Integration of Multimodal Data for Deciphering Brain Disorders

Jingqi Chen,1,2,3 Guiying Dong,1 Liting Song,1 Xingzhong Zhao,1 Jixin Cao,1 Xiaohui Luo,1 Jianfeng Feng,1,2,3,4 and Xing-Ming Zhao1,2,3

1Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai 200433, China; email: xmzhao@fudan.edu.cn, jffeng@fudan.edu.cn
2MOE Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence and MOE Frontiers Center for Brain Science, Fudan University, Ministry of Education, Shanghai 200433, China
3Zhangjiang Fudan International Innovation Center, Fudan University, Shanghai 200433, China
4Department of Computer Science, University of Warwick, Coventry CV4 7AL, United Kingdom

Abstract

The accumulation of vast amounts of multimodal data for the human brain, in both normal and disease conditions, has provided unprecedented opportunities for understanding why and how brain disorders arise. Compared with traditional analyses of single datasets, the integration of multimodal datasets covering different types of data (i.e., genomics, transcriptomics, imaging, etc.) has shed light on the mechanisms underlying brain disorders in greater detail across both the microscopic and macroscopic levels. In this review, we first briefly introduce the popular large datasets for the brain. Then, we discuss in detail how integration of multimodal human brain datasets can reveal the genetic predispositions and the abnormal molecular pathways of brain disorders. Finally, we present an outlook on how future data integration efforts may advance the diagnosis and treatment of brain disorders.
INTRODUCTION

With the development and wide application of modern technologies such as high-throughput sequencing and magnetic resonance imaging (MRI), we have observed in recent years a great increase in the number of datasets collected for the human brain for both normal and disordered conditions. Two significant features of the recently accumulated datasets stand out—their large sample sizes and the diverse data types within the same dataset. On the one hand, the large sample sizes have been shown to greatly improve the statistical power and thus the number of significant findings, as repeatedly proven by the Psychiatric Genomics Consortium (PGC) (1). On the other hand, different types of data collected from the same or different subjects provide a unique way to dissect a condition (usually a disease) from different angles that may corroborate one another. For example, PsychENCODE collected transcriptomics and epigenetics data for schizophrenia (SCZ) and autism spectrum disorder (ASD) and identified novel etiological molecular pathways confirmed by both gene expression and gene regulation (2). With large sample sizes and diverse data types, it is now possible to interrogate the mechanisms underlying brain disorders with unprecedented resolution and confidence.

Among the large brain datasets, the most typical ones are molecular omics data, such as genomics, transcriptomics, epigenetics, and chromatin 3D structures, as well as brain imaging data such as structural MRI (sMRI) and functional MRI (fMRI). We have provided a resource summarizing popular genetics, epigenetics, and imaging datasets available for the human brain, along with their essay types, diseases of interest, sample sizes, and web links (Supplemental Table 1).

Genomics and Transcriptomics

In the past few decades, high-throughput DNA chips have enabled researchers to cost-effectively collect information on variations such as single nucleotide polymorphisms (SNPs) for a preselected large set of genomic loci, yielding studies known as genome-wide association studies (GWAS). GWAS have identified hundreds and thousands of genomic loci that are significantly associated with various brain disorders. For example, PGC identified 108 significant genomic loci for SCZ with a sample size of about 150,000 (∼37,000 cases) (3). More recently, whole-exome and whole-genome sequencings have been used to gain information on rarer variants and larger variants (such as structural variants). As whole-exome sequencing and whole-genome sequencing are much more expensive and computationally demanding than DNA chips, the sample sizes of such studies are usually not as large as those of GWAS. For example, the Autism Sequencing Consortium (ASC) planned to sequence the exome of about 10,000 ASD patients (4). At the same time, microarrays and high-throughput sequencing are both commonly used ways to collect transcriptomics data. The Allen Brain Atlas provided transcriptomic illustrations across multiple brain regions through microarrays, while more recent efforts such as PsychENCODE collected more detailed transcriptomic information through next-generation sequencing (2, 5). In the past 5–10 years, single-cell sequencing has developed very swiftly and been applied to investigate multiple brain regions in both normal and disease conditions, as concluded in the STAB (Spatiotemporal Cell Atlas of the Human Brain) single-cell RNA-Seq (RNA sequencing) database (6).

