Details of the research work duly signed by the applicant, for which the Sun Pharma Science Foundation Research Award is claimed, including references and illustrations.

The main contributions of Prof. Sanghamitra Bandyopadhyay for which the Sub Pharma Science Foundation Research Award is claimed lie in the intersection of Artificial Intelligence and Biology. In particular, she has developed innovative algorithms for single cell RNA-seq data analyses, provided efficient and fast methods for mapping long, noisy SMRT reads to the reference genome and studied the exosomal microRNA content from HIV1 infected cells to provide deeper insights into the viral pathogenic mechanisms. In addition she has made most of her methods freely available for use by the scientific community.

Transcriptomic data analysis is a fundamental step in numerous biological studies. The conventional approach to obtain transcriptomic data using microarrays pools a large number of cells resulting in an averaging effect on the data, leading to loss of detailed information/signals in the process. In contrast, single cell RNA sequencing data (scRNA-seq) can help assess the transcriptomic profile on a genomic scale at the granularity of a single cell data, thereby holding a virtual trove of information for deeper biological analysis.

In an earlier work [NAR, 46(6), 2018], Prof. Bandyopadhyay and her group developed dropClust, an algorithm for clustering large scale single cell RNA-seq data capable of detecting very small clusters comprising rare cell types. This algorithm utilizes the features of Locality Sensitive Hashing (LSH) to speed up clustering of large-scale single cell expression data. Later on, Prof. Bandyopadhyay and her co-workers have vastly improved dropClust, and provided a complete R package that is, fast, interoperable and minimally resource intensive. One of the works for which the SunPharma Research Award is being claimed is the improved version of dropClust that was published in [Bioinformatics, 36(6), 2021]. The entire source code is available in https://github.com/debsin/dropClust while https://debsinha.shinyapps.io/dropClust/ has the online version of the webserver. In the following paragraphs, details of the improvements made to dropClust are mentioned.

In the existing version of dropClust, principal components are rank-ordered based on the number of Gaussian mixtures detected from projection of single cells on individual components. In the enhanced version, the Gaussian Kernel Density estimator has been used to identify the number of modes in a distribution. This improvement led to a 3.5X speed-up over the earlier Gaussian mixture model based approach.

Removal of batch effect is a very important improvement to the original dropClust technique. A simple, rank based method has been introduced during clustering for this purpose. The batch correction pipeline is composed of two steps. First, the expression matrices from all individual batches are merged into a single matrix with a common set of informative genes. Second, the merged matrix undergoes a rank transformation and low dimensional embedding. The integrative analysis pipeline is provided in Fig. 1, while flowchart of the improved dropClust algorithm is provided in Fig. 2. The description of each step is provided below.

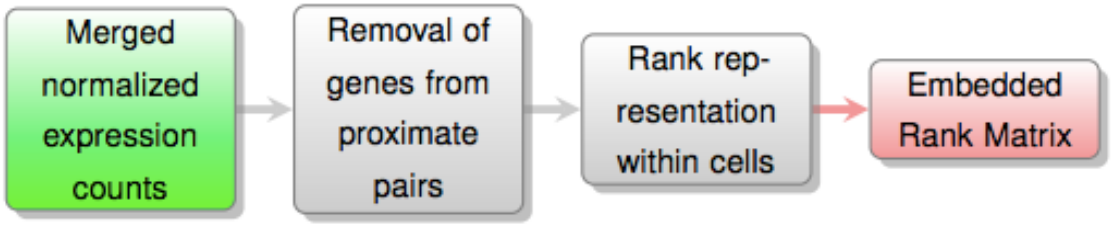

Fig. 1: Integrative analysis workflow

Fig. 2: The flowchart of the enhanced version of dropClust with batch effect removal

The algorithm takes a merged expression matrix as input. As features, it selects, the union of cluster-specific differentially up-regulated genes obtained from batch-wise dropClust analyses. As part of an additional gene filtering step, some more genes are dropped to ensure that the remaining genes do not display similar expression levels in a large fraction of the cells under study. Cell-wise expression ranks are then used for the final clustering. For expediting the clustering step, structure preserving locality sensitive hashing is used. Moreover, instead of only one option to cluster using the hierarchical clustering, the improved version offers users three different clustering approaches, topological clustering using Louvian, agglomerative hierarchical clustering and k- means.

