Published online 21 May 2018 *Nucleic Acids Research, 2018, Vol. 46, Web Server issue* ***W141–W147***
doi: 10.1093/nar/gky421

CellAtlasSearch: a scalable search engine for single cells

Divyanshu Srivastava[1,†], Arvind Iyer[1,†], Vibhor Kumar[1,*] and Debarka Sengupta[1,2,*]

[1]Center for Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India and [2]Department of Computer Science and Engineering, Indraprastha Institute of Information Technology, New Delhi, India

Received February 21, 2018; Revised May 03, 2018; Editorial Decision May 03, 2018; Accepted May 04, 2018

ABSTRACT

Owing to the advent of high throughput single cell transcriptomics, past few years have seen exponential growth in production of gene expression data. Recently efforts have been made by various research groups to homogenize and store single cell expression from a large number of studies. The true value of this ever increasing data deluge can be unlocked by making it searchable. To this end, we propose CellAtlasSearch, a novel search architecture for high dimensional expression data, which is massively parallel as well as light-weight, thus infinitely scalable. In CellAtlasSearch, we use a Graphical Processing Unit (GPU) friendly version of Locality Sensitive Hashing (LSH) for unmatched speedup in data processing and query. Currently, CellAtlasSearch features over 300 000 reference expression profiles including both bulk and single-cell data. It enables the user query individual single cell transcriptomes and finds matching samples from the database along with necessary meta information. CellAtlasSearch aims to assist researchers and clinicians in characterizing unannotated single cells. It also facilitates noise free, low dimensional representation of single-cell expression profiles by projecting them on a wide variety of reference samples. The web-server is accessible at: http://www.cellatlassearch.com.

BACKGROUND

Single cell transcriptomics provides a powerful means for delineating subtle phenotypic differences among seemingly similar cells (1). Over the past few years single cell RNA-Sequencing (scRNA-seq) has emerged as a popular choice for studying tissue heterogeneity in the context of development and disease. Moreover, continuous upgradation of the throughput capabilities has made scRNA-seq a reliable tool for systematic discovery of rare cell types (2,3). Owing to its promises and popularity significant resources have lately been deployed through community-level initiatives such as Human Cell Atlas (4) and Oxford Single Cell Biology Consortium.

How to characterize individual cells? How to ward off noise while clustering transcriptomes? How to ensure if a seemingly novel transcriptomic pattern indeed corresponds to a new and unreported cell type? These are among the most frequent and persistent questions when it comes to downstream analysis of single cell expression data. We built CellAtlasSearch to address these important questions by exploiting the massive amount of pre-existing messenger RNA sequencing data.

Oftentimes a single cell manifests its identity through multiple previously known phenotypes. For example, glioblastomas have traditionally been stratified into four categories: classical, neural, pro-neural and mesenchymal (5). However, single cell studies revealed transcriptomes that have mixed representation of these phenotypes (6). The ability to compare a query single cell transcriptome with a large number of reference expression data directly benefits characterization of single cells, as it assists in zeroing down on the potential phenotypes.

Efforts have been made in archiving both single cell and bulk expression data. Single Cell Portal, Recount2 (7) and JingleBells (8) are notable among these. A few webservers have also been developed for online search of matching microarray and bulk-RNA-seq based expression profiles (9–12). CellAtlasSearch, for the first time, allows user query single-cell expression profiles to retrieve matching single cell or bulk expression data from over 2000 different studies.

Besides discerning tissue heterogeneity, large-scale single-cell studies often lead to the discovery of rare cells (2). CellAtlasSearch can be used to cross-validate if a suspected rare cell is indeed unreported. Upon submission of a rare cell transcriptome as a query, it reports zero hits.

Single cell assays are usually fragile due to the paucity of input RNA. As a result, clustering single-cell expres-

[*]To whom correspondence should be addressed. Tel: +91 11 26907446; Email: debarka@iiitd.ac.in
Correspondence may also be addressed to Vibhor Kumar. Email: vibhor@iiitd.ac.in
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

sion profiles is often challenging in presence of high levels of noise, technical variation and batch effect (BioRxiv: https://doi.org/10.1101/025528). In a recent article, it has been shown that the best way to deal with noise in single-cell data is to project it on a wide variety of reference samples (13). However, due to data curation and computation related challenges, the authors had to limit their scope to the BioGPS Primary Cell Atlas. CellAtlasSearch breaks the barrier by allowing comparison of query cells with a vast pool of reference expression data. Users can download the resulting similarity matrix and use it as a replacement for the expression matrix for noise-free clustering of the individual transcriptomes.

