**** Key Contributions to the Scientific Community ****

(1) Showed that multi-omic data can be recast as networks and be compared, usefully, to networks in other contexts

Contribution: The Gerstein lab demonstrated that many types of "linear" genomic data can be recast into a network representation and that various models and analogies can be applied from other disciplines. In particular, the lab showed that (1) various network prediction techniques could be transferred to a biological context from other disciplines; (2) the transcriptional regulatory network could be understood in terms of a regulatory hierarchy, with more dominant regulators at the top; and (3) the constraints imposed by the three-dimensional protein structure often require a different organization of the protein-protein interaction network than a conventional scale-free topology, creating different types of hubs (e.g., permanent vs. transient).

Impact: This work directly connected the world of biological molecules to other areas in network science and allowed the computational biology community to leverage approaches from the broader network science community. Conversely, the work showed that the physical constraints of protein structure and molecular evolution require somewhat different connectivity constraints than more common network growth models.

-- Architecture of the human regulatory network derived from ENCODE data.
M Gerstein, A Kundaje, M Hariharan, SG Landt, KK Yan, C Cheng, XJ Mu, E Khurana, J Rozowsky, R Alexander, R Min, P Alves, A Abyzov, N Addleman, N Bhardwaj, AP Boyle, P Cayting, A Charos, DZ Chen, Y Cheng, D Clarke, C Eastman, G Euskirchen, S Frietze, Y Fu, J Gertz, F Grubert, A Harmanci, P Jain, M Kasowski, P Lacroute, JJ Leng, J Lian, H Monahan, H O'Geen, Z Ouyang, EC Partridge, D Patacsil, F Pauli, D Raha, L Ramirez, TE Reddy, B Reed, M Shi, T Slifer, J Wang, L Wu, X Yang, KY Yip, G Zilberman-Schapira, S Batzoglou, A Sidow, PJ Farnham, RM Myers, SM Weissman, M Snyder (2012). *Nature* 489: 91-100.

-- Comparing genomes to computer operating systems in terms of the topology and evolution of their regulatory control networks.
KK Yan, G Fang, N Bhardwaj, RP Alexander, **M Gerstein** (2010). *Proc Natl Acad Sci U* S A 107: 9186-91.

-- Relating three-dimensional structures to protein networks provides evolutionary insights.
PM Kim, LJ Lu, Y Xia, **M Gerstein** (2006). *Science* 314: 1938-41.

-# A Bayesian networks approach for predicting protein-protein interactions from genomic data.
R Jansen, H Yu, D Greenbaum, Y Kluger, NJ Krogan, S Chung, A Emili, M Snyder, JF Greenblatt, **M Gerstein** (2003). *Science* 302: 449-53.

(2) Applied concepts from basic human genome annotation to disease genomics and showed how this can enable successful machine-learning models of variant-effect prediction

Contribution: The Gerstein lab has performed several studies that harness insights from basic genome annotation to interpret genomes of diseased individuals, particularly in the context of cancer and psychiatric illness. Related to cancer, the lab developed a number of variant interpretation tools (under the "FunSeq" umbrella) that have focused mostly on the annotation of non-coding variants. In this context, the lab has leveraged ENCODE annotations as well as many of the cross-species and population genetics conservation metrics developed by 1,000 Genomes and other projects. In addition, the lab built an integrative model that predicts the risk of

neuropsychiatric disease considerably better than existing polygenetic risk scores by incorporating functional genomics data. This was instantiated as an interpretable deep learning model by embedding a representation of the actual gene regulatory network into the neural network architecture.

Impact: The tools developed by the Gerstein lab – including methods to prioritize variants in a somatic context and methods that focus on particular types of annotations – are widely used and address practical real-world problems relevant to precision medicine. Moreover, the interpretable deep learning model highlighted new genes and pathways associated with schizophrenia.

-# Comprehensive functional genomic resource and integrative model for the human brain.
D Wang, S Liu, J Warrell, H Won, X Shi, FCP Navarro, D Clarke, M Gu, P Emani, YT Yang, M Xu, MJ Gandal, S Lou, J Zhang, JJ Park, C Yan, SK Rhie, K Manakongtreecheep, H Zhou, A Nathan, M Peters, E Mattei, D Fitzgerald, T Brunetti, J Moore, Y Jiang, K Girdhar, GE Hoffman, S Kalayci, ZH Gumus, GE Crawford, PsychENCODE Consortium, P Roussos, S Akbarian, AE Jaffe, KP White, Z Weng, N Sestan, DH Geschwind, JA Knowles, **M Gerstein** (2018). *Science* 362: eaat8464.

