BIOGRAPHICAL SKETCH

Provide the following information for the Senior/key personnel and other significant contributors.
Follow this format for each person. DO NOT EXCEED FIVE PAGES.

NAME: Mark Gerstein

eRA COMMONS USER NAME (credential, e.g., agency login): MGERSTEIN

POSITION TITLE: Albert L. Williams Professor of Biomedical Informatics

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable. Add/delete rows as necessary.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	MM/YY	FIELD OF STUDY
Harvard College	AB	1989	Physics
Cambridge University	PhD	1993	Bioinformatics/Chemistry
Stanford University	post-doc	1993-1996	Bioinformatics

A. Personal Statement

This proposal involves work in bioinformatics and computational genomics. Prof Gerstein is a leader in these fields and thus well-suited to be part of this the proposal. He has many peer-reviewed publications (as of '19, >575 total with an H-index via Google scholar of >160). He has served as the "analysis" lead on many NIH-funded projects (e.g. ENCODE, psychENCODE & 1000 Genomes). Most recently, he has developed quantitative approaches and practical tools for the processing of next-generation sequencing data, including those related to chIP-seq, RNA-seq and the detection of DNA structural variation. He has also developed data-science approaches for analyzing molecular networks, evaluating genomic privacy, and performing integrative data mining across a wide variety of data contexts. 4 recent publications of note include:

B Wang, C Yan, S Lou, P Emani, B Li, M Xu, X Kong, W Meyerson, YT Yang, D Lee, **M Gerstein** (2019). "Building a Hybrid Physical-Statistical Classifier for Predicting the Effect of Variants Related to Protein-Drug Interactions." Structure 27: 1 [PMID: 31279629]

KK Yan, S Lou, **M Gerstein** (2017). "MrTADFinder: A network modularity based approach to identify topologically associating domains in multiple resolutions." PLoS Comput Biol 13: e1005647. [PMC5546724]

J Rozowsky, RR Kitchen... (6 authors)..., A Milosavljevic, **M Gerstein** (2019). "exceRpt: A Comprehensive Analytic Platform for Extracellular RNA Profiling." Cell Syst 8: 352. [PMID: 30956140]

FC Navarro, J Hoops, L Bellfy, E Cerveira, Q Zhu, C Zhang, C Lee, **M Gerstein** (2019). "TeXP: Deconvolving the effects of pervasive and autonomous transcription of transposable elements." PLoS Comput Biol 15: e1007293. [PMID: 31425522]

B. Positions and Honors

Positions and Employment

1997- Prof. Molecular Biophysics & Biochemistry, Yale (asst., '97-'01; assoc '01-'06)
1999- Prof. of Computer Science, Yale (asst., '99-'01; assoc. '01-'06)
2002- co-director Yale Computational Biology & Bioinformatics Program
2006- Williams Prof. Biomedical Informatics, Yale
2017- co-director Yale Center of Biomedical Data Science
2017- Prof. of Statistics & Data Science, Yale

Honors

1989-1993 Herchel-Smith Scholarship funding for PhD at Cambridge
1993-1996 Damon Runyon-Walter Winchell post-doctoral Fellowship
1997-2001 Young Investigator Awards from Navy & IBM, PhRMA, Donaghue, & Keck foundations
2009 AAAS Fellow
2015 ISCB Fellow

Other Experience and Professional Memberships
Editorial boards: Genome Res., PLoS Comp Bio, GenomeBiology, Fac. of 1000 (Big Data & Analytics)
Analysis Working Group co-chair: modENCODE ('07-'14), exRNA Consortium ('13-), 1000 Genomes Functional Interpretation Group ('10-'15), PsychENCODE Consortium (’14-), Pan-Cancer Analysis Working Group #2 (PCAWG regulatory drivers) ('14-), ENCODE (’17-)

C. Contribution to Science

Genome Annotation & Variant Interpretation

We have made significant efforts to annotate the human genome, be it through active participation in worldwide collaborations including ENCODE, modENCODE, Gencode and the 1000 Genomes, or through the development of computational analyses. Our work targets coding and noncoding genomic regions and ranks somatic and germline variants in relation to their potential functional impact. Additionally, we contributed to the annotation of noncoding RNA and pseudogenes.