We have recently shown how Locality Sensitive Hashing (LSH) improves speed and accuracy of cell type clustering (14). CellAtlasSearch implements LSH on the powerful GPU architecture to attain an unmatched speed in archiving and querying expression data. Hashing based low dimensional encoding of expression profiles makes data transactions efficient and inexpensive, thus future-proof.

Here, we first assess the effectiveness of GPU in speeding up expression data archival and query. We also show the accuracy of information retrieval using cell line data. Notably, CellAtlasSearch shows substantial tolerance to high dropout rates, which is common in scRNA-seq data. Further, we furnish two case-studies depicting the potential applications of CellAtlasSearch. In the first case study, we query the transcriptomes of a few circulating tumor cells (CTCs) along with a large number of non-cancerous immune cells to assess the efficacy of CellAtlasSearch in distinguishing the rare cells from the previously known abundant cell types. In the second case study, we show how CellAtlasSearch manages to bypass batch effects in grouping single-cell expression profiles from two different cell lines, each processed in two independent batches.

IMPLEMENTATION DETAILS

Data curation and warehousing

CellAtlasSearch currently features 304,769 expression profiles from 2044 different studies (Supplementary Table S1). These include both single cell and bulk RNA-Seq samples. To compile the reference database, we downloaded expression profiles in the form of raw-count data from three sources: Gene Expression Omnibus (GEO) (15), Recount2 (7) and the 10xGenomics website (https://www.10xgenomics.com). For meta-data, we relied on the study abstracts and sample descriptions. A vast majority of the sample descriptions were sourced from the GEO submission pages of the respective studies. For extracting information pertaining to lineage and phenotype we used Extract-2.0 (BioRxiv: https://doi.org/10.1101/111088). Extract-2.0 was able to populate the necessary details for ∼70% of the studies. We employed Locality Sensitive Hashing (LSH) (16,17) to generate relative-proximity preserving, low dimensional bit vectors (hash-codes) for the individual reference expression profiles in the CellAtlasSearch database.

LSH projects high dimensional data points (transcriptomes in this case) onto a set of randomly defined hyperplanes. For each data point, a binary hash code is generated based on its location with respect to these hyperplanes. For the *embarrassingly parallel* computational task of hash code generation, we exploited the GPU architecture with the support of the CUDA APIs, distributed by NVIDIA. We then performed key-value pairing between the hash codes and the individual expression profiles for efficient warehousing. A hash-code can be imagined as a bucket. Due to this special encoding strategy, similar expression profiles tend to share their bucket (Figure 1).

When queried, the input expression profiles (read-counts) are projected onto the pre-defined set of hyperplanes. Hash codes thus generated are compared with the archived hash codes and hamming distances are computed between the bit vectors associated with the query and the buckets (hash-codes). Of note, hamming distance computed on such hash codes approximate cosine similarity between the associated high dimensional data-points (18). Cosine similarity is known to be ideal for expression data analysis as it is agnostic to scaling (library size) related issues (19). Given a query, only a few proximal buckets are considered for an exhaustive search of nearest neighbors. LSH reduces the nearest neighbor search time dramatically (Supplementary Figure S1a, S1b).

CellAtlasSearch reports the estimated cosine similarity values for the top matches. The approximate nearest neighbor search implemented in CellAtlasSearch competes well with its exhaustive counterpart (Supplementary Figure S1c).

Estimating significance and accuracy of cell-type match

Expression data collected from diverse sources inherit varied levels of noise and technical bias (7,20). Such bewildering level of variability poses a significant challenge on computing expression similarity. This, motivated us to compute the statistical significance of the search results as a countermeasure. To do this, we created 2000 randomized query single cell transcriptomes by averaging 5 expression profiles at a time collected from independent studies. Then we calculated cosine similarity between all possible pairs of random and pre-archived expression profiles. These cosine values are taken together to create the null distribution. For every cosine similarity value corresponding to a certain search-hit, CellAtlasSearch displays a *P*-value, estimated empirically using the Monte Carlo method (21). These *P*-values are then subjected to multiple test correction using the Benjamini–Hochberg proposed method (22). Besides significant matches, CellAtlasSearch also provides heat-map depicting the similarity between the query expression profiles and the union set of their top hits.