Epigenetics and Chromatin 3D Structures

Along with sequencing, both DNase-Seq [DNase I hypersensitive sites (DHS) sequencing] and ATAC-Seq (assay for transposase-accessible chromatin using sequencing) have been widely used to identify open chromatin regions, an important indication of whether chromatin regions are actively involved in transcription or regulation; examples include the DHS datasets (generated
through DNase-Seq) hosted by the Roadmap Epigenomics Project (7) and the ATAC-Seq datasets collected by PsychENCODE (2). In addition to open chromatin regions, chromatin modifications and transcription factor binding sites that can be identified using methods like ChIP-Seq (chromatin immunoprecipitation and sequencing) also provide valuable information on the transcriptional regulation status in a given condition of the brain. For example, PsychENCODE collected chromatin modifications for H3K27Ac for three brain regions (prefrontal cortex, temporal cortex, and cerebellum) in SCZ patients and healthy controls (2). Moreover, DNA and chromatin methylation profiles have become increasingly addressed in the investigation of brain disorders, for example, the Bisulfite-Seq (bisulfite sequencing), MeDIP-Seq (methylated DNA immunoprecipitation sequencing), and MRE-Seq (methylation-sensitive restriction enzyme sequencing) data for multiple fetal and adult brain regions, which are collected by the Roadmap Epigenomics Project (7). The 3D interaction profiles among chromatin regions can be identified using high-throughput assays, such as Hi-C (high-throughput chromosome conformation capture) and CHi-C (capture Hi-C), and provide important information on the global regulation relationships between genes and various regulators (such as enhancers) (8, 9).

Typical chromatin 3D interaction datasets for the brain include the fetal cortex Hi-C dataset by Won et al. (10), the dorsolateral prefrontal cortex (DLPFC) and hippocampus Hi-C datasets uniformly processed by Yang et al. [3DIV (3D genome interaction viewer and database)] (11), the high-resolution DLPFC Hi-C dataset by PsychENCODE (2), and the DLPFC and hippocampus CHi-C datasets by Jung et al. (12) (see Supplemental Table 1).

**Structural and Functional Magnetic Resonance Imaging**

Apart from molecular omics data, brain imaging data comprise another essential part of recently published large brain datasets (Supplemental Table 1). MRI is the most commonly used technique to obtain images of the brain in both normal and disease conditions. Generally, there are three main branches of MRI, namely, sMRI, fMRI, and diffusion tensor imaging (DTI). sMRI presents multiple structural measures across the brain, DTI shows the connections of different regions through white matter fiber tracts, and fMRI can be used to measure the functional associations among regions of interest within the brain (e.g., the connectome). Researchers may choose to perform fMRI on subjects in a resting state (resting-state fMRI) or on subjects before and after carrying out certain tasks (task-based fMRI). sMRI has been used to show abnormal structural variations in multiple large brain datasets, such as the Alzheimer’s Disease Neuroimaging Initiative (ADNI) (13) and IMAGEN (14). DTI has been used to measure and monitor the abnormalities of brain white matter, such as in the ENIGMA (enhancing neuroimaging genetics through meta-analysis) Consortium (15). fMRI has also shown great potential in elucidating the functional abnormalities in brain disorders, especially in psychiatric disorders, as exemplified by datasets such as the 1000 Functional Connectomes Project (16) and SchizConnect (17).

Integrating different types of brain datasets has so far shown great promise in the battle against various brain disorders. PsychENCODE presented a successful example of integrating multi-omics molecular datasets of the brain to interpret how genomic loci level findings for SCZ can be confirmed by and translated into transcriptional level signals by jointly interrogating the genomic variants (GWAS), the gene expression changes (RNA-Seq), the transcriptional factor binding sites (ChIP-Seq), and the 3D interactions of regions (Hi-C) (2, 18). Previously we also successfully performed an analysis to integrate transcriptome and fMRI-derived connectomes to identify gene signatures for age-specific cortex development, which were found to be significantly associated with brain disorders (19). In another study, we integrated neuroimaging and genomics data from several large cohorts to find genetic loci associated with brain structural variations indicated...
in SCZ (20). In the following sections, we discuss many examples of how integrating different types of brain datasets can shed light on the investigation of brain disorders, driven by two types of scientific issues: the genetic predispositions of brain disorders and the molecular pathways underlying the development of brain disorders. We do not intend to cover all important works in this field; instead, we provide a few examples to illustrate the main directions. At the end of the review we discuss potential obstacles and promising future directions.

