Briefings in Functional Genomics, 00(00), 2021, 1–12

doi: 10.1093/bfgp/elab024
Review paper

Review of multi-omics data resources and integrative analysis for human brain disorders

Xianjun Dong, Chunyu Liu and Mikhail Dozmorov

Corresponding author: Xianjun Dong, Tel: 857-307-5423; Fax: 857-307-5476; E-mail: xdong@rics.bwh.harvard.edu; Mikhail Dozmorov, Tel: 804-827-2055; Fax: 804-828-8900; E-mail: mikhail.dozmorov@vcuhealth.org

Abstract

In the last decade, massive omics datasets have been generated for human brain research. It is evolving so fast that a timely update is urgently needed. In this review, we summarize the main multi-omics data resources for the human brains of both healthy controls and neuropsychiatric disorders, including schizophrenia, autism, bipolar disorder, Alzheimer's disease, Parkinson's disease, progressive supranuclear palsy, etc. We also review the recent development of single-cell omics in brain research, such as single-nucleus RNA-seq, single-cell ATAC-seq and spatial transcriptomics. We further investigate the integrative multi-omics analysis methods for both tissue and single-cell data. Finally, we discuss the limitations and future directions of the multi-omics study of human brain disorders.

Key words: multi-omics; neuropsychiatric diseases; single-cell omics; human brains; integrative analysis

Introduction

Global burden of neuropsychiatric disorders and research efforts

According to the data from the latest Global Burden of Diseases (GBD) Study [1, 2], the death rate per 100 k population caused by neuropsychiatric disorders has increased by 76% in the last 30 years (Figure 1A). The disability-adjusted life-years (DALYs; the sum of years lived with disability and years of life lost) also show significantly different patterns of age groups among the various neuropsychiatric disorders (Figure 1B). For example, headache and depressive disorders are leading in young and mid-aged adults, while neurodegenerative disorders such as Alzheimer's and Parkinson's diseases are greatest in the age group 75 years to older than 95 years. Alzheimer's disease and other dementias have increased the most among other neuropsychiatric disorders, partially due to the aging population structure of many countries such as China and the United States. Neuropsychiatric disorders are becoming a significant burden for the whole world. In the United States, total expenses in 2020 alone for health care, long-term care and hospice services for people aged 65 and older with dementia are estimated to be US$305 billion [3].

Understanding the human brain is one of the critical steps to finding cures for neuropsychiatric diseases eventually. Over the past decades, many countries have recognized the urgent need and launched large-scale projects, including the US BRAIN Initiative, the Human Brain Project at European Union, China Brain Project, Canadian Brain Research Strategy, Australian Brain Alliance, Japan Brain/MINDS Project and the recently launched International Brain Initiative [4]. While these brain initiatives may have various goals and different foci, for example, the EU Human Brain Project initially aimed to build a brain stimulator, they now all have a common component—using the latest omics technologies to understand the molecular functions of brain cells and their roles in neuropsychiatric diseases.

Xianjun Dong is an Assistant Professor of Neurology at Harvard Medical School, head of the Genomics and Bioinformatics Hub at Brigham and Women's Hospital. He is interested in developing and applying computational methods to study the transcriptional regulation in the human brain and how the regulation is dysfunctional in neurological diseases.

Chunyu Liu is a Professor of Psychiatry and Behavioral Sciences, Professor of Neuroscience and Physiology at SUNY Upstate Medical University. His research focuses on identifying the molecular mechanisms of psychiatric disorders using comprehensive approaches including genetics, bioinformatics, genomics, cellular and animal models.

Mikhail Dozmorov is an Associate Professor at the Department of Biostatistics, Virginia Commonwealth University. His training includes computer sciences and neurobiology. He is interested in the 3D structure of the human genome in health and disease and develops biostatistics methods and bioinformatics software to analyze chromatin conformation capture data.

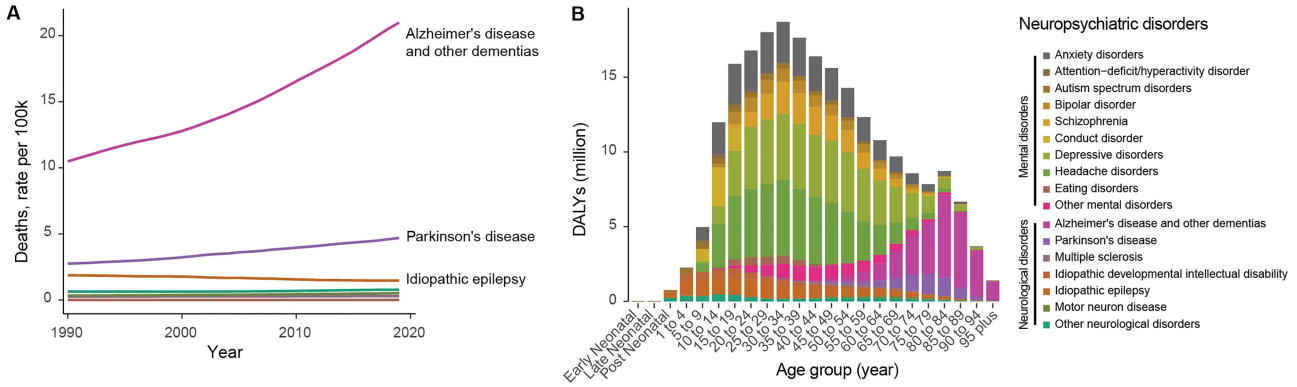

Figure 1. The global death rate of various neuropsychiatric disorders between 1990 and 2019, and DALYs by age in 2019. The neuropsychiatric disorders here include neurological disorder (B.5) and mental disorders (B.6) defined in the Global Burden of Diseases study. The plots were made based on data retrieved from the GBD Compare Viz Hub website [2] in December 2020

Multi-omics and its use in research of neuropsychiatric disorders

Two major types of research are commonly used to study neuropsychiatric disorders in humans: genetic study and biomarker study. The genetic study is based on the fact that most neuropsychiatric disorders have genetic roots. Genetic studies can detect common single-nucleotide polymorphisms (SNPs) associated with disorders, de novo mutations in affected offspring and rare copy number variants that are enriched in patients. While these genetic findings provided limited mechanistic insights for diseases, the major problems in the genetic studies include: (1) incomplete annotation of gene functions; (2) imprecise mapping of causal variants; (3) gap between association and causation; (4) poor understanding of the role of genetic variants in noncoding parts of the genome [5]. Another more translation-oriented approach is to discover biomarkers. Biomarker study targets the biological changes in patients in various states. Various biomaterials have been explored in comparing patients to controls, treated versus drug-naïve patients, or responders versus nonresponders. Biomarker-based disease subtyping and patient subgrouping may be valuable for optimizing treatment.

Both genetic and biomarker studies rely heavily on multi-omics data to achieve their goals. Multi-omics data refer to system-wide data of multiple 'omes', such as the genome, methylome, transcriptome, epigenome, proteome, metabolome and microbiome. Multi-omics data are frequently used to annotate disease-associated SNPs and genomic regions, to construct putative regulatory networks, and to assess potential causal or mediation relationships. Multiple omics can be used to interpret biological mechanisms of neuropsychiatric disorders in a concerted way. Multi-omics data offer unique support for independent biological validation of mechanistic findings. Following the central dogma, multi-omics data can help us track the molecular changes at different biological levels associated with the genetic variants, environmental changes and affected status.

In the past 10 years, massive data of multi-omics have been generated. Some of them were specifically in the brain, which is critical for the study of neuropsychiatric disorders. The other non-brain data are also useful references (e.g. the impact of intrauterine environment on the risk of schizophrenia [6]). Many multi-omics data were generated in matched tissues from both patients and controls, while the rest were only from healthy controls. Some of the data are from tissues of a few donors or a few cultured cells, while the others are from population cohorts. Data from sizable samples can present population variations, which is important for genetic studies. Besides the original quantification data of each -omics, derivatives such as reference maps, quantitative trait loci (QTL) are commonly used. They are important resources for both genetic study and biomarker discovery. This review summarized the major resource of available multi-omics data that can be used for the study of neuropsychiatric disorders, along with tools.

