## 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, Cambridge, MA | AB | 06/1989 | Physics |
| Cambridge University, Cambridge, UK | PhD | 05/1993 | Bioinformatics/Chemistry |
| Stanford University, Palo Alto, CA | post-doc | 09/1996 | 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 '16, >500 total with an H-index via Google scholar of >135). He has served as the "computational" lead on many previous NIH-funded projects (e.g. ENCODE & 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 approaches to analyze molecular networks and perform integrative data mining in a wide variety of contexts.


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]

C Cheng, E Andrews, K 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]

L Lochovsky, J Zhang, Y Fu, E Khurana, **M Gerstein** (2015). "LARVA: an integrative framework for large-scale analysis of recurrent variants in noncoding annotations." *NAR* 43: 8123. [PMC4787796]

A Abyzov, S Li, D Kim, M Mohiyuddin, A Stutz, N Parrish, X Mu, W Clark, K Chen, M Hurles, JO Korbel, H Lam, C Lee, **M Gerstein** (2015). "Analysis of deletion breakpoints from 1,092 humans reveals details of mutation mechanisms." *Nat Commun* 6: 7256. [PMC4451611]


### B. Positions and Honors

**Positions and Employment**

2006- AL Williams Prof. Biomedical Informatics, Yale
2002- Co-Director Yale Computational Biology and Bioinformatics Program
1999- Prof. of Computer Science, Yale (asst., '99-'01; assoc. '01-'06)
1997- Prof. Molecular Biophysics & Biochemistry, Yale (asst., '97-'01; assoc '01-'06)

**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., MSB, J Struc Func Gen., PLoS Comp Bio, GenomeBiology
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 (regulatory drivers)('14-)

**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. [PMC4834871]

**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]

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]

**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

**D. Research Support**

**Ongoing Research Support**

U01 MH103365-01 - 06/15/14-05/31/17 - NIH
Gene regulatory elements and transcriptome in iPSCs and embryonic human cortex
- Role: Multi-PI (PIs: Vaccarino, Gerstein, Weissman) This application is in response to FOA [MH-14-020 ("PsychENCODE")]. The major aims are to provide a comprehensive catalogue of all types of RNA and their regulatory elements in the cerebral cortex of the mid-gestational human fetal brain as compared to induced pluripotent stem cells (iPSCs) derived from the same fetuses.

R01 DA030976-05 - 09/30/10-05/31/16 - University of North Carolina - NIH (NCE)
Deep sequencing studies for cannabis and stimulant dependence
- Role: Co-I (PI: Wilhelmsen) The major goals of this project are to determine structural variants in the genome from deep sequencing studies for cannabis and stimulant dependence.

DE-AC02-98CH10886 - 10/17/11-07/31/16 - Brookhaven National Laboratory - DOE
Kbase: An Integrated Knowledgebase for Predictive Biology and Environmental Research
- Role: Co-PI (PI: Maslov) The major goal of this project is to assist in the construction of the DOE Knowledgebase. Our role is to provide support to the plant and microbial subcomponents.

U41 HG007000-03 - 09/21/12-07/31/16 - University of Massachusetts - NIH
EDAC: ENCODE Data Analysis Center
- Role: Co-I (PI: Weng) The major goal of this project is to perform global and integrative data analysis for the ENCODE project.

U41 HG007355-02 - 09/20/13-07/31/17 - University of Washington - NIH
Creating Comprehensive Maps of Worm and Fly Transcription Factor Binding Sites
- Role: Co-I (PI: Waterston) Our role on the project is the determination of binding sites for transcription factors in worm and the fly. We will analyze large-scale Chip-Seq experiments to identify regions in the genome that are bound by transcription factors.

U43 DA036134-01 - 08/01/13-07/31/18 - Baylor College of Medicine - NIH
Data Management and Resource Repository for the exRNA Atlas
- Role: Multi-PI (PIs: Gerstein, Galas, Milosavljevic) Our role on the project is administering the DIAC (data integration and analysis center) for ex-RNA data.

U41 HG007234-02 - 04/01/13-03/31/17 - Wellcome Trust - NIH
GENCODE: Comprehensive gene annotation for human and mouse
- Role: Co-I (PI: Harrow) Our role in the project is to identify pseudogenes comprehensively in human and mouse genomes and provide a systematic annotation of them.

U41 HG007497-02 - 09/20/13-08/31/16 - Jackson Laboratory - NIH
An Integrative Analysis of Structural Variation for the 1000 Genomes Project
- Role: Co-I (PI: Lee) Our role on the project is analyzing the 1000 genomes data set to determine structural variation on a large scale.

