UNIVERSITY OF CALIFORNIA,

IRVINE

Computational Biomedicine via Single-Cell Analysis

THESIS

submitted in partial satisfaction of the requirements

for the degree of

MASTER OF SCIENCE

in Computer Science

by

Che Yu Lee

Thesis Committee:

Assistant Professor Jing Zhang, Chair

Assistant Professor Matthew J Girgenti, Yale University

Professor Xiaohui Xie

2023

DEDICATION

To

Prof Jing Zhang

and the Other Mentors I have met during my M.S. Degree

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

I thank all my colleagues at the Zhang Lab. I sincerely also want to thank all the mentors I have met during my Master of Science degree research experience.

I thank PLOS Computational Biology for permission to include copyrighted photographs in my own first-author publication as part of my master's thesis. Figures 8, 9, 10, & 11 were previously published CC by 4.0 in PLOS Computational Biology, used with permission from PLOS (https://creativecommons.org/licenses/by/4.0/). The co-authors listed in this publication are Yuhang Chen, Ziheng Duan, Min Xu, Matthew J. Girgenti, Ke Xu, Mark Gerstein, and Jing Zhang. Jing Zhang directed and supervised research which forms the basis for the thesis. Financial support was provided by the University of California, Irvine, Graduate Student Researcher (GSR) GSR-Tuit & Fee Rem (3284).

ABSTRACT OF THE THESIS

Computational Biomedicine via Single-Cell Analysis

by

Che Yu Lee

Master of Science in Computer Science

University of California, Irvine, 2023

Assistant Professor Jing Zhang, Chair

The advent of single-cell sequencing has allowed us to simultaneously capture transcripts in millions of cells, providing the opportunity to dissect important biological regulatory mechanisms at an unprecedented resolution. Unfortunately, computational modeling of single-cell data has faced several challenges. Specifically, it is sparse with many zeros, sensitive to numerous experimental confounding factors, and complicated with many non-linear biological interactions, making it hard for computational analysis. In the following three studies — across the tissue-cellular-DNA levels, we utilized biological information and mathematical models to address these computational challenges. Firstly, at the tissue level, we leveraged the resolution of single-cell sequencing to perform a novel cell-to-cell communication analysis to discover dysregulated communicating cell types. Secondly, at the cellular level, we performed a cell-type-specific analysis to identify key driver genes in Alcohol Use Disorder. Thirdly, at the DNA level, we developed a computational pipeline that studies virus infection that pinpoints retroviral integration sites at the genetic base pair resolution within specific cell types. By synergizing single-cell sequencing with tailored computational analyses, we pave the way for a new era in medicine, enabling physicians to practice with unparalleled insight and precision.

INTRODUCTION

In medicine, studying the brain is of utmost importance. Psychiatric conditions are widespread, with depression alone affecting over 264 million people globally, according to the World Health Organization (WHO). Schizophrenia is estimated to affect about 20 million people worldwide (Kahn et al., 2015). Substance abuse, including alcohol and illicit drugs, contributes to over 11 million deaths each year, and the global burden of disease attributable to substance abuse is substantial, with opioids and alcohol being significant contributors (Newton, 2018). Neurodegenerative diseases are also a growing concern, with Alzheimer's disease and other dementias affecting an estimated 50 million people worldwide—a number that is expected to triple by 2050 due to aging populations (DeTure & Dickson, 2019). These statistics underline the urgency of advancing our understanding and treatment of brain disorders. In computational biomedicine, the analysis of large-scale brain -omics using advanced algorithms could help identify patterns and risk factors, potentially leading to earlier intervention and more personalized approaches to treatment, ultimately aiming to mitigate the extensive individual and societal impacts of these conditions.

The recent single-cell sequencing technology has revolutionized genetic and genomic studies by simultaneously profiling molecular signatures across thousands to millions of cells (Zheng et al., 2017). It enables scientists to explore cellular diversity, gene expression patterns, and cellular interactions in complex tissues and health conditions, allowing us to identify unique cell types, discover disease-specific cellular signatures, and unravel the intricate mechanisms underlying genetic disorders. As a result, several single-cell genomic research studies have been conducted to investigate neurological disease pathology and provide new molecular insights. The complex and multidirectional interplay between these factors (and their properties) plays crucial roles in

tissue development, cellular responses, disease progression, and therapeutic interventions. Understanding and manipulating this relationship can provide insights into disease mechanisms and guide the development of novel therapeutic strategies.

At the tissue level, our research has embraced the high resolution of single-cell sequencing to forge a novel analytical approach to cell-to-cell communication for psychiatric diseases. Most of the existing studies solely focused on molecular perturbations within each individual cell. However, cells are not isolated entities but live in a microenvironment, or cell niche, composed of dynamically interacting entities, including extracellular matrix (ECM), neighboring cells, and soluble factors (Bloom & Zaman, 2014; Spill, Reynolds, Kamm, & Zaman, 2016). We leveraged the large-scale and publicly available single-nucleus RNA sequencing (snRNA-seq) in the human brain to investigate cell-to-cell communication patterns and their perturbations in diseased phenotypes. This has unearthed previously obscured cellular dialogues, shedding light on the dysregulated communication networks that may underpin complex psychiatric diseases.

At the cellular level, our focus shifted to deciphering the cryptic language of genes within the context of a subtype of substance abuse, Alcohol Use Disorder (AUD). Given the highly diversified nature of the affected biological processes, it is unlikely that one particular cell type is responsible for AUD pathology. The human brain is made up of a myriad of cell types and subtypes and several have been implicated in substance abuse pathology including both excitatory and inhibitory neurons, endothelial cells, and microglia that could be responsible for the changes observed (Hodge et al., 2019; Lake et al., 2018). By identifying the key driver genes, we have opened a window into the cellular mechanics that could be leveraged for therapeutic interventions.

Finally, moving even deeper, at the DNA level, our work has honed a computational pipeline with the precision to study viruses and pinpoint retroviral integration sites, including HIV-associated

Dementia. The once-in-a-century COVID-19 pandemic has shown the importance of studying viral infections (Ahmad, Haroon, Baig, & Hui, 2020). Venus takes advantage of single-cell sequencing for virus detection and integration site discovery. Specifically, Venus addresses two main questions: whether a tissue/cell type is infected by viruses or a virus of interest? And if infected, whether and where has the virus inserted itself into the human genome? This tool offers a magnified view into the viral landscape of infected cells, with implications that stretch far beyond the immediate study.

CHAPTER 1: Cell-to-cell Communication

Firstly, at the tissue level, we leveraged the resolution of single-cell sequencing to perform a novel cell-to-cell communication (C2C) analysis to discover dysregulated communicating cell types. We utilized the gene expression patterns of known ligand-receptor pairs from the snRNA-seq data to infer the C2C networks via popular software packages CellChat and NeuronChat (**Figure 1**) (Jin et al., 2021; Zhao, Johnston, Ren, Xu, & Nie, 2023). Specifically, we constructed a three-dimensional matrix representing the communication strength between any sender and receiver cell type pair via a specific ligand-receptor pair. Finally, we connected them with downstream risk genes via NicheNet (Browaeys, Saelens, & Saeys, 2020). As a result, this allowed us to aggregate the C2C communication patterns in diseased brains, measure C2C changes between conditions, infer disease-driving signal pathways, and connect ligand genes to downstream risk genes in a cell-type-specific manner. We will discuss the detailed results in the following sections.

Communication pattern analysis reveals inter-mixing of cell types and signaling pathways in brains affected with psychiatric disorders

With the 3D C2C matrix constructed, we first explored how multiple cell types coordinate intercellular communications using certain pathways in an unsupervised manner. To achieve this goal, we first flattened the 3D communication matrix into a 2D sender-by-LigandReceptorPair matrix and performed non-negative matrix factorization (NMF) to identify latent communication groups and their key ligand-receptor signaling contributors (Brunet, Tamayo, Golub, & Mesirov, 2004; D. D. Lee & Seung, 1999). We demonstrated our outgoing C2C network results in the alluvial plot, where the middle bar represents the latent patterns, and the flow indicates how different signaling pathways (or cell types) belong to each pattern. Interestingly, we found normal prefrontal cortices employ three distinct outgoing communication latent patterns in three major

cell groups, excitatory neurons, inhibitory neurons, and supporting cells. All of the outgoing supporting cells are characterized by pattern 1, dominated by biologically relevant pathways named after genes such as ANGPT, BMP, SPP1, and TGFβ (**Figure 2**). Inhibitory neurons are represented by pattern 2, driven by expected signaling pathways such as VIP, SST, CCK, and CRH while excitatory neurons are characterized by pattern 3, driven by signaling pathways such as CSF, SEMA3, and NRG. In contrast, we found that this pattern has been disrupted in Alzheimer's Disease (AD) prefrontal cortices. For instance, the inhibitory and excitatory neurons demonstrated mixed latent communication patterns (e.g., Chandelier cells have been grouped into excitatory patterns in Alzheimer's). In addition, the major driving signal pathways for different cell types also changed noticeably. For example, the WNT pathways became one major contributor to the excitatory group, while ANGPT switched from major contributors in supporting cells to the inhibitory group. Together, these results suggested extensive alterations in global C2C communication patterns and signaling usage in the outgoing network.

Cell type-centric cell-to-cell comparison highlights disturbed communication strength across various cell types in brains affected with psychiatric disorders

After checking the global C2C pattern perturbations, we focused on cell-type-centric communication changes by aggregating all Ligand receptor pairs in our 3D C2C matrix. In Post-Traumatic Stress Disorder (PTSD), we found that the INH SST cells have significant downregulation of neurotransmitter synthesis and transport enzymes resulting in a decrease in sender communication when compared to other neuronal cell types (**Figure 3A**). We found that the differential strength of communication from INH SST to every other neuronal cell type was downregulated, and we observed modest decreases with astrocytes, endothelial cells, and OPCs

(**Figure 3B**). Noticeably, the most downregulated communication occurs from INH SST to INH KCNG1 cells. We speculate that downregulated communication signaling is related to GABAergic transmission decreases from INH SST cells throughout the PFC and is consistent with previous findings in the PTSD brain.