Feature selection

CellAtlasSearch reduces data dimensionality as part of the Locality Sensitive Hashing. As we evaluated the perfor-

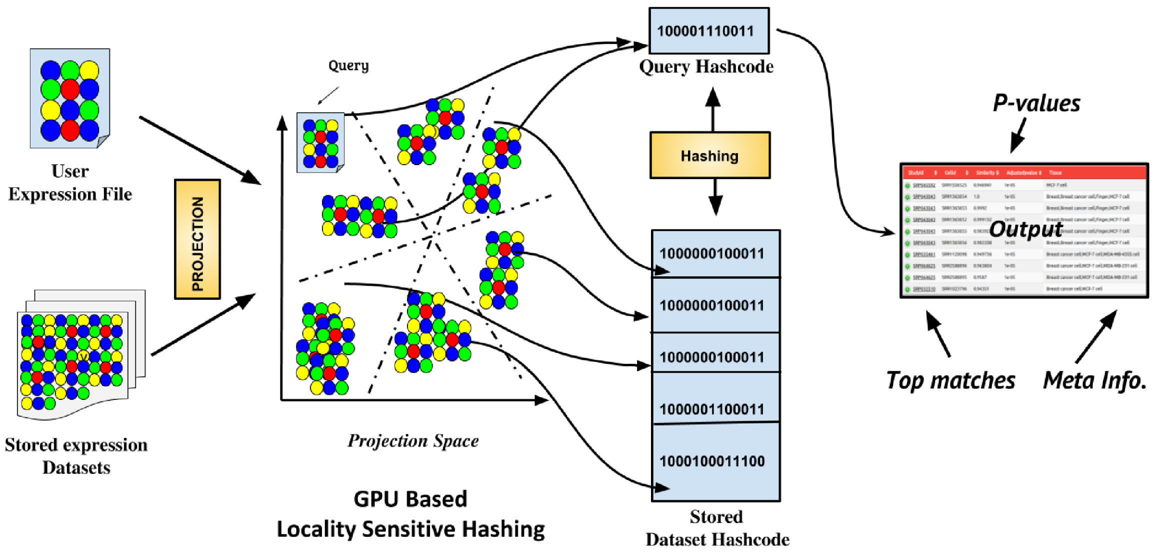


Figure 1. CellAtlasSearch Pipeline. The entire web-server is based on GPU framework. Expression profiles are stored as hash codes obtained through LSH. Like-samples are archived in the same bucket. Query expression data is first converted into hash code and then mapped to one of the buckets. User can query one or more single cell transcriptomes.

mance of CellAtlasSearch in batch effect removal we realized that feature selection also plays an important role (see Results section). CellAtlasSearch currently provides an option for using a predefined set of informative (feature) genes. We created the list of informative genes by using the all-tissue bulk expression data published by the GTEx consortium (23). For this, we first applied quantile normalization on the data. Average expression profiles were then computed for each tissue type. Genes, having high values of expression fold change over their respective median expression in at least one of the tissue types were considered as informative. A fold change cutoff of 8.75 returned nearly 500 feature genes. Feature selection option is currently available purely for experimentation and exploration purposes.

User-friendly graphical user interface

CellAtlasSearch has an easy to use user interface. Figure 2 shows a schematic of user input and output result page of CellAtlasSearch. The server comes with a simple search form in which user furnishes some basic details about the query and uploads the expression data as an excel or CSV file (Figure 2A). The user can choose between the scRNA-seq and the bulk-seq databases. After submission of the query, the user receives a notification about the job status and a result URL, which he can bookmark for future reference (Figure 2B). As a query gets processed, the user receives a table listing the top matches corresponding to the individual queried cells (Figure 2C) and a D3-js based interactive visualization of the descriptions corresponding to the matching samples (Figure 2D). Result page also displays a heat-map of the cosine similarity values between the query expression profiles and the matching reference samples (Figure 2E). In addition, a spectral t-SNE based 2D map provides a view of the query cells on the basis of reference panel projections (Figure 2F). The CellAtlasSearch website provides details about the usage and the utilities through its various content pages such as 'How to Use' and 'Frequently Asked Questions'.