IDENTIFYING THE GENETIC PREDISPOSITIONS AND BIOMARKERS OF BRAIN DISORDERS

Understanding the genetic predispositions and biomarkers of brain disorders—especially common neuropsychiatric disorders such as SCZ—has been essential for understanding how brain disorders arise. With large cohorts of genomic data becoming available, more genetic risk loci have been identified over the years (21). It is then critical to identify real signals among the many findings that could contain noises, as well as to increase the reproducibility of real findings. Moreover, many of the genetic risk loci are found to have very small effects, and they together contribute to the onset of disorders (21). As discussed below in detail and shown in Figure 1, researchers have addressed these questions by integrating the same (e.g., multiple GWAS studies) or different (e.g., GWAS and neuroimaging studies) types of brain datasets.

Identifying Disorder Risk Loci Through the Integration of Multiple GWAS Cohorts

Integrating multiple cohorts (e.g., GWAS) of the same disorder through meta-analysis can usually increase the credibility of the results, while integrating cohorts of multiple disorders may help find the distinct and convergent risk factors of different disorders. For example, Demontis et al. integrated 12 GWAS cohorts for attention deficit/hyperactivity disorder (ADHD) from iPSYCH (The Lundbeck Foundation Initiative for Integrative Psychiatric Research) and PGC and identified the first 12 independent loci that are genome-wide significant for this disorder, most of which were further supported by 3 additional independent datasets (22). In another study, Lam et al. performed a fixed-effect meta-analysis on SCZ cohorts of East Asian and European ancestries (from PGC), followed by a trans-ancestry fine-mapping that focused the findings onto 44 potentially causal loci (23). These examples demonstrated the power of meta-analysis of multiple cohorts for the same disorder.

In order to perform integrative analysis on cohorts for different disorders, researchers need methods for comparing and quantifying the findings, especially when many small-effect loci exist for many neuropsychiatric disorders. To this end, widely used methods have been developed for estimating polygenic risks, such as LDpred (linkage disequilibrium predictor; 24), and genetic correlation, such as LDSC (linkage disequilibrium score regression; 25) and Popcorn (26), as reviewed by Martin et al. (27). In a cross-disorder meta-analysis of eight psychiatric disorders, researchers from PGC identified 109 loci that are pleiotropic (associated with at least two disorders) and employed genetic correlation (using LDSC) to categorize the disorders into three mechanistically meaningful groups, characterized respectively by compulsive behaviors [anorexia nervosa, obsessive–compulsive disorder, and Tourette syndrome (TS)], mood and psychotic disorders (major depression, bipolar disorder, and SCZ), and early-onset neurodevelopmental disorders (ASD, ADHD, and TS) (28). Another cross-disorder analysis by the Brainstorm Consortium showed that psychiatric disorders such as ADHD, major depressive disorder (MDD), and SCZ share significantly correlated common risk loci and are strongly correlated with the personality trait...
Figure 1
Identification of disease-associated genome loci or molecular biomarkers through the integration of multimodal data. In the clockwise direction, the three arrows respectively indicate the integration of neuroimaging and genomics data; the integration of multiple GWAS cohorts; and the integration of genomics, epigenomics, chromatin 3D structure, and transcriptomics data. Database logos are courtesy of their official websites. Abbreviations: CHi-C, capture Hi-C; GWAS, genome-wide association study; Hi-C, high-throughput chromosome conformation capture.

neuroticism and early-life cognitive measures, while neurological disorders [Alzheimer’s disease (AD), epilepsy, and migraine] are more distinct from each other and less correlated with the three aforementioned psychiatric disorders, neuroticism, or early-life cognitive measures (29). These findings have repeatedly underscored the shared heritability and etiology among some of the brain disorders (especially the psychiatric disorders) and will promote a rethinking of comorbidities, clinical subtypes, and effective managements of such disorders from a genetic point of view.

