**C. Contribution to Science**

**Human Genome Annotation & Interpretation of Variants**

The Gerstein lab has made a number of contributions to developing large-scale human genome annotation, ranging from noncoding RNAs to enhancers to pseudogenes, and using these annotations to interpret variants in personal genomes in a functional context. Our tools address both germline and somatic variants. Our interpretation scheme ranks these variants in relation to their deleteriousness in causing disease, and also interprets potential functional effects.

Y Fu, Z Liu, S Lou, J Bedford, X Mu, KY Yip, E Khurana, **M Gerstein** (2014). "FunSeq2: A framework for prioritizing noncoding regulatory variants in cancer." *Genome Biol* 15: 480. [PMC4203974]

E Khurana, Y Fu, V Colonna, XJ Mu... (42 authors)... H Yu, MA Rubin, C Tyler-Smith, **M Gerstein** (2013). "Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics." *Science* 342: 1235587 [PMC3947637].

E Khurana, Y Fu, J Chen, **M Gerstein** (2013). "Interpretation of genomic variants using a unified biological network approach." *PLoS Comput Biol* 9: e1002886. [PMC3591262]

E Khurana, Y Fu, D Chakravaty, F Demichelis, MA Rubin, **M Gerstein** (2016). “Role of non-coding sequence variants in cancer.” *Nat Rev Genet.* 2:93-98. [PMID26781813]

**Personal Genomics & Privacy**

The unique character of each individual’s genome has potential impacts ranging from disease propensity to physical appearance to intelligence. We have developed tools to build personal genomes from DNA-sequencing data and to link molecular phenotypes such as gene expression to differences in parental alleles. We have also developed tools to address the critical question of whether it is possible to share molecular data without compromising the identities or the highly personal genetic information of sample donors.

J Rozowsky, A Abyzov, J Wang, P Alves, D Raha, A Harmanci, J Leng, R Bjornson, Y Kong, N Kitabayashi, N Bhardwaj, M Rubin, M Snyder, **M Gerstein** (2011). "AlleleSeq: analysis of allele-specific expression and binding in a network framework." *Mol Syst Biol* 7: 522. [PMC3208341]

L Habegger, A Sboner, TA Gianoulis, J Rozowsky, A Agarwal, M Snyder, **M Gerstein** (2011). RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries. *Bioinformatics* 27: 281. [PMC3018817]

D Greenbaum, A Sboner, X J Mu, **M Gerstein** (2011). “Genomics and Privacy: Implications of the New Reality of Closed Data for the Field” *PLoS Comput Biol* 7: e1002278 [PMC3228779]

A Harmanci, **M Gerstein** (2016). “Quantification of private information leakage from phenotype-genotype data: linking attacks.” *Nat Methods* 13:251. [PMID26828419]

**Comparative & Integrative Genomics**

We have developed a number of approaches for comparing the human genome to the genomes of model organisms. Our comparative analyses, particularly for the transcriptome, have yielded conserved principles of regulation. We have also developed integrative models that relate the transcriptome to the epigenome, and for combining these together to improve regulatory region annotation.

**M Gerstein**, J Rozowsky, KK Yan, D Wang...(89 authors)... TR Gingeras, R Waterston (2014). "Comparative analysis of the transcriptome across distant species." *Nature*512: 445. [PMC4155737]

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). "Comparative analysis of pseudogenes across three phyla." *PNAS* 111: 13361. [PMC4169933]

KK Yan, D Wang, J Rozowsky, H Zheng, C Cheng, **M Gerstein** (2014). "OrthoClust: an orthology-based network framework for clustering data across multiple species." *Genome Biology* 15:R100 [PMC4289247]

**M Gerstein**, ZJ Lu... (128 authors)... L Stein, JD Lieb, RH Waterston (2010). "Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project." *Science* 330: 1775. [PMC3142569].

**Analysis of Diverse Networks**

Network representations can be applied consistently to many different types of biological data. We have developed tools to build and analyze regulatory networks, protein-protein interactions and metabolic pathways, identifying key nodes such as hubs and bottlenecks. Moreover, we have integrated networks with dynamic gene-expression data (identifying transient hubs), 3D-protein structures, and other regulatory data to find large-scale regulatory principles for biological systems.

**M Gerstein**, A Kundaje... (50 authors)... R Myers, S Weissman, M Snyder (2012). " Architecture of the human regulatory network derived from ENCODE data." *Nature* 489: 91 [PMC4154057]

D Wang, KK Yan, C Sisu, C Cheng, J Rozowsky, W Meyerson, **M Gerstein** (2015). “Loregic: a method to characterize the cooperative logic of regulatory factors.” *PLoS Comput Biol* 11: e1004132. [PMC4401777]

PM Kim, LJ Lu, Y Xia, **M Gerstein** (2006). “Relating three-dimensional structures to protein networks provides evolutionary insights.” *Science* 314:1938-41. [PMID17185604]

C Cheng, E Andrews, KK Yan, M Ung, D Wang, **M Gerstein** (2015). “An approach for determining and measuring network hierarchy applied to comparing the phosphorylome and the regulome.” *Genome Biol* 16: 63. [PMC4404648]

**Tools for Processing Next-Gen Sequencing Data**

Next-gen sequencing has been one of the most exciting advances in the biological sciences, producing data on an unprecedented scale. This has given rise to the need to create new tool sets that can process very large-scale data very efficiently. We have developed tool sets that address a wide range of biological problems from sequencing data, including calling structural genetic variants and annotating specific regions of biological activity.

KY Yip, C Cheng, N Bhardwaj, JB Brown, J Leng, A Kundaje, J Rozowsky, E Birney, P Bickel, M Snyder, **M Gerstein** (2012). "Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors." *Genome Biol* 13: R48. [PMC3491392]

A Abyzov, AE Urban, M Snyder, **M Gerstein** (2011). "CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing." *Genome Res* 21: 974-84. [PMC3106330]

A Harmanci, J Rozowsky, **M Gerstein** (2014). "MUSIC: Identification of Enriched Regions in ChIP-Seq Experiments using a Mappability-Corrected Multiscale Signal Processing Framework." *Genome Biol* 15: 474. [PMC4234855]

A Abyzov, R Iskow, O Gokcumen, DW Radke, S Balasubramanian, B Pei, L Habegger, The 1000 Genomes Project Consortium, C Lee, **M Gerstein** (2013). "Analysis of variable retroduplications in human populations suggests coupling of retrotransposition to cell division." *Genome Res* 23: 2042 [PMC3847774]

**Complete List of Publications**

http://www.ncbi.nlm.nih.gov/sites/myncbi/mark.gerstein.1/bibliography/44005333/public