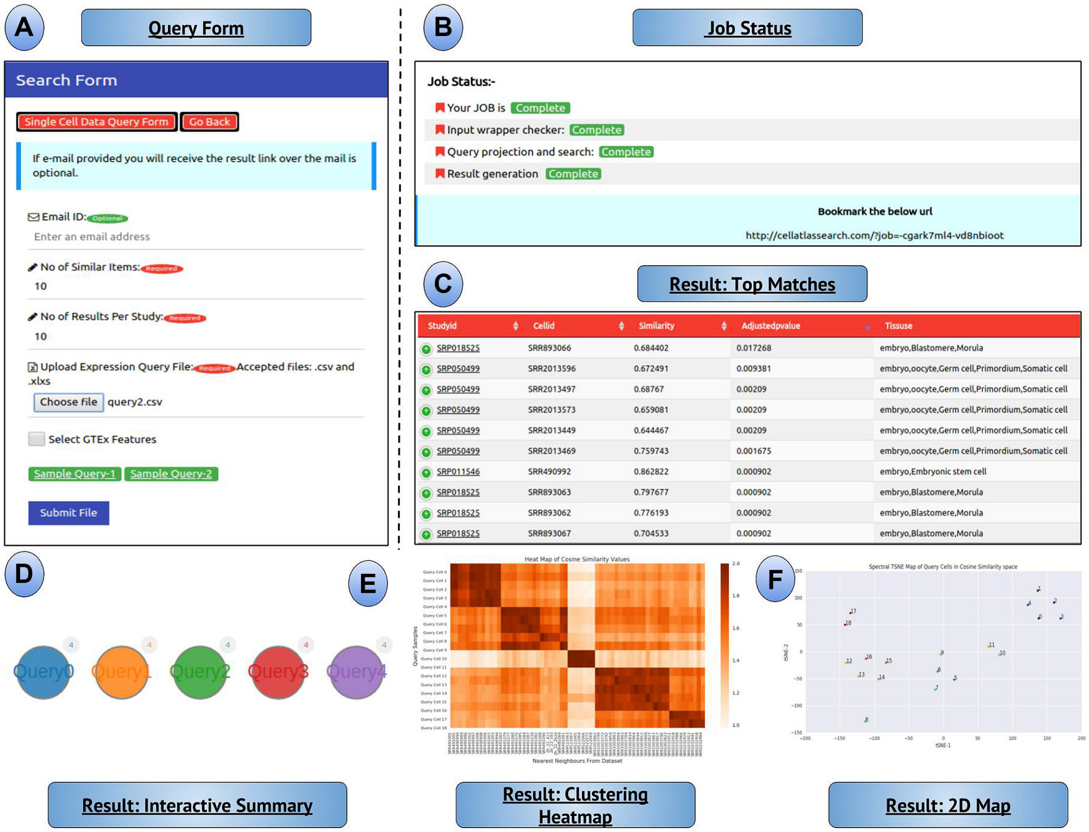
RESULTS

Speedup due to GPU

We first checked if adopting the GPU architecture brings down the computation time drastically. With the use of GPU, CellAtlasSearch achieved significant speedup in the most compute-intensive step, i.e. codifying the reference expression profiles by projecting them on a set of randomly defined hyperplanes (Figure 3A, Supplementary Figure S1a). This is a desirable feature as it helps keep up with the exponential growth in the production of single cell expression data. We found the GPU architecture to be instrumental in accelerating the processing of the user queries. Also, we observed that GPU becomes even more powerful as query sample size increases (Supplementary Figure S1b).

Gene dropouts and accuracy of search

Single cell expression data suffer from excessive levels of zero expression values owing to the paucity of input RNA. To this end, we checked the accuracy of search results against varied drop-out rates. Although dropouts are often modeled as a Poisson process, it has not so far been proved in a systematic manner. Hence, to make our evaluation more challenging, we simulated dropouts by mutting gene expression (making the expression zero) randomly. From our reference database, we randomly picked cells and introduced artificial dropouts at different rates. For a specific simulated dropout rate, 20 randomly picked cells were queried. Our analysis revealed that even with ∼40% genes missing, CellAtlasSearch reported the exact same cell within among the top five hits (Figure 3B, Supplementary Figure S2a, S2b).



Search Form

Single Cell Data Query Form | Go Back

If e-mail provided you will receive the result link over the mail is optional.

Email ID: Optional
Enter an email address

No of Similar Items: Required
10

No of Results Per Study: Required
10

Upload Expression Query File: Required Accepted files: .csv and .xlxs
Choose File query2.csv

□ Select GTEx Features

Sample Query-1 | Sample Query-2

Submit File

Job Status:-

Your JOB is Complete
Input wrapper checker: Complete
Query projection and search: Complete
Result generation Complete

Bookmark the below url
http://cellatlassearch.com/?job=-cgark7ml4-vd8nbioot

Studyid	Cellid	Similarity	Adjustedpvalue	Tissuse
SRP018525	SRR893066	0.684402	0.017268	embryo,Blastomere,Morula
SRP050499	SRR2013596	0.672491	0.009381	embryo,oocyte,Germ cell,Primordium,Somatic cell
SRP050499	SRR2013497	0.68767	0.00209	embryo,oocyte,Germ cell,Primordium,Somatic cell
SRP050499	SRR2013573	0.659081	0.00209	embryo,oocyte,Germ cell,Primordium,Somatic cell
SRP050499	SRR2013449	0.644467	0.00209	embryo,oocyte,Germ cell,Primordium,Somatic cell
SRP050499	SRR2013469	0.759743	0.001675	embryo,oocyte,Germ cell,Primordium,Somatic cell
SRP011546	SRR490992	0.862822	0.000902	embryo,Embryonic stem cell
SRP018525	SRR893063	0.797677	0.000902	embryo,Blastomere,Morula
SRP018525	SRR893062	0.776193	0.000902	embryo,Blastomere,Morula
SRP018525	SRR893067	0.704533	0.000902	embryo,Blastomere,Morula