-- Integrative annotation of variants from 1092 humans: application to cancer genomics.
E Khurana, Y Fu, V Colonna, XJ Mu, HM Kang, T Lappalainen, A Sboner, L Lochovsky, J Chen, A Harmanci, J Das, A Abyzov, S Balasubramanian, K Beal, D Chakravarty, D Challis, Y Chen, D Clarke, L Clarke, F Cunningham, US Evani, P Flicek, R Fragoza, E Garrison, R Gibbs, ZH Gumus, J Herrero, N Kitabayashi, Y Kong, K Lage, V Liluashvili, SM Lipkin, DG MacArthur, G Marth, D Muzny, TH Pers, GRS Ritchie, JA Rosenfeld, C Sisu, X Wei, M Wilson, Y Xue, F Yu, 1000 Genomes Project Consortium, ET Dermitzakis, H Yu, MA Rubin, C Tyler-Smith, **M Gerstein** (2013). *Science* 342: 1235587.

(3) Developed computational models of the chromatin in non-coding regions and related this to gene expression

Contribution: The Gerstein lab developed models and tools relating chromatin structure and gene expression. Most notably, the lab developed a number of early machine-learning models that can accurately predict the expression of genes through their upstream chromatin. Finally, the lab developed several additional tools to characterize the chromatin landscape at different length scales.

Impact: These computational models transfer well between contexts, in particular between different organisms, highlighting the deep conservation of gene regulation between organisms.

-# Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project.
M Gerstein, ZJ Lu, EL Van Nostrand, C Cheng, BI Arshinoff, T Liu, KY Yip, R Robilotto, A Rechtsteiner, K Ikegami, P Alves, A Chateigner, M Perry, M Morris, RK Auerbach, X Feng, J Leng, A Vielle, W Niu, K Rhrissorrakrai, A Agarwal, RP Alexander, G Barber, CM Brdlik, J Brennan, JJ Brouillet, A Carr, MS Cheung, H Clawson, S Contrino, LO Dannenberg, AF Dernburg, A Desai, L Dick, AC Dose, J Du, T Egelhofer, S Ercan, G Euskirchen, B Ewing, EA Feingold, R Gassmann, PJ Good, P Green, F Gullier, M Gutwein, MS Guyer, L Habegger, T Han, JG Henikoff, SR Henz, A Hinrichs, H Holster, T Hyman, AL Iniguez, J Janette, M Jensen, M Kato, WJ Kent, E Kephart, V Khivansara, E Khurana, JK Kim, P Kolasinska-Zwierz, EC Lai, I Latorre, A Leahey, S Lewis, P Lloyd, L Lochovsky, RF Lowdon, Y Lubling, R Lyne, M MacCoss, SD Mackowiak, M Mangone, S McKay, D Mecenas, G Merrihew, DM Miller, A Muroyama, JI Murray, SL Ooi, H

Pham, T Phippen, EA Preston, N Rajewsky, G Ratsch, H Rosenbaum, J Rozowsky, K Rutherford, P Ruzanov, M Sarov, R Sasidharan, A Sboner, P Scheid, E Segal, H Shin, C Shou, FJ Slack, C Slightam, R Smith, WC Spencer, EO Stinson, S Taing, T Takasaki, D Vafeados, K Voronina, G Wang, NL Washington, CM Whittle, B Wu, KK Yan, G Zeller, Z Zha, M Zhong, X Zhou, modENCODE Consortium, J Ahringer, S Strome, KC Gunsalus, G Micklem, XS Liu, V Reinke, SK Kim, LW Hillier, S Henikoff, F Piano, M Snyder, L Stein, JD Lieb, RH Waterston (2010). *Science* 330: 1775-87.