Once clustering the sampled data is complete, the remaining cells are assigned to one of their closest clusters based on the most frequently observed cluster identities among its nearest neighbours. Clustering outcomes improved significantly upon the introduction of an acceptance threshold to discard ambiguous cluster assignments.

To increase modularity and cross package interaction, the latest implementation of dropClust adopts the widely used SingleCellExperiment container class. The container class allows the user to store raw and the normalized single-cell expression profiles along with multiple layers of information (annotations) regarding individual cells and genes, accessible via the rowData/colData modules. The use of SingleCellExperiment enables seamless interaction with other best practice software utilities.

Results are demonstrated on a large scRNA-seq dataset containing about 68000 peripheral blood mononuclear cell (PBMC) transcriptomes to benchmark various methods. Unsupervised clusters obtained using Seurat, Scanpy and the improved dropClust offered ARI of 0.33, 0.34 and 0.40, respectively. The enhanced version of dropClust attained a 4.5 X speedup as compared to the original implementation of the software, whereas 1.3X and 26X speedups were achieved over Scanpy and Seurat respectively. The latest implementation of the software consumed 8.3 times less memory compared to the original version. The integrative analysis pipeline of the improved version of dropClust is also compared with three recently published batch-correction methods on two gold-standard datasets, CellBench and RCA. While CellBench dataset has data in three batches, RCA has data in two batches. Results are demonstrated in Fig. 3 and Fig. 4 for CellBench and RCA, respectively. Visual inspection clearly reveals the superior performance of improved dropClust.