Figure 2. CellAtlasSearch web application interface. (**A**) Query submission form, where the user insert database preference, uploads the query file and submits processing request. (**B**) A custom URL is generated for the result page, even before the result gets compiled. The user can bookmark it for future references. (**C**) Result page, showing the top hits in a tabular form, with necessary meta information. (**D**) The interactive summary shows graphical view of the frequently occurring descriptions (or phenotypes) corresponding to each query transcriptome. Each big circle represents a query cell whereas the small ones the corresponding frequently occurring descriptions (or phenotypes). The descriptions are displayed when the bubbles are hovered upon by the cursor. (**E**) A heat map of cosine similarity values between pairs of query cells and reference samples. (**F**) Spectral-tSNE plot of the query cells made using cosine similarities as feature variables. Elements (E) and (F) are produced when the query has at least 5 samples.

So far we have been considering a match to be correct even if it comes from the same study from which the query cells are taken. However, in reality, we expect the user queries to be independent of our reference database. Hence we subjected CellAtlasSearch to another round of accuracy testing. We queried HCT116 (cell-line) transcriptomes that were not in our reference-dataset. In the absence of expression profiles from the same study in the reference database, CellAtlasSearch was still able to report correct cell type for 60–70% of the query cells even at a simulated dropout rate of 25% (Figure 3C; Supplementary Table S2). Our findings were similar for HEK293T cells (Figure 3D).

It is possible that gene names in the input data match partially with CellAtlasSearch genes. In such cases, CellAtlasSearch assigns zero expression values to the missing genes. Simulations show that such artificial dropouts do not cause any remarkable difference in the result quality (Supplementary Figure S3).

Case study: Detecting rare cells—application on circulating tumor cells (CTCs)

Encountering a suspiciously rare expression signature gives rise to both excitement and skepticism. CellAtlasSearch allows researchers to check if a newly found transcriptomic signature is indeed rare. For example, circulating tumor cells or CTCs are rare in blood. Marker agnostic identification and characterization of CTCs is considered to be the holy grail of cancer biology. A general strategy for the same is the size based enrichment of CTCs in peripheral blood, followed by single cell expression profiling. However, in the absence of pan CTC molecular marker, their identification in a vast pool of immune cells remains challenging, only by looking at the transcriptomes (24,25). We hypothesized that CellAtlasSearch would either report primary or metastatic cancer cells as top hits or report zero hits as our database does not yet contain CTC transcriptomes. To verify this we prepared a query consisting of expression profiles of 10 circulating tumor cells (CTCs) and 90 peripheral mononuclear cells (PBMC). CTCs were sorted from the blood of a mouse xenograft model of human lung cancer (26) (GEO id: GSE74639). The PBMC scRNA-seq counts were downloaded from the 10xGenomics portal. CellAtlasSearch reported correct matches (with Padj < 0.1) for all PBMC cells for which like expression profiles were found (93%). However, among the 10 query CTCs 5 cells did not have any matching result and the remaining five cells matched with

