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: Kundaje, Anshul B.

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

POSITION TITLE: Assistant Professor of Genetics and Computer Science

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	Completion Date MM/YYYY	FIELD OF STUDY
VJTI, Mumbai University, Mumbai, India	B.E.	06/2001	Electrical Engineering
Columbia University, New York, NY, USA	M.S.	02/2003	Electrical Engineering
Columbia University, New York, NY, USA	Ph.D	10/2008	Computer Science
Stanford University, Stanford, CA, USA	Postdoc	01/2012	Computational Biology
Massachusetts Institute of Technology, Boston, MA, USA	Research Scientist	07/2013	Computational Biology

A. Personal Statement

My lab develops machine learning (ML) methods that integrate functional genomic and genetic data across diverse contexts to learn models of gene regulation and decipher regulatory genetic variation. We specialize in developing interpretable deep neural networks for integrative analysis of bulk and single-cell regulatory genomics data. I have led computational analysis efforts of the Encyclopedia of DNA Elements (ENCODE) consortium and the Roadmap Epigenomics Project. We developed probabilistic models and machine learning methods for deciphering comprehensive maps of cell-type specific regulatory elements, deconvolving sequence, structural and functional heterogeneity of elements, modeling three-dimensional long-range regulatory interactions, learning predictive regulatory network models, modeling the impact of natural genetic variation on the epigenome and predicting the downstream molecular effects of disease-associated genetic variation. We also have developed uniform processing processing and QC pipelines for a variety of functional genomic data. For the modENCODE and mouseENCODE projects, we developed integrative analysis methods to understand conservation and divergence of regulatory chromatin state across worms, flies, mice and humans. For the Genomics of Gene Regulation (GGR U01) collaborative initiative, we developed machine learning methods to learn dynamic regulatory networks from differentiation time courses. We have leveraged ML models of gene regulation to dissect functional genetic variation in the context of rare and complex diseases and traits (IGVF Consortium) from large biobanks and genome sequencing projects spanning colorectal cancer (GECCO consortium), cardiometabolic (AMP-CMD, CZI Seed networks), neurodegenerative (ADSP consortium) and neuropsychiatric disorders (PsychENCODE). Finally, we have significant experience developing software and web portals for mining and visualization of large-scale regulatory genomics data.

B. Positions, Scientific Appointments and Honors

Positions

2003 – 2008 Research Assistant, Computational Biology Group, Computer Science, Columbia University
2003 – 2003 Research Software Engineer, Functional genomics and Systems Biology group, IBM T. J. Watson Research Center
2008 – 2012 Postdoctoral Research Associate, Computer Science, Stanford University
2010 – 2010 Consultant, DNAnexus
2012 – 2013 Research Scientist, Massachusetts Institute of Technology, Broad Institute of MIT & Harvard
2012 – 2013 Consultant, Silicon Valley Biosystems
2015 – 2017 Scientific Advisory Board, Deep Genomics Inc.
2015 – 2018 Scientific Advisory Board, Epinomics Inc.

2019 – 2020 Consultant, Biogen Inc.
2017 – 2020 Scientific Advisory Board, Freenome Inc.
2020 – 2021 Scientific Advisory Board, ImmunAI Inc.
2013 – present Assistant Professor, Dept. of Genetics, Dept. of Computer Science, Stanford University
2019 – present Scientific co-founder, Ravel Biotechnology Inc.
2021 – present Scientific advisory board, PatchBio Inc.
2021 – present Scientific advisory board, SerImmune Inc.
2021 – present Scientific advisory board, OpenTargets
2021 – present Consultant (Fellow), Illumina Inc.
2022 – present Scientific advisory board, TensorBio Inc.
2022 – present Scientific advisory board, AINovo Inc.