In this review, we focus on the omics data from human brain samples of healthy controls and patients with neuropsychiatric disorders. The omics include measurements of abundance and variations from genomic DNA, various DNA methylation, coding mRNA and noncoding RNA (microRNA, long noncoding RNA or lncRNA), actively translated RNA measured by Ribo-seq and protein, along with other epigenomic features, including chromatin accessibility measured by ATAC-seq or DNase-seq, 3D chromosome and genome structure by Hi-C. Most of the data were generated by next-generation sequencing. Small amounts of data were from microarray or Illumina BeadChip. These different omics data covered major biological components of the central dogma. Various brain regions and developmental stages were covered at different degrees. Single-cell data are described separately as it captures cell-type-specific features that can be missed by bulk tissue data.

Due to space limitations, we did not cover many other phenotypic data useful for studying neuropsychiatric diseases. Examples include brain imaging data, physiological and behavioral traits and animal models. All of these can be informative in modeling disease mechanisms and in formulating prediction algorithms.

Multi-omics data resources for neuropsychiatric disorders

We investigated the major cohorts or projects producing multi-omics data either general for human biology or specific to neuropsychiatric diseases (see summary in Table 1). We have to emphasize that the list of consortia listed below is likely incomplete. There are many other important omics resources unlisted here due to space limitation. We might also miss some individual omics datasets for other neuropsychiatric disorders, for example, 24 total RNA-seq for the cortex of autistic brains [7], RNA-seq for 15 brains for multiple sclerosis [8] and >110 K single-cell neuronal transcriptomes for epilepsy [9]. This field is fast evolving. To keep tracking the up-to-date human brain

Table 1. Summary of main multi-omics consortia

General consortia				
Consortium	**Access**	**Diseases** [α]	**Subject size** [β]	**Omics assay type** [γ]
ENCODE	Open	Healthy, cancer	variable	(sc)RNA-seq, ChIP-seq, DNase-seq, ATAC-seq, eCLIP-seq, ChIA-PET, Hi-C, CAGE, methylation etc.
GTEx	Controlled	Healthy	948	WGS, WES, polyA+ RNA-seq
Roadmap Epigenomics	Open	Healthy	variable	polyA+ RNA-seq, ChIP-seq, DNase-seq, methylation
FANTOM	Open	Healthy	variable	CAGE, RNA-seq, RADICL-seq
OmicsDI	Open	Various	variable	Transcriptomics, genomics, proteomics, metabolomics, metagenomics, etc.
UK Biobank	Controlled	Various	500 000	Genotyping, WES, WGS
Brain-related consortia				
Consortium	**Access**	**Diseases** [α]	**Subject size** [β]	**Omics assay type** [γ]
psychENCODE	Controlled	Healthy, SCZ, BP, ASD	2793	(sc)RNA-seq, ChIP-seq, Hi-C, ATAC-seq, methylation, genotyping
CommonMind	Controlled	Healthy, SCZ, BP	1143	WGS, RNA-seq, ATAC-seq
AMP-PD	Controlled	Healthy, PD	10 247	WGS, total RNA-seq
AMP-AD	Controlled	Healthy, AD, PSP	4470	WGS, genotyping, RNA-seq
BRAINcode	Controlled	Healthy, PD, AD	200	Total RNA-seq, genotyping
BrainSeq	Open	Healthy, SCZ, MDD, BP	861	RNA-seq, genotyping, methylation
Aging, Dementia and TBI Study	Controlled	Healthy, TBI, AD	107	Total RNA-seq
NIAGADS - ADSP umbrella	Controlled	Healthy, AD, CBD, PSP	23 868	WGS, WES
NIAGADS - NG00057	Open	Healthy, PD, AD	18	Total RNA-seq
NIAGADS - NG00038	Open	Healthy, AD	492	Expression array, genotyping
BrainSpan	Open	Healthy	42	polyA+ RNA-seq, expression array, methylation
IMAGE-CpG	Open	Healthy, EP	17	Methylation
BOCA	Open	Healthy	5	ATAC-seq
Braineac	Dead link	Healthy	134	Expression array, genotyping
HBT	Open	Healthy	57	Expression array, genotyping
BrainCloud	Open	Healthy	270	Expression array, genotyping, methylation
NeMO	Open	Healthy	418	scRNA-seq, scATAC-seq

[α] SCZ: Schizophrenia; BP: Bipolar disorder; ASD: Autism Spectrum Disorders; MDD: Major Depression Disorder; TDI: Traumatic Brain Injury; EP: Epilepsy; PD: Parkinson's disease; AD: Alzheimer's disease; PSP: Progressive supranuclear palsy; CBD: Corticobasal degeneration
[β] Note we try to count the subject size (not the sample size) in the cohort, except for NeMO.
[γ] We try our best to specify the RNA-seq type of a cohort: polyA+ RNA-seq (if all are polyA+ RNA-seq), total RNA-seq (if all are total RNA-seq), or RNA-seq (if mixture of both). WGS: Whole-genome sequencing; WES: Whole-exome sequencing.

multi-omics datasets, we keep the data source table [10] publicly editable so that everyone can contribute to the list.

General consortia

ENCODE

The Encyclopedia of DNA Elements (ENCODE) project was initially launched in 2003 as an international effort to characterize the regulatory function of approximately 1% of the human genome [11]. Further development included collecting whole-genome data generated by sequencing-based technologies, such as RNA-seq, DNA methylation, chromatin immunoprecipitation sequencing (ChIP-seq for histone and TF), DNase I hypersensitive sites, and these assays were performed across multiple cell lines [12]. Transcription factor binding site data represents a unique resource not readily available from other projects [13]. Recent years have seen the expansion of data types, such as chromatin interaction analysis by paired-end tagging (ChIA-PET), high-throughput chromatin conformation capture technology (Hi-C). As of December 2020, the ENCODE Experimental Matrix showed 10 314 human multi-omics datasets. Although ENCODE has primarily focused on cell lines, recent updates include some omics data from the human brain, primary neurons or neuronal cell lines (i.e. 449 experiments in human brain organ as of December 2020 [14]). At the current Phase 4, the ENCODE project continues data assembly, computational analysis and functional characterization of the human genome [15].

Roadmap Epigenomics

Roadmap Epigenomics project collected RNA-seq, ChIP-seq (histone), DNase-seq and methylation data primarily from human blood and 22 tissue types [16]. Although the program ended in 2018, the open-access epigenomic data, now included in the ENCODE data portal, remain a viable resource for multi-omics integrative analyses [17]. As of December 2020, the ENCODE Experimental Matrix showed 2154 multi-omics human datasets released in Roadmap, including 158 datasets associated with adult and fetal human brains.

FANTOM

Over the last 20 years, the Functional AnnoTation Of the Mammalian genome(FANTOM) Consortium has collected an extensive resource to understand transcriptional regulation across healthy human/mouse cells and tissues, including noncoding transcripts [18, 19]. Cap Analysis of Gene Expression (CAGE) data are the very unique data from this project, critical for defining transcription start sites. FANTOM5 created the most comprehensive catalog of enhancers and promoters, which included brain-specific annotations [20]. FANTOM6 focuses on annotating noncoding RNA, which has been a major weakness in the current genome annotation. FANTOM data have been used in numerous multi-omics tools, such as HACER (Human Active Enhancers to interpret Regulatory variants) that integrates FANTOM5, expression quantitative trait locus (eQTL) databases, transcription factor binding sites (TFBSs) from ENCODE, nascent RNA sequencing (GRO/PRO-seq) and Hi-C data, including brain cell-specific analysis [21].

GTEx

The Genotype-Tissue Expression (GTEx) project aims to characterize variation in gene expression levels across individuals and diverse tissues of the human body [22]. Besides the identification of cis-eQTLs, tissue-specific trans-eQTLs were discovered, as well as eQTL interactions across cell types, sex chromosome-specific eQTLs. The processed data are available for download and accessible through a web interface. It should be noted that GTEx has transcriptome profiles of multiple brain regions with a fairly large sample size ($n < 255$).

OmicsDI

Omics Discovery Index (OmicsDI) is a platform for searching multi-omics datasets [23]. It integrates proteomics, genomics, metabolomics and transcriptomics datasets from multiple databases, ranging from GEO and European Genome-phenome Archive (EGA) to the Library of Integrated Network-Based Cellular Signatures (LINCS), dbGaP and more. It enables searching for data by the organism, disease, tissue, gene identifiers and keywords. As of December 2020, the search for 'brain' yields 116 261 results, out of which 10 are multi-omics datasets.