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]

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]

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].

B Pei, C Sisu, A Frankish... (11 authors)... **M Gerstein** (2012). "The GENCODE pseudogene resource." *Genome Biol* 13: R51. [PMC6323946]

Personal Genomics & Privacy

All human beings share the vast majority of the genome, yet only a small fraction of each individual’s genome sequence shapes her/his unique combination of traits. We have developed tools that study personal genomics and link molecular phenotypes such as gene expression to differences in parental alleles. To address challenges stemming from the uniqueness of each individual’s “genomic footprint,” we also developed tools to assess the feasibility of sharing molecular data without jeopardizing the privacy of sample donors.

J Chen, J Rozowsky, T Galeev, A Harmanci, R Kitchen, J Bedford, A Abyzov, Y Kong, L Regan, **M Gerstein** (2016). "A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals." *Nat Commun* 7: 11101 [PMC4837449]

A Harmanci, **M Gerstein** (2018). "Analysis of sensitive information leakage in functional genomics signal profiles through genomic deletions" *Nat Commun* 9: 2453. [PMC6015012]

A Harmanci, **M Gerstein** (2016). “Quantification of private information leakage from phenotype-genotype data: linking attacks.” *Nat Methods* 13:251. [PMC4834871]

Rozowsky J, Abyzov A... **M Gerstein** (2011). "AlleleSeq: analysis of allele-specific expression and binding in a network framework." *Mol. Syst. Bio.* 2011 7:522. [PMC3208341]

Disease Genomics (Neurogenomics & Cancer)

The declining cost of next-gen sequencing has greatly impacted the study of genomic contributions to disease. In the Gerstein Lab, we have contributed in this direction through comprehensive studies and computational tools that aim to establish connections between genomic variation and disease development. We have studied a considerable number of diseases with a focus on cancers and brain disorders. In tandem, we co-led an effort to establish a comprehensive functional genomic resource for the human brain.

D Wang, S Liu, J Warrell, H Won, ...(34 named authors)... D Geschwind, J Knowles, **M Gerstein** (2018). "Comprehensive functional genomic resource and integrative model for the human brain." *Science* 362: eaat8464. [PMC6413328]

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].

S Li, BM Shuch, **M Gerstein** (2017). "Whole-genome analysis of papillary kidney cancer finds significant noncoding alterations." *PLoS Genet* 13: e1006685. [PMC5391127]

Data Science & Biological Networks

The increase in biomedical data during the last two decades engendered the need for computational tools that can process large-scale datasets and study a variety of data representations. We have developed tools to build and analyze multi-omics data sets, regulatory networks, protein-protein interactions and metabolic pathways, identifying key nodes, such as hubs and bottlenecks. We have also integrated networks with dynamic gene-expression data, 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]

K Yan, G Fang, N Bhardwaj, R Alexander, **M Gerstein** (2010). "Comparing genomes to computer operating systems in terms of the topology and evolution of their regulatory control networks." *PNAS* 107: 9186. [PMC2889091]

Machine Learning & Informatics Tools

Significant advances in Artificial Intelligence and supercomputing paralleled a still-ongoing rapid growth in biomedical data generation. In machine learning, a branch of AI that integrates algorithmic and statistical techniques, we have developed a number of tools to perform predictive tasks that provide insights on the genome. Our tools are able to process large-scale datasets to find genomic variants, predict protein binding, and annotate the human genome among other tasks.

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]

PP Kuksa, MR Min, R Dugar, **M Gerstein** (2015). High-order neural networks and kernel methods for peptide-MHC binding prediction. Bioinformatics 31: 3600-7. [PMID26206306]

J Rozowsky, G Euskirchen, RK Auerbach, ZD Zhang, T Gibson, R Bjornson, N Carriero, M Snyder, **M Gerstein** (2009). "PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls." Nat Biotechnol 27: 66-75. [PMC2924752]

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]

Complete List of Publications - http://www.ncbi.nlm.nih.gov/sites/myncbi/mark.gerstein.1/bibliography/44005333/public