Honors

2001 Prof. P.R. Dandavate Memorial Award: Highest GPA in the Bachelor's Program
2001 D.D. & L.H. Prize: Consistent Academic Career in the Bachelor's Program
2001 M.B.P. Memorial Foundation Award: Highest GPA in final Year of Bachelors Program
2014 Alfred Sloan Foundation Research Fellowship
2016 NIH Director's New Innovator Award
2019 Human Genome Organization (HUGO) Chen Award of Excellence

C. Contribution to Science (* - co-first/equal contribution, + -co-corresponding)

1. **Computational methods for quality control and denoising of large-scale functional genomic data:** High-throughput experiments are riddled with various types of noise, artifacts and systematic biases and the first step to successful data integration is the effective filtering and normalization of data. As part of ENCODE and Roadmap, we have developed robust statistical pipelines for automated normalization, thresholding and quality control of 1000s of datasets. The methods we have developed are a key part of ENCODE's ChIP-seq data standards. We have used these methods to evaluate all publicly available ChIP-seq data in GEO and found extensive heterogeneity and suggested key areas for improvement of data standards. We have developed deep learning approaches to denoise ChIP-seq data and diffusion-based methods to evaluate the reproducibility of chromosome conformation capture data.
    a. Landt SG*, Marinov GK*, **Kundaje A***, Kheradpour P*, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, Chen Y, DeSalvo G, Epstein C, Fisher-Aylor KI, Euskirchen G, Gerstein M, Gertz J, Hartemink AJ, Hoffman MM, Iyer VR, Jung YL, Karmakar S, Kellis M, Kharchenko PV, Li Q, Liu T, Liu XS, Ma L, Milosavljevic A, Myers RM, Park PJ, Pazin MJ, Perry MD, Raha D, Reddy TE, Rozowsky J, Shoresh N, Sidow A, Slattery M, Stamatoyannopoulos JA, Tolstorukov MY, White KP, Xi S, Farnham PJ, Lieb JD, Wold BJ, Snyder M. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res*. 2012 Sep;22(9):1813-31. PMID: 22955991
    b. Koh PW, Pierson E, **Kundaje A**. Denoising genome-wide histone ChIP-seq with convolutional neural networks Bioinformatics (2017) 33 (14): i225-i233. PMID: 28881977

2. **Comprehensive catalogs of putative regulatory elements across cell-types and species:** Different combinations of epigenomic marks (chromatin states) have been found to define different types of functional domains in the genome. Chromatin states of genomic domains are modified during cellular differentiation giving rise to different cell-types and these states are often disrupted in different diseases. We have trained multivariate hidden Markov models on 1000s of epigenomic datasets to learn a limited repertoire of hidden chromatin states and automatically segment the human genome into cell-type specific regions annotated with different chromatin state labels. These dynamic chromatin-state maps are not only revealing a staggering number of novel regulatory domains but are also allowing us to infer detailed similarities and differences of epigenomic regulation between the different cell-types. By correlating the dynamic transitions of chromatin state labels of regulatory elements with transcriptional activity of genes across cell-types, using novel probabilistic models, we have been able to infer long-range regulatory interactions between distal regulatory elements and their target genes. We are currently exploring interpretable deep learning approaches to reveal the regulatory sequence grammars underlying ~2.3 million dynamic regulatory elements discovered in the human genome. We have also analyzed conservation and divergence of chromatin state across distant species namely worm (*C. elegans*), fly (*D. melanogaster*) and human by integrating chromatin ChIP-seq data from the ENCODE and modENCODE

consortia. We have discovered that with a few exceptions (such as heterochromatin) all 3 species share a remarkable similarity of combinatorial chromatin states although there are significant differences in the distance distribution and sizes of chromatin domains across the species.