-= Comparative analysis of the transcriptome across distant species.
M Gerstein, J Rozowsky, KK Yan, D Wang, C Cheng, JB Brown, CA Davis, L Hillier, C Sisu, JJ Li, B Pei, AO Harmanci, MO Duff, S Djebali, RP Alexander, BH Alver, R Auerbach, K Bell, PJ Bickel, ME Boeck, NP Boley, BW Booth, L Cherbas, P Cherbas, C Di, A Dobin, J Drenkow, B Ewing, G Fang, M Fastuca, EA Feingold, A Frankish, G Gao, PJ Good, R Guigo, A Hammonds, J Harrow, RA Hoskins, C Howald, L Hu, H Huang, TJ Hubbard, C Huynh, S Jha, D Kasper, M Kato, TC Kaufman, RR Kitchen, E Ladewig, J Lagarde, E Lai, J Leng, Z Lu, M MacCoss, G May, R McWhirter, G Merrihew, DM Miller, A Mortazavi, R Murad, B Oliver, S Olson, PJ Park, MJ Pazin, N Perrimon, D Pervouchine, V Reinke, A Reymond, G Robinson, A Samsonova, GI Saunders, F Schlesinger, A Sethi, FJ Slack, WC Spencer, MH Stoiber, P Strasbourger, A Tanzer, OA Thompson, KH Wan, G Wang, H Wang, KL Watkins, J Wen, K Wen, C Xue, L Yang, K Yip, C Zaleski, Y Zhang, H Zheng, SE Brenner, BR Graveley, SE Celniker, TR Gingeras, R Waterston (2014). *Nature* 512: 445-8.

(4) Comprehensively identified pseudogenes, highlighting their somewhat fuzzy demarcation from genes

Contribution: The Gerstein lab has worked in annotating the dead genes and genetic elements in the non-coding genome, both in the human and in other organisms. These elements give a sense of the history of the genome, as well as its function. They also highlight the complexities in defining exactly what a functioning gene is.

Impact: The lab was the first to perform large-scale surveys, finding almost as many pseudogenes as genes in many mammalian genomes – including the human genome – with many having interesting relationships to the parent genes from which they derive. The lab has shown that a number of pseudogenes have hints of regulatory activity.

-- What is a gene, post-ENCODE? History and updated definition.
M Gerstein, C Bruce, JS Rozowsky, D Zheng, J Du, JO Korbel, O Emanuelsson, ZD Zhang, S Weissman, M Snyder (2007). *Genome Res* 17: 669-81.

-= Comparative analysis of pseudogenes across three phyla.
C Sisu, B Pei, J Leng, A Frankish, Y Zhang, S Balasubramanian, R Harte, D Wang, M Rutenberg-Schoenberg, W Clark, M Diekhans, J Rozowsky, T Hubbard, J Harrow, **M Gerstein** (2014). *Proc Natl Acad Sci U S A* 111: 13361-6.

(5) Explored the large-scale structure of functional genomic data, illustrating non-obvious privacy leaks

Contribution: One of the achievements of the Gerstein lab has been in the area of data science and genomic privacy. A big issue with high-dimensional data is the creation of cryptic quasi-identifiers

and the leakage of private information. The lab has shown that functional genomics potentially have serious and subtle privacy leaks.

Impact: Tools developed by the Gerstein lab provide ways to remove private variants while allowing the data to be shared. In addition, the lab has written influential conceptual and thought pieces about genomic privacy, genomics as a data science discipline, and the ever-increasing scale of genomic and biomedical data.

-- Quantification of private information leakage from phenotype-genotype data: linking attacks.
A Harmanci, **M Gerstein** (2016). *Nat Methods* 13: 251-6.

(6) Developed a database classification system for macromolecular flexibility and used it to explain aspects of motions and in terms of simple packing geometry

Contribution: Work in the Gerstein lab has focused on classifying and characterizing macromolecular flexibility both from a physical and informatics perspective. The lab pioneered the use of databases to collect and classify these motions and explain them in terms of simple packing rules. In addition, the lab showed how these simple rules can lead to useful variant interpretation frameworks for coding regions.

Impact: The lab has developed many physically based variant impact tools focused on various aspects of structures – such as their level of frustration or their contact communities – to assess the overall variant impact on coding regions.

-- MolMovDB: analysis and visualization of conformational change and structural flexibility.
N Echols, D Milburn, **M Gerstein** (2003). *Nucleic Acids Res* 31: 478-82.

** Ranking of 12 publications listed

-# : 3 most key

-- : 7 key

-= : 2 additional