Pathway-centric analysis of neuroinflammation and neuroprotection signaling in brains with psychiatric disorders are dis-regulated in a cell-type-specific manner

Our previous analyses mainly focused on the cell-type-level communication strength perturbations in the C2C network comparison without considering the impact of their communication pathways. To fill this gap, we also performed a signaling-pathway-centric analysis by evaluating the contribution of all involved ligand-receptor pairs. Simply, for each ligand-receptor interaction, we conducted a paired sample Wilcoxon signed-rank test comparing all possible sender-receiver cell type combinations between diseased and control groups (Conover, 1971). A significant P-value occurs when all the interactions belonging to one diagnosis rank lower than the interactions from the other diagnosis (**Figure 4A**). We chose to focus on 4 canonical, literature-driven ligand-receptor interactions for further analysis, namely the WNT, CSF, TGFβ, and CX3C pathways, which were all statistically significant.

Neuronal inflammation plays a significant role in the AD pathology. For instance, immune cells such as microglia respond to the accumulation of beta-amyloid plaques, a hallmark of Alzheimer's, by triggering an inflammatory response. Also, prolonged microglia activation can result in chronic inflammation, leading to neuronal damage and the exacerbation of plaque buildup, thus creating a vicious cycle. Consistently, we found that two inflammation-related pathways WNT and CSF are dysregulated in the C2C communication process in our analysis. For example, the WNT

signaling pathway plays multifaceted roles in CNS diseases by modulating neuroimmune interactions. We found that the WNT pathway has significantly reduced its involvement in C2C communication (30% of control, P=2.086e-7, **Figure 4B**), which has been primarily driven by the global reduction of communication usage from the sender endothelial cells to both inhibitory and excitatory neuron receivers. Mechanistically, the downregulation of the WNT ligand gene may cause overactivity of the lithium-targeted GSK3β enzyme, leading to changes in neurogenesis, inflammation, oxidative stress, and circadian dysregulation in neuronal cell types. Additionally, lines of literature reported that the CSF pathway is a well-known disease-related signaling pathway primarily involved in microglia. which can activate the recruitment of microglia and worsen inflammatory response. Consistently, we found that the CSF pathway has been significantly upregulated in AD patients (2.5x of control, P=0, **Figure 4C**). Such increased involvement is mainly driven by the increased communication from the excitatory neurons L6b to Microglia cells.

Next, we move on to neuroprotective signaling pathways. We observed the downregulation of TGFβ signaling in Alzheimer's in the communication to Micro/PVM cell type (60% of control, P=0, **Figure 4D**). A decreased TGFβ1 has been associated with a higher burden of Aβ in the parenchyma, which correlates with an increased microglia activation. The suppression of the neuroprotective role of the signaling pathway TGFβ1 against Aβ toxicity in the diseased cell types may be the molecular mechanism underlying the symptoms of Alzheimer's disease. Adding on, we also found the decrease of another neuroprotective signaling pathway, the CX3C pathway (70% of control, P=4.883e-2, **Figure 4E**). CX3CL1 has been demonstrated to play a neuroprotective role in CNS by reducing neurotoxicity and microglial activation. Our C2C analysis agrees with the literature as we see all communication in the CX3C pathway is directed to the Micro/PVM cell type. Moreover, with single-cell resolution, we can further see that this decrease happens primarily from the excitatory neurons to Micro/PVM. In summary, we discover that both

the signaling pathways that cause neuroinflammation and those that protect against it are regulated in a cell-type-specific manner.

Intracellular cell-to-cell communication analysis reveals a strong connection to neuroinflammatory psychiatric risk genes

Lastly, we extend our extracellular cell-to-cell communication analysis by considering related disruptions to intracellular signaling pathways. Specifically, for each ligand gene, a database of ligands regulating downstream genes is constructed with a regulatory potential score (Kanehisa & Goto, 2000). We perform a correlation test of each target gene's regulatory potential score with the actual gene expression to determine whether that ligand gene is important. By utilizing known risk genes and setting support cells (i.e., non-neurons) as the senders and neurons as the receivers, we find ligand-receptor links connecting risk genes to potential upstream effectors, such as FOXP1 and its ligand EBI3 in bipolar disorder and MECP2 and its ligand PDGFB in schizophrenia (**Figure 5A**, **B**).

CHAPTER 2: Single-cell Multiomic Analysis

Secondly, at the cellular level, we performed a cell-type-specific analysis to identify key driver genes in Alcohol Use Disorder. Alcohol Use Disorder (AUD) is a multigenic disorder occurring in the substance abuse of alcohol. Recent studies have begun to detail the molecular biology of the postmortem AUD brain using bulk-tissue transcriptomic and epigenetic analyses. However, given the array of AUD-perturbed molecular pathways identified thus far, it is unlikely that a single cell type is responsible. It is therefore necessary to uncover the individual cell types contributing to the molecular pathology of AUD (Akbarian et al., 2015). We performed a single-cell resolute transciptomic and epigenetic analysis (**Figure 6**). For gene expression (RNA), we tested whether differentially expressed genes and their pathways are enriched for specific biological functions in

each cell population (Li et al., 2020). For chromatin accessibility (ATAC), we will measure chromatin peaks in AUD that may affect gene expression (Granja et al., 2021; Stuart, Srivastava, Madad, Lareau, & Satija, 2021; Y. Zhang et al., 2008).

Covariate-corrected differential analysis of gene expression between Alcohol Use Disorder cases and controls

To better understand the cell type-specific biological processes affected by AUD, we first performed differential gene expression analysis systematically across all 7 brain cell type (17 sub cell types) clusters in the snRNA-seq dataset. We employed a method commonly used in the field: MAST with covariate correction (Finak et al., 2015). The covariates we employed in MAST included: age, sex, ancestry, PMI, and RIN. Specifically, we utilized a generalized linear model in which the first dimension was the condition of interest (AUD or Control) and the other dimensions were the covariates (McCullagh & Nelder, 1989). For each cluster, we report DEGs that were identified as overlapping between the two tests and shared directional fold change (FC) > 1.2 and FDR < 0.01. For our DEG analysis, we analyzed the canonical cell types but also examined gene expression in specific neuronal subtypes. Specifically, through the excitatory cell type, we found an upregulation of the ethanol metabolic enzyme, Aldehyde Dehydrogenase (**Figure 7**).

CHAPTER 3: Venus

Thirdly, at the DNA level, we developed a computational pipeline that studies virus infection that pinpoints retroviral integration sites at the genetic base pair resolution within specific cell types (Dobin & Gingeras, 2016; C. Y. Lee et al., 2022). Recent advances in single-cell RNA sequencing technologies have allowed us to simultaneously capture transcripts in millions of cells, providing the opportunity to dissect the transcriptome at a single-cell resolution. While several recent computational methods were developed to study viruses at a single-cell resolution, they failed to

identify the many integration-able viruses and report virus integration sites (**Figure 8**) (Chen et al., 2013; Yasumizu, Hara, Sakaguchi, & Ohkura, 2021). To address the aforementioned challenges, we developed Venus, an efficient Virus infection and fusion site detection method for both bulk-tissue and single-cell RNA-seq data. Venus consists of two main modules: virus detection and integration site discovery (**Figure 9**).

Computational parameters of Venus's Detection Module

Venus utilized a sequential analysis to detect viruses (**Fig 9A**). It first aligned reads to the human genome and then aligned the leftover unmapped reads to a mega-viral genome. Finally, the virusThreshold parameter removed viral species with low number of supporting reads (**S1 File**). What is most important will be the threshold set for transcript filtering. We recommend starting with a threshold of zero first and then deciding on a new threshold with the results. For single-cell data, barcode and UMI were specified while a whitelist was inputted if available.

Human genome (version GRCh38.p13) and annotation file (version GRCh38.p13) were downloaded from the GENCODE website. 7571 viral genomes were downloaded from NCBI and then concatenated to make the mega-virus index (annotation files were unavailable). Indices and reads were built and mapped using STAR version 2.7.9a (Dobin & Gingeras, 2016). To perform a limited amount of benchmarking on the detection module, we dropped out a certain portion of the reads and found that viral detection decreased with increasing dropout percentage (**S4 File**).

Computational parameters of Venus's Integration Module

After detecting the virus of interest (target virus), we further developed efficient pipelines for integration site discovery. Specifically, Venus contained three steps for accurate integration site detection, as shown (**Fig 9B**). Parameters used are described and bolded (**S2 File**). What is most

important in the integration module will be the integrSeq.fna file, which contains biological sequences Venus should specifically look for in its fusion sites to classify meaningful integration sites. For HIV and other retroviruses, this will be the LTR sequences. Firstly, Venus selected the reads mappable to the target virus genome as the starting point for maximum processing efficiency because viruses have smaller genomes than humans and mapping first to the virus genome without splicing increases detection sensitivity. Secondly, the virus-mappable reads were then mapped with splicing to a custom hybrid genome, made from concatenating human and target viral fasta/gtf files. Thirdly, chimeric fusion transcripts were sorted and classified based on the integrSeq parameter to provide biologically relevant integration sites.

