10 relevant papers (importance and contribution)

1. Poonia S, Goel A, Chawla S, Bhattacharya N, Rai P, Lee YF, Yap YS, West J, Bhagat AA, Tayal J, Mehta A, Ahuja G, Majumdar A*, Ramalingam N*, **Sengupta D***. Marker-free characterization of full-length transcriptomes of single live circulating tumor cells. **Genome Res.** 2023 Jan;33(1):80-95. Doi: 10.1101/gr.276600.122. Epub 2022 Nov 22. PMID: 36414416; PMCID: PMC9977151.

   Circulating tumor cells (CTCs) harbor a wealth of molecular information about the tumor of origin. However they are rare in blood (1 or fewer cells in 105 - 106 PBMCs). This makes the detection and characterisation of CTCs challenging. Size an/or marker based CTC enrichment techniques might potentially miss out on CTCs of unusual phenotype or suffer from WBC contamination. We developed unCTC, a tailored computational framework for label free characterisation of single CTC transcriptomes, while avoiding WBC contaminants. As a remarkable achievement, our software could accurately recognise single circulating Triple Negative Breast Cancer cells (TNBCs). I conceived and led the study in consultation with Dr. Naveen Ramalingam (Senior Director, Fluidigm Corp., USA), and Prof. Angshul Majumdar (my colleague at IIIT-D). Dr. Ramalingam produced the CTC transcriptomes, while Prof. Majumdar assisted in the algorithm development. The first author of the study is a Ph.D. student affiliated with our laboratory (soon to join Cleveland Clinic as a postdoc).

2. Chawla S, Rockstroh A, Lehman M, Ratther E, Jain A, Anand A, Gupta A, Bhattacharya N, Poonia S, Rai P, Das N, Majumdar A, Jayadeva, Ahuja G, Hollier BG, Nelson CC*, **Sengupta D***. Gene expression based inference of cancer drug sensitivity. **Nat Commun**. 2022 Sep 27;13(1):5680. Doi: 10.1038/s41467-022-33291-z. PMID: 36167836; PMCID: PMC9515171.

   Precision oncology research aims to utilize molecular profiles of cancers for personalized therapeutic choices, with the assistance of AI. To enhance the generalizability of AI approaches to unseen drug compounds, we developed a chemo-transcriptomics approach called Precily. This approach incorporates cancer molecular profiles and drug descriptors as explanatory variables. Precily demonstrated significant improvements compared to existing solutions. We evaluated Precily using cell lines, in-house mouse xenografts (in collaboration with Prof. Colleen Nelson's lab at QUT, Brisbane), and TCGA human cancer data. We also demonstrated how Precily can accurately infer dose response of unseen drugs. Our predictions were supported by in-house experimental data. I conceived the study with Prof. Nelson. The first author of the study is a Ph.D. student affiliated with our laboratory (currently a postdoc at the Harvard Medical School).

3. Goswami C, Chawla S, Thakral D, Pant H, Verma P, Malik PS, Jayadeva, Gupta R*, Ahuja G*, **Sengupta D***. Molecular signature comprising 11 platelet-genes enables accurate blood-based diagnosis of NSCLC. **BMC Genomics**. 2020 Oct 27;21(1):744. doi: 10.1186/s12864-020-07147-z. PMID: 33287695; PMCID: PMC7590669.

We performed feature engineering to synthesize a 11 platelet gene panel which detects early onset of cancer with 97% accuracy, using state of the art classification algorithms. The panel has been experimentally validated at AIIMS-Delhi on NSCLC patients (Goswami et al., BMC Genomics 2020). IIIT-D successfully transferred the technology to CareOnco Biotech Pvt. Ltd., where I serve as the Chief Scientific Advisor and a partner. This is an indigenously developed liquid biopsy assay (Indian patent application no. 202011042049), currently undergoing large scale clinical validation under the commercial name TEPScan™ (300 + patients across India and Norway). For this work, I received the 2022 INAE Young Innovator & Entrepreneur Award. The whole work was conceived by self, and executed with help of multiple collaborators from AIIMS Delhi, and IIT Delhi. Both the first authors have been my Ph.D. students. Chitrita is currently working as a postdoc at Roche, NY, and Smriti is a postdoc at the Harvard Medical School.

4. Gupta P, Jindal A, Ahuja G, Jayadeva*, **Sengupta D***. A new deep learning technique reveals the exclusive functional contributions of individual cancer mutations. **J Biol Chem**. 2022 Aug;298(8):102177. doi: 10.1016/j.jbc.2022.102177. Epub 2022 Jun 24. PMID: 35753349; PMCID: PMC9304782.

   It is hard to distinguish cancer-related somatic mutations from germline variants and noncancerous somatic mutations. In this study, we introduced Continuous Representation of Codon Switches (CRCS), a deep learning-based method that generates numerical vector representations of mutations. CRCS enables the detection of cancer-related somatic mutations without matched normal samples, identification of potential driver genes, and prediction of patient survival based on mutation scores in bladder urothelial carcinoma, hepatocellular carcinoma, and lung adenocarcinoma. These findings establish CRCS as a valuable computational tool for analyzing the functional significance of individual cancer mutations, addressing critical limitations in mutation characterization and patient outcome prediction. I conceived this study. Prof. Jayadeva provided critical advice on model architectures. The first author Prashant Gupta has been jointly supervised by Prof. Jayadeva and myself (working as a postdoc at the Wellcome Sanger Institute).

