Jianyu Yang

+1 8149968462 | yztxwd@gmail.com | yztxwd | Jianyu

Skills

Programming	Python, JAVA, R
Machine learning & Statistics	Scipy, Scikit-Learning, CuML. General machine learning skills, specialized in Bayesian method
Deep learning	Tensorflow, DALI, Pytorch, Webdataset, Pytorch-lightning, Huggingface, Captum, TF-Modisco CNN, Transformer style model design, training, and interpretation
Genomics	ChIP-seq, ATAC-seq, RNA-seq, MNase-seq, BS/TAB-seq analysis
Pipelines & Infra	Snakemake, HDF5, Singularity/Apptainer, Docker, Conda, Slurm, HPC clusters, Proxmox
Wetlab	Molecular Cloning, CRISPR-Cas9, Cell Culture

Experience

Interpret Genomics Deep Learning Models via Concept Attribution

Skills: Pytorch, Captum/DeepLiftSHAP, TF-Modisco, Scikit-Learn *Oct. 2024 - Now*

- Improved global concept attribution method Testing with Concept Activation Vector (TCAV) and adapted it to Genomics models
- Highly generalizable and flexible, proved reliability on large scale foundation models and various type of model inputs.

Nucleosome Calling with Bayesian Gaussian Mixture Models (SEM Algorithm)

Skills: Java, MNase-seq pipeline, Bayesian Methods, Cell Culture, Molecular Biology, CRISPR-Cas9 *Sep. 2017 - April. 2024*

- Constructed TET family mutant cell lines by CRISPR-Cas9.
- Designed and implemented the **Size-based Expectation Maximum (SEM) algorithm (github.com/YenLab/SEM) in Java** to classify nucleosome subtypes using expectation-maximization and Gaussian mixtures. Explored the properties of fragile nucleosome in mES.

Training and interpreting deep learning model for FOXA1 binding partner in A549

Skills: Pytorch, Pytorch-Lightning, Captum/DeepLiftSHAP, TF-Modisco *Jan. 2022 - Oct. 2023*

- Designed and Applied a bimodal neuron network to dissect the influence of sequence and chromatin on FOXA1 in A549. Applied attribution (DeepLIFT/SHAP) + motif discovery (TF-MoDISco) to reveal cooperativity between FOXA1 and AP-1

Multimodal Deep Learning for induced Fox Factor Binding Prediction

Skills: Tensorflow, Pytorch, Pytorch-Lightning, Captum/DeepLiftSHAP, TF-Modisco *Jan. 2022 - Sep. 2023*

- Built multi-modal CNN/Transformer-based neural networks that integrate sequence + chromatin features to predict induced Fox family factor binding. Interpreted the trained models to reveal motif and pre-existing chromatin preferences of Fox factors

Regulatory Network Analysis on RUNX1

Skills: R, DESeq, Snakemake *Sep. 2020 - April. 2021*

- Used automated Snakemake workflows to analyze differential expression and TF binding to identify RUNX1 downstream targets in leukemia cells. Integrated multi-omics datasets to find CENPE as a RUNX1-regulated gene affecting proliferation.

SOFTWARE & TOOLING PROJECTS

Seqchromloader (Training Data Toolkit for Genomic DL)

Skills: Python, Pytorch, Webdataset *github.com/seqcode/seqchromloader*

- Built a production-ready toolkit to construct training datasets for sequence/chromatin DL models. Optimized for high-throughput, distributed dataset streaming. Has been widely adopted by lab members.

HDF5-Backed Genome Coverage & Heatmap Engine

Related skills: Python, HDF5 *github.com/yztxwd/chiptoolkit*

- Developed a Deeptools-like plotting engine using HDF5 to pre-store genome-wide tracks, enabling extremely fast data retrieving and heatmap/composite plot generation for thousands of regions.

General Snakemake Pipelines for NGS Data

Skills: R, Python, Snakemake, Common packages used in NGS analysis, Slurm, HPC *github.com/yztxwd/snakemake-pipeline-general*

- Built modular snakemake pipelines for ATAC-seq, ChIP-seq, RNA-seq, BS-seq, and MNase-seq. Designed for reproducibility, portability, and HPC batch environments (Slurm). Adapted by lab members for routine preprocessing workflows.

Education

Ph.D., Bioinformatics and Genomics Program, Pennsylvania State University *Aug. 2020 - May. 2026 (Expected)*

M.S., Developmental Biology, Southern Medical University *Sep. 2017 - Jun. 2020*

B.S., Preclinical Medicine, Southern Medical University *Sep. 2012 - Jun. 2017*

Publications

Published research in **Molecular Cell, Genome Research, NAR Genomics & Bioinformatics, and Frontiers in Molecular Biosciences**, focusing on deep learning and statistical modeling for regulatory genomics, transcription factor binding, and chromatin architecture. Full list available on Google Scholar (scholar.google.com/citations?user=r7sRhzoAAAAJ)