Brain-specific consortia

Allen Brain Atlas

The Allen Institute for Brain Science was among the first efforts for creating a systematic resource of brain-specific omics data linked to anatomic structures [24]. The Allen Human Brain Atlas collected microarray expression profiling and MRI measures of approximately 900 neuroanatomical slices of the brain from two individuals, effectively demonstrating that gene expression correlates with spatial localization [25]. Further development included collecting temporal and spatial gene expression programs of developing human and mouse brains, aging, dementia, traumatic brain injury (TBI) and IVY glioblastoma atlas project. Two such programs include BrainSpan (RNA-seq for up to 16 brain regions from 42 developing human brains) and the Aging, Dementia and TBI study (total RNA-seq for 107 subjects). The latest additions include single-cell RNA sequencing and *in situ* hybridization from various parts of the human, mouse, and rhesus macaque cortex, hippocampus, spinal cord, comparative cellular anatomy in the thalamus. Patch-seq is another new addition [26]. The data are freely available online or via programmatic access (AllenSDK, Brain Modeling Toolkit, DiPDE simulation platform). Although gene expression is the primary focus of the Allen Brain Atlas, the availability of electrophysiological and morphological data makes it a unique resource for spatial transcriptomics of the brain.

PsychENCODE

PsychENCODE Consortium was created in 2015 to focus on genomics and epigenomics data of the human brain for studying neuropsychiatric disorders [27]. PsychENCODE data are featured with the largest collection of brains (2793 unique donors) of controls and major psychiatric disorders, including schizophrenia, bipolar disorder and autism. The majority of the data are from postnatal, adult brains. But developmental aspects have also been consistently covered for the interests of developmental disorders [28, 29]. Almost all brains have genotype data, making them the best-powered data for mapping molecular quantitative trait loci. The first release of data is primarily from bulk tissue. RNA-seq transcriptome is available for all tissues. Histone markers ChIP-seq, ATAC-seq, Ribo-seq, proteomics, DNA methylation, Hi-C data are available for some tissues. The frontal cortex is the major brain region studied by this consortium. The analyses of part of the first release data delivered 11 research papers [30]. Besides controlled-access raw data, psychENCODE provides a constant growing list of processed, derived and integrative results, such as lists of brain-expressed genes, disease-associated genes, co-expression modules, brain single-cell expression profiles, histone modification data, over 79 K enhancers, chromatin loops, and topologically associating domains, over 2.5 M eQTLs and SNPs associated with splicing, cell specificity and chromatin activity. Linking transcription factors and enhancers to target genes enabled the creation of a network of 321 genes. These data have been used to train machine- and deep-learning models to predict psychiatric phenotypes, achieving over a 6-fold increase in performance over additive polygenic risk scores [31]. PsychENCODE data are one of the best data for evaluating population variations of omics in the human brain, as complementary to ENCODE, Roadmap or Brain Atlas data. GTEx, which is popular for its multiple-tissue eQTL, is dwarfed by brain-specific data from PsychENCODE for its size and comprehensiveness. PsychENCODE is moving into its second phase with an emphasis on single-cell data.

AMP-AD

The Accelerating Medicines Partnership (AMP) is a precompetitive partnership among NIH, pharmaceutical companies and nonprofit organizations that focuses on identifying and validating promising biological targets for diagnostics and drug development. AMP-AD, as one of three initial programs under the AMP umbrella, was budgeted for 5 years with US$185.2 million. The goal of AMP-AD is to apply cutting-edge systems and network biology approaches to integrate multidimensional human molecular data (genomic, epigenomic, RNA, proteomic) from more than 2000 human brains at all stages of Alzheimer's disease (AD) with clinical and pathological data. The three largest AMP-AD studies that contributed the most multi-omics data are ROSMAP, MSBB and MayoRNAseq. For example, the Mount Sinai Brain Bank (MSBB) study provided total RNA sequencing for over 1700 brain samples from four brain regions (superior temporal gyrus, frontal pole, parahippocampal gyrus

and inferior frontal gyrus). Most of the subjects had also done whole-exome sequencing (WES), whole-genome sequencing (WGS) and mass spectrometry for proteomics. Additional to the multi-omics data, comprehensive clinical and neuropathological assessments including Braak AD-staging score, the density of neuritic plaques, clinical dementia rating scale (CDR), mini-mental state examination (MMSE), CERAD score [32], and histology staining were also provided for most participants. The other AMP-AD study ROSMAP provided genomics, transcriptomics, proteomics, epigenomics (e.g. H3K9ac ChIP-seq, DNA methylation) and metabolomics for up to 2214 brain samples. ROSMAP also generated single-cell transcriptomics data for over 48 brains from AD and subjects with no cognitive impairment [33, 34]. The rich set of AMP-AD data, together with other AD projects, is hosted in the AD Knowledge Portal (https://adknowledgeportal.synapse.org), one of the largest data portals for AD research.

AMP-PD

Another program under the AMP umbrella for neurological disease is AMP-PD (https://amp-pd.org). Launched in 2018, AMP-PD aims to identify and validate diagnostic, prognostic and/or disease progression biomarkers for Parkinson's disease (PD) via an in-depth molecular characterization and longitudinal clinical profiling of PD patient data and biosamples in existing cohorts. The initial release of AMP-PD included 4298 participants from four cohorts [35]: the Michael J. Fox Foundation (MJFF) and National Institutes of Neurological Disorders and Stroke (NINDS) BioFIND study, Harvard Biomarkers Study (HBS), the NINDS Parkinson's disease Biomarkers Program (PDBP) and MJFF Parkinson's Progression Marker Initiative (PPMI). Its latest release expanded to include 10 247 participants from seven cohorts, providing clinical, genomics and longitudinal transcriptomics data for PD and control. Note that this longitudinal program collected whole blood samples at different visits (e.g. baseline, 0.5, 6, 12, 18, 24 and 36 months) for total RNA sequencing. In total, 8461 RNA-seq datasets (from n = 3274 participants) along with harmonized clinical measurement are available in the current release at the AMP-PD website, providing a great resource to study the progression of Parkinson's. AMP-PD program hosts its data in Google Cloud Platform (GCP) and recommends users to access, run analysis tools, and collaborate via the Terra platform (http://app.terra.bio/). Considering the trend that large omics cohorts like this will generate more and more data, bringing codes and tools to the data, instead of downloading the data to the local server, will be a more practical solution.

BRAINcode

The BRAINcode project (http://www.humanbraincode.org) has been established to decode the function of brain neurons vulnerable to neurodegenerative diseases (e.g. dopamine neurons for PD, pyramidal neurons for AD) by integrative analysis of the human neuronal genome, transcriptome and epigenome. In the initial release, it included ultra-deep ribo-depleted total RNA-seq from 50 000 neurons laser-captured from about 100 human post-mortem brains, including dopamine neurons from substantial nigra (n = 86 brains) and pyramidal neurons from cortex (n = 10 brains). Over 71 000 novel transcribed noncoding elements (TNEs) were discovered in dopamine neurons, many of which are consistent with the epigenetic features of active enhancers (or eRNAs) [36]. eQTL analysis and GWAS colocalization analysis found that these eRNA elements are enriched for genetic variants of both dopamine-hypoactive (e.g. PD) and dopamine-hyperactive traits (e.g. schizophrenia and addiction). The second phase of BRAINcode included 100 more brains from both AD and PD patients and identified >11 000 neuronal circRNAs in the brain transcriptome (data to be released soon). BRAINcode is moving to its new phase by focusing on single-nucleus RNA-seq, single-nucleus ATAC-seq, spatial transcriptome and single-cell eQTL for additional 200 human brains.

CommonMind

CommonMind Consortium (CMC) generated a public resource of functional genomic data from the dorsolateral prefrontal cortex (DLPFC; Brodmann areas 9 and 46) of ~1000 individuals from four separate brain banks, including 501 control individuals, 353 diagnosed with schizophrenia and 120 with bipolar disorder [37]. The genomic data include RNA-seq and SNP genotypes on 980 individuals and ATAC-seq on 269 individuals, of which 264 are a subset of individuals with RNA-seq. Both raw and processed data are available on the Synapse platform at http://CommonMind.org.

