Harpreet Singh

+1 (778) 682-2155 | 808harpreet@gmail.com | linkedin.com/in/hrrysprk | github.com/hrrysprk | hrrysprk.com

Professional Summary

Computational biologist with 4+ years of research experience and 3 peer-reviewed publications in npj Aging, BMC Genomics, and Genetics. Co-developed **DiSCO**, a published **Hi-C** normalization algorithm (BMC Genomics 2020), and developed novel chromatin accessibility algorithm (**CSAA**) deployed in production pipelines. Expertise spans single-cell and bulk multi-omics (**scRNA-seq**, **scATAC-seq**, **scHi-C**, **ATAC-seq**, **ChIP-seq**, **Hi-C**), cancer genomics via **TCGA**, 3D genome architecture, and cloud-native pipeline development on **AWS**. Wet lab background in epigenomics and chromatin biology informs computational analysis design. Currently completing an MDS at UBC.

Professional Experience

Graduate Teaching Assistant – Data Visualization | University of British Columbia, Vancouver, BC 2025 – 2026

- Hold twice-weekly office hours for 80+ students, debugging **D3.js** and **JavaScript** code, providing feedback on visualization design, and grading assignments

Bioinformatician (Research Associate) | IISER Mohali 2017 – 2020

- Co-developed **DiSCO** (Distance Sorted Contact Optimization), a novel **Hi-C** normalization algorithm correcting distance-dependent interaction frequency biases in condensed and decondensed chromatin domains across 45 Hi-C datasets spanning 6 protocol variants (in-situ, in-solution, single-cell, **DNase**, native, Drosophila **Hi-C**); published in **BMC Genomics** (2020) [link]
- Deployed **Tanay Lab scHi-C** pipeline on **SGE** and **Torque/PBS** HPC clusters using early-access Babraham Institute protocols; refactored a published cell cycle phase-scoring algorithm in **R/misha** enabling automated G1/S/G2/M classification of single cells based on chromatin interaction decay profiles; identified TADs and loops using **HOMER**, **TADbit**, insulation scores, and directionality indices
- Performed comparative genomics across 34 vertebrate genomes linking convergent chromosomal rearrangements at the 3p21.31 tumour-suppressor locus to cancer resistance; analyzed **TCGA** Pan-Cancer CNV and RNA-seq data to characterize co-deletion patterns and expression divergence near breakpoints; mapped germline and somatic **DNA breakpoints** from congenital disorder and cancer datasets to infer developmental tolerance of genomic rearrangements; applied multi-resolution **Hi-C** contact map analysis, **A/B compartment** calling, and insulation score domainograms; published in **Nature NPJ** (2021) [link]
- Built production-scale NGS pipelines using **Snakemake** processing **1+ TB** of multi-omics data (bulk and single-cell **RNA-seq**, **ATAC-seq**, **ChIP-seq**, **Hi-C**); full FASTQ-to-discovery stack including QC (**FastQC**, **BBDuk2**), alignment (**BWA**, **Bowtie2**, **SAMtools**), and variant processing (**VCFtools**); 4x throughput increase and 50% runtime reduction through HPC parallelization
- Generated virtual **4C** interaction profiles from **Hi-C** matrices using TSS as bait with Loess regression and 3-SD significance thresholds; integrated **BrainSpan** developmental RNA-seq, **ChromHMM** chromatin state annotations (ENCODE/Epigenome Roadmap), **H3K4me1 ChIP-seq**, **Repli-seq** replication timing, and **GWAS SNP** linkage disequilibrium data across 5 mammalian species; published in **Genetics** (2019) [link]
- Applied **DESeq2** differential expression, **GO/MPO** enrichment (**GREAT**, **ToppGene**), and **KEGG** pathway analysis; integrated DNA methylation (CpG) profiles with **Hi-C** for cell-cycle-specific epigenomic characterization; identified conserved non-coding elements using **PHAST**, **Ancora**, and **UCNEbase**; applied motif enrichment via **RSAT/JASPAR**
- Managed acquisition of multi-terabyte datasets from **NCBI SRA**, **GEO**, and **ENA** via **SRA Toolkit** and **Aspera Connect**; applied phylogenetic modeling (**BAMM**, Ornstein-Uhlenbeck, Pagel's lambda via **phytools**) across 312 rodent species; developed custom **Python** scraping utilities for automated annotation extraction from **FoldIndex**, **Ensembl**, and **GO** databases

Core Skills

Programming: Python, R, SQL, JavaScript, Bash, D3.js, Three.js, SvelteKit, Tailwind CSS

Bioinformatics: scRNA-seq, scATAC-seq, scHi-C, RNA-seq, ATAC-seq, ChIP-seq, Hi-C, ChIA-PET, 4C, variant calling, DESeq2, Seurat, Scanpy, CellTypist, GSEA, HiCUP, HOMER, TADbit, misha, BWA, Bowtie2, SAMtools, FastQC, BBDuk2, Bioconductor, FASTQ/BAM/VCF

Genomic Resources: TCGA, COSMIC, NCBI SRA, GEO, ENA, Ensembl, ENCODE, Epigenome Roadmap, 4DN Portal, WashU Epigenome Browser