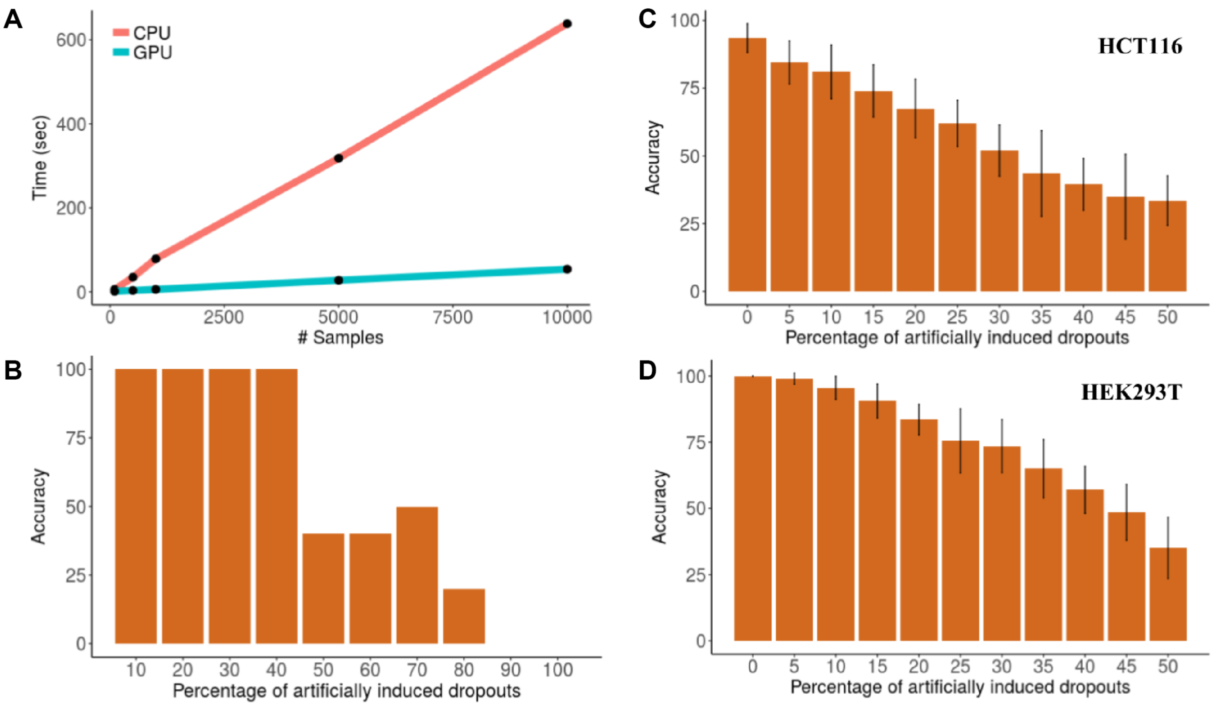


Figure 3. Speed Up and accuracy analysis of CellAtlasSearch. (**A**) Time taken to generate hash codes with varying sample sizes. The blue curve shows the time taken by our parallelized implementation, which is approximately 10 times lesser than the CPU based serialized version. (**B**) Gene dropout analysis for randomly chosen cells from the dataset. An accurate match was considered when CellAtlasSearch was able to recover the exact same cell within top five results. (**C**) Accuracy of CellAtlasSearch in finding samples of the same source cell line from independent studies, upon submission of the query scRNA-seq data of HCT116 cell line, produced by a certain research group. A retrieval is deemed successful if expression data of the same cell line, contributed by an independent group, appears within top five hits. (**D**) Similar analysis for HEK293T cell line.

samples from cancer and stem cell related studies (Supplementary Table S3). This shows that CellAtlasSearch can be used in segregating rare cells from major cell subpopulations.

Case study: fighting batch effect

Single cell expression datasets contributed from different groups come with their own technical biases which could be due to the difference in chemistry, variable reagents concentration, sequencing errors etc. Inspired by our previous findings (13), we hypothesized that an ideal approach to reducing the noise level in single cell data is to project it on a wide variety of reference samples. To test our hypothesis, we considered scRNA-seq data of GM12878 and H1 cell lines processed in two different batches (13) (GEO id : GSE81861). A spectral tSNE visualization of GM12878 and H1 cells expression, after quantile normalization and log transformation, indeed depicted the presence of two batches (Figure 4A).

Batch effect got partially resolved when the spectral t-SNE was applied after replacing the genes by reference panel projection vectors featuring cosine distances between queries and top hits (Figure 4B). On the next iteration, we introduced feature selection. We chose 500 top variable genes using GTEx tissue expression data and repeated the procedure (Implementation Details). The inherent batch specific biases got convincingly resolved owing to the careful selection of feature genes (Figure 4C).

DISCUSSION

The emergence of single cell genomics has inspired the development of a spectrum of new generation computational techniques. Many of these techniques involve considerable ingenuity, mainly aimed at addressing the scalability issues. The key contribution of the present work lies in the introduction of GPU computing to the field of single cell transcriptomics. CellAtlasSearch marries LSH, a popular big data technique with GPU to attain unprecedented efficiency in archival and query of expression data. Besides this, CellAtlasSearch exploits its vast pool of expression data to amplify biological variability while controlling for technical variabilities. This is particularly beneficial for single cell clustering. There are very few studies presenting in-depth analysis on data curated from multiple heterogeneous sources, largely due to the technical variability and batch effect. CellAtlasSearch attempts to overcome this challenge by making the combined data reliably searchable. The approach discussed in this article will encourage researchers for non-redundant experiment designing and use existing datasets for finding greater insights.