Inferring Risk Genes Through the Integration of Multi-Omics Molecular Datasets
Researchers have also demonstrated through multiple successful cases that integration of genomic information with other molecular omics data, including transcriptomic data [such as expression quantitative trait loci (eQTL)] and chromatin 3D structure data, can help pinpoint credible risk genes or biomarkers for disorders. For example, Matoba et al. identified DDHD2 as a candidate risk gene for ASD by combining risk loci found in a meta-analysis of genomic datasets and existing knowledge of eQTL for adult and prenatal brains (30). Based on the MAGMA (31) method for aggregating GWAS summary information to identify risk genes, Gerring et al. designed eMAGMA
to identify risk genes by integrating GWAS summary statistics and eQTL relationships, and with this method they successfully identified 58 novel MDD risk genes (32, 33). Sey et al. established H-MAGMA to further incorporate brain chromatin interaction profiles and identified many risk genes that were distally regulated by risk loci (34). We also contributed a method named nMAGMA, which integrated eQTL, chromatin interaction profiles, and gene–gene networks to more thoroughly identify risk genes, and demonstrated its usefulness in finding meaningful novel risk genes in SCZ (35). The integration of multi-omics data is still quite new and has not yet been widely used in GWAS analyses, but we consider it highly promising that these methods will help better integrate multi-omics data and uncover more disease biomarkers.

**Identifying Loci Associated with Brain Structures Through the Integration of Genomic and Neuroimaging Datasets**

In addition to integration of multi-omics molecular datasets, neuroimaging features have also proven useful in prioritizing risk loci and genes for multiple brain disorders. Scelsi et al. combined neuroimaging phenotypes of cortical amyloid burden and bilateral hippocampal volume based on PET (positron emission tomography) and T1 MRI data to compute a disease progression score, which was then used for identifying significantly associated genomic loci from the genomic dataset (all datasets from ADNI) (36). Also integrating data in ADNI, Du et al. designed a different approach using multitask sparse canonical correlation analysis to identify SNPs that affect brain imaging quantitative traits over time (37). Elliot et al. used the matched genomic, fMRI, and sMRI data from UK Biobank, carried out a GWAS for the imaging phenotypes, and identified 148 clusters of significant genetic loci, some of which were interpretable through existing knowledge; for example, loci near genes for mid-line axon development were associated with the organization of the pontine-crossing tract (38). With sMRI and genomic data for almost 34,000 subjects from the ENIGMA consortium and UK Biobank, Grasby et al. identified 306 genome-wide-significant loci that are associated with cortical structural measures such as cortical surface area and thickness and showed that the genetic loci profile of cortical surface area is significantly correlated with that of many brain disorders, e.g., Parkinson’s disease, MDD, and ADHD (39). Through a multicenter study on SCZ patients and healthy subjects from the IMAGEN cohort, UK biobank, and three other cohorts, Luo et al. performed a voxel-wise GWAS to identify genetic loci associated with adolescent brain structure (20); they successfully found a nonsynonymous SNP in the SCZ risk gene \textit{SLC39A8} that was associated with greater gray matter volume of the putamen and displayed different association strength in SCZ patients and healthy subjects. These recent publications highlighted the importance of integrating large genomic and neuroimaging datasets towards finding meaningful risk loci and biomarker genes for brain disorders.

**ELUCIDATING THE MOLECULAR MECHANISMS UNDERLYING BRAIN DISORDERS**

Beyond the genetic predispositions and biomarkers for brain disorders, hypothesis-driven data integration of different types of brain datasets can further help reveal the molecular alterations before or after disorders arise—for example, biological processes, pathways, and neuronal circuits that alter due to genetic predispositions, environmental factors, or a combination of both. Such molecular alterations, which underlie the development of brain disorders, will then lead to a deepened understanding of how the human brain functions and dysfunctions and pave the way for better, more effective treatment, which is still largely lacking for many brain disorders. In this section, as shown in Figure 2, we discuss recent progress in this area according to two categories: the integration of molecular-omics data and the integration of neuroimaging and molecular-omics data.
Pinpointing the molecular mechanisms underlying brain disorders based on the integration of imaging and omics data. Unraveling the functional context of genomic loci, including SNPs, CNVs, and SVs, identified through genomic and neuroimaging studies of brain disorders, may help explain how the neural circuits are affected in these disorders. (●) The epigenetic features associated with aberrant brain structure or function identified with the integration of neuroimaging and epigenetics data. (●) The genes or variants associated with aberrant brain structures or functions identified through the joint analysis of neuroimaging and genomics data. Abbreviations: CADD, combined annotation-dependent depletion; CNV, copy number variant; eQTL, expression quantitative trait loci; fMRI, functional MRI; GTEx, genotype-tissue expression; Hi-C, high-throughput chromosome conformation capture; MRI, magnetic resonance imaging; PPI, protein–protein interaction; PRS, polygenic risk score; sMRI, structural MRI; SNP, single-nucleotide polymorphism; SV, structural variant.