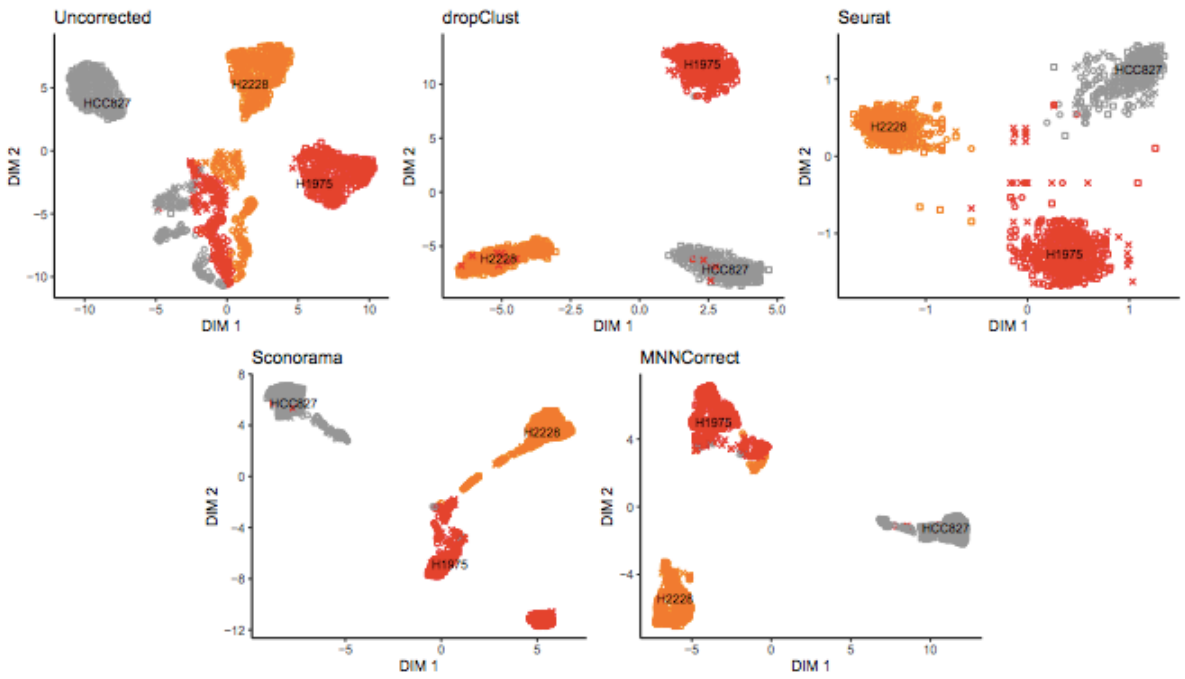

Fig. 3: Comparing the performance of the imporved version of dropClust integrative analysis outcome with MNNCorrect, Scanorama and Seurat on CellBench (lung adenocarcinoma cell lines) data. Batches are indicated by different shapes. The panel labeled “Uncorrected” refers to the figure generated by the dropClust pipeline without performing any batch correction.

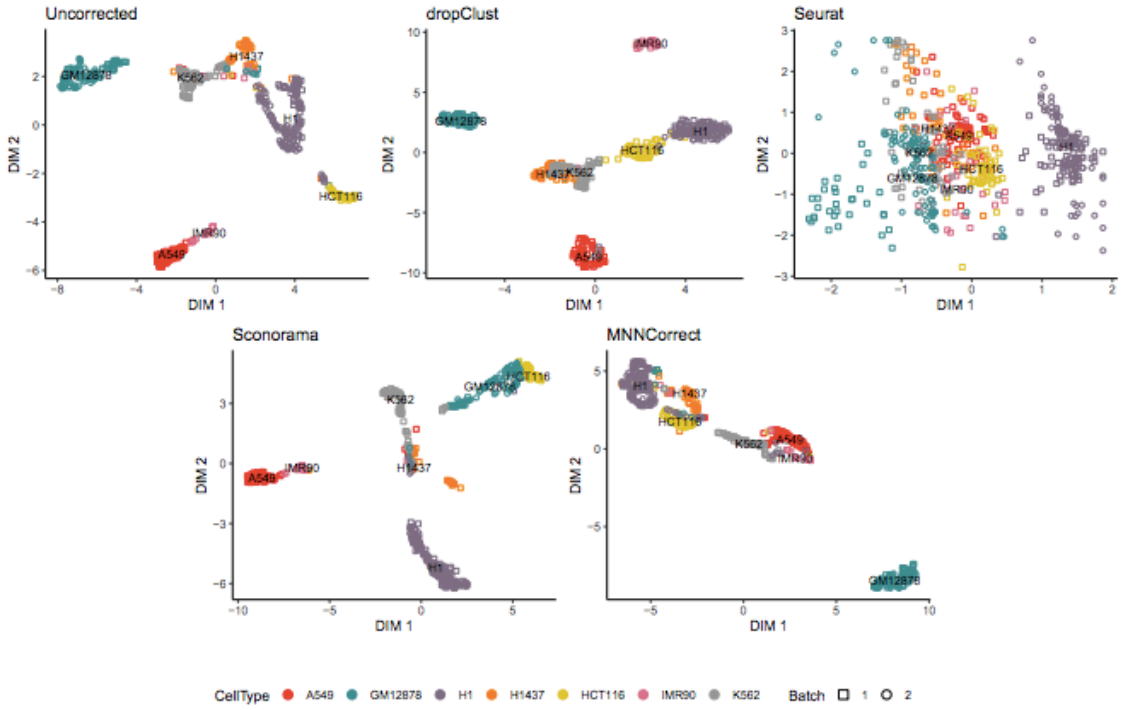

Fig. 4: Comparing the performance of the latest dropClust integrative analysis outcome with MNNCorrect, Scanorama and Seurat on the RCA data. Batches are indicated by different shapes. The panel labeled "Uncorrected" refers to the figure generated by the dropClust pipeline without performing any batch correction

Accurately modelling gene expression and identifying differentially expressed genes from scRNA-seq data presents a challenge that is different from similar analysis in bulk RNA-seq data. In the past few years, considerable effort has been made to identify appropriate parametric models for single cell expression data. The zero-inflated version of Poisson/Negative Binomial and Log-Normal distributions have emerged as the most popular alternatives due to their ability to accommodate high dropout rates, as commonly observed in single cell data. The potential of modeling expression-ranks, as robust surrogates for transcript abundance, is explored by Bandyopadhyay and her collaborators in [Genome Research, 2021] for developing ROSeq, a rank based test to determine differential expression from scRNA-seq data. Since ranks are known to be more robust than the expression values, it is expected that a rank based analysis will yield more stable results.