Complexity analysis of Venus

We performed runtime and memory analyses on a downsampled HIV-infected T-cell dataset with 16 CPUs and 64 GB RAM. Runtime linearly depended on the number of reads, while memory remained constant at 30 GB, the size of the human genome (**S3 File**). A short list of Venus's software dependencies includes STAR, Samtools, and Numpy, but a full list can be found on our GitHub page. For hardware dependencies, Venus needs to have a writing disk space of 100GB while around 30GB for RAM, ideally with at least 8 parallel threads for timely analysis.

Venus precisely identified HIV-infected cells at a single-cell resolution in monocytes at various stages of maturity

We first tested Venus's detection module (**Figure 9A**). We demonstrated Venus's single-cell capability by analyzing a HIV-infected single-cell dataset, which had 8 uninfected samples as controls, 24 HIV-infected as treatment one, and another 24 HIV-infected but AntiRetroviral Therapy-treated (ART) as treatment two (**Figure 10A**, **B**) (León-Rivera, Morsey, Niu, Fox, & Berman, 2020). As expected, Venus found no viral load in all control samples, high viral load in

treatment one, and low viral load in treatment two. Non-ART treated patients had a range of 531 to 2670 HIV transcripts, significantly higher than those from ART-treated patients with 7 to 198 HIV transcripts. Expectedly, ART treatment significantly suppressed viral load, exhibiting Venus's accurate detection capability in a single-cell setting.

To visualize Venus's single-cell capability, we labeled each infected cell with Venus-generated output to produce a UMAP plot in Seurat (**Figure 10C**) (Butler, Hoffman, Smibert, Papalexi, & Satija, 2018). Out of the 25,211 cells that had passed Seurat's default filters, 1056 cells harbored HIV transcripts. And after clustering, 12 different gene-expression groups of monocytes were found. While there was no preference for infection toward any of the 12 different clusters, it exhibits Venus's capability to provide a single-cell resolution picture of viral infection.

Venus discovered HIV integration sites with varying biological significance and confidence in T-cells

We then tested Venus's integration site discovery module (**Figure 9b**). Lines of literature have highlighted the importance of virus integration sites due to their strong linkage to viral persistence, especially in the incessant HIV/AIDs epidemic. Despite this, integration sites are often falsely concluded due to library preparation and sequencing artifacts. To address these challenges, Venus classified HIV fusion transcripts into three categories based on biological relevance: Class I) fusion sites with human sequence reading into HIV's U3 sequence, HIV's U5 reading into human sequence, or splice donor-acceptor pairs; Class II) fusion sites with the aforementioned sequences but reading into noncoding human regions; Class III) fusion sites mapped to the middle of HIV genes (**Figure 11**).

In the HIV-infected T-cells dataset, Venus found 17 Class I, 2 Class II, and 6116 Class III integration sites (Liu et al., 2020). We were confident that the first two classes of fusion sites were integration sites because of three telltale signs: 1) Unmatched sequences overlay perfectly onto the opposite specie's reference; 2) Reads switch sharply in the middle between species, labeled by the red triangle breakpoints; 3) Nucleotides match the canonical U3 and U5 sequences used in HIV's integration events (**Figure 11A**, **B**) (J. Zhang & Crumpacker, 2022). Indeed, all three signs together showed that biologically-accurate integration sites were detected. Integration sites are inherently very difficult to detect, requiring a sequencing depth of 10X coverage. While it may be interesting to compare across datasets, of the three HIV datasets studied, namely brain, monocytes, and T cells, only T cells were sequenced deeply enough to detect such integration sites.

While both Venus's integration site classification algorithm and visualization capability were used to obtain high-confidence integration sites, they were also used to discard biologically irrelevant fusion sites. In contrast to Class I and IIs, Class IIIs likely signified partial integrations and sequencing artifacts due to their HIV gene disruptions (**Figure 11C**). With the guide integrSeq parameter and subsequent visualization in IGV, Venus reduced the large amount of noise inherent to viral integration site discovery (Robinson et al., 2011). We have provided a visualization capability in Venus because we understood viral integration events may vary from virus to virus, thus wishing to rest the final decision to each user. In conclusion, not only could Venus detect chimeric fusion transcripts but also was it able to classify them into biologically meaningful integration sites.

CONCLUSION

The leap forward afforded by single-cell sequencing has revolutionized our understanding of cellular complexity and the intricacies of gene expression. This technological marvel has provided us with the tools to delve into the biological labyrinth at a level of detail that was once beyond our reach. However, the path has been challenging. Computational modeling of such intricate data sets has been hampered by sparsity, confounding experimental variables, and the inherently non-linear nature of biological systems. Despite these hurdles, the studies presented herein have not only navigated these challenges but have also turned them into opportunities for innovation and discovery.

At the tissue level, our research has embraced the high resolution of single-cell sequencing to forge a novel analytical approach to cell-to-cell communication. This has unearthed previously obscured cellular dialogues, shedding light on the dysregulated communication networks that may underpin complex psychiatric disorders. At the cellular level, our focus shifted to deciphering the cryptic language of genes within the context of Alcohol Use Disorder. By identifying the key driver genes, we have opened a window into the cellular mechanics that could be leveraged for therapeutic interventions. Moving even deeper, at the DNA level, our work has honed a computational pipeline with the precision to pinpoint retroviral integration sites. This tool offers a magnified view into the viral landscape of infected cells, with implications that stretch far beyond the immediate study.

In sum, the confluence of single-cell sequencing with sophisticated computational strategies holds the promise of a transformative shift in medical practice. The power to dissect and understand the cellular and molecular underpinnings of disease at such a granular level stands to usher in an era of precision medicine unlike any before. Physicians armed with this knowledge can tailor

treatments to the individual, not just the illness, turning the tide in the fight against myriad diseases. As we stand on the brink of this new medical horizon, it is clear that the integration of advanced sequencing technologies and computational analysis will be the cornerstone of future biomedical breakthroughs, promising better outcomes for patients and a more nuanced understanding of the living tapestry that is human health.

REFERENCES

Ahmad, T., Haroon, Baig, M., & Hui, J. (2020). Coronavirus Disease 2019 (COVID-19) Pandemic and Economic Impact. *Pak J Med Sci, 36*(COVID19-S4), S73-S78. doi:10.12669/pjms.36.COVID19-S4.2638

Akbarian, S., Liu, C., Knowles, J. A., Vaccarino, F. M., Farnham, P. J., Crawford, G. E., . . . Consortium, P. (2015). The PsychENCODE project. *Nat Neurosci, 18*(12), 1707-1712. doi:10.1038/nn.4156

Bloom, A. B., & Zaman, M. H. (2014). Influence of the microenvironment on cell fate determination and migration. *Physiol Genomics, 46*(9), 309-314. doi:10.1152/physiolgenomics.00170.2013

Browaeys, R., Saelens, W., & Saeys, Y. (2020). NicheNet: modeling intercellular communication by linking ligands to target genes. *Nat Methods, 17*(2), 159-162. doi:10.1038/s41592-019-0667-5

Brunet, J. P., Tamayo, P., Golub, T. R., & Mesirov, J. P. (2004). Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci U S A, 101*(12), 4164-4169. doi:10.1073/pnas.0308531101

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., & Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol, 36*(5), 411-420. doi:10.1038/nbt.4096

Chen, Y., Yao, H., Thompson, E. J., Tannir, N. M., Weinstein, J. N., & Su, X. (2013). VirusSeq: software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue. *Bioinformatics, 29*(2), 266-267. doi:10.1093/bioinformatics/bts665

Conover, W. J. (1971). *Practical nonparametric statistics*. New York: Wiley.

DeTure, M. A., & Dickson, D. W. (2019). The neuropathological diagnosis of Alzheimer's disease. *Mol Neurodegener, 14*(1), 32. doi:10.1186/s13024-019-0333-5

Dobin, A., & Gingeras, T. R. (2016). Optimizing RNA-Seq Mapping with STAR. *Methods Mol Biol, 1415*, 245-262. doi:10.1007/978-1-4939-3572-7_13

Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., . . . Gottardo, R. (2015). MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol, 16*, 278. doi:10.1186/s13059-015-0844-5

Granja, J. M., Corces, M. R., Pierce, S. E., Bagdatli, S. T., Choudhry, H., Chang, H. Y., & Greenleaf, W. J. (2021). ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat Genet, 53*(3), 403-411. doi:10.1038/s41588-021-00790-6

Hodge, R. D., Bakken, T. E., Miller, J. A., Smith, K. A., Barkan, E. R., Graybuck, L. T., . . . Lein, E. S. (2019). Conserved cell types with divergent features in human versus mouse cortex. *Nature, 573*(7772), 61-68. doi:10.1038/s41586-019-1506-7

Jin, S., Guerrero-Juarez, C. F., Zhang, L., Chang, I., Ramos, R., Kuan, C. H., . . . Nie, Q. (2021). Inference and analysis of cell-cell communication using CellChat. *Nat Commun, 12*(1), 1088. doi:10.1038/s41467-021-21246-9

Kahn, R. S., Sommer, I. E., Murray, R. M., Meyer-Lindenberg, A., Weinberger, D. R., Cannon, T. D., . . . Insel, T. R. (2015). Schizophrenia. *Nat Rev Dis Primers, 1*, 15067. doi:10.1038/nrdp.2015.67

Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res, 28*(1), 27-30. doi:10.1093/nar/28.1.27

Lake, B. B., Chen, S., Sos, B. C., Fan, J., Kaeser, G. E., Yung, Y. C., . . . Zhang, K. (2018). Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat Biotechnol, 36*(1), 70-80. doi:10.1038/nbt.4038

Lee, C. Y., Chen, Y., Duan, Z., Xu, M., Girgenti, M. J., Xu, K., . . . Zhang, J. (2022). Venus: An efficient virus infection detection and fusion site discovery method using single-cell and bulk RNA-seq data. *PLoS Comput Biol, 18*(10), e1010636. doi:10.1371/journal.pcbi.1010636

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature, 401*(6755), 788-791. doi:10.1038/44565

León-Rivera, R., Morsey, B., Niu, M., Fox, H. S., & Berman, J. W. (2020). Interactions of Monocytes, HIV, and ART Identified by an Innovative scRNAseq Pipeline: Pathways to Reservoirs and HIV-Associated Comorbidities. *mBio, 11*(4). doi:10.1128/mBio.01037-20

Li, B., Gould, J., Yang, Y., Sarkizova, S., Tabaka, M., Ashenberg, O., . . . Regev, A. (2020). Cumulus provides cloud-based data analysis for large-scale single-cell and single-nucleus RNA-seq. *Nat Methods, 17*(8), 793-798. doi:10.1038/s41592-020-0905-x

Liu, R., Yeh, Y. J., Varabyou, A., Collora, J. A., Sherrill-Mix, S., Talbot, C. C., . . . Ho, Y. C. (2020). Single-cell transcriptional landscapes reveal HIV-1-driven aberrant host gene transcription as a potential therapeutic target. *Sci Transl Med, 12*(543). doi:10.1126/scitranslmed.aaz0802

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). Boca Raton ; London ; New York: Chapman and Hall.