BrainSeq

The BrainSeq is a project leading by the Lieber Institute and pharmaceutical companies, aiming to characterize the genetic and epigenetic regulation of transcriptome in distinct brain regions across the human lifespan in samples of major neuropsychiatric disorders and controls [38, 39]. Phase 1 of the consortium included DLPFC polyA+ RNA-seq on 738 subjects spanning the lifespan and three main psychiatric diagnostic groups (schizophrenia, major depression disorder and bipolar disorder). Phase 2 expanded to include the hippocampus region and performed RiboZero-treated total RNA-seq for 551 subjects spanning the lifespan and adult patients with schizophrenia. There were in total of 861 unique donors across the first two phases of BrainSeq. There was also Illumina 450 k microarray data from Phase 2 on both hippocampus and DLPFC to profile DNA methylation [40, 41]. They also provided development- and schizophrenia-associated gene profiles and eQTL resources at http://eqtl.brainseq.org. It's worth noting that Lieber Institute is also launching the African Ancestry Neuroscience Research Initiative, aiming to make publicly available the largest omics dataset in the world on human post-mortem brains of individuals of recent African ancestry.

NeMO

The Neuroscience Multi-omic Archive (NeMO Archive) is a data repository specifically focused on the storage and dissemination of omics data generated from the BRAIN Initiative and related brain research projects. NeMO data include transcription activity, methylation, histone modification profiles and chromatin accessibility for human, mouse and marmoset. The current search of human data on the public BRAIN Initiative Cell Census Network (BICCN) showed 418 samples with single-cell RNA-seq (n = 412) and single-cell ATAC-seq (n = 6) with open access. More human brain single-cell omics data (e.g. single-cell PLAC-seq, ATAC-seq, RNA-seq data to define cell-type-specific 3D epigenomes [42]) are available on NeMO through restricted access.

Integrative multi-omics analyses in brain research

Multi-omics integration at single-cell level

The high cellular complexity of the brain prompts the application of single-cell omics approaches to understand genomic regulation on a single-cell level. Darmanis *et al.* [43] were among the first to provide single-cell transcriptomic data from 466 cells of the healthy human cortex. Subsequent efforts included time course scRNA-seq profiling during neurogenesis, revealing lineage-specific trajectories and the dynamics of neurogenic transcription factors [44]. Single-cell methylation data were also used to reveal neuronal subpopulation in the human cortex [45]. Latest studies, such as single-cell transcriptomic analysis of 48 Alzheimer's patients and healthy individuals in the ROSMAP cohort, have scaled up to 80 K cells, revealing unprecedented insights into disease pathophysiology [46]. Recently STAB, a spatio-temporal cell atlas of the human brain, defines 42 cell subtypes across 20 brain regions and 11 developmental periods by analyzing 13 available human brain scRNA-seq datasets [47]. Although human single-cell data remain scarce, various resources provide mouse scRNA-seq data, with DropViz (http://dropviz.org/, 690 K cells) and 10X Genomics (1.3 M cells [48]) currently being the largest.

scRNA-seq technologies have evolved to incorporate other layers of multi-omics information, with open chromatin being one of the recent additions [49]. Lake *et al.* [50] integrated single-nucleus RNA sequencing (snDrop-seq) with single-cell open chromatin profiling (scTHS-seq) in over 60 000 cells from the human adult visual cortex, frontal cortex and cerebellum, demonstrating better resolution of cell subpopulation and the ability to predict one omics data from the other. Similarly, Trevino *et al.* [51] integrated human forebrain-specific RNA-seq and ATAC-seq data over the time course, revealing detailed enhancer-gene activity correlations, the temporal activity of neurogenesis-specific transcription factors and cell types and time periods associated with susceptibility to neuropsychiatric disorders. Li *et al.* [52] integrated scRNA-seq, small RNA-seq, histone-seq and methylation data from psychENCODE and BrainSpan to outline functional genomics of human brain development and cell type-specific gene expression modules associated with neuropsychiatric disorders. Details about these and other studies are available at https://github.com/mdozmorov/scRNA-seq_notes#brain.

By integrating with genetic variation from genotyping array or WGS, scRNA-seq also allows us to map eQTLs across different cell types and in dynamic processes, many of which are obscured when using bulk-tissue methods. To apply this technology to large-scale population genetics studies, Luke Franke and his colleagues [53] have founded the single-cell eQTLGen consortium (sc-eQTLGen), aimed at pinpointing the cellular contexts in which disease-causing genetic variants affect gene expression.

Recent development in spatial transcriptomics, such as 10X Genomics Visium [54, 55], Slide-seq [56], HDST [57], MERFISH [58] and LCM-seq [36, 59], enabled the unambiguous identification of location-specific single-cell gene expression programs [60]. These technologies are starting to be applied to reveal the layered structure of the human DLPFC marked by distinct gene expression programs [55]. Importantly, integration of spatial transcriptomics with other data, such as neuropsychiatric gene sets, demonstrated location-specific relevance of disease-associated signal [55], opening the new chapter in integrative multi-omics. More brain-specific single-cell spatial transcriptomics datasets are becoming available [61]. A few other methods used scRNA-seq data to resolve spatial expression [62, 63].

Another addition is the 3D chromatin organization in brain cells. Several studies used Hi-C and its variants to integrate 3D genomics of the human brain with gene expression, histone modifications (ChIP-seq), open chromatin (ATAC-seq) and GWAS signals, demonstrating the importance of spatial organization of the genome [5, 64–66]. Chromosome conformation capture technologies have been extended on a single cell level [67], and, integrated with gene expression, revealed associations between the 3D structures and gene expression [68]. Recent developments included technologies for simultaneous chromatin conformation capture and methylation in single cells [69, 70].

Methods and tools for multi-omics data integration

As summarized in Subramanian *et al.* [71], the goals for multi-omics data integration can be approximately classified into three categories: 1. Disease subtyping and classification based on multi-omics profiles; 2. Prediction of biomarkers for various applications; 3. Deriving biological insights. For example, in disease genetic study, we usually combine genomic and transcriptomic data via eQTL analysis to improve the detection of common variants with functional effects (e.g. GTEx study [22]). However, this approach is not easy because environmental factors and disease state can also affect the transcriptome. Mohammadi *et al.* [72] developed ANEVA-DOT to identify heterozygous DNA variants with unusually strong effects on gene expression by comparing the expression activity of individuals' maternal and paternal alleles. Montaner *et al.* [73] recently reviewed the integrated analysis of multi-omics data (incl. proteomics, genomics, transcriptomics and metabolomics) and provided useful insight into stroke pathogenesis, identification of therapeutic targets and biomarker discovery. The methods to achieve that can be characterized as early and late integration approaches, with the former combining the omics matrices into one and then analyzing it, and the latter analyzing each omics modality separately and then combining the results, reviewed in Ref. [74]. Alternatively, the integration methods can be classified as unsupervised (Matrix Factorization, correlation-based, Bayesian methods, network-based methods) or supervised (network-based methods, multiple kernel learning), reviewed in Refs. [75–77], benchmarked in Tini *et al.* [78], and many are implemented in the mixOmics R package [79]. The latest development includes neural network architectures, such as variational autoencoders (VAE), for data integration [80]. As an application of these disease classification methods, Zhang et al. [81] recently developed an end-to-end VAE-based model called OmiVAE to extract low dimensional features and classify samples from multi-omics data. By integrating genome-wide DNA methylation and gene expression profiles together with 450 804 molecular features, they evaluated this model with 9081 samples of 33 tumor types and normal ones using the pan-cancer multi-omics datasets from The Cancer Genome Atlas (TCGA) [81]. OmiVAE was shown to achieve an average classification accuracy of 97.49% after 10-fold cross-validation among 33 tumor types and normal samples, which shows better performance than other existing methods.

Comparing to bulk tissue data, single-cell omics data provide more accurate transcriptome profiling of highly expressed genes in high-resolution subtypes of cells, but at the time suffers the problems of high cost, low coverage, shallow depth and high missing data rate. Therefore, single-cell omics requires new algorithms while adopting techniques developed for bulk

omics data analysis. We provide examples of some of the most representative single-cell integration methods and refer to the aforementioned reviews for a more detailed overview. Methods for integrating single-cell omics data include the use of non-negative matrix factorization (NMF) [82] or similar dimensionality reduction or low-dimension embedding methods [83]. An example method for integrating scRNA-seq with other single-cell data is LIGER, an NMF-based method for integrating and analyzing multiple single-cell datasets, across conditions, technologies (scRNA-seq, methylation, spatial transcriptomics) or species (human and mouse) [84]. When applied to human and mouse brain cells, it resolved otherwise unobservable spatial cell states. A shared embedding-based method, Harmony, has been used for the integration of scRNA-seq and spatial transcriptomics data [85]. Integration of scRNA-seq and scATAC-seq using Latent Semantic Indexing (LSI) and the modified term frequency-inverse document frequency (TF-IDF) procedure has been implemented in the Signac extension of the Seurat R package [86]. Using dimensionality reduction and clustering, the ArchR R package outperformed Signac in integrating scATAC-sec and scRNA-seq data [87]. A network similarity-based CellWalker method has been shown to be more robust to the sparsity and noise of scRNA-seq and scATAC-seq data [88]. It has been applied to the developing human brain and identified cells transitioning between transcriptional states, resolved cell-specific enhancers and mapped neurological trait-associated genes to specific cell types via enhancers [88]. The MAESTRO suite of tools utilizes best practices of integrative data analysis (e.g. graph-based and density-based clustering, modeling gene regulatory potential from chromatin accessibilities) for comprehensive integration of scRNA-seq and scATAC-seq data, and provides blood- and brain-specific cell signatures for annotating cell clusters [89]. 3D genomics integration methods also started to appear. An NMF-based approach for subpopulation-specific deconvolution of bulk Hi-ChIP signal using scRNA-seq and scATAC-seq has recently been proposed [90]. These methods demonstrate the potential of single-cell multi-omics data integration to reveal novel biological insights into complex cellular systems, such as the brain and neuropsychiatric diseases.