For gene expression modeling, ROSeq accepts normalized read count data as input. For each gene, ROSeq first defines its range by identifying the minimum and the maximum values by pooling the normalized expression estimates across both the cell-groups under study. Next, the range is split into k * σ sized bins, where k is a scalar with a default value of 0.05, and σ is the standard deviation of the pooled expression estimates across the cell-groups. Each of these bins is assigned a rank, based on the sequential order of its expression range. At the level of a cell-group, this leads to mapping of bin-wise cell frequencies to ranks, such that the bin with the highest cellular frequency is assigned the least rank (i.e., 1). The Discrete Generalized Beta Distribution (DGBD) is used as a probability mass function to express a normalized bin-wise cell-frequency y_r, as a function of its corresponding rank r using two real parameters a and b. In other words, the DGBD formulation can

be thought of as a discrete distribution of the rank-frequencies. If N be the total number of bins for a given gene, then the DGBD specifies the probability p_r for the r-th rank to have a (relative) size of y_r, can be expressed as

$p_r = A\frac{(N+1-r)^b}{r^a}$, r= *1, ..., N,*

where A is the normalization constant ensuring $\sum_r p_r$=1. Note that the sum of the normalized frequencies also equals one ($\sum_r y_r$=1).

For a given gene and a specific cell-group, the best-fitting parameter values ($\hat{a}$, $\hat{b}$) are determined by maximizing, with respect to (a,b), the Log-Likelihood corresponding to the model given by the above equation. Considering the discrete probability distribution structure of the DGBD formulation of (relative) rank sizes, the resulting likelihood function is given by

$$L = \prod_{r=1}^{N} {p_r}^{y_r} = A\prod_{r=1}^{N}\frac{(N+1-r)^{by_r}}{r^{ay_r}}.$$

Now, taking logarithm, the required Log-Likelihood function, *log L*, can be computed as in the following Equation.

$$log\ L(a,b) = -a * \sum_{r=1}^{N} y_r log(r) + b * \sum_{r=1}^{N} y_r log(N+1-r) + log\ (A).$$

The resulting estimates ($\hat{a}$, $\hat{b}$) correspond to the DGBD under which the observed data is most likely to be generated. Such maximum likelihood estimates (MLE) are the most efficient (least standard error) and enjoy several optimum properties on large sample sizes.

To test differential expression of a gene between two cell-groups we additionally need estimates of their standard errors or their variance. From the theory of maximum likelihood, the asymptotic variance of ($\hat{a}$, $\hat{b}$) is given by the inverse of the associated Fisher information matrix *I(a,b),* which can be consistently estimated by $I(\hat{a}, \hat{b})$. For the log likelihood function of the DGBD model given in the previous equation, the form of the Fisher information matrix I may be simplified in a more succinct form as described in [Genome Research, 2021].

Further, in order to statistically test if a gene is differentially expressed between two sub-populations, ROSeq uses the (asymptotically) optimum two-sample Wald test based on the MLE of the parameters and their asymptotic variances, given by the inverse of the Fisher information matrix.

The ROSeq R package (R Core Team 2020) is available at the Bioconductor portal (http://www.bioconductor.org/packages/release/bioc/ html/ROSeq.html). A more frequently updated version of the software is accessible at the GitHub (https://github.com/krishan57gupta/ROSeq).

Transfer of exosomal miRNAs to other recipient cells is recognized as a mode of communication between different cells and tissues. Prof. Bandyopadhyay and her collaborators conducted an innovative research on microRNA profiles in exosomes of cells infected with HIV. The HIV Nef

protein is a multifunctional virulence factor that perturbs intracellular membranes and signalling and is secreted into exosomes. While Nef-containing exosomes are known to have a distinct proteomic profile, the first study involving the miRNA cargo in the exosomes was carried out in [J Extracellular Vesicles, 3, 2014].