Newton, D. E. (2018). *The opioid crisis : a reference handbookContemporary world issues* (pp. 1 online resource (xvi, 358 pages)).

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nat Biotechnol, 29*(1), 24-26. doi:10.1038/nbt.1754

Spill, F., Reynolds, D. S., Kamm, R. D., & Zaman, M. H. (2016). Impact of the physical microenvironment on tumor progression and metastasis. *Curr Opin Biotechnol, 40*, 41-48. doi:10.1016/j.copbio.2016.02.007

Stuart, T., Srivastava, A., Madad, S., Lareau, C. A., & Satija, R. (2021). Single-cell chromatin state analysis with Signac. *Nat Methods, 18*(11), 1333-1341. doi:10.1038/s41592-021-01282-5

Yasumizu, Y., Hara, A., Sakaguchi, S., & Ohkura, N. (2021). VIRTUS: a pipeline for comprehensive virus analysis from conventional RNA-seq data. *Bioinformatics, 37*(10), 1465-1467. doi:10.1093/bioinformatics/btaa859

Zhang, J., & Crumpacker, C. (2022). HIV UTR, LTR, and Epigenetic Immunity. *Viruses, 14*(5). doi:10.3390/v14051084

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., . . . Liu, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol, 9*(9), R137. doi:10.1186/gb-2008-9-9-r137

Zhao, W., Johnston, K. G., Ren, H., Xu, X., & Nie, Q. (2023). Inferring neuron-neuron communications from single-cell transcriptomics through NeuronChat. *Nat Commun, 14*(1), 1128. doi:10.1038/s41467-023-36800-w

Zheng, G. X., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., . . . Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nat Commun, 8*, 14049. doi:10.1038/ncomms14049

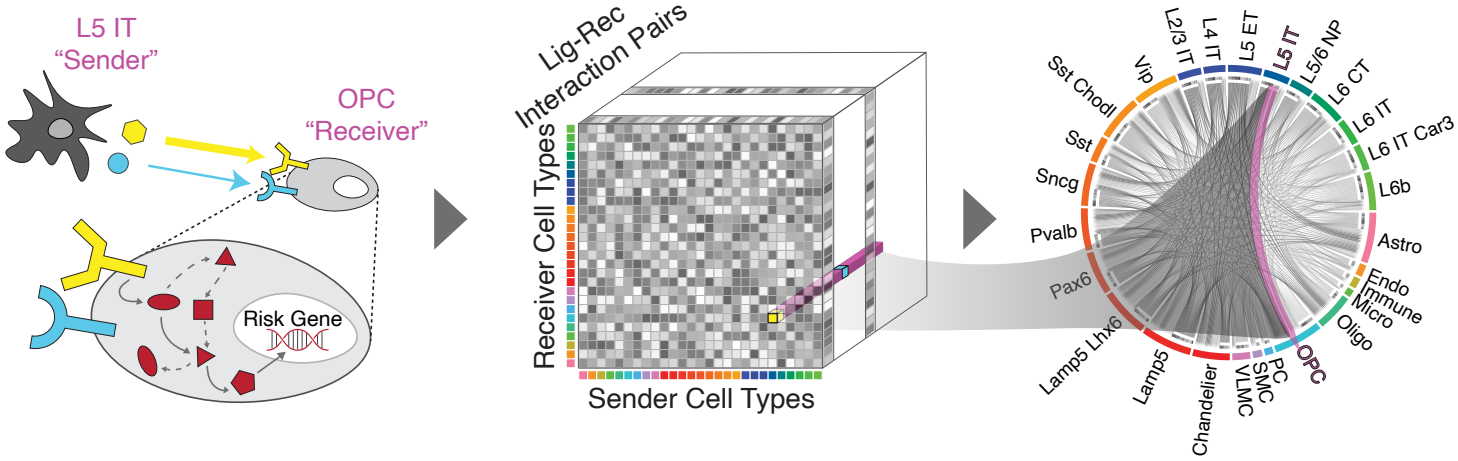

L5 IT
"Sender"
OPC
"Receiver"
Risk Gene
Lig-Rec
Interaction Pairs
Receiver Cell Types
Sender Cell Types
L2/3 IT
L4 IT
L5 ET
L5 IT
L5/6 NP
L6 CT
L6 IT
L6 IT Car3
L6b
Astro
Endo
Immune
Micro
Oligo
OPC
PC
SMC
VLMC
Chandelier
Lamp5
Lamp5 Lhx6
Pax6
Pvalb
Sncg
Sst
Sst Chodl
Vip

Figure 1

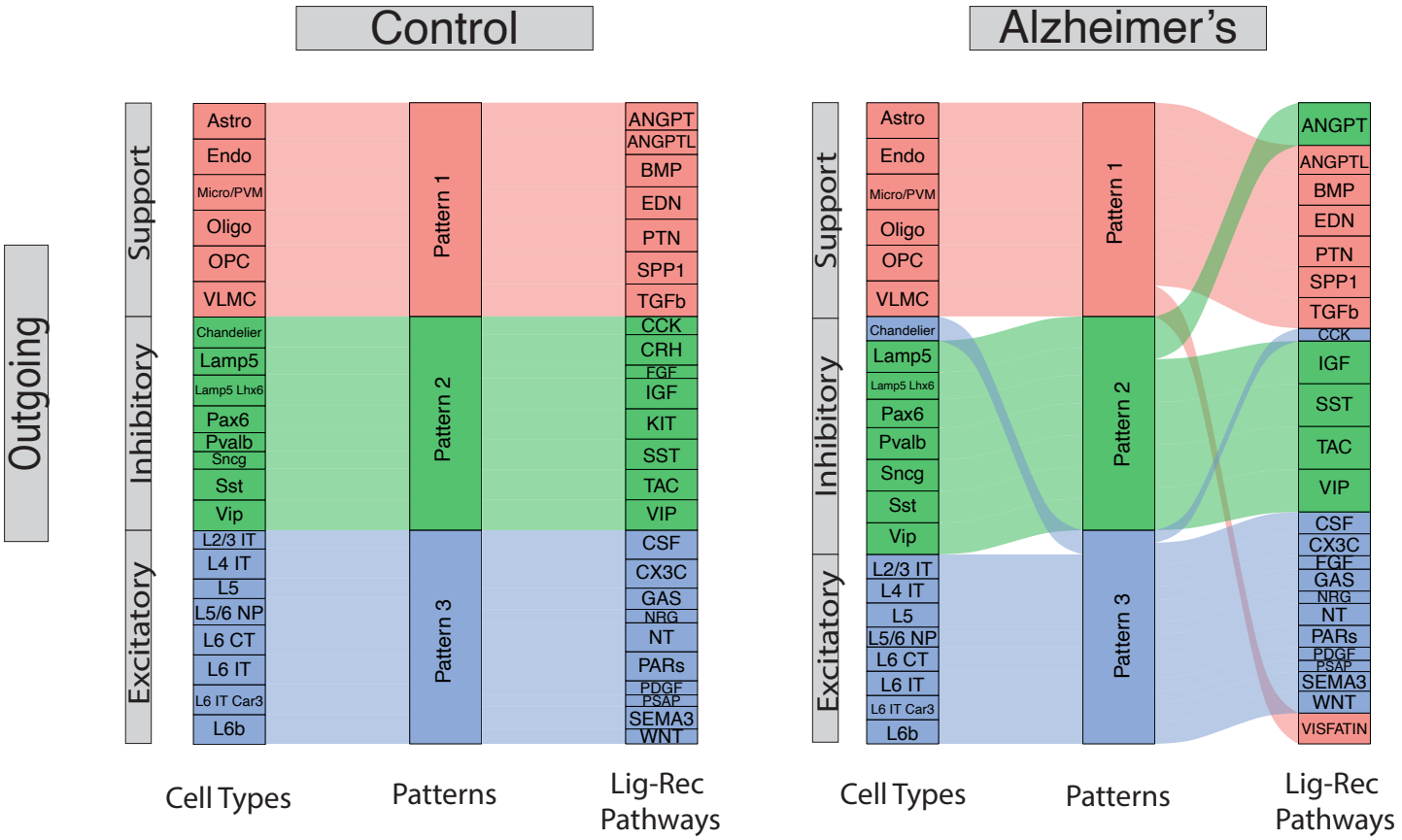

Control
Alzheimer's
Outgoing
Support
Inhibitory
Excitatory
Astro
Endo
Micro/PVM
Oligo
OPC
VLMC
Chandelier
Lamp5
Lamp5 Lhx6
Pax6
Pvalb
Sncg
Sst
Vip
L2/3 IT
L4 IT
L5
L5/6 NP
L6 CT
L6 IT
L6 IT Car3
L6b
Pattern 1
Pattern 2
Pattern 3
ANGPT
ANGPTL
BMP
EDN
PTN
SPP1
TGFb
CCK
CRH
FGF
IGF
KIT
SST
TAC
VIP
CSF
CX3C
GAS
NRG
NT
PARs
PDGF
PSAP
SEMA3
WNT
VISFATIN
Cell Types
Patterns
Lig-Rec Pathways