With an example query containing CTC transcriptomes, we have shown that CellAtlasSearch can be used to identify previously un-characterized cells. Currently, researchers are trying to develop different kinds of assays to find multiple types of CTCs without relying on antibody based sorting(25). This is because some well known epithelial markers like EpCAM may not necessarily be present on the surface

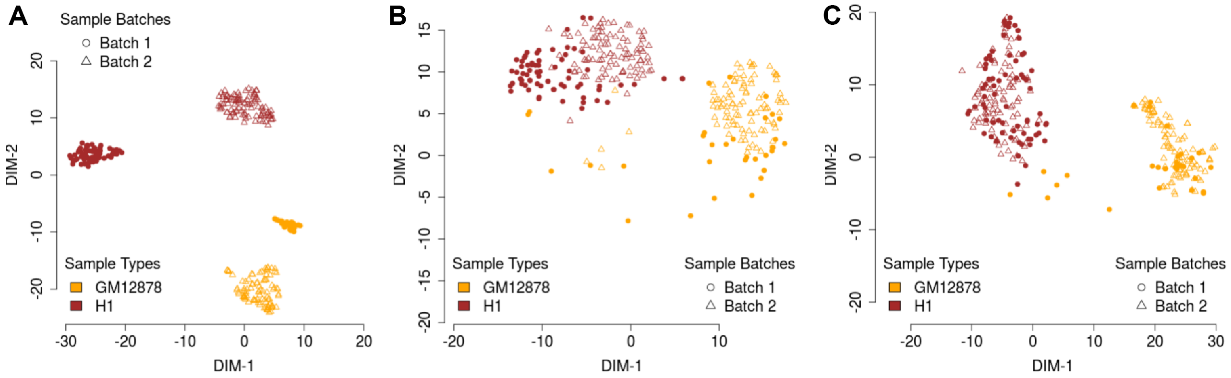


Figure 4. Dealing with batch-effect. (**A**) Spectral tSNE based visualization of GM12878 and H1 cells using log transformed scaled counts of genes. For spectral tSNE, top 10 principal components of gene-count matrix were used. For both cell lines, batches are observed to form separate clusters. (**B**) Visualization based on cosine similarities between queries (GM12878 and H1 cells) and matching references from single cell dataset, as returned by CellAtlasSearch. Here top 10 principal components of cosine similarity matrix were used with tSNE. Cells from different batches of same cell type tend to come closer to each other, yet they did not intermix. (**C**) Batches of same cell type get intermixed when projection for cosine similarity calculation was done using only GTEx-selected-features in the transcriptomes.

of all CTCs (24,25). Even after enriching potential CTCs using different methods, researchers find it challenging to zero in on the identity of the CTCs. In an unsupervised setting, CellAtlasSearch can help recognizing CTCs in a large pool of blood PBMCs, either as cells having no matching transcriptomes or ones with solid cancer transcriptomes as top hits.

CellAtlasSearch is meant for all human tissue/ cell types. However, depending on the data availability, frequency of the matches may vary from cell type to cell type. Due to the stochastic nature of Locality Sensitive Hashing (LSH), sometimes it may not find the true nearest neighbors for a query expression profile. Another major challenge is to fight the curse of batch effect sourced due to the integration of transcriptomic datasets of diverse origin. To address the above difficulties, CellAtlasSearch will continue to explore novel strategies and methodological advances. With further improvements and additions, CellAtlasSearch can potentially provide more value added services such as tissue specific search, RNA-seq based real time infection surveillance, transcriptome based drug resistance prediction in cancer etc.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We are thankful to Indraprastha Institute of Information Technology - Delhi (IIIT-D) for the support of computational facilities. We are grateful to Ojaswa Sharma, Assistant Professor, IIIT-D for his expert advice on the GPU implementation of LSH.

FUNDING

INSPIRE Faculty [DST/INSPIRE/04/2015/003068 to D.S.] by Department of Science and Technology, Govt. of India (in part). Funding for open access charge: Institutional Funding.

Conflict of interest statement. None declared.