Understanding the Molecular Mechanisms Underlying Disorders by Exploring the Functional Context of Genetic Risk Loci

Milestone findings have emerged in recent years that integrate many types of molecular-omics data to understand the molecular processes underlying brain disorders. Consortium efforts such as the GTEx project (40), the Roadmap Epigenomics Project (7), and the Allen Human Brain Atlas (5) have provided illustrations and benchmarks of the epigenetic and transcriptomic molecular features in the normal human brain, which have laid a foundation for disorder-oriented research. Large datasets on the disordered brain (usually for both case and control), such as the PGC (1), PsychENCODE (2) and ASC (4), in contrast, focus more on specific conditions and often simultaneously incorporate information from large consortium datasets for the normal brain. Based on the integration of such large molecular-omics datasets, existing knowledge on genetics and genomics, and sometimes local smaller, more specific datasets, researchers have made much progress in tackling the questions underlying the development of many brain disorders. For example, with the genetic loci identified from a meta-analysis on neuroticism in 449,484 subjects from UK Biobank, 23andme, and Genetic Personality Consortium, Nagel et al. performed comprehensive functional enrichment analysis by integrating loci with epigenetics...
annotations, chromatin interactions, tissue/cell-specific gene expression, and gene set (such as biological processes) annotations and revealed the involvement of dopaminergic neuroblasts, medium spiny neurons, serotonergic neurons, and six brain regions in neuroticism (41). Similarly, Howard et al. identified the involvement of the frontal cortex and anterior cingulate cortex in depression via analysis of risk loci with tissue-specific expression data from the GTEx project (42). Xu et al. analyzed the functional implications of risk loci for depression through gene set enrichment, protein–protein interaction, and cell type–specific gene expression and found that they may regulate early development of the hippocampus and affect the amyloid-beta binding process (43). This analysis was used to interpret the mechanisms by which depression contributes to the risk of conversion from amnestic mild cognitive impairment to AD.

In the three aforementioned examples, the researchers all started from risk loci for brain disorders and then performed follow-up analyses to identify the molecular implications. There are also many examples where researchers have focused on investigating a normal process (e.g., development or aging) of the brain and then bridged the findings to understand brain disorders. Oh et al. analyzed a large microarray gene expression dataset (which included the North American Brain Expression Consortium dataset and the UK Human Brain Expression Database) and a DNA modification dataset from the North American Brain Expression Consortium and found that, interestingly, the brains of older people are more similar than those of younger people at both the epigenetic and transcriptomic levels (44). They then found clues for accelerated epigenetic assimilation in AD, which provided a novel angle for understanding the origins of neurodegenerative disorders (44). Using whole-genome Bisulfite-Seq for NeuN+ neurons in the first two decades of human cortex development, Price et al. identified developmentally dynamic and neuron-specific DNA methylation patterns with enriched heritability for neuropsychiatric disorders, indicating the importance of DNA methylation in the development of brain disorders (45). Song et al. performed CHi-C, ATAC-Seq, and RNA-Seq in four distinct neural cell types, identifying a large number of long-range interaction relationships between promoter and distal regions (such as enhancers and repressors) and revealing putative biological processes that might be disrupted by the effects of disorder risk loci on such long-range relationships (6).

Translating Genetic Findings into Pathophysiological Mechanisms of Brain Disorders Through the Integration of Neuroimaging and Molecular-Omics Datasets

Recently, much effort has been made to integrate neuroimaging and molecular-omics data and recapitulate the disorder-associated changes at the molecular level, as well as in the neuronal circuits. On the one hand, based on genetic findings, researchers have integrated neuroimaging data to understand how genetic findings translate into alterations in neuronal circuits. For example, Erk et al. investigated the effects of SCZ risk loci on brain activation in five Research Domain Criteria neurocognitive domains in unaffected subjects by jointly analyzing their genomic and fMRI data, and they found that one of the SNPs may confer SCZ risk by affecting amygdala recruitment during emotion processing (46). Similarly, Cao et al. integrated the genetic and multiparamadigm fMRI data of healthy subjects from the Human Connectome Project and found that the SCZ polygenic risk scores were correlated with lower functional connectivity involving the visual, default-mode, and frontoparietal systems (47). Combining the genetic and sMRI information hosted in UK Biobank, Neilson et al. (48) and Warland et al. (49) identified cortical structural features that might be affected by SCZ risk SNPs and copy number variants (CNVs), respectively. Neilson et al. (48) found that higher polygenic risk scores for SCZ inferred from SNPs were associated with lower cortical thickness, and Warland et al. (49) found that SCZ risk CNV carriers
showed reduced subcortical volumes in SCZ-relevant subcortical regions, both of which provide an extra level of explanation for how genetic risks translate into pathophysiological conditions.