Exosomes were purified from human monocytic U937 cells that stably expressed HIV-1 Nef. The RNA from cells and exosomes was profiled for 667 miRNAs using a Taqman Low Density Array. Selected miRNAs and their mRNA targets were validated by quantitative RT-PCR. It was identified that Nef expression affected a significant fraction of miRNAs in U937 cells. The analysis showed 47 miRNAs to be selectively secreted into Nef exosomes and 2 miRNAs to be selectively retained in Nef-expressing cells. The exosomal miRNAs were predicted to target several cellular genes in inflammatory cytokine and other pathways important for HIV pathogenesis, and an overwhelming majority had targets within the HIV genome. This was the first study to report miRnome analysis of HIV Nef expressing monocytes and exosomes. The results demonstrated that Nef causes large-scale dysregulation of cellular miRNAs, including their secretion through exosomes. This suggested a novel viral strategy to affect pathogenesis and to limit the effects of RNA interference on viral replication and persistence.

The other work for which the Sunpharma Research award is being claimed is the development of algorithms for aligning reads to the reference genome. While next gen sequencing (NGS) data resulted in vast amounts of sequence data being collected very quickly, thus bringing down the sequencing cost significantly, it suffers from the problem of short read lengths, thus making alignment to the reference genome difficult and prone to errors. Single molecule real time (SMRT) sequencing developed by Pacific Biosciences alleviated this problem to a large extent with its longer read length of 20KB on an average and low GC bias. The increased read length is particularly useful for alignment of repetitive regions during genome. However, one major concern with SMRT reads is that they come with a higher error probability of 13% to 15% per base, errors being mostly indels. Thus the state-of-the-art aligners tuned for NGS data turned out to be inappropriate for aligning SMRT reads to the reference genome. In [Comp Biol. Chem., 85, 2020], Prof. Bandyopadhyay and her student incorporated a novel *con*textual Locality Sensitive Hashing (conLSH) based algorithm for aligning the noisy SMRT reads to the reference genome. In conLSH, sequences are hashed together not only based on their similarity, but also if they share similar contexts. The probability of collision of two sequences is proportional to the number of common contexts between them. The *context*s play an important role in grouping sequences having noisy bases like in the PacBio data.

At the indexing phase, the reference genome is virtually split into several overlapping windows. For each window, a set of conLSH values is computed using the contexts from K (*concatenation factor*) different locations, where each context is of size $2 \times \lambda + 1$, λ being the *context factor*. This increases the seed length, which helps to resolve repeats and thereby prevents false positive targets. Therefore, even if a k-mer is repetitive, conLSH aligner can map it properly with the help of its contexts. The conLSH values are stored as a B-Tree index to facilitate logarithmic index search. The aligner hashes the query SMRT reads using the same conLSH functions and retrieves the window-list from B-Tree index that produced the same hash values as the read. This forms the list of candidate sites for possible alignment after extension. Finally, Sparse Dynamic Programming (SDP) is used to obtain the best possible alignment(s) for each read. The entire workflow of indexing the reference genome and alignment of SMRT reads using conLSH is portrayed in Fig. 5 below.