Figure 2

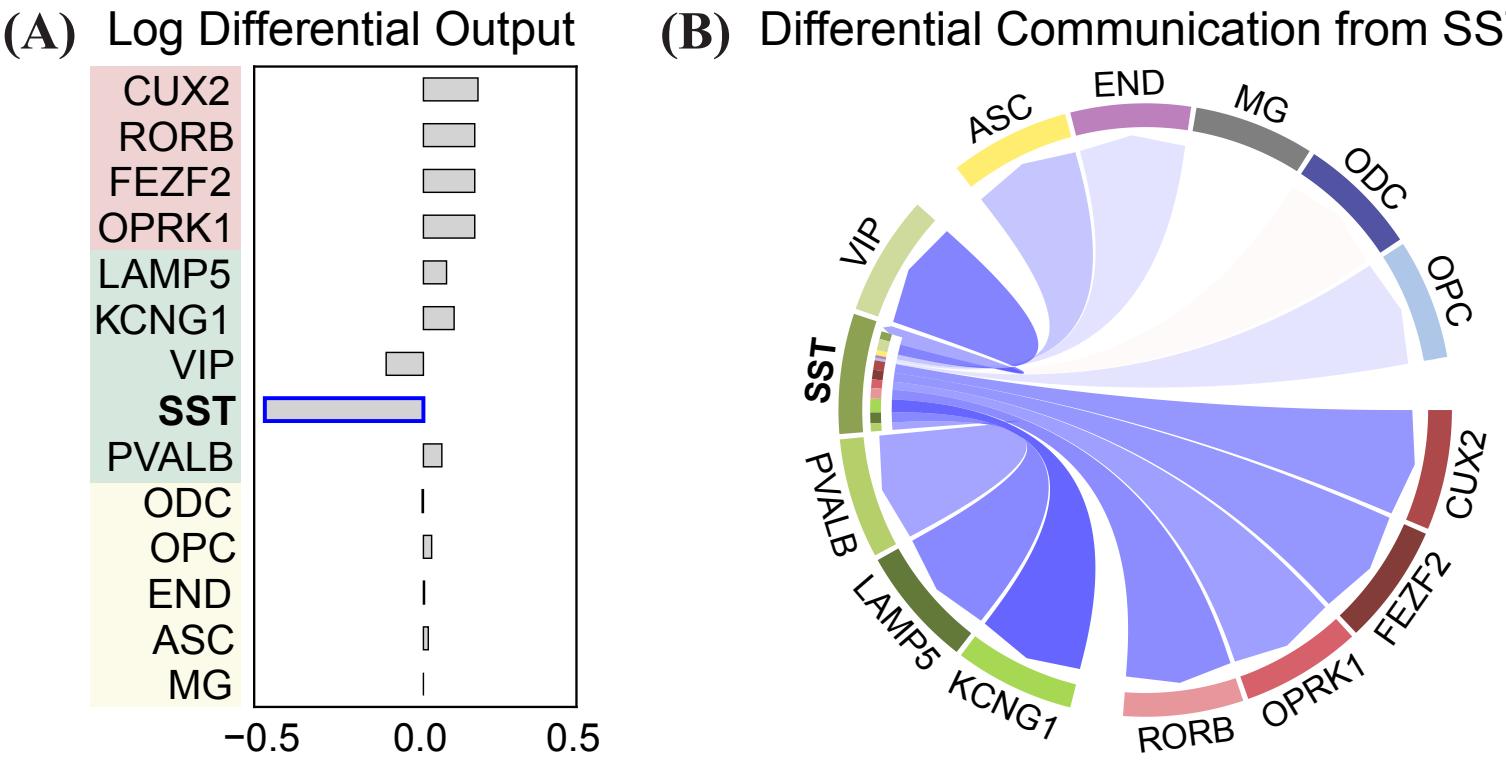

(A) Log Differential Output
CUX2
RORB
FEZF2
OPRK1
LAMP5
KCNG1
VIP
SST
PVALB
ODC
OPC
END
ASC
MG
−0.5
0.0
0.5
(B) Differential Communication from SST
END
MG
ODC
OPC
CUX2
FEZF2
OPRK1
RORB
KCNG1
LAMP5
PVALB
SST
VIP
ASC

Figure 3

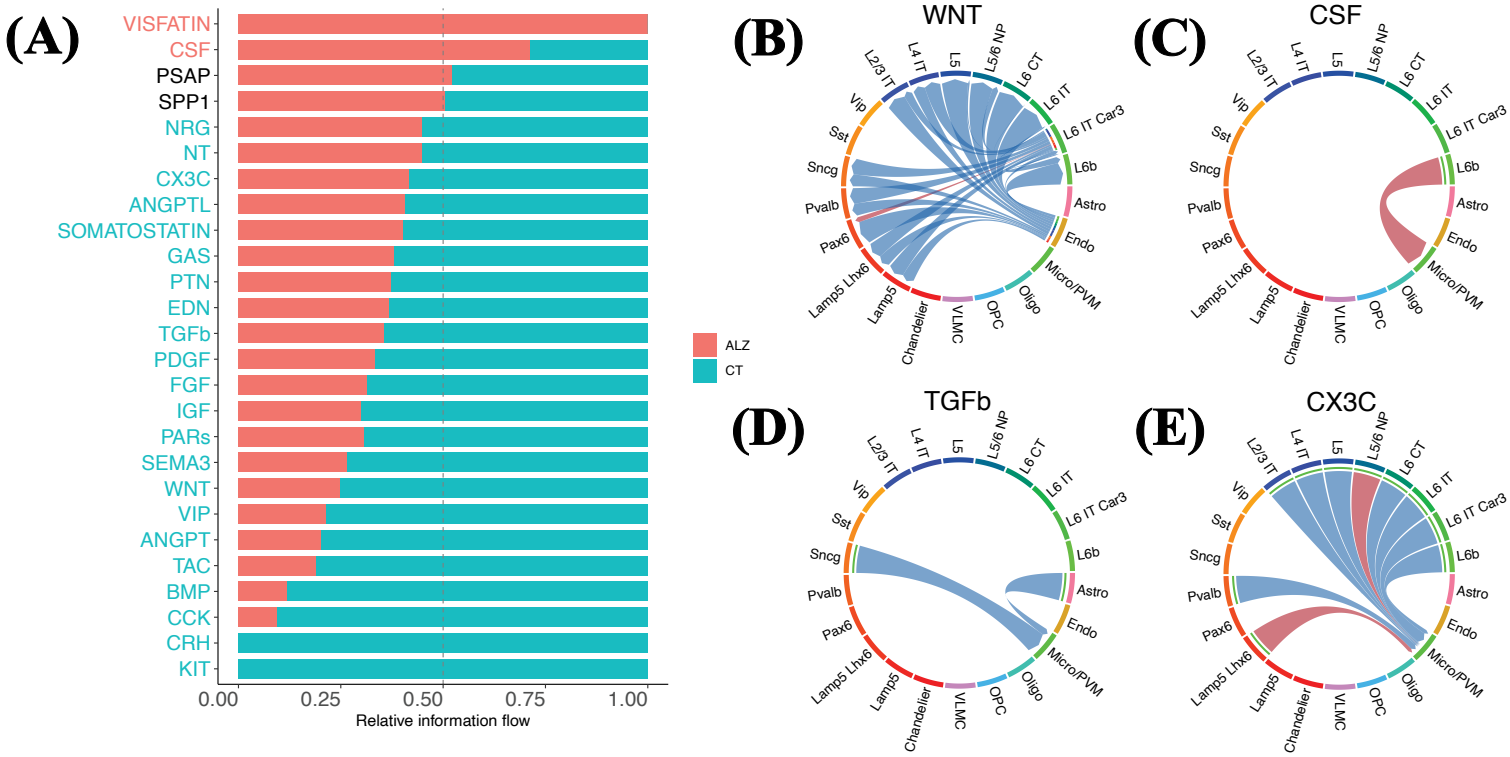

(A)
VISFATIN
CSF
PSAP
SPP1
NRG
NT
CX3C
ANGPTL
SOMATOSTATIN
GAS
PTN
EDN
TGFb
PDGF
FGF
IGF
PARs
SEMA3
WNT
VIP
ANGPT
TAC
BMP
CCK
CRH
KIT
0.00
0.25
0.50
0.75
1.00
Relative information flow
ALZ
CT
(B)
WNT
(C)
CSF
(D)
TGFb
(E)
CX3C
L2/3 IT
L4 IT
L5
L5/6 NP
L6 CT
L6 IT
L6 IT Car3
L6b
Astro
Endo
Micro/PVM
Oligo
OPC
VLMC
Chandelier
Lamp5
Lamp5 Lhx6
Pax6
Pvalb
Sncg
Sst
Vip