On the other hand, researchers have also investigated the relationships (such as correlation) between molecular multi-omics features and neuroimaging features and then in turn learned useful mechanisms underlying disorder conditions. Gaiteri et al. used matched gene expression, DNA methylation, and sMRI data of the postmortem brains of 222 subjects from the Religious Orders Study and the Rush Memory and Aging Project and identified many features of gene expression and loci methylation that were correlated with brain microstructural characteristics, with some of the genes associated with AD phenotypes (50). As part of the ENIGMA Consortium, Jia et al. integrated 11 international cohorts and identified significant correlations between certain blood DNA methylation loci and hippocampus volume. They showed that DNA methylation at these loci may affect genes involved in processes known to be related to neuropsychiatric disorders, such as learning and memory (51). Based on a combination of three sMRI cohorts, Morgan et al. evaluated the structural similarity between brain regions in psychosis by an approach they called morphometric similarity mapping and found reduced similarity in patients’ frontal and temporal regions; they also found that genes that were expressionally correlated with such anatomical patterns were enriched for nervous system–related functions and were associated with SCZ (52). These findings, as well as future ones, will eventually help us decipher pathological mechanisms at both the molecular and the subphenotypic levels, making it possible to construct fuller pictures of how brain disorders arise.

**ADVANCING THE TREATMENT OF BRAIN DISORDERS WITH LARGE OMICS AND IMAGING DATA**

The next steps after identifying biomarkers and molecular mechanisms for brain disorders, ultimately, would be to optimize the current treatment strategies. For many complex brain disorders, there are no or very few options of effective interventions, the effects of which often vary from patient to patient. To this end, here we briefly discuss three possible ways that large brain datasets could help with advancing the treatment of brain disorders.

Firstly, the integration of current knowledge and existing data could help prioritize molecules and pathways that may serve as novel drug targets. As discussed in previous sections, researchers have successfully identified many biomarkers and pathways indicated in brain disorders through the analysis of large brain datasets. Taking AD as an example, current prescription drugs such as cholinesterase inhibitors and memantine only slightly improve the condition, and the difference between these drugs and placebo is actually not very significant (53). Therefore, researchers have been testing novel treatments using many strategies informed by findings from large omics datasets. For example, immunotherapies are being evaluated in clinical trials that target amyloid-beta and tau proteins and some neural pathways that have been shown by many omics data–driven studies to be associated with AD (54). Moreover, microglia therapeutic targets have gained some attention, as omics data integration studies have found that microglia may play an important role in AD (55). In both examples, the accumulation and analysis of large brain datasets have provided promising directions for research, although we are still far from the development of effective drugs for AD.

Secondly, the integration of molecular and imaging brain datasets may help better define brain disorders and their subtypes so as to better guide treatment. For example, Kalin has discussed how to integrate neuroimaging and genomics data to help guide the diagnosis and treatment of SCZ (56). By integrating GWAS datasets for MDD and bipolar disorder with individual drug response information, Amare et al. found that bipolar disorder patients who had a low polygenic risk for
MDD were more likely to have a good response to lithium, a drug for treating bipolar disorder, which helped to explain why only one-third of bipolar disorder patients respond well to lithium and highlighted the importance of disease subtypes (57).

Thirdly, brain multimodal datasets may shed light on novel intervention strategies, in addition to drug treatment. For example, Chung et al. successfully employed fMRI to evaluate the treatment effect of cognitive behavior therapy in the form of a prosocial online game (58). Mehler et al. performed a randomized controlled trial to determine in detail how real-time fMRI neurofeedback training can help reduce depressive symptoms (59). Both cases demonstrate alternative interventions that could help treat brain disorders.

**DISCUSSION AND OUTLOOK**

In this review, we have provided a brief summary of recent large brain datasets and efforts to integrate multimodal brain datasets to decipher brain disorders. We have shown that through integration of molecular multi-omics and neuroimaging datasets, great advances have been achieved in the identification of genetic risks and biomarkers of brain disorders, as well as in the understanding of their underlying molecular mechanisms. Below we discuss potential better ways of generating and utilizing brain datasets that may promote future directions of research on brain disorders.