A case study of schizophrenia with multi-omics data

As a case study for analyzing omics data, we would like to highlight the study of schizophrenia by integrating multi-omics data in PsychENCODE and other consortia [31]. The integrative analysis [31] has merged multi-omics data from PsychENCODE, GTEx, ENCODE, CommonMind, Roadmap Epigenomics and single-cell analyses into a pyramidal structure, e.g. a base of raw data files, a middle layer of uniformly processed and easily shareable results (such as open chromatin regions and gene expression quantifications) and a top-level 'cap' of an integrative, deep-learning model, based on regulatory networks and QTLs.

- At the transcriptomics level, they deconvoluted the bulk-tissue RNA-seq data using an **NMF**-based approach and compared the NMF top components (NMF-TC) with the single-cell-derived cell type signature. Upon validation, they then deconvoluted the bulk-tissue RNA-seq with the single-cell signatures to estimate cell fractions across individuals. They showed that the weighted combinations of single-cell signatures could account for most of the population-level expression variation in bulk tissue, with an accuracy of >88%. They then compared the cell fraction changes of each cell type in various neuropsychiatric traits and reported disease-specific cell proportion changes (see figure S15 in [31]).
- At the epigenomics level, they annotated 79 056 enhancers in the PFC by integrating the H3K27ac and H3K4me4 ChIP-seq data in PsychENCODE and DNase-seq and ChIP-seq data from Roadmap PFC samples. They then identified QTLs affecting gene expression and chromatin activity by performing expression, splicing-isoform, chromatin and cell fraction QTLs (eQTLs, isoQTLs, cQTLs and fQTLs, respectively). They revealed 33 000 eGenes in PFC, approaching the total number of genes estimated to express in brains, reflecting their large sample size. They also tested the enrichment of various QTLs in different genomic annotations. They showed the greatest intersection between eQTLs and cQTLs. They revealed 2477 multi-QTLs. By integrating with Hi-C interaction data, they found that QTLs involving SNPs distal to eGenes but linked by Hi-C interactions showed significantly stronger associations than those with SNPs directly in the eGene promoter or exons.
- They further constructed the gene regulatory network by linking enhancers, TFs and target genes. For example, they used the coefficients of **elastic net regression** (e.g. assuming that target gene expression is determined by a linear combination of the expression levels of its regulating TFs) to infer the linkage between TF and target genes. Based on the regulatory network, they further connected the noncoding GWAS variants to disease genes. They identified a set of 1111 putative SCZ-associated genes (the SCZ genes), 321 of which were supported by more than two evidence sources (e.g. QTLs and Hi-C). Interestingly, they found that most SCZ genes were not even in LD with the index SNPs (~67%, with $r^2 < 0.6$). Last, they looped back these SCZ genes with single-cell profiles and found that they are highly expressed in neurons, particularly excitatory ones.
- Last, to fully capture the interaction between genotype and phenotype beyond their regulatory network, they incorporated an interpretable deep-learning framework, the **Deep Structured Phenotype Network** (DSPN). Unlike traditional classification methods such as logistic regression which predicts phenotype from genotype without intermediate layers, DSPN can include intermediate layers for molecular phenotypes (e.g. gene expression, enhancers, co-expression modules, cell fraction) with sparse connectivity. The DSPN was able to gain a larger, 6-fold improvement in predicting traits, which may reflect its ability to incorporate nonlinear interactions. For schizophrenia, the variance explained by the full DSPN model exceeds that explained by common SNPs (e.g. 32.8% versus 25.6%).

Limitations and future directions

Newly generated data

Ideally, we will need multi-omics data covering all biological levels, intermediate steps from DNA to protein, all developmental stages from stem cells to death, all cell types from neurons to glial cells and all states from drug-naive to patients under various treatments. In reality, only a small fraction of the desired data has been produced and is ready for use. Besides the data summarized above, we know that more single-cell transcriptome, ATAC-seq data will be available in the coming 1 or 2 years. Spatial transcriptome, Hi-C data will be useful for a better understanding of brain transcriptome and its regulation. More and more eQTL and other molecular QTL will be generated on

the brain, brain cells at different developmental stages, different racial backgrounds and sexes. Encouragingly, technologies for assessing multi-omics signals from the same cells started to appear, e.g. SNARE-seq assessing scRNA-seq and scATAC-seq data [91].

Some specific omics data are still underrepresented. DNA methylation data, microRNA expression and proteomics data are examples that can be better covered. Ribo-seq (or ribosome profiling) data will unlikely increase after Illumina ceased to produce the kit. ChIP-seq data of transcription factors in brain cells is one major category of data that is unfortunately largely missing. Mitochondria-related genomics and epigenomics have not obtained sufficient attention they may deserve yet.

We also should note that -omics are sensitive to sex, racial genetic background and other variables, and the effects are not presented well in most public databases. Racially diverse data are still broadly not available. In 2019, researchers found that ~78% of GWAS individuals are of European ancestry [92]. The brain omics data have even less diversity. For example, less than 4% of participants are non-white/Caucasian in the current AMP-PD release. It is a massive problem for the research aiming to be more inclusive. Initiatives like the African Ancestry Neuroscience Research are expected to close the gap.

The 'true' single-cell transcriptome for human brains

The majority of single-cell transcriptome data in human brain research is actually single-nuclei RNA-seq, not single-cell RNA-seq, due to the current technology difficulty in extracting intact neuronal cells from post-mortem frozen brain tissue without breaking the cell membrane. Moreover, except for a few experimental tryouts of total RNA sequencing in single cells (e.g. SuPeR-seq [93], MATQ-seq [94], RamDA-seq [95] and DART-seq [96]), most current single-cell RNA-seq studies are based on polyA-enriched RNA sequencing methods, leaving many interesting non-polyA RNAs (e.g. miRNAs, piRNAs, circRNAs, eRNAs) out of attention. The integration of single-cell multi-omics is affected by the issues in the single modality data, such as the dropouts in the single-cell data and the resolution in cell cluster definition and annotation. These will also be a huge challenge in integration study. The validation and benchmarking of single-cell/integration analytic tools are also urgently needed.

Human brain specimens versus cell lines

Typical approaches to study neurological diseases involved human brain specimens from healthy and diseased individuals. In contrast to widely accessible blood specimens, brain specimens are typically obtained from post-mortem tissues, which has its limitations: tissue degradation is the major one. RNA is particularly sensitive to time after death. Another limitation is that post-mortem tissue can only offer a snapshot of biological systems, which may not be sufficient for revealing the dynamics of symptoms and treatment responses. Cultured cells and newly developed brain organoids are an important alternative for generating multi-omics data with the advantages of relatively homogeneous environmental factors and cell composition.

Host-microbiome multi-omics integration

Other than the host itself, integration with the omics data in the microbiome is also emerging as an interesting direction. Because the gut microbiome is an important activator of inflammatory substances, researchers have observed that increased expression of immune-modulating microbiota such as *Clostridia* in the guts leads to higher microglial density and IL-1β expression in the brains of stress vulnerable rats [97]. Recent research in Alzheimer's showed that gut infection could trigger the production of amyloid clumps in Alzheimer's brains (see review [98]). While mechanisms behind these gut-brain associations are largely unclear, multi-omics integration between host and microbe could shed light on new insights [99, 100].