a. Roadmap Epigenomics Consortium, **Kundaje A***, Meuleman W, Ersnt J, Bilenky M, et al. Integrative analysis of 111 reference human epigenomes. Nature 518, 317–330 (19 February 2015) doi:10.1038/nature14248 PMID: 25693563
b. Dunham I*, **Kundaje A***, et al., ENCODE Project Consortium. An integrated Encyclopedia of DNA Elements in the human genome. *Nature*. 2012 Sep 6;489(7414):57-74. PMID: 22955616
c. Ho JW, Jung YL, Liu T, Alver BH, Lee S, Ikegami K, Sohn KA, Minoda A, Tolstorukov MY, Appert A, Parker SC, Gu T, **Kundaje A***, Riddle NC, et al. Comparative analysis of metazoan chromatin organization. *Nature*. 2014 Aug 28;512(7515):449-52. doi: 10.1038/nature13415 PMID: 25164756

3. **Interpretable machine learning approaches to decipher the regulatory code of the genome:** The functional effect of transcription factor binding is often determined by and correlated with co-association of other regulatory proteins and chromatin context. We developed a machine learning approach that used quantitative TF binding profiles to dissect the variability of co-association patterns of TFs within and between cell-types. Many novel associations were learned giving an unprecedented view of the complexity of regulatory grammars. We have developed deep neural networks to learn cis-regulatory sequence syntax encoded in regulatory DNA sequences associated with transcription factor binding and chromatin accessibility. We developed a novel feature attribution method called DeepLIFT for estimating the predictive importance of individual features (nucleotides or motifs) in any input DNA sequence to its associated regulatory activity. We developed a novel Fourier based attribution prior to stabilize attribution scores from deep learning models of regulatory DNA sequence. We also developed Deep Feature Interaction Maps (DFIM), a new method to efficiently estimate interactions between all pairs of features in any input DNA sequence.
    a. Gerstein MB*, **Kundaje A***, Hariharan M, Landt SG, et al. Architecture of the human regulatory network derived from ENCODE data. *Nature*. 2012 Sep 6;489(7414):91-100. PMID: 22955619
    b. Avsec Ž, Weilert M, Shrikumar A, Alexandari A, Krueger S, Dalal K, Fropf R, McAnany C, Gagneur J, **Kundaje A+**, Zeitlinger J+. Base-resolution models of transcription factor binding reveal soft motif syntax. Nat Genet. 2021 Feb 18 DOI: 10.1038/s41588-021-00782-6. PMID: 33603233
    c. Shrikumar A, Greenside P, **Kundaje A**. Learning Important Features Through Propagating Activation Differences. Proceedings of the 34th International Conference on Machine Learning (ICML), PMLR 70:3145-3153, 2017
    d. Tseng AM, Shrikumar A, **Kundaje A**. Fourier-transform-based attribution priors improve the interpretability and stability of deep learning models for genomics. Proceedings of the 2020 Advances In Neural Information Processing Systems (NeurIPS) Conference

4. **Deciphering regulatory impact of natural and disease-associated genetic variation:** We have developed statistical methods to understand the relationship between natural genetic variation and the regulatory variation across individuals from diverse populations. We have found extensive chromatin state variation especially at enhancer elements driven by variants largely affecting transcription factor binding. We have also developed new machine learning methods that train on bulk and single cell chromatin and expression profiling experiments in disease-relevant tissues to fine map and interpret regulatory impact of genetic variants associated with complex diseases. We have also been able to predict cell types and tissues that are likely to manifest the regulatory effects of GWAS variants from 100s of diseases and traits. We are now actively collaborating with several clinicians and GWAS consortia to uncover genetic and regulatory mechanisms underlying cardiometabolic, neurodegenerative, neuropsychiatric disease as well as colorectal cancer.
    a. Grubert F*, Zaugg J*, Kasowski M*, Ursu O*, Spacek DV, Greenside P, Srivas R, Martin A, Phanstiel D, Pekowska A, Heidari N, Euskirchen G, Huber W, Pritchard JP, Bustamante C, Steinmetz L, **Kundaje A**, and Snyder M. Genetic control of chromatin states in humans involves local and distal chromosomal interactions. *Cell*. 2015 Aug 19. pii: S0092-8674(15)00964-2. doi: 10.1016/j.cell.2015.07.048. PMID: 26300125
    b. Corces MR, Shcherbina A, Kundu S, Gloudemans MJ, Fresard L, Granja JM, Louie BH, Eulalio T, Shams S, Bagdatli ST, Mumbach MR, Liu B, Montine KS, Greenleaf WJ, **Kundaje A**, Montgomery SB, Chang HY, Montine TJ. Single-cell epigenomic analyses implicate candidate causal variants at

