, Andrew Miranker, Mark B. Gerstein

bioRxiv 2021.07.18.452853; doi: https://doi.org/10.1101/2021.07.18.452853

1. Gamze Gürsoy, Charlotte M Brannon, Sarah Wagner, Mark Gerstein

bioRxiv 2020.03.03.975334; doi: https://doi.org/10.1101/2020.03.03.975334

1. G Gursoy, T Li, S Liu, E Ni, CM Brannon, MB Gerstein (2021). *Nat Rev Genet* 23: 245-258.
2. G Gursoy, E Chielle, CM Brannon, M Maniatakos, M Gerstein (2021). *Cell Syst* 13: 173-182e3.
3. G Gursoy, N Lu, S Wagner, M Gerstein (2021). *Genome Biol* 22: 263.
4. G Gursoy, P Emani, CM Brannon, OA Jolanki, A Harmanci, JS Strattan, JM Cherry, AD Miranker, M Gerstein (2020). *Cell* 183: 905-917e16.
5. G Gursoy, CM Brannon, FCP Navarro, M Gerstein (2020). *Bioinformatics* 36: 5145-5150.
6. G Gursoy, R Bjornson, ME Green, M Gerstein (2020). *BMC Med Genomics* 13: 78.
7. W Meyerson, J Leisman, FCP Navarro, M Gerstein (2020). *BMC Bioinformatics* 21: 227.
8. G Gursoy, CM Brannon, M Gerstein (2020). *BMC Med Genomics* 13: 74.
9. A Harmanci, M Gerstein (2018). *Nat Commun* 9: 2453.
10. A Harmanci, M Gerstein (2016). *Nat Methods* 13: 251-6.
11. D Greenbaum, A Harmanci, M Gerstein (2014). *IEEE* IEEE International Symposium on Ethics in Science, Technology and Engineering.
12. D Greenbaum, M Gerstein (2013). *The Scientist* 27:26 (1 Oct.)
13. D Greenbaum, A Sboner, XJ Mu, M Gerstein (2011). *PLoS Comput Biol* 7: e1002278.
14. D Greenbaum, M Gerstein (2011). *Am J Bioeth* 11: 39-41.
15. L Habegger, A Sboner, TA Gianoulis, J Rozowsky, A Agarwal, M Snyder, M Gerstein (2011). *Bioinformatics* 27: 281-3.
16. Yale Law School Roundtable on Data and Code Sharing (2010). *Computing in Science & Engineering* 12(5): 8-13 (Sept/Oct).
17. D Greenbaum, M Gerstein (2009). *Am J Bioeth* 9: 15-9.
18. D Greenbaum, J Du, M Gerstein (2008). *Am J Bioeth* 8: 71-4.
19. A Smith, D Greenbaum, SM Douglas, M Long, M Gerstein (2005). *Genome Biol* 6: 119.
20. D Greenbaum, A Smith, M Gerstein (2005). *Nucleic Acids Res* 33: D3-4.
21. D Greenbaum, SM Douglas, A Smith, J Lim, M Fischer, M Schultz, M Gerstein (2004). *Nat Biotechnol* 22: 771-2.

**Nomenclature, naming & building genomic tool resources**

We are currently working on developing a framework for generating privacy preserving Universal IDs for genotype samples across different cohorts. The framework aims at tagging every individual genotype to an ID to be used universally across cohorts which is generated by multiple layers of hashing using Bloom Filters in a privacy preserving manner where the resulting ID cannot be traced back to the individual without additional identifying information. The aim is to prevent sample swaps resulting in disconnected experiments without compromising the security of private data.