REFERENCES

1. Wagner,A., Regev,A. and Yosef,N. (2016) Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol.*, **34**, 1145.
2. Grün,D., Lyubimova,A., Kester,L., Wiebrands,K., Basak,O., Sasaki,N., Clevers,H. and van Oudenaarden,A. (2015) Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature*, **525**, 251.
3. Macosko,E.Z., Basu,A., Satija,R., Nemesh,J., Shekhar,K., Goldman,M., Tirosh,I., Bialas,A.R., Kamitaki,N. and Martersteck,E.M. (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, **161**, 1202–1214.
4. Regev,A., Teichmann,S.A., Lander,E.S., Amit,I., Benoist,C., Birney,E., Bodenmiller,B., Campbell,P., Carninci,P., Clatworthy,M. *et al.* (2017) The human cell atlas. *Elife*, **6**, e27041.
5. Verhaak,R.G., Hoadley,K.A., Purdom,E., Wang,V., Qi,Y., Wilkerson,M.D., Miller,C.R., Ding,L., Golub,T., Mesirov,J.P. *et al.* (2010) Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*, **17**, 98–110.
6. Patel,A.P., Tirosh,I., Trombetta,J.J., Shalek,A.K., Gillespie,S.M., Wakimoto,H., Cahill,D.P., Nahed,B.V., Curry,W.T., Martuza,R.L. *et al.* (2014) Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, **6**, 1254257.
7. Collado-Torres,L., Nellore,A., Kammers,K., Ellis,S.E., Taub,M.A., Hansen,K.D., Jaffe,A.E., Langmead,B. and Leek,J.T. (2017) Reproducible RNA-seq analysis using recount2. *Nat. Biotechnol.*, **35**, 319.
8. Ner-Gaon,H., Melchior,A., Golan,N., Ben-Haim,Y. and Shay,T. (2017) JingleBells: a repository of immune-related single-cell RNA–sequencing datasets. *J. Immunol.*, **198**, 3375–3379.
9. Fujibuchi,W., Kiseleva,L., Taniguchi,T., Harada,H. and Horton,P. (2007) CellMontage: similar expression profile search server. *Bioinformatics*, **23**, 3103–3104.
10. Zinman,G.E., Naiman,S., Kanfi,Y., Cohen,H. and Bar-Joseph,Z. (2013) ExpressionBlast: mining large, unstructured expression databases. *Nat. Methods*, **10**, 925.
11. DeFreitas,T., Saddiki,H. and Flaherty,P. (2016) GEMINI: a computationally-efficient search engine for large gene expression datasets. *BMC Bioinformatics*, **17**, 102.
12. Duan,Q., Reid,S.P., Clark,N.R., Wang,Z., Fernandez,N.F., Rouillard,A.D., Readhead,B., Tritsch,S.R., Hodos,R., Hafner,M. *et al.* (2016) L1000CDS 2: LINCS L1000 characteristic direction signatures search engine. *NPJ Syst. Biol. Applic.*, **2**, 16015.
13. Li,H., Courtois,E.T., Sengupta,D., Tan,Y., Chen,K.H., Goh,J.J.L., Kong,S.L., Chua,C., Hon,L.K., Tan,W.S. *et al.* (2017) Reference

component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat. Genet.*, **49**, 708.

14. Sinha,D., Kumar,A., Kumar,H., Bandyopadhyay,S. and Sengupta,D. (2018) dropClust: efficient clustering of ultra-large scRNA-seq data. *Nucleic Acids Res.*, **46**, e36.
15. Edgar,R., Domrachev,M. and Lash,A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research*, **30**, 207–210.
16. Indyk,P. and Motwani,R. (1998) Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM Symposium on Theory of computing*, ACM pp. 604–613.
17. Gionis,A., Indyk,P., Motwani,R. *et al.* (1999) Similarity search in high dimensions via hashing. *VLDB*, **99**, 518–529.
18. Lee,H., Ng,R.T. and Shim,K. (2011) Similarity join size estimation using locality sensitive hashing. *Proc. VLDB Endowment*, **4**, 338–349.
19. Jaskowiak,P.A., Campello,R.J. and Costa,I.G. (2014) On the selection of appropriate distances for gene expression data clustering. In *BMC Bioinformatics*, BioMed Central, Vol. **15**, p. S2.
20. Zheng,G.X., Terry,J.M., Belgrader,P., Ryvkin,P., Bent,Z.W., Wilson,R., Ziraldo,S.B., Wheeler,T.D., McDermott,G.P., Zhu,J. *et al.* (2017) Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.*, **8**, 14049.
21. North,B.V., Curtis,D. and Sham,P.C. (2002) A note on the calculation of empirical P values from Monte Carlo procedures. *Am. J. Hum. Genet.*, **71**, 439.
22. Benjamin,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B*, **57**, 289–300.
23. Consortium,G. *et al.* (2015) The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, **348**, 648–660.
24. Alix-Panabières,C. and Pantel,K. (2014) Challenges in circulating tumour cell research. *Nat. Rev. Cancer*, **14**, 623.
25. Song,Y., Tian,T., Shi,Y., Liu,W., Zou,Y., Khajvand,T., Wang,S., Zhu,Z. and Yang,C. (2017) Enrichment and single-cell analysis of circulating tumor cells. *Chem. Sci.*, **8**, 1736–1751.
26. Zheng,Y., Miyamoto,D.T., Wittner,B.S., Sullivan,J.P., Aceto,N., Jordan,N.V., Yu,M., Karabacak,N.M., Comaills,V., Morris,R. *et al.* (2017) Expression of β-globin by cancer cells promotes cell survival during blood-borne dissemination. *Nat. Commun.*, **8**, 14344.