inherited risk loci for Alzheimer's and Parkinson's diseases. Nat Genet (2020). https://doi.org/10.1038/s41588-020-00721-x. PMID: 33106633)

c. Trevino AE, Muller F, Andersen J, Sundaram L, Kathiria A, Shcherbina A, Farh K, Chang HY, Pasca AM, **Kundaje A**, Pasca SP, Greenleaf WJ. Chromatin and gene-regulatory dynamics of the developing human cerebral cortex at single-cell resolution. Cell. 2021 Aug 11 DOI: 10.1016/j.cell.2021.07.039 (PMID: 34390642)

Complete List of Published Work:
http://www.ncbi.nlm.nih.gov/myncbi/browse/collection/42219758/?sort=date&direction=descending

D. Research Support

Ongoing Research Support

5R01HD094513 National Institutes of Health Molecular images and machine learning to extract placental function from maternal cfDNA Major Goals: Develop technology to utilize cfDNA as means to assess placental function non-invasively.	Baker (PI), Kundaje (MPI)	03/20/2018 – 02/28/2023
1U2CCA233311-01 National Institutes of Health Precancer Atlas of Familial Adenomatous Polyposis Major Goals: Develop machine learning methods to dissect regulatory models and causal mutations in cancer	Snyder (PI), Kundaje (co-I)	09/30/2018-06/30/2023
1U01MH11652901A1 National Institute of Health Integrated, cell type specific functional genomics analyses of regulatory sequence elements and their dynamic interaction networks in neuropsychiatric brain tissues Major Goals: Learning regulatory networks and fine mapping variants associated with psychiatric disorders from bulk and single cell multi-omic profiling of brain regions (PsychENCODE consortium)	Urban (PI), Kundaje (co-I)	06/20/2019 - 03/31/2024
UM1 DK126185-01 National Institutes of Health Bridging the gap between Type 2 Diabetes GWAS and therapeutic targets Major Goals: Develop machine learning methods to predict causal cis and trans regulatory networks in T2D	Gloyn (PI), Kundaje (co-I)	07/01/2020-06/30/2025
R01 HG011466-01 National Institutes of Health Methods for imputing regulatory genomic and 3D nucleome data across cell types, tissues and organisms Major Goals: Develop deep learning models to impute functional genomics data	Noble (PI), Kundaje (MPI)	07/01/2020-06/30/2024
U01AG072573 National Institutes of Health Multi-omic functional assessment of novel AD variants using high-throughput and single-cell technologies Major Goals: Develop machine learning approaches for fine mapping Alzheimer's associated variants	Montine (PI), Kundaje (MPI)	04/01/2021 - 03/31/2026
5R01HL13481704 National Institutes of Health Causal variant association mechanisms in TCF21 binding coronary disease loci Major Goals: Develop computational approaches to infer causal variants associated with coronary heart disease	Quertermous (PI), Kundaje (co-I)	01/01/2021-12/31/2025
1R01MH125244 National Institutes of Health Identifying causal genetic variants and molecular mechanisms impacting mental health Major Goals: Develop deep learning approaches to infer causal variants associated with mental health	Montgomery (PI), Kundaje (MPI)	03/31/2021-08/31/2025
SPO#215866 Milky Way Research Foundation Reprogramming of Aging Major Goals: Develop aging clocks for the brain based on epigenomic and transcriptome profiling data	Brunet (PI), Kundaje (co-I)	04/01/2021-03/31/2024
1U01HG011762-01 National Institutes of Health Stanford Mendelian Genomics Research Center Major Goals: Develop machine learning approaches for scoring mendelian disease mutations	Montgomery (PI), Kundaje (co-I)	08/01/2021-05/31/2025
2U24HG007234 National Institutes of Health GENCODE: comprehensive reference genome annotation for human and mouse	Flicek (PI), Kundaje (co-I)	08/01/2021-05/31/2025