Longitudinal multi-omics analysis

Most multi-omics analyses in human neuropsychiatric diseases are cross-sectional (e.g. case versus control, sub-types of diseases). Profiling omics longitudinally coupled with clinical measures and treatment outcomes could provide a more comprehensive assessment to improve disease risk prediction, early detection and better treatment. Previous longitudinal multi-omics efforts had successfully identified disease markers for few diseases [101–103], but not much in neuropsychiatric diseases. One of the multi-omics cohorts we reviewed here, AMP-PD, includes longitudinal blood RNA-seq data and clinical data for Parkinson's patients. We expect more longitudinal omics data in neuropsychiatric research. Such longitudinal data are typically from peripheral tissues. Therefore, multi-omics comparative analysis comparing brain and peripheral tissue is needed to validate the brain relevance.

Correlation versus causality analysis

Many multi-omics studies generated correlation results. For example, eQTL analysis is to identify the correlation between genetic variation and gene expression. Many so-called 'biomarkers' are actually biomolecular signals that are associated/correlated with diseases, traits or states. We know correlation does not prove causation. A typical example is that GWAS leading variants are not necessarily the trait/disease causal variants [104]. Several statistical fine-mapping methods have been developed to suggest underlying causality from GWAS output [105]. Machine learning and deep learning have been used to find patterns and correlations in the multi-omics data, which might work well enough in many cases (e.g. tumor recognition, disease prediction). However, if a model could capture causal relationships, it will be more generalizable. Additionally, if we can tell the cause from the effect, we are better positioned to find cures for disease. Several complementary approaches (e.g. Mendelian randomization, structural equations modeling, Bayesian networks) have been applied to discover novel causal effects of genomic and epigenomic variation on lipid phenotypes [106].

High-dimension reduction challenge

Another challenge in multi-omics integration is the high dimension. Although many multi-omics cohorts we reviewed here have provided large sample sizes, the number of samples is still much less than the number of features ($N \ll P$). This is becoming a trend as subjects are assessed with more and more features (For example, UK Biobank has 500 000 participants and each participant has 96 million SNPs, thousands of clinical/lifestyle phenotypes and around 4000 imaging-derived phenotypes). This situation brings several problems when building a model, such as overfitting, multicollinearity of the features and infinite solutions for coefficients [107]. Reducing dimension is recommended before integrating the multi-omics. Kegerreis *et al.* [108] showed that classification based on WGCNA co-expression modules [109] can better cope with variation among datasets

compared to the classification based on raw gene expression. Other dimension-reduction techniques, such as support vector machines (SVMs), random forest (RF) and singular value decomposition (SVD) are also commonly used to reduce the overfitting issue. In single-cell omics, methods such as PCA, t-SNE and UMAP were used to reduce the dimension. Advanced deep learning methods such as variational autoencoder (VAE) can also output the lower dimensional latent representation for high-dimensional data. Multiple testing inflation and significance criteria are accompanying problems [110].

Heterogeneity and harmonization of the data sources

Samples from different cohorts or consortia could actually come from the same subjects. For example, the integrative analysis of DLPFC RNA-seq data from over 1800 brains in the PsychENCODE consortium [31, 111] includes ~500 brains from the first phase of the BrainSeq study [112]. Most of our BRAINcode brain samples are from Banner Health brain bank, which might also be studied in other cohorts. Connecting different cohorts via source ID (e.g. SOURCE_SUBJECT_ID in dbGap) or universal ID (e.g. GUID in PDBP) could potentially reduce bias by removing duplicated samples and increase power by connecting samples from the same subjects. Sample identification and matching are critical for some data integration analyses when the analysis relies on omics from the same subjects, such as QTL mapping. The method DRAMS [113] offers a genotype-based solution to ensure data alignment. Last but not least, the current single-cell omics data are spread around in various publications. With more and more single-cell omics data coming, a centralized data repository for single-cell omics data with harmonized QC will be helpful for cross-cohort comparison and integration.

Open data sharing

Open genomic data sharing has been an essential ingredient for successful research, long-rooted back to the Human Genome Project. The broad sharing of data generated by genomic research studies has maximized the utility of the data and the public benefit of such projects [114]. In the last decade, both public and private funding agencies recognized the importance of data sharing and urged to share the data after they are generated, even before the first use by data producers. Brain-related consortia such as psychENCODE, AMP-AD, CommonMind and AMP-PD are good advocates and practitioners for that policy. Centralized data repositories such as Synapse (https://www.synapse.org) and NIGADS (https://www.niagads.org) have made data sharing and downloading easy. Open sharing policy is also being applied to the protocols, methods and codes, to enhance the research reproducibility [115].

Key Points

- Global efforts in brain research have made many multi-omics data resources available.
- Open data sharing policy will make the best use of the data.
- Integrative analysis across multiple cohorts is needed.
- The fast-evolving single-cell omics technologies are revolutionizing brain research.
- A public-editable webpage for the brain multi-omics resources is provided: http://bit.ly/brain_omics

Funding

XD was supported by American Parkinson's Disease Association, Aligning Science Across Parkinson's ASAP-000301 through the Michael J. Fox Foundation for Parkinson's Research (MJFF), and NIH U01NS120637. CL was supported by NIH U01MH116489, U01MH122591, 1R01MH110920, U01MH103340. MD was supported by the PhRMA Foundation Research Informatics Award and the George and Lavinia Blick Research Scholarship.

Conflicts of Interest

No conflict of interest is claimed by the authors. Funding body did not have any role in planning of the study.

References

1. Vos T, Lim SS, Abbafati C, *et al*. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease study 2019. *The Lancet* 2020;**396**:1204–22.
2. Institute for Health Metrics and Evaluation. GBD Compare|IHME Viz Hub. https://vizhub.healthdata.org/gbd-compare/
3. 2020 Alzheimer's disease facts and figures. *Alzheimers Dement* 2020;**16**:391–460. https://pubmed.ncbi.nlm.nih.gov/32157811/
4. Adams A, Albin S, Amunts K, *et al*. International brain initiative: an innovative framework for coordinated global brain research efforts. *Neuron* 2020;**105**:212–6.
5. Nott A, Holtman IR, Coufal NG, *et al*. Brain cell type-specific enhancer-promoter interactome maps and disease-risk association. *Science* 2019;**366**:1134–9.
6. Ursini G, Punzi G, Chen Q, *et al*. Convergence of placenta biology and genetic risk for schizophrenia. *Nat Med* 2018;**24**:792–801.
7. Irimia M, Weatheritt RJ, Ellis JD, *et al*. A highly conserved program of neuronal microexons is Misregulated in autistic brains. *Cell* 2014;**159**:1511–23.
8. Frisch T, Elkjaer ML, Reynolds R, *et al*. MS atlas - a molecular map of brain lesion stages in progressive multiple sclerosis. *Netw Syst Med* 2019;**3**:122–9.
9. Pfisterer U, Petukhov V, Demharter S, *et al*. Identification of epilepsy-associated neuronal subtypes and gene expression underlying epileptogenesis. *Nat Commun* 2020;**11**:5038.
10. Dong X. List of multi-omics datasets for human brain disorders. http://bit.ly/brain_omics
11. ENCODE Project Consortium. The ENCODE (ENCyclopedia of DNA elements) project. *Science* 2004;**306**:636–40.
12. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;**489**:57–74.
13. Wang J, Zhuang J, Iyer S, *et al*. Factorbook.Org: a wiki-based database for transcription factor-binding data generated by the ENCODE consortium. *Nucleic Acids Res* 2013;**41**:D171–6.
14. Human brain data in ENCODE. https://www.encodeproject.org/matrix/?type=Experiment&status=released&replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&biosample_ontology.organ_slims=brain&award.project=ENCODE
15. ENCODE Project Consortium, Snyder MP, Gingeras TR, *et al*. Perspectives on ENCODE. *Nature* 2020;**583**:693–8.
16. Satterlee JS, Chadwick LH, Tyson FL, *et al*. The NIH common fund/roadmap Epigenomics program: successes