Figure 4

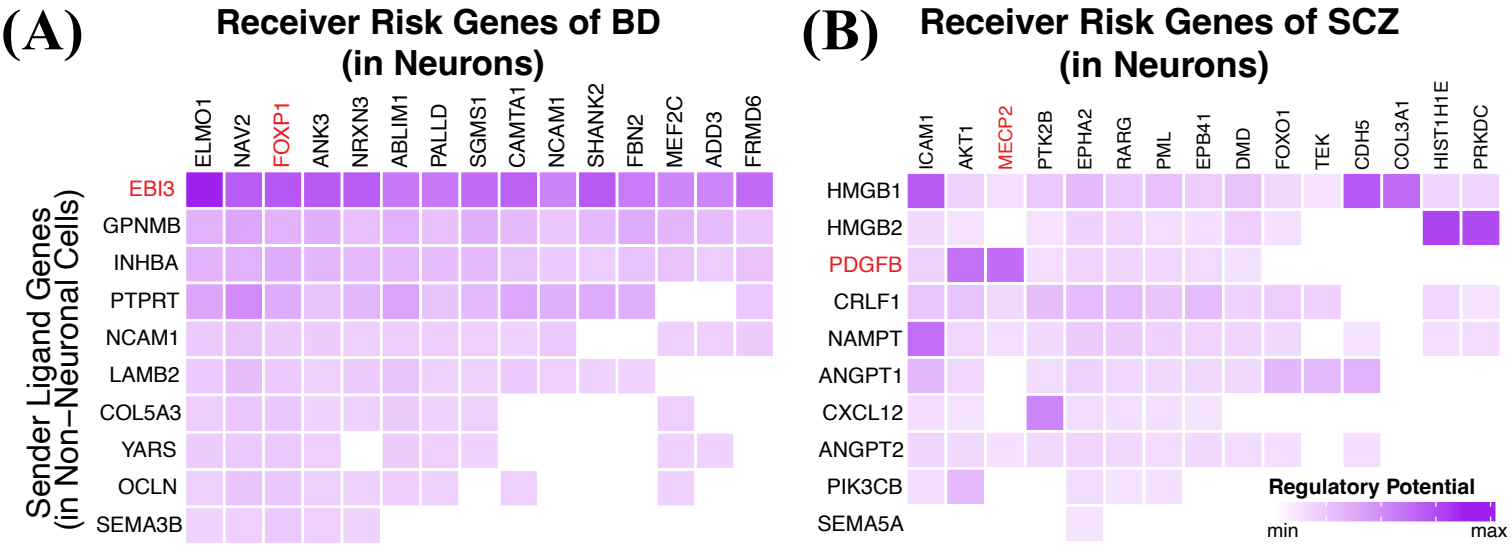

(A)
Receiver Risk Genes of BD
(in Neurons)
ELMO1
NAV2
FOXP1
ANK3
NRXN3
ABLIM1
PALLD
SGMS1
CAMTA1
NCAM1
SHANK2
FBN2
MEF2C
ADD3
FRMD6
Sender Ligand Genes
(in Non−Neuronal Cells)
EBI3
GPNMB
INHBA
PTPRT
NCAM1
LAMB2
COL5A3
YARS
OCLN
SEMA3B
(B)
Receiver Risk Genes of SCZ
(in Neurons)
ICAM1
AKT1
MECP2
PTK2B
EPHA2
RARG
PML
EPB41
DMD
FOXO1
TEK
CDH5
COL3A1
HIST1H1E
PRKDC
HMGB1
HMGB2
PDGFB
CRLF1
NAMPT
ANGPT1
CXCL12
ANGPT2
PIK3CB
SEMA5A
Regulatory Potential
min
max

Figure 5

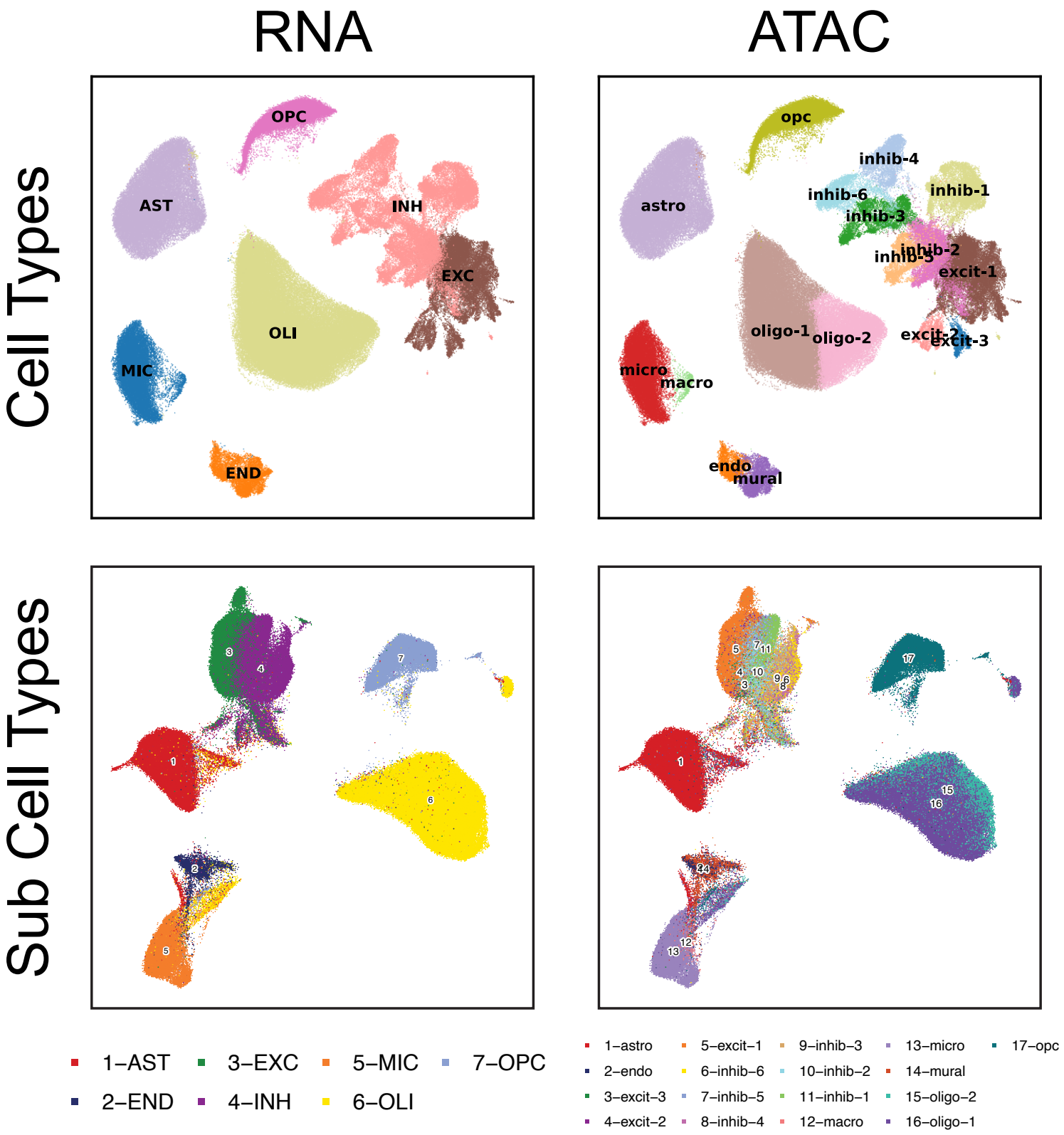
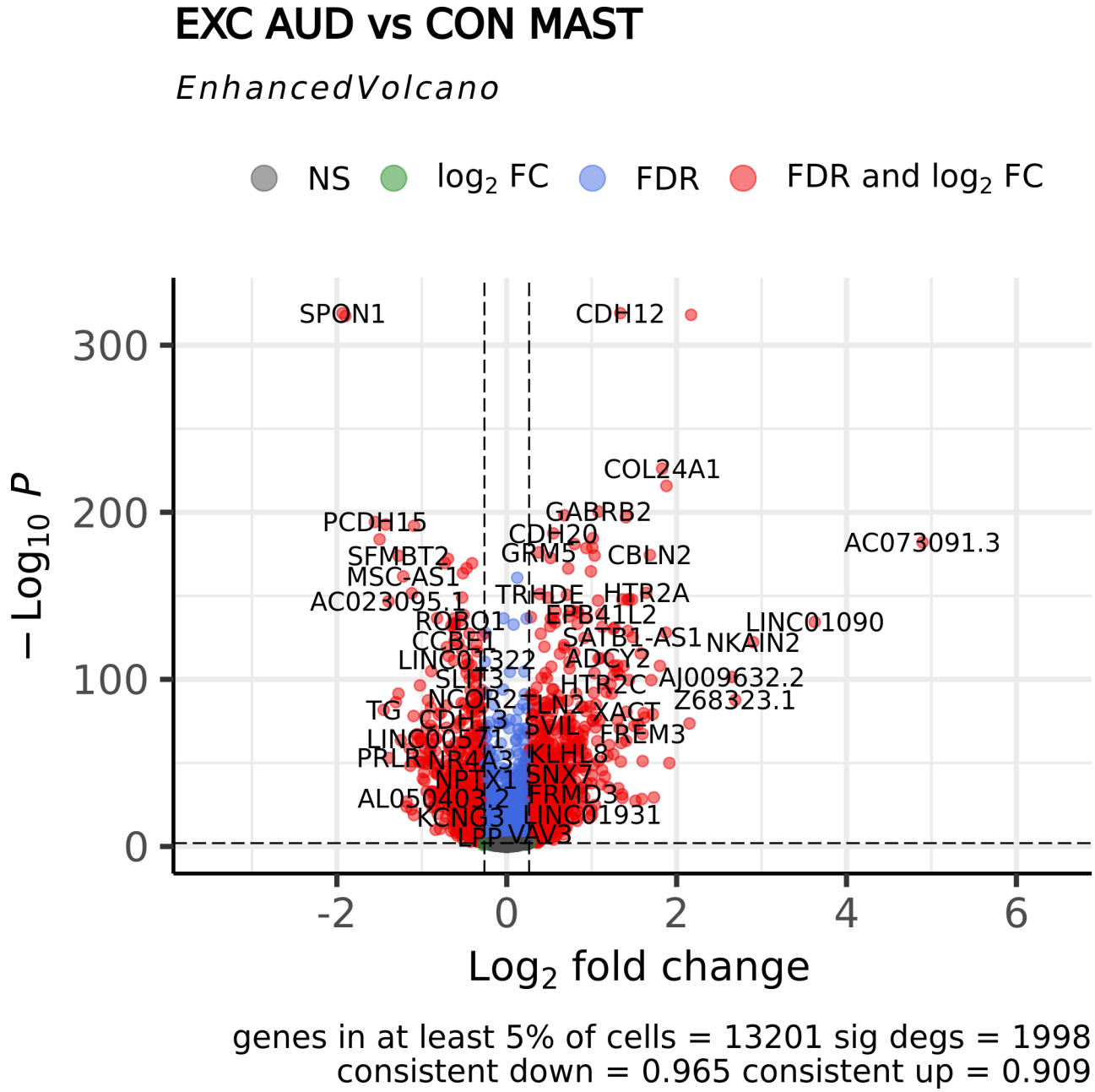