First, in-depth collaboration between neuroscientists, biologists, medical professionals, and data scientists will enable well-designed, hypothesis-driven research projects, where the data can be collected in cost-effective ways. Successful examples include large databases such as the Allen Brain Atlas (5). Although large amounts of data have accumulated so far, such collaborative efforts are still very much needed, especially for addressing specific biological questions concerning the mechanisms of brain disorders. We have been constructing a multimodal brain database, named the Zhangjiang International Brain Biobank (ZIB), which is designed to collect matched genomic, transcriptomic, metabolic, and neuroimaging data for six brain disorder cohorts. To do so, we have recruited an international team of scientists from biology, medicine, computer science, physics, statistics, and mathematics, who have greatly contributed to the design and establishment of this database. We believe that in-depth interdisciplinary collaborations will be the new norm in this area.

Second, more good questions need to be asked, and more good methods need to be designed, to better mine the wealth of the accumulated data. Although we have reviewed many important efforts in this area, this does not mean there is little left to do—rather, this is just the beginning of future research. This observation is based on the following evidence: (a) Most current works have established the important associations among genomic, transcriptomic, metabolic, and neuroimaging data but have not explained their causal relationships or why these associations exist; and (b) most research focuses on only one part of the whole picture, so much effort is still needed to connect the dots. In addition, more fundamental and specific questions need to be asked, such as questions about specific neurocircuits, which can only be asked by experienced neuroscientists, or questions about hidden data patterns, which can only be recognized by experienced statisticians and mathematicians. To better master the unprecedented large amounts of data, we will need better tools for data-driven mining, noise reduction, and meta-analysis, which will likely arise from fast-developing areas such as artificial intelligence.

Third, other types of omics data may help uncover the unknown mechanisms of brain disorders. For example, the human gut has more than 1,000 types of bacteria, which have been proven to be associated with brain health (60)—in fact, gut microbiota are so important and relevant to the human brain that they are sometimes called the second brain (61). There are many ways that gut microbes could impact the brain, including secreting chemicals that enter the brain through...
the blood vessels; stimulating nerve-related cells in the gut lining; altering the immunal balance of the gut, which then indirectly affect the immune system inside the brain; and so on (62). Gut microbiome data can be obtained through 16S ribosomal RNA (rRNA) sequencing or whole-genome sequencing and have been used to find bacteria species as biomarkers for specific brain disorders and related phenotypes. For example, Shen et al. performed 16S rRNA sequencing to compare the differences in gut microbiota between SCZ patients and healthy controls and identified 12 microbiota biomarkers that could be used as diagnostics (63). Liu et al. also used 16S rRNA sequencing to identify gut bacteria species associated with the severity or the progression of AD and established gut microbiota models that could distinguish mild cognitive impairment and AD from controls (64). Researchers have also found metabolites generated by the gut microbiome that may help regulate the brain. For instance, through a large cohort dataset of the Flemish Gut Flora Project, Valles-Colomer et al. showed that microbial γ-aminobutyric acid produced by the gut microbiome may play a role in depression (65). Gut microbiota can also help identify molecular connections between brain disorders that co-occur with gut disorders (66) and can be integrated with neuroimaging data to help identify how gut bacteria may affect brain structure (67). In addition to their usefulness in understanding brain disorders, gut microbiota may offer alternative ways of treatment or interventions for brain disorders. For example, Olson et al. identified in mice that some gut bacteria could mediate and confer the antiseizure effects of the ketogenic diet (68). Taken together, we believe that incorporating gut microbiota datasets into existing analysis frameworks of multimodal data may offer additional information that can deepen our understanding of brain disorders and even help find novel treatments for them.

Last but not least, with the accumulation of important findings through integrative analysis of various types of brain disorders, the question of how to translate these findings into clinical usage will become ever more pressing. We hope to see more research in the future that integrates current knowledge on the biomarkers and the molecular mechanisms of brain disorders to facilitate intelligent, fast, and early diagnosis and to inform disease subtyping followed by subtype-specific treatment. To this end, current and future multimodal brain datasets will continue to be a treasure to be mined by researchers tackling brain disorders.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This work is supported by the National Key R&D Program of China (2019YFA0709502, 2018YFC1312904, 2018YFC0910500), the 111 Project (B18015) of China, the Shanghai Municipal Science and Technology Major Project (2018SHZDZX01), the National Natural Science Foundation of China (61932008, 61772368), ZJ Lab, and Shanghai Center for Brain Science and Brain-Inspired Technology. This work is also supported by the Zhangjiang International Brain Biobank (ZIB) Consortium.