of a comprehensive consortium. *Sci Adv* 2019;**5**: eaaw6507.
17. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, *et al*. Integrative analysis of 111 reference human epigenomes. *Nature* 2015;**518**:317–29.
18. Lizio M, Harshbarger J, Abugessaisa I, *et al*. Update of the FANTOM web resource: high resolution transcriptome of diverse cell types in mammals. *Nucleic Acids Res* 2017;**45**:D737–43.
19. Abugessaisa I, Ramilowski JA, Lizio M, *et al*. FANTOM enters 20th year: expansion of transcriptomic atlases and functional annotation of non-coding RNAs. *Nucleic Acids Res* 2021;**49**:D892–8.
20. Andersson R, Gebhard C, Miguel-Escalada I, *et al*. An atlas of active enhancers across human cell types and tissues. *Nature* 2014;**507**:455–61.
21. Jansen IE, Savage JE, Watanabe K, *et al*. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat Genet* 2019;**51**: 404–13.
22. GTEx Consortium, Laboratory. Data Analysis & Coordinating Center (LDACC)—analysis working group, statistical methods groups—analysis working group, et al. genetic effects on gene expression across human tissues. *Nature* 2017;**550**:204–13.
23. Perez-Riverol Y, Bai M, da Veiga Leprevost F, *et al*. Discovering and linking public omics data sets using the omics discovery index. *Nat Biotechnol* 2017;**35**:406–9.
24. Jones AR, Overly CC, Sunkin SM. The Allen brain atlas: 5 years and beyond. *Nat Rev Neurosci* 2009;**10**:821–8.
25. Hawrylycz MJ, Lein ES, Guillozet-Bongaarts AL, *et al*. An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* 2012;**489**:391–9.
26. Berg J, Sorensen SA, Ting JT, *et al*. Human cortical expansion involves diversification and specialization of supragranular intratelencephalic-projecting neurons. *bioRxiv* 2020; doi: 10.1101/2020.03.31.018820.
27. Akbarian S, Liu C, Knowles JA, *et al*. The psych ENCODE project. *Nat Neurosci* 2015;**18**:1707–12.
28. Amiri A, Coppola G, Scuderi S, *et al*. Transcriptome and epigenome landscape of human cortical development modeled in organoids. *Science* 2018;**362**:eaat6720.
29. Zhu Y, Sousa AMM, Gao T, *et al*. Spatiotemporal transcriptomic divergence across human and macaque brain development. *Science* 2018;**362**:eaat8077.
30. Psych ENCODE publication collection. https://www.sciencemag.org/collections/psychencode
31. Wang D, Liu S, Warrell J, *et al*. Comprehensive functional genomic resource and integrative model for the human brain. *Science* 2018;**362**:eaat8464.
32. Mirra SS, Heyman A, McKeel D, *et al*. The consortium to establish a registry for Alzheimer's disease (CERAD): part II. Standardization of the neuropathologic assessment of Alzheimer's disease. *Neurology* 1991;**41**:479–9.
33. Olah M, Patrick E, Villani A-C, *et al*. A transcriptomic atlas of aged human microglia. *Nat Commun* 2018;**9**:539.
34. Olah M, Menon V, Habib N, *et al*. A single cell-based atlas of human microglial states reveals associations with neurological disorders and histopathological features of the aging brain. *bioRxiv* 2018; doi: 10.1101/343780.
35. Iwaki H, Leonard HL, Makarious MB, *et al*. Accelerating Medicines Partnership: Parkinson's disease. Genetic resource. *medRxiv* 2020; doi: 10.1101/2020.11.19.20235192.
36. Dong X, Liao Z, Gritsch D, *et al*. Enhancers active in dopamine neurons are a primary link between genetic variation and neuropsychiatric disease. *Nat Neurosci* 2018;**21**:1482–92.
37. Hoffman GE, Bendl J, Voloudakis G, *et al*. Common mind consortium provides transcriptomic and epigenomic data for schizophrenia and bipolar disorder. *Scientific Data* 2019;**6**:180.
38. Schubert CR, O'Donnell P, Quan J, *et al*. Brain Seq: Neurogenomics to drive novel target discovery for neuropsychiatric disorders. *Neuron* 2015;**88**:1078–83.
39. Jaffe AE, Straub RE, Shin JH, *et al*. Developmental and genetic regulation of the human cortex transcriptome illuminate schizophrenia pathogenesis. *Nat Neurosci* 2018;**21**:1117–25.
40. Jaffe AE, Gao Y, Deep-Soboslay A, *et al*. Mapping DNA methylation across development, genotype and schizophrenia in the human frontal cortex. *Nat Neurosci* 2016;**19**:40–7.
41. Mandell KAP, Eagles NJ, Wilton R, *et al*. Widespread methylation quantitative trait loci and their role in schizophrenia risk. *bioRxiv* 2020. doi: 10.1101/2020.09.24.311878.
42. Song M, Pebworth M-P, Yang X, *et al*. Cell-type-specific 3D epigenomes in the developing human cortex. *Nature* 2020;**587**:644–9.
43. Darmanis S, Sloan SA, Zhang Y, *et al*. A survey of human brain transcriptome diversity at the single cell level. *Proc Natl Acad Sci* 2015;**112**:7285–90.
44. Nowakowski TJ, Bhaduri A, Pollen AA, *et al*. Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. *Science* 2017;**358**: 1318–23.
45. Luo C, Keown CL, Kurihara L, *et al*. Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. *Science* 2017;**357**:600–4.
46. Mathys H, Davila-Velderrain J, Peng Z, *et al*. Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* 2019;**570**:332–7.
47. Song L, Pan S, Zhang Z, *et al*. STAB: a spatio-temporal cell atlas of the human brain. *Nucleic Acids Res* 2021;**49**:1029–37.
48. 10x Genomics. 1.3 Million Brain Cells from E18 Mice. https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.3.0/1M_neurons
49. Stuart T. Single-cell ATAC + RNA co-assay methods. https://timoast.github.io/blog/single-cell-atac-rna/
50. Lake BB, Chen S, Sos BC, *et al*. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat Biotechnol* 2018;**36**:70–80.
51. Trevino AE, Sinnott-Armstrong N, Andersen J, *et al*. Chromatin accessibility dynamics in a model of human forebrain development. *Science* 2020;**367**:eaay1645.
52. Li M, Santpere G, Imamura Kawasawa Y, *et al*. Integrative functional genomic analysis of human brain development and neuropsychiatric risks. *Science* 2018;**362**:eaat7615.
53. van der Wijst M, de Vries D, Groot H, *et al*. The single-cell eQTLGen consortium. *Elife* 2020;**9**:e52155.
54. Ståhl PL, Salmén F, Vickovic S, *et al*. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* 2016;**353**:78–82.
55. Maynard KE, Collado-Torres L, Weber LM, *et al*. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *bioRxiv* 2020; February 28, 2020, doi 10.1101/2020.02.28.969931.