RNA
ATAC
Cell Types
Sub Cell Types
OPC
AST
INH
EXC
OLI
MIC
END
opc
astro
inhib-4
inhib-6
inhib-1
inhib-3
inhib-2
inhib-5
excit-1
oligo-1
oligo-2
excit-2
excit-3
micro
macro
endo
mural
1–AST
3–EXC
5–MIC
7–OPC
2–END
4–INH
6–OLI
1–astro
5–excit–1
9–inhib–3
13–micro
17–opc
2–endo
6–inhib–6
10–inhib–2
14–mural
3–excit–3
7–inhib–5
11–inhib–1
15–oligo–2
4–excit–2
8–inhib–4
12–macro
16–oligo–1

Figure 6



genes in at least 5% of cells = 13201 sig degs = 1998
consistent down = 0.965 consistent up = 0.909

Figure 7

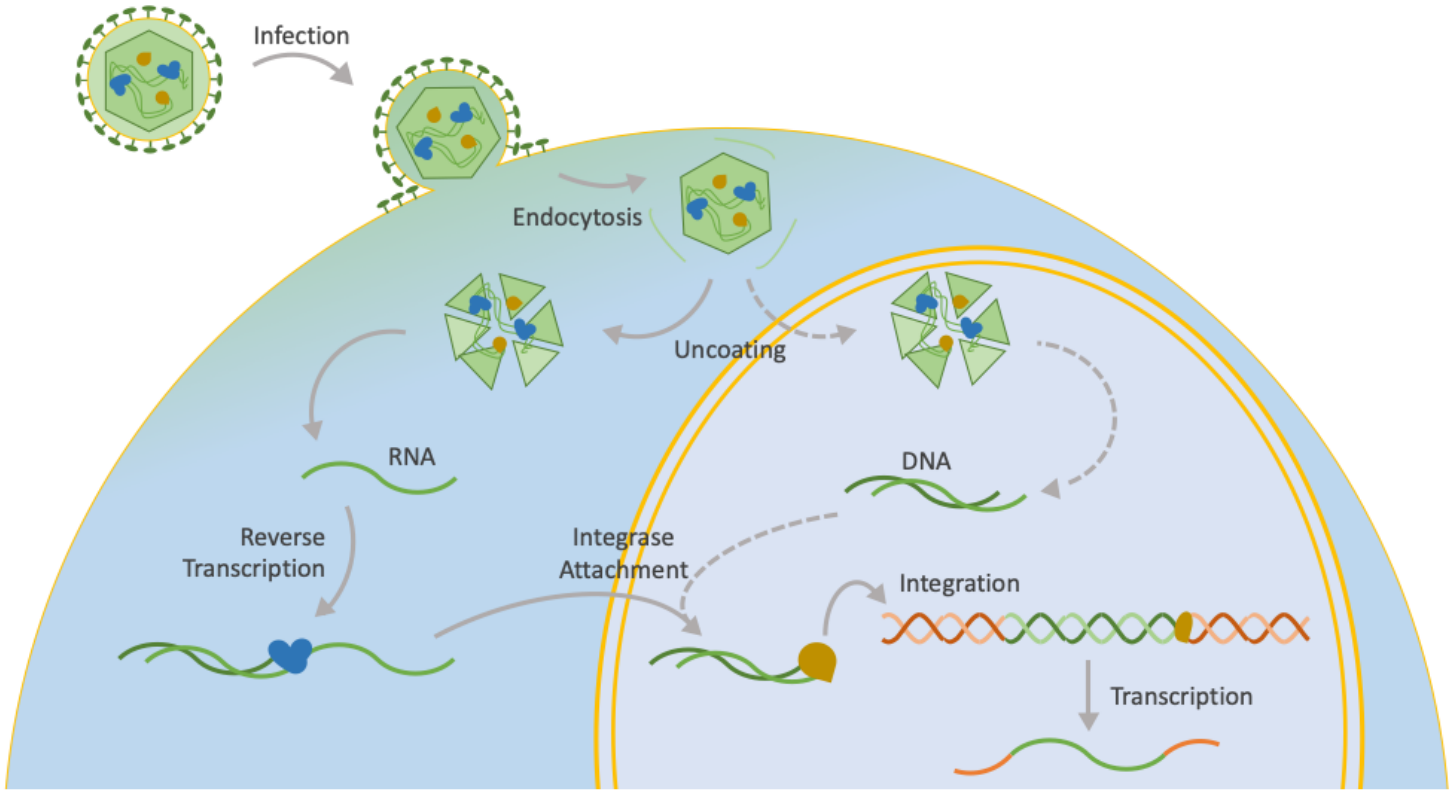


Figure 8 by Che Yu Lee et. al, 2022, is licensed CC by 4.0, available from https://doi.org/10.1371/journal.pcbi.1010636.