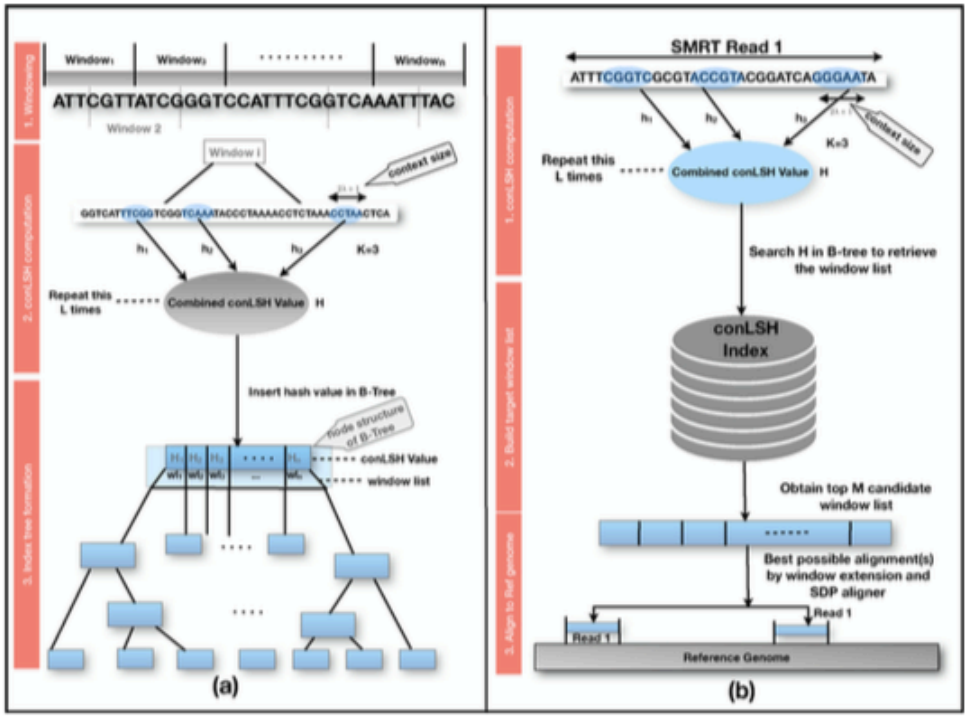

Fig. 5: (a) Describes the three different stages: Windowing, conLSH computation and Index tree formation of conLSH-indexer. conLSH- aligner computes the hash value for the reads and uses the index tree build by conLSH-indexer to map reads back to the reference genome. (b) The workflow of conLSH-aligner, which goes through the phases of conLSH computation, Build target window list, and alignment to reference genome.

The algorithm conLSH is shown to have $O(n^{\rho+1})$ space requirement, where n is the number of sequences in the corpus and ρ is a constant. The indexing time and querying time are bounded by $O(\frac{n^{\rho+1}\ln n}{\ln\frac{1}{P_2}})$ and $O(n^{\rho})$ respectively, here $P_2 > 0$ is a probability value. The proposed conLSH based aligner is compared with rHAT popularly used for aligning SMRT reads, and is found to comprehensively beat it in speed as well as in memory requirements. In particular, it takes approximately 24.2% less processing time, while saving about 70.3% in peak memory requirement for *H.sapiens* PacBio dataset.

Moving on from alignment based approach, Prof. Bandyopadhyay, her student and her collaborator recently developed a novel alignment-free method of sequence mapping [BMC Bioinformatics, 22(64), 2021]. The new mapper called S-conLSH uses Spaced context based Locality Sensitive Hashing. It employs multiple spaced-seeds or patterns to find gapped mappings of noisy SMRT reads to reference genomes. The spaced-seeds are strings of 0's and 1's where '1' represents the match position and '0' denotes don't care position where matching in the symbols is not mandatory. The substring formed by extracting the symbols corresponding to the '1' positions in the pattern is defined as the *spaced-context* of a sequence. Therefore, a spaced-context can minimize the effect of erroneous bases thereby enhancing the quality of mapping because it does not check all the bases for a match. This differentiates the proposed

method from conLSH that looks into the entire context to compute the hash values. S-conLSH achieves a sensitivity of 99%, without using any traditional base-to-base alignment, on human simulated sequence data, while it is at least 2 times faster than the recently developed method lordFAST.

The full workflow of S-conLSH is provided in Fig. 6.