Major Goals: Develop machine learning approaches for semi-automated genome annotation

U01HG012069 Kundaje (PI) 09/01/2021 - 06/30/2026
National Institutes of Health
Predicting context-specific molecular and phenotypic effects of genetic variation through the lens of the cis-regulatory code
Major Goals: Develop machine learning approaches to decode regulatory elements and regulatory variation from multi-omics assays

1UM1HG011972-01 Engreitz (PI), Kundaje (co-I) 09/01/2021 - 06/30/2026
National Institutes of Health
Stanford Center for Connecting DNA Variants to Function and Phenotype
Major Goals: Develop functional characterization assays and computational methods to decipher functional genomic variation

R01HG010140 Montgomery (PI), Kundaje (MPI) 09/01/2021 - 09/31/2022
National Institutes of Health
Software for large-scale inference of genetics of lifestyle measures, biomarkers and common and rare disease
Major Goals: Develop machine learning methods for analysis of summary statistic data from population biobanks, and disease-focused genome sequencing programs

Completed Research Support (past 3 years)

1UM1HG009436 Greenleaf (PI), Kundaje (co-I) 02/01/2017 – 01/31/2022
National Institutes of Health
High-throughput systematic characterization of regulatory element function
Major Goals: To develop high-throughput validation of functional genomic elements

5U01HG009431 Pritchard (PI), Kundaje (co-I) 02/01/2017 - 01/31/2022
National Institutes of Health
Decoding the regulatory architecture of the human genome across cell types, individuals and disease
Major Goals: Develop machine learning methods to decipher regulatory elements and variants

1U24HG009446 Weng (PI), Kundaje (co-I) 02/01/2017–01/31/2022
National Institutes of Health
EDAC: ENCODE Data Analysis Center
Major Goals: Develop data processing methods for the Encyclopedia of DNA Elements (ENCODE4)

5R01CA201407 Peters (PI), Kundaje (co-I) 09/23/2016 – 08/31/2021
The Fred Hutchinson Cancer Research Center
National Institutes of Health
Using Functional Genomics to Inform Gene Environment Interactions for Colorectal Cancer
Major Goals: Machine learning methods to infer functional gene-environment interactions

1DP2GM123485 Kundaje (PI) 09/30/2016-05/31/2021
National Institutes of Health
Deep Learning frameworks for regulatory genomics
Major Goals: To develop interpretable, integrative deep learning frameworks for genomics

1R01HG00967401 Kundaje (PI) 08/09/2017 - 12/30/2020
National Institutes of Health
Learning Regulatory Drivers of Chromatin and Expression Dynamics during Nuclear Reprogramming
Major Goals: Learn dynamic models of transcription regulation of cellular reprogramming.

2P01AG036695 Rando (PI), Kundaje (co-I) 06/01/2017 - 04/31/2022
National Institutes of Health
Molecular Regulation of Stem Cell Aging
Major Goals: Develop machine learning methods for integrative models of gene regulation in stem cell aging

CZI Seed Networks for Human Cell Atlas Quertermous (PI), Kundaje (MPI) 06/01/2019-05/31/2022
Chan Zuckerberg Initiative (Silicon Valley Community Foundation)
Single Cell Transcriptomic and Epigenomic features of the human vasculature
Major Goals: Mapping cis-trans regulatory networks of cell types in the human vasculature by integrating single cell RNA-seq and ATAC-seq data