Figure 8

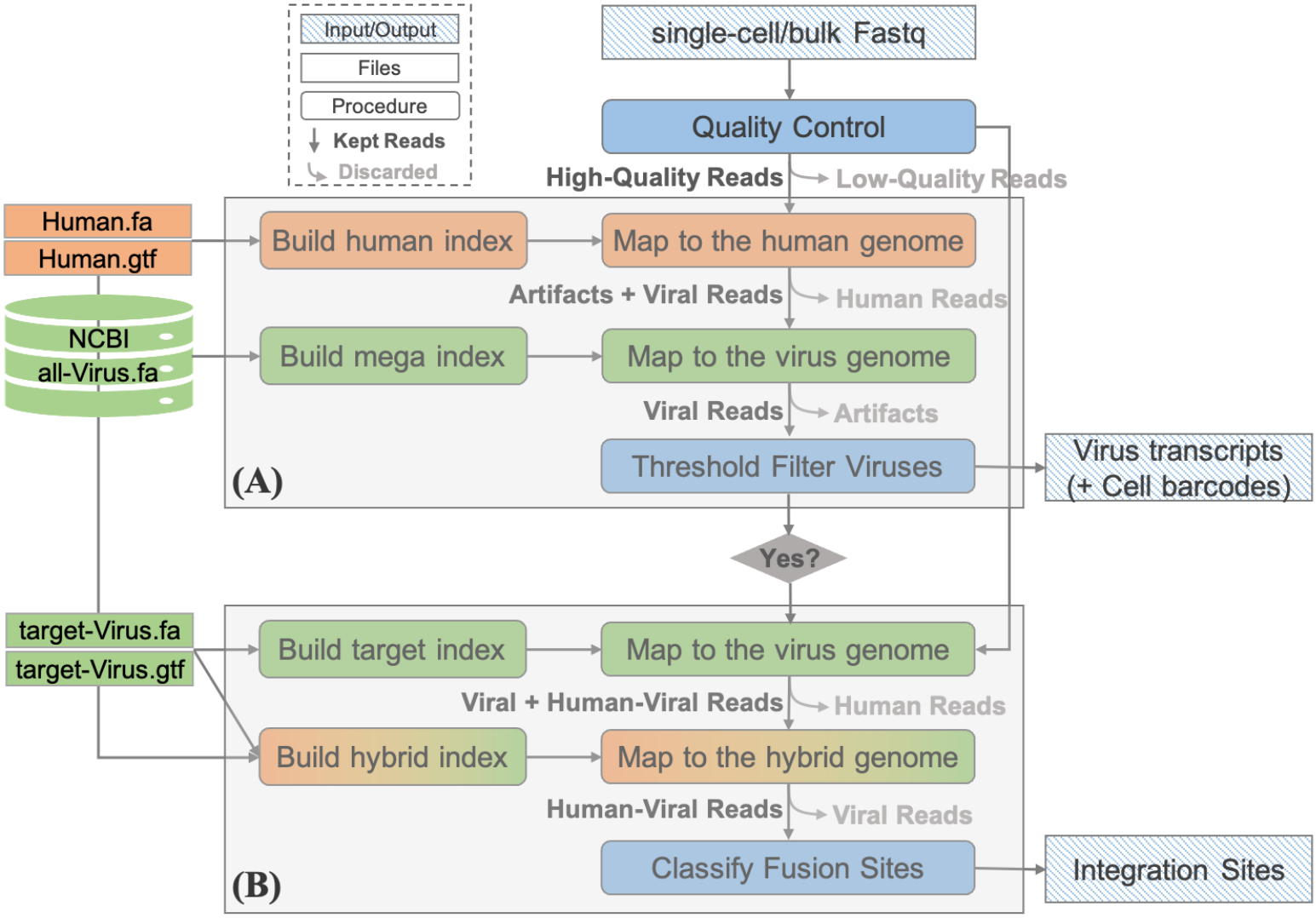


Figure 9

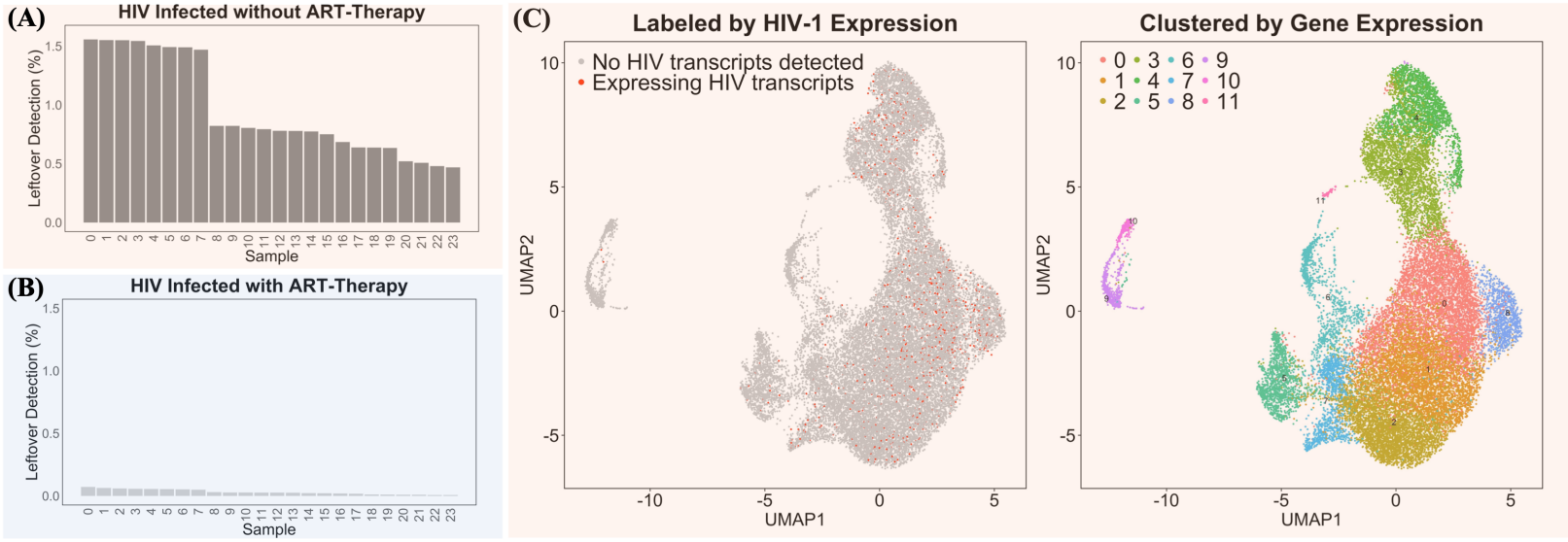


Figure 10

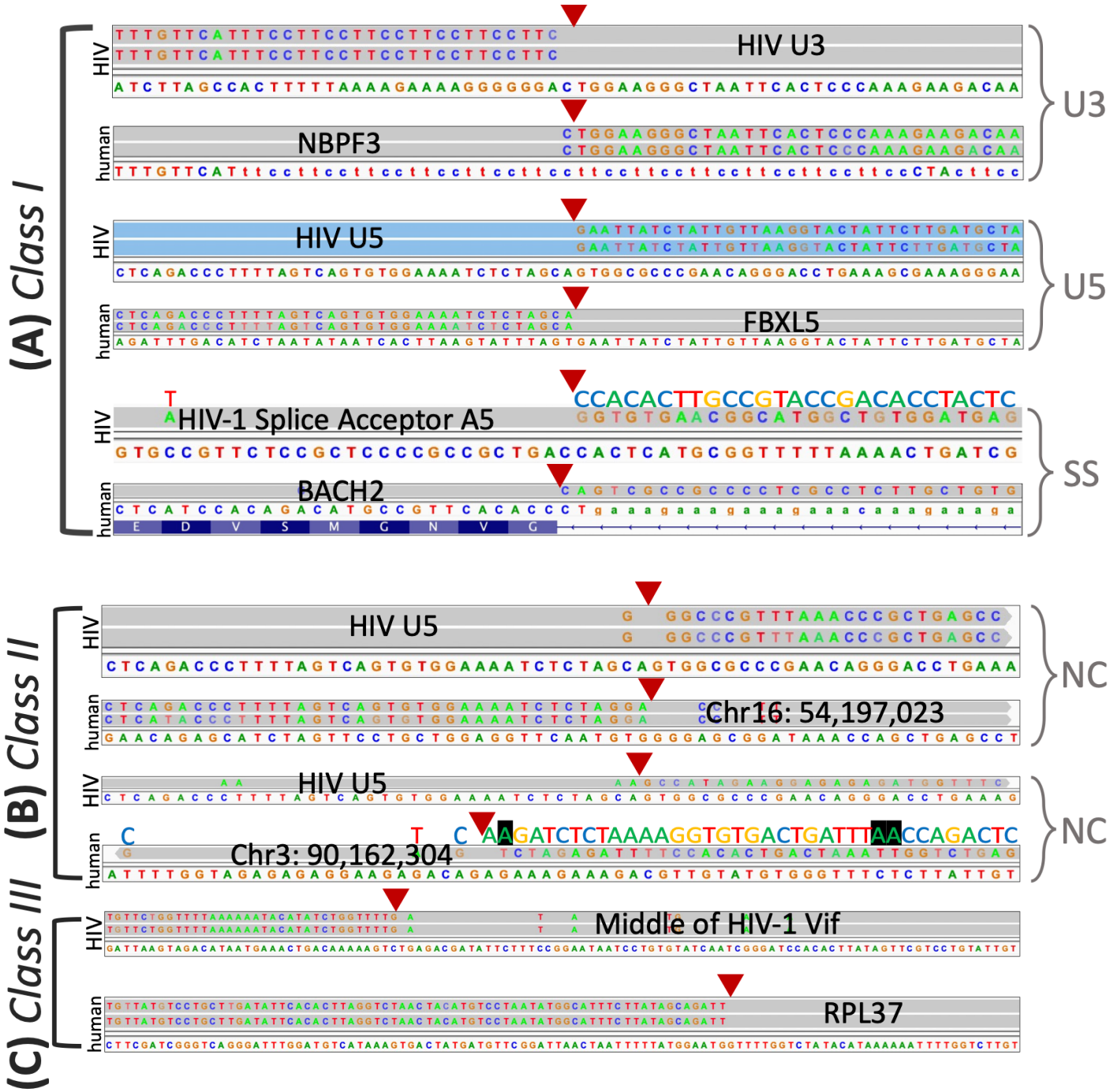


Figure 11 by Che Yu Lee et. al, 2022, is licensed CC by 4.0, available from https://doi.org/10.1371/journal.pcbi.1010636.

Figure 11

APPENDIX

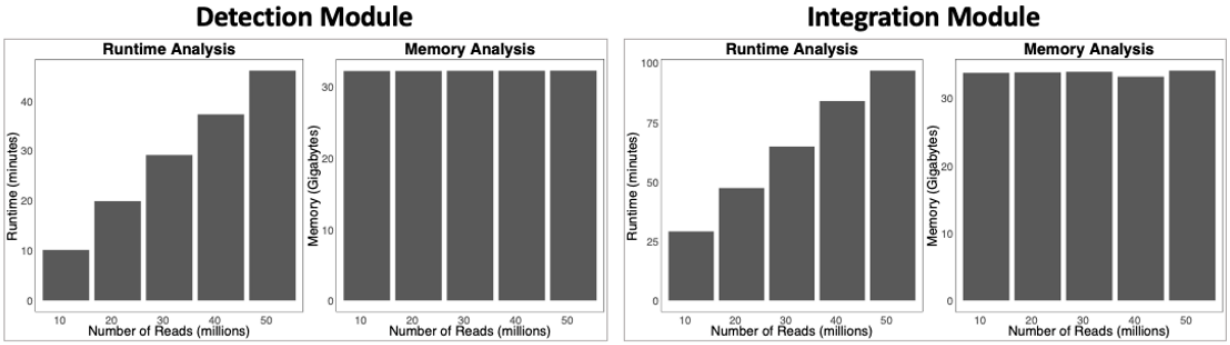
Supplementary File 1. Venus's detection module parameters.

Option	Description
--reads read_1.fastq read_2.fastq	Reads
--virusThreshold 5	Virus threshold for filtering
--virusChrRef virus_chr-ref.tsv	NCBI accession to species metadata file
--virusGenome virus.genomeDir --humanGenome human.genomeDir	Genome indices directories created to map our reads
--singleCellBarcode 1, 16 --singleUniqueMolIdent 17, 12 --singleWhitelist whitelist.txt	Specifications for single-cell data Numbers represent position, length, respectively
--out path/to/output/dir --readFilesCommand zcat --thread 32	General parameters

Supplementary File 2. Venus's integration site discovery module parameters.

Option	Description
--reads read_1.fastq read_2.fastq	Reads, should only be cDNA reads (no barcodes/UMI)
--guideFASTA integrSeq.fa	*integrSeq.fa* are sequences for fusion site classification.
--geneBed genes.bed	*genes.bed* converts genomic coordinates to genes
--virusChr NC_001802.1	NCBI virus accession id
--virusGenome virus.genomeDir --hybridGenome hybrid.genomeDir	Genome indices directories created to map our reads
--out path/to/output/dir --readFilesCommand zcat --thread 32	General parameters

Supplementary File 3. Venus's integration site discovery module parameters.



Supplementary File 4. Venus's Benchmark