R01 GM108663-02 - 02/01/14-01/31/18 - NIH
Deciphering mechanisms governing functional partitioning of the C. elegans genome
- Role: Co-I (PI: Reinke) Our role on the project is assisting the PI with bioinformatic analyses related to the worm genome.

R01 MH100914-01A1 - 01/01/14-12/31/18 - NIH
Genomic mosaicism in developing human brain
- Role: Multi-PI (PIs: Vaccarino, Sestan, Gerstein) Our role on the project is to analyze somatic variations in the human genome.

U01 HL126495-01 - 08/01/14-04/30/19 - U Massachusetts - NIH
Racial and Ethnic Diversity in Human Extracellular RNA
- Role: Multi-PI (PIs: Freedman, Gerstein, Mukamal, O'Donnell) The primary goal of this proposal is the generation of exRNA profiles in healthy individuals in two large and well-defined cohorts, the Framingham Heart Study and the Multi-Ethnic Study of Atherosclerosis, to be used as a reference to facilitate disease diagnosis and discovery.

P50 MH106934-01 - 09/19/14-07/31/19 - NIH
Functional Genomics of Human Brain Development
- Role: Co-I (PI: Sestan) This grant will apply functional genomics to study human brain development. Our role is do analyses of these datasets.

3P50 MH106934-02S1 - 08/01/15 - 07/31/17 - NIH (concurrent with parent)
-Role: Co-I (PI: Sestan) Functional Genomics of Human Brain Development
Supplement to P50 MH106934 (above), effort on parent grant.

P30 DA018343-11A1 - 07/01/15-05/31/20 - Yale/NIDA Neuroproteomics Research Center - NIH
- Role: Co-I (PI: Williams) The overall grant is funding for a neuroproteomics center at Yale. The Gerstein lab contribution is to develop approaches for comparing and correlating protein abundance and mRNA levels in relation to neuroproteomics.

2UM1HG006504-05 - 01/14/16 - 11/30/19 - NIH
Yale Center for Mendelian Genomics
- Role: Multi PI (PIs: Lifton, Gerstein, Gunel, Mane) The majority of genomic variation in Mendelian disorders is due to variation in protein coding regions in the genome. Sequencing of these regions allow for rapid identification of disease causing mutations, which will allow us to understand and dissect the biology of these disorders leading to better diagnostic and therapeutic tools.

**Completed Research Support in the Last Three Years**

5U54 HG004558-05 - 01/01/10-06/30/13
Production Center for Global Mapping of Regulatory Elements
- Role: Co-I (PI: Snyder) The major goal of this project is to comprehensively probe transcription factor binding throughout the human genome.

5U54 HG004555-04S1 - 09/27/07-03/31/13
Integrated human genome annotation: generation of a reference gene set
- Role: Co-I (PI: Hubbard) The major goal of this project is to construct a pseudogene annotation of the human genome.

5U01 HG004267-05S1 - 01/01/10-03/31/13
Global Identification of Transcription Factor Binding Sites in *C. elegans*
- Role: Co-I (PI: Snyder) The major goal of this project is to build a genome-wide map of the binding sites for every *C. elegans* transcription factor.

5U01 HG005718-02 - 09/13/10-06/30/14 - NIH
Loss-of-function variants in the 1000 genomes data set and implications to GWAS
- Role: Co-I (PI: Zhao). The goals of this project are to survey loss-of-function (LOF) variants in the 1000 genomes data set and make this analysis available to the community as a useful resource.

5R01CA152057-03 - 08/01/11-07/31/15 - Weill Medical College of Cornell - NIH
Comprehensive Prostate Cancer Characterization by Genomic and Transcriptomic Profiling
- Role: Co-I (PI: Rubin) The major goal of this project is to identify biomarkers for prostate cancer through analysis of various types of 'omics data.

DE-SC0004856 - 08/15/10-08/14/15 - DOE
Tools and Models for Integrating Multiple Cellular Network
- Role: PI The major goals of this project are the development of tools for the analysis of network and pathways in micro-organisms for the Systems Biology Knowledgebase proposed by the DOE.

5U54HG006504-03 - 12/05/11-11/30/15 - NIH
Yale Center for Mendelian Disorders
- Role: Multi-PI (PIs: Gerstein, Lifton, Gunel, Mane) The major goal of this project is to develop informatics approaches to characterize rare variants in the framework of the Centers for Mendelian Genomics.