LITERATURE CITED

35. Yang A, Chen J, Zhao X-M. 2020. nMAGMA: a network enhanced method for inferring risk genes from GWAS summary statistics and its application to schizophrenia. bioRxiv 2020.08.15.250282. [https://doi.org/10.1101/2020.08.15.250282](https://doi.org/10.1101/2020.08.15.250282)


Contents

Using Phecodes for Research with the Electronic Health Record:
From PheWAS to PheRS
Lisa Bastarache ................................................................. 1

The 3D Genome Structure of Single Cells
Tianming Zhou, Ruochi Zhang, and Jian Ma ........................................ 21

Integration of Multimodal Data for Deciphering Brain Disorders
Jingqi Chen, Guiying Dong, Liting Song, Xingzong Zhao, Jixin Cao,
Xiaobui Luo, Jianfeng Feng, and Xing-Ming Zhao .................................. 43

African Global Representation in Biomedical Sciences
Nicola Mulder, Lyndon Zass, You Hamdi, Houcemeddine Othman,
Sumir Panji, Imane Allali, and Yasmina Jaweela Fakim ............................. 57

Phenotyping Neurodegeneration in Human iPSCs
Jonathan Li and Ernest Fraenkel .................................................. 83

Perspectives on Allele-Specific Expression
Siobhan Cleary and Cathal Seoighe .............................................. 101

Ethical Machine Learning in Healthcare
Irene Y. Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman,
and Marzyeh Ghassemi .......................................................... 123

The Ethics of Consent in a Shifting Genomic Ecosystem
Sandra Soo-Jin Lee .......................................................................... 145

Modern Clinical Text Mining: A Guide and Review
Bethany Percha ............................................................................. 165

Mutational Signatures: From Methods to Mechanisms
Yoo-Ab Kim, Mark D.M. Leiserson, Priya Moorjani, Roded Sharan,
Damian Wojtowicz, and Teresa M. Przytycka ....................................... 189

Single-Cell Analysis for Whole-Organism Datasets
Angela Oliveira Pisco, Bruno Tojo, and Aaron McGeever ....................... 207
Neoantigen Controversies
Andrea Castro, Maurizio Zanetti, and Hannah Carter .................. 227

The Exposome in the Era of the Quantified Self
Xinyue Zhang, Peng Gao, and Michael P. Snyder .................. 255

Metatranscriptomics for the Human Microbiome and Microbial
Community Functional Profiling
Yancong Zhang, Kelsey N. Thompson, Tobyn Branck, Yan Yan, Long H. Nguyen,
Eric A. Franzosa, and Curtis Huttenhower .................. 279

Artificial Intelligence in Action: Addressing the COVID-19 Pandemic
with Natural Language Processing
Qingyu Chen, Robert Leaman, Alexis Allot, Ling Luo, Chih-Hsuan Wei,
Shankai Yan, and Zhiyong Lu .................. 313

Data Science in the Food Industry
George-John Nychas, Emma Sims, Panagiotis Tsakanikas, and Fady Mobarek ........ 341

Illuminating the Virosphere Through Global Metagenomics
Lee Call, Stephen Nayfach, and Nikos C. Kyrpides .................. 369

Probabilistic Machine Learning for Healthcare
Irene Y. Chen, Shalmali Joshi, Marzyeh Ghassemi, and Rajesh Ranganath .................. 393

Satellite Monitoring for Air Quality and Health
Tracey Holloway, Daegan Miller, Susan Anenberg, Minghui Diao,
Bryan Duncan, Arlene M. Fiore, Daven K. Henze, Jeremy Hess, Patrick L. Kinney,
Yang Liu, Jessica L. Neu, Susan M. O’Neill, M. Talat Odman,
R. Bradley Pierce, Armistead G. Russell, Daniel Tong, J. Jason West,
and Mark A. Zondlo ............................................. 417

Errata
An online log of corrections to Annual Review of Biomedical Data Science articles may be
found at http://www.annualreviews.org/errata/biodatasci