ML & Statistics: scikit-learn, PyTorch, XGBoost, regression, clustering, PCA, ANOVA, hypothesis testing, Bayesian methods, bootstrap resampling

Cloud & Infrastructure: AWS (Batch, S3, ECR), Nextflow, Snakemake, Docker, Linux, SGE, Torque/PBS, HPC, CI/CD, Git, MLflow

AI-Assisted Dev: **Claude**: architecture, research, planning, code review | **Kiro IDE**: project specs, system design, complex multi-file context | **Cursor Agent**: structured task execution via `updates.md` queue with rollback | **GitHub Copilot**: CI/CD setup, boilerplate | **Gemini** & **Grok**: real-time research, trending papers, external source discovery | **Multi-model validation**: cross-referencing Claude, Grok, Gemini, ChatGPT to guardrail facts and design decisions | **Cross-model prompt engineering**: designing prompts in Claude optimized for deployment in Grok, Gemini, Cursor, and Kiro

Selected Projects

ChromApipe – Chromosome Accessibility Pipeline github.com/hrrysprk/chromapipe

- Engineered cloud-native **Nextflow** pipeline on **AWS Batch** analyzing 3D chromosome structures across all 22 autosomes using **Wave containers** and **Fusion file system** for direct **S3** access, eliminating file staging overhead
- Implemented novel **CSAA** algorithm for genome-wide chromatin accessibility quantification as an ATAC-seq alternative; parallel **Ensembl REST API** annotation fetching; outputs analysis-ready **Parquet** files via **Polars**

GenBrowser – 3D Chromosome Visualization hrrysprk.com/genBrowser

- Built interactive 3D dashboard mapping biological metrics (GC content, solvent accessibility, radial distance) onto chromosome 1 using **Three.js** and **Vite**; deployed via **GitHub Actions** to **GitHub Pages**
- Processed raw **PDB** structural data through a 4-stage **Python** pipeline using **NumPy**, **SciPy**, and **Polars** to compute per-residue surface accessibility and spatial metrics at chromosome scale

spaceGen – Single-Cell Multiome ML Pipeline github.com/hrrysprk/spaceGen

- Investigating how microgravity-induced changes in chromatin accessibility and gene expression converge to alter neuronal cell identity, linking 3D genome reorganization to transcriptional dysregulation in spaceflight-exposed brain tissue
- Building end-to-end **snRNA-seq** and **snATAC-seq** multiome pipeline on NASA OSDR spaceflight data (Rodent Research-3) using **Scanpy**, **AnnData**, **CellTypist**, and **MuData/Muon**; unsupervised clustering, supervised cell type classification, and gene regulatory network analysis with **MLflow** experiment tracking; developed using AI-assisted workflows with **Kiro IDE** and **Claude** for architecture design and iterative refinement

PolicyLens – LLM-Powered Course Policy QA github.com/tanav2202/PolicyLens

- Built full-stack QA application with custom **RAG architecture** using **FastAPI**, **React**, and **Ollama**; intent classification with word-by-word streaming; scraped and structured course policy data via **BeautifulSoup** with **pytest** test coverage

Publications

- **Convergent evolution of genomic rearrangement may explain cancer resistance in hystrico- and sciuromorpha rodents.** Jain Y, Chandradoss KR, Anjoom AV, Bhattacharya J, Lal M, Bagadia M, **Singh H**, Sandhu KS. – npj Aging, 2021
- **Biased visibility in Hi-C datasets marks dynamically regulated condensed and decondensed chromatin states genome-wide.** Chandradoss KR, Guthikonda PK, Kethavath S, Dass M, **Singh H**, Nayak R, Kurukuti S, Sandhu KS. – BMC Genomics, 2020
- **Evolutionary loss of genomic proximity to CNEs impacted gene expression dynamics during mammalian brain development.** Bagadia M, Chandradoss KR, Jain Y, **Singh H**, Lal M, Sandhu KS. – Genetics, 2019

Education

- **Master of Data Science** – University of British Columbia, expected May 2026
  Relevant coursework: Supervised Learning I & II, Unsupervised Learning, Feature and Model Selection, Statistical Inference I & II (Frequentist and Bayesian), Algorithms and Data Structures, Web and Cloud Computing, Databases and Data Retrieval
  Electives: Biostatistics, Genomics, Systems Biology, Graphs and Network Theory
- **BS-MS Biological Sciences** – IISER Mohali, 2017
  Specialization in computational biology and 3D genome organization; Government of India Research Grant recipient
  Thesis (1): Computational study of conserved non-coding elements and genomic position effects on gene regulation, replication timing, and histone marks across mammalian evolution (published in Genetics 2019)
  Thesis (2): Multivariate network analysis of social dynamics in academic communities using bipartite graphs, unsupervised **PCA** and **KNN** clustering, and Big Five personality data integrated with academic performance and demographic variables

Certifications & Awards

2023: Nextflow Workshop – Advanced Workflow Management & Pipeline Development

2023: Google Data Analytics Professional Certificate – Coursera

2019: Government of India Research Grant – 3D Genome Organization Study

2013–2017: DST-INSPIRE Fellowship – Top 1% national science students