5. Mittal A, Mohanty SK, Gautam V, Arora S, Saproo S, Gupta R, Sivakumar R, Garg P, Aggarwal A, Raghavachary P, Dixit NK, Singh VP, Mehta A, Tayal J, Naidu S, **Sengupta D***, Ahuja G*. Artificial intelligence uncovers carcinogenic human metabolites. **Nat Chem Biol**. 2022 Nov;18(11):1204-1213. Doi: 10.1038/s41589-022-01110-7. Epub 2022 Aug 11. PMID: 35953549.

   In this collaboration work with Dr. Gaurav Ahuja's laboratory, we developed Metabokiller, an ensemble classifier that accurately identifies carcinogens by quantifying their electrophilicity, proliferative potential, oxidative stress induction, genomic stability impact, epigenetic alterations, and anti-apoptotic response. Metabokiller is highly interpretable, outperforming current methods for carcinogenicity prediction. My key contribution was supervising key parts of the AI modeling tasks.

6. Flores BCT, Chawla S, Ma N, Sanada C, Kujur PK, Yeung R, Bellon MB, Hukari K, Fowler B, Lynch M, Chinen LTD, Ramalingam N*, **Sengupta D***, Jeffrey SS*. Microfluidic live tracking and transcriptomics of cancer-immune cell doublets link intercellular proximity and gene regulation. **Commun Biol.** 2022 Nov 12;5(1):1231. doi: 10.1038/s42003-022-04205-y. PMID: 36371461; PMCID: PMC9653407.

   In this collaborative work with Standard BioTools (Dr. Naveen Ramalingam) and Stanford University (Prof. Stefanie Jeffrey), we tracked the temporal variations in natural killer—triple-negative breast cancer cell distances and compared them with terminal cellular transcriptome profiles. Our analyses highlighted time-bound activities of regulatory modules and alluded to the existence of transcriptional memory. Smriti Chawla, a co-first author in the study, has been my Ph.D. student, and currently doing a postdoc at the Harvard Medical School.

7. Gupta K, Lalit M, Biswas A, Sanada CD, Greene C, Hukari K, Maulik U, Bandyopadhyay S, Ramalingam N, Ahuja G, Ghosh A*, **Sengupta D***. Modeling expression ranks for noise-tolerant differential expression analysis of scRNA-seq data. **Genome Res.** 2021 Apr;31(4):689-697. doi: 10.1101/gr.267070.120. Epub 2021 Mar 5. PMID: 33674351; PMCID: PMC8015842.

   Single cell transcriptomics data is noisy. We developed a ROSeq, a parametric differential expression test that is based on expression ranks. ROSeq outperforms a good number of popular single cell differential expression tests such as SCDE, and MAST. Dr. Abhik Ghosh from ISI, Kolkata advised us on select components of the underlying theory. First author of the study Dr. Krishan Gupta was my Ph.D. student, and currently doing a postdoc at the Harvard Medical School.

8. Jindal A, Gupta P, Jayadeva*, **Sengupta D***. Discovery of rare cells from voluminous single cell expression data. **Nat Commun.** 2018 Nov 9;9(1):4719. Doi: 10.1038/s41467-018-07234-6. PMID: 30413715; PMCID: PMC6226447.

   We introduced a novel algorithm called Finder of Rare Entities (FiRE) for single cell messenger RNA sequencing (scRNA-seq) data. FiRE efficiently assigns a rareness score to each expression profile, even in large datasets containing tens of thousands of cells. This algorithm enables us to quickly identify and focus on rare cell subpopulations within ultra-large scRNA-seq data. In particular, we applied FiRE to a large scRNA-seq dataset of mouse brain cells and successfully identified a previously unknown subtype of the pars tuberalis lineage. At the time of publication FiRE was the only algorithm that could analyze large scale scRNA-seq profiling data obtained from drop-seq experiments. I conceived this study, in consultation with Prof. Jayadeva. The first author Aashi Jindal has been jointly supervised by Prof. Jayadeva and myself (currently working as a Data Scientist in a private company in India).

9. Srivastava D, Iyer A, Kumar V*, **Sengupta D***. CellAtlasSearch: a scalable search engine for single cells. **Nucleic Acids Res.** 2018 Jul 2;46(W1):W141-W147. doi: 10.1093/nar/gky421. PMID: 29788498; PMCID: PMC6030823.

CellAtlasSearch was the first algorithm for scalable single cell transcriptome search. We applied elements of big data algorithms and leveraged the power of GPU computing to attain unprecedented speed in real-time cell transcriptome search based on query gene expression profiles. This was a collaborative work with Dr. Vibhor Kumar at our institute. Both the students are amidst their doctoral studies in Switzerland.

10. Sinha D, Kumar A, Kumar H, Bandyopadhyay S*, **Sengupta D***. dropClust: efficient clustering of ultra-large scRNA-seq data. **Nucleic Acids Res.** 2018 Apr 6;46(6):e36. doi: 10.1093/nar/gky007. PMID: 29361178; PMCID: PMC5888655.

    The introduction of droplet-based techniques revolutionized single cell profiling, resulting in a sharp increase in the number of cells studied per analysis. Existing clustering algorithms for single cell data were unable to keep up with the growing demands. To address this, we implemented Big Data algorithmic techniques to analyze large-scale single cell expression data. By employing a modified version of locality sensitive hashing for approximate near neighbor search, we achieved significant speed improvements. Our clustering algorithm, dropClust, outperformed Seurat and Ranger in both speed and accuracy, with a notable advantage. Notably, dropClust facilitated the identification of a rare population of regulatory T cells. This study was conceived in collaboration with Prof. Sanghamitra Bandyopadhyay, and I served as the joint thesis supervisor for Debajyoti Sinha, the study's first author (current working as a postdoc at the University of Nantes).