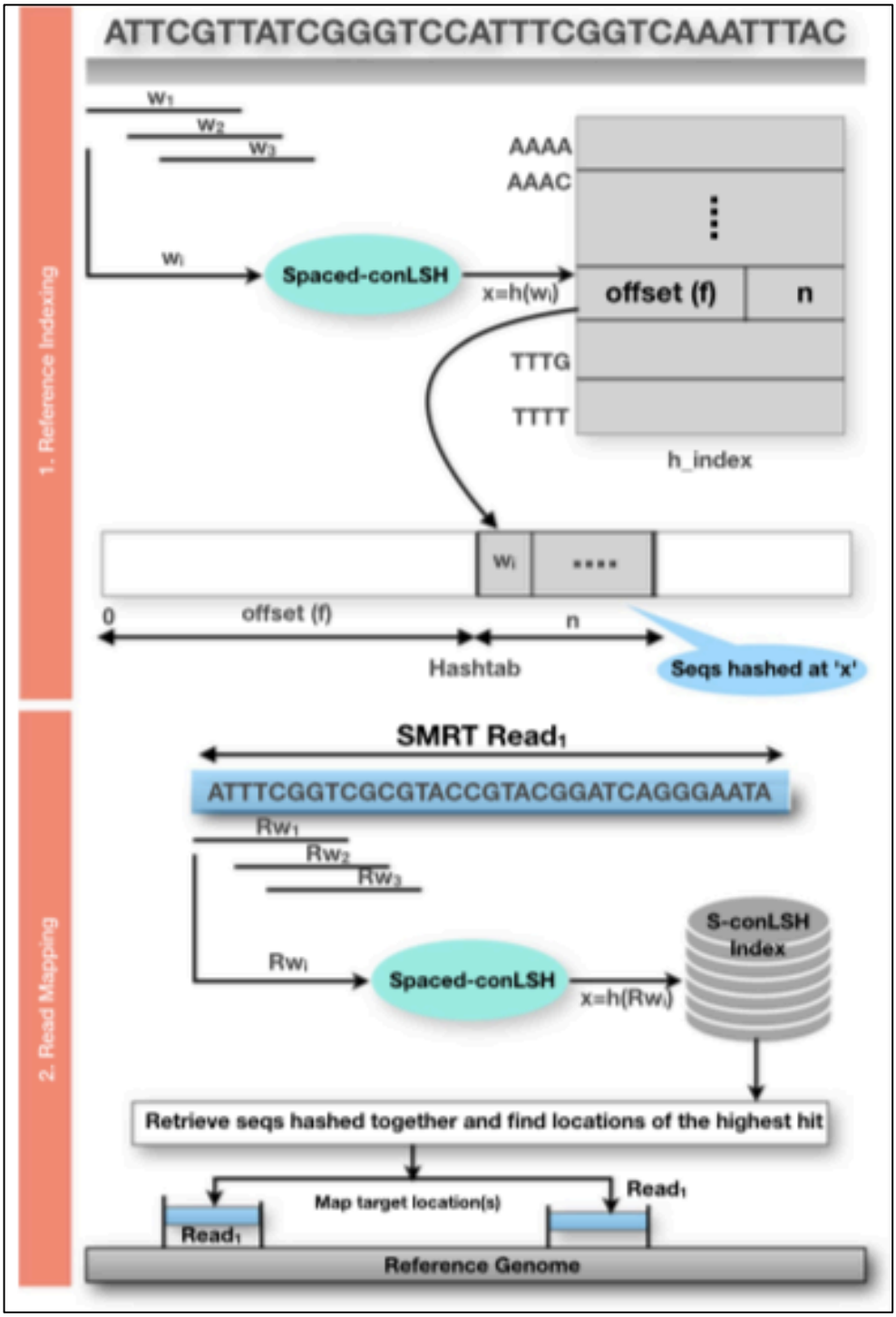

Fig. 6: A schematic workflow of indexing and mapping using S-conLSH

Finally, making all the developed techniques available for use by the research community, either as webservers or as source codes in github constitutes a major contribution of Prof. Bandyopadhyay. The microRNA target prediction webservers TargetMiner (https://www.isical.ac.in/~bioinfo_miu/targetminer20.htm) and MultiMiTar (https://www.isical.ac.in/~bioinfo_miu/multimitar.htm), a database of transcription factors of human microRNAs PuTmiR (https://www.isical.ac.in/~bioinfo_miu/TF-miRNA/TF-miRNA.html), a database for putative microRNA-microRNA regulations called PmmR (https://www.isical.ac.in/~bioinfo miu/pmmr.php), disease specific TF-miRNA-gene subnetworks called DisTMGneT (https://www.isical.ac.in/~bioinfo_miu/dscsgen.php) are some such examples.

---------- **XXX** ----------