56. Rodriques SG, Stickels RR, Goeva A, *et al*. Slide-seq: a scalable technology for measuring genome-wide expression at high spatial resolution. *Science* 2019;**363**:1463–7.
57. Vickovic S, Eraslan G, Salmén F, *et al*. High-definition spatial transcriptomics for in situ tissue profiling. *Nat Methods* 2019;**16**:987–90.
58. Chen KH, Boettiger AN, Moffitt JR, *et al*. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* 2015;**348**:aaa6090.
59. Jaffe AE, Hoeppner DJ, Saito T, *et al*. Profiling gene expression in the human dentate gyrus granule cell layer reveals insights into schizophrenia and its genetic risk. *Nat Neurosci* 2020;**23**:510–9.
60. Fan Z, Chen R, Chen X. SpatialDB: a database for spatially resolved transcriptomes. *Nucleic Acids Res* 2020;**48**:D233–7.
61. 10x Genomics. Spatial Gene Expression Datasets. https://support.10xgenomics.com/spatial-gene-expression/datasets.
62. Zhu Q, Shah S, Dries R, *et al*. Identification of spatially associated subpopulations by combining scRNAseq and sequential fluorescence in situ hybridization data. *Nat Biotechnol* 2018;**36**:1183–90.
63. Abdelaal T, Mourragui S, Mahfouz A, *et al*. SpaGE: spatial gene enhancement using scRNA-seq. *Nucleic Acids Res* 2020;**48**:e107–7.
64. Won H, de la Torre-Ubieta L, Stein JL, *et al*. Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature* 2016;**538**:523–7.
65. Giusti-Rodríguez P, Lu L, Yang Y, *et al*. Using three-dimensional regulatory chromatin interactions from adult and fetal cortex to interpret genetic results for psychiatric disorders and cognitive traits. *Bio Rxiv* 2019; doi: 10.1101/406330.
66. Song M, Yang X, Ren X, *et al*. Mapping cis-regulatory chromatin contacts in neural cells links neuropsychiatric disorder risk variants to target genes. *Nat Genet* 2019;**51**:1252–62.
67. Ulianov SV, Tachibana-Konwalski K, Razin SV. Single-cell Hi-C bridges microscopy and genome-wide sequencing approaches to study 3D chromatin organization. *Bioessays* 2017;**39**: 1700104.
68. Tan L, Ma W, Wu H, *et al*. Experience-independent transformation of single-cell 3D genome structure and transcriptome during postnatal development of the mammalian brain. *bioRxiv* 2020; doi: 10.1101/2020.04.02.022657.
69. Li G, Liu Y, Zhang Y, *et al*. Joint profiling of DNA methylation and chromatin architecture in single cells. *Nat Methods* 2019;**16**:991–3.
70. Lee D-S, Luo C, Zhou J, *et al*. Simultaneous profiling of 3D genome structure and DNA methylation in single human cells. *Nat Methods* 2019;**16**:999–1006.
71. Subramanian I, Verma S, Kumar S, *et al*. Multi-omics data integration, interpretation. *and Its Application Bioinform Biol Insights* 2020;**14**:1177932219899051.
72. Mohammadi P, Castel SE, Cummings BB, *et al*. Genetic regulatory variation in populations informs transcriptome analysis in rare disease. *Science* 2019;**366**: 351–6.
73. Montaner J, Ramiro L, Simats A, *et al*. Multilevel omics for the discovery of biomarkers and therapeutic targets for stroke. *Nat Rev Neurol* 2020;**16**:247–64.
74. Rappoport N, Shamir R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Res* 2018;**46**:1044.
75. Huang S, Chaudhary K, Garmire LX. More is better: recent progress in multi-omics data integration methods. *Front Genet* 2017;**8**:84.
76. Meng C, Zeleznik OA, Thallinger GG, *et al*. Dimension reduction techniques for the integrative analysis of multi-omics data. *Brief Bioinform* 2016;**17**:628–41.
77. Bersanelli M, Mosca E, Remondini D, *et al*. Methods for the integration of multi-omics data: mathematical aspects. *BMC Bioinformatics* 2016;**17**(Suppl 2):15.
78. Tini G, Marchetti L, Priami C, *et al*. Multi-omics integration-a comparison of unsupervised clustering methodologies. *Brief Bioinform* 2019;**20**:1269–79.
79. Rohart F, Gautier B, Singh A, *et al*. Mix omics: an R package for'omics feature selection and multiple data integration. *PLoS Comput Biol* 2017;**13**:e1005752.
80. Simidjievski N, Bodnar C, Tariq I, *et al*. Variational autoencoders for cancer data integration: design principles and computational practice. *Front Genet* 2019;**10**:1205.
81. Weinstein JN, Collisson EA, Mills GB, *et al*. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* 2013;**45**:1113–20.
82. Duren Z, Chen X, Zamanighomi M, *et al*. Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorizations. *Proc Natl Acad Sci U S A* 2018;**115**:7723–8.
83. Stuart T, Butler A, Hoffman P, *et al*. Comprehensive integration of single-cell data. *Cell* 2019;**177**:1888, e21–902.
84. Welch JD, Kozareva V, Ferreira A, *et al*. Single-cell multiomic integration compares and contrasts features of brain cell identity. *Cell* 2019;**177**:1873, e17–87.
85. Korsunsky I, Millard N, Fan J, *et al*. Fast, sensitive and accurate integration of single-cell data with harmony. *Nat Methods* 2019;**16**:1289–96.
86. Stuart T, Srivastava A, Lareau C, *et al*. Multimodal single-cell chromatin analysis with Signac. *bioRxiv* 2020; doi: 10.1101/2020.11.09.373613.
87. Granja JM, Corces MR, Pierce SE, *et al*. Arch R: an integrative and scalable software package for single-cell chromatin accessibility analysis. *bioRxiv* 2020; doi: 10.1101/2020.04.28.066498.
88. Przytycki PF, Pollard KS. Cell Walker integrates single-cell and bulk data to resolve regulatory elements across cell types in complex tissues. *bioRxiv* 2020; doi: 10.1101/847657.
89. Wang C, Sun D, Huang X, *et al*. Integrative analyses of single-cell transcriptome and regulome using MAESTRO. *Genome Biol* 2020;**21**:198.
90. Zeng W, Chen X, Duren Z, *et al*. DC3 is a method for deconvolution and coupled clustering from bulk and single-cell genomics data. *Nat Commun* 2019;**10**:4613.
91. Chen S, Lake BB, Zhang K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat Biotechnol* 2019;**37**:1452–7.
92. Sirugo G, Williams SM, Tishkoff SA. The missing diversity in human genetic studies. *Cell* 2019;**177**:26–31.
93. Fan X, Zhang X, Wu X, *et al*. Single-cell RNA-seq transcriptome analysis of linear and circular RNAs in mouse preimplantation embryos. *Genome Biol* 2015;**16**:148.
94. Sheng K, Cao W, Niu Y, *et al*. Effective detection of variation in single-cell transcriptomes using MATQ-seq. *Nat Methods* 2017;**14**:267–70.
95. Hayashi T, Ozaki H, Sasagawa Y, *et al*. Single-cell full-length total RNA sequencing uncovers dynamics of recursive splicing and enhancer RNAs. *Nat Commun* 2018;**9**: 619.

96. Saikia M, Burnham P, Keshavjee SH, *et al*. Simultaneous multiplexed amplicon sequencing and transcriptome profiling in single cells. *Nat Methods* 2019;**16**:59–62.
97. Pearson-Leary J, Zhao C, Bittinger K, *et al*. The gut microbiome regulates the increases in depressive-type behaviors and in inflammatory processes in the ventral hippocampus of stress vulnerable rats. *Mol Psychiatry* 2020;**25**: 1068–79.
98. Abbott A. Are infections seeding some cases of Alzheimer's disease? *Nature* 2020;**587**:22–5.
99. Heintz-Buschart A, May P, Laczny CC, *et al*. Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. *Nat Microbiol* 2017;**2**:16180.
100. Wang Q, Wang K, Wu W, *et al*. Host and microbiome multi-omics integration: applications and methodologies. *Biophysical Reviews* 2019;**11**:55–65.
101. Chen R, Mias GI. Li-Pook-than J, et al. personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* 2012;**148**:1293–307.
102. Price ND, Magis AT, Earls JC, *et al*. A wellness study of 108 individuals using personal, dense, dynamic data clouds. *Nat Biotechnol* 2017;**35**:747–56.
103. Schüssler-Fiorenza Rose SM, Contrepois K, Moneghetti KJ, *et al*. A longitudinal big data approach for precision health. *Nat Med* 2019;**25**:792–804.
104. Farh KK-H, Marson A, Zhu J, *et al*. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* 2015;**518**:337–43.
105. Schaid DJ, Chen W, Larson NB. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat Rev Genet* 2018;**19**:491–504.
106. Auerbach J, Howey R, Jiang L, *et al*. Causal modeling in a multi-omic setting: insights from GAW20. *BMC Genet* 2018;**19**:74.
107. Liao JG, Chin K-V. Logistic regression for disease classification using microarray data: model selection in a large p and small n case. *Bioinformatics* 2007;**23**:1945–51.
108. Kegerreis B, Catalina MD, Bachali P, *et al*. Machine learning approaches to predict lupus disease activity from gene expression data. *Sci Rep* 2019;**9**:9617.
109. Langfelder P, Horvath SWGCNA. An R package for weighted correlation network analysis. *BMC Bioinformatics* 2008;**9**: 559.
110. Asif H, Alliey-Rodriguez N, Keedy S, *et al*. GWAS significance thresholds for deep phenotyping studies can depend upon minor allele frequencies and sample size. *Mol Psychiatry* 2020. doi: 10.1038/s41380-020-0670-3
111. Gandal MJ, Zhang P, Hadjimichael E, *et al*. Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. *Science* 2018;**362**:eaat8127.
112. Collado-Torres L, Burke EE, Peterson A, *et al*. Regional heterogeneity in gene expression, regulation, and coherence in the frontal cortex and hippocampus across development and schizophrenia. *Neuron* 2019;**103**:203, e8–16.
113. Jiang Y, Giase G, Grennan K, *et al*. DRAMS: a tool to detect and re-align mixed-up samples for integrative studies of multi-omics data. *PLoS Comput Biol* 2020;**16**:e1007522.
114. Pereira S, Gibbs R, McGuire A. Open access data sharing in genomic research. *Gen* 2014;**5**:739–47.
115. Turkyilmaz-van der Velden Y, Dintzner N, Teperek M. Reproducibility starts from you today. *Patterns* 2020; **100099**:1.