Biobank-scale Brain Imaging Genetics: Clinical and Methodological Advances

University of North Carolina at Chapel Hill

Hongtu Zhu

Joint works with all members of the UNC BIG-S2 Lab, Bingxin Zhao, Tengfei Li, Yun Li, Stephen Smith, and Jason Stein

https://www.med.unc.edu/big-s2
Part 1
Methodological Challenges
AI Ecological Layout

Google, Amazon, Facebook, BAT, DiDi, ...

- Healthcare
- Retail
- Education
- Transportation
- Home
- Robotics
- Agriculture
- Government affairs
- Food
- Human Resources
- Security
- Judicial
- Internet Security
- Intelligence Cities
- Autonomous Driving
- Manufacturing
- Speech Recognition
- NLP
- Computer Vision
- Prediction and Decision
- IOT

Source: DiDi AI Labs, DiDi Strategic Department
EHR is an information resource that takes residents' personal health as the core, runs through the entire life process, covers various health-related factors, realizes multi-channel information dynamic collection, and meets the needs of residents' self-care, health management and health decision-making.

**PM**: Personalization, precision (time and plan), and health management. High-level medical technology is formed on the basis of in-depth understanding of people, diseases, and medicines. Analyze the health status of the entire population and improve the health of the general public.

Record the changes of all vital signs of an individual from birth to death, including personal living habits, past medical history, diagnosis and treatment, family medical history, current medical history, previous diagnosis and treatment history, previous physical examination results and other information, and accurately record digitally, so as to construct an integrated health service of prevention, diagnosis, treatment, rehabilitation, and health management.
Multi-modal Data

Clinical/Behavioral

Imaging

Genetics
Large-scale Medical Studies

PING - 900 Pediatric Imaging, Neurocognition, and Genetics
BCP - 300 Baby Connectome Project
ADNI - 2000 Alzheimer’s Disease Neuroimaging Initiative
PNC - 1400 Philadelphia Neurodevelopmental Cohort
HCP - 1200 Human Connectome Project
ABCD - 10000 Adolescent Brain Cognitive Development
UKB - 500,000 UK Biobank Project
TCIA – 37,600 The Cancer Imaging Archive
NLST - 19,000 National Lung Screening Trial
OAI – 4800 Osteoarthritis Initiative
AllOfUs-1000,000+ All of us project
Data Challenges

- Prevention
- Treatment
- Diagnosis
- Prognosis

EHR

PM
Data Challenges

• Over 15M labeled high resolution images
• Roughly 80K categories
• Collected from web and labeled by Amazon Mechanical Turk

Lack of a large number of annotated data with high-quality
Method Challenges

Healthcare

Genetics

Imaging

Clinical

Shallow Information

Speech Recognition

NLP

Computer Vision

Prediction and Decision

IOT

Source: 滴滴AI Labs, 滴滴战略部

https://qz.com/989137/when-a-robot-ai-doctor-misdiagnos
Method Challenges

Ritchie et al. (2015).
Nature Review Genetics
Method Challenges

Heterogeneity at the subject, group, and study levels
Ecological Layout

Deconvolution

Integration

Learning

Prediction

Large-scale Database


**Tumor heterogeneity**

**Workflow**
- Deconvolution of mixed transcriptomes and genomes
- Gene/pathway selection/ranking
- Pan-cancer annotation

**Future goals:**
- Pre-clinical models
- New therapeutic targets

**Somatic mutation:** allele-specific and cell-type-specific expression (sACE)

**sACE-associated regulatory pathways**

**Chromatin States**
- Inactive
- Poised
- Active

**Cancer treatment and prognosis**

**Pan-cancer variation of sACE**

**Genetic variation**
- Sample 1
- Sample 2
- Sample 3
- \( \ldots \)
- Sample \( n-1 \)
- Sample \( n \)

**Gene expression**
- Sample 1
- Sample 2
- Sample 3
- \( \ldots \)
- Sample \( n-1 \)
- Sample \( n \)
We developed a CliP (Clonal and subclonal structure identification through Pairwise difference penalization), to distinguish sub-clones.

We tested CliP on 965 simulated samples generated by the Broad Institution, all samples are generated using copy number profiles from actual patients samples.
Results on ICGC samples

The International Cancer Genome Consortium has collected whole genome sequencing for over 2,700 samples. The clonality study shows that the clone/subclonality compositions are quite different across cancer types.

Figure: clonality composition of selected types of cancer. Both the number of subclones and subclonal fractions are different across tumor types.
Baby Connectome Project
Population based Structural Connectomes
Brain Function-based Structural Connectome Atlas

Stage 1:
Whole-brain Structural Connectome

Stage 2:
Creation of Fiber Skeleton

Stage 3:
Sparse Representation
Brain Function-based Structural Connectome Atlas

Fiber atlas
Fiber skeleton atlas

(B)

Legend:
- Posterior-Multimodal
- Cingulo-Opercular
- Dorsal-Attention
- Somatomotor
- Language
- Default
- Visual1
- Visual2
- Auditory
- Frontoparietal
- Ventral-multimodal
- Orbito-Affective
CAMELYON17

Normal

Tumor

H&E

IGC

H&E

BRAF

Raw images → Preprocessing & Segmentation → Feature Extraction → Prediction Model

Dictionary Learning

XGBoost
Breast Cancer Risk Prediction

Shape features
Statistics features
Wavelet feature
Textural features

Raw images
Preprocessing
Feature Extraction
Prediction Model

Image 1

Uni-resolution Bias Correction...
Tumor Seg.......

Image n

Shape, texture... Intensity

High risk

Low risk

Accuracy
AUC
F1 score

High vs. Low
High vs. Intermediate vs. Low
High vs. Intermediate vs. High
Low vs. Intermediate vs. High
High vs. Intermediate vs. High

MORE THAN A JOURNEY | didiglobal.com
Ecological Layout

Integration

http://www.pnas.org/content/105/13/5213/F1.expansion.html
Image Genetics

Genome-wide association study (GWAS) of hundreds of imaging phenotypes with more than 50,000 subjects from five publicly available datasets (largest brain imaging GWAS so far)
Part II
Big Data in Imaging Genetics
Capture the brain structure and function changes associated with major brain-related disorders and normal development.

Alzheimer’s disease (AD) is associated with brain shrinkage.
Genetics of Brain Disorders

Most major brain disorders (like AD) are **heritable complex traits/diseases**

Together 50%-70% of AD risk  
75%-90% of ADHD risk  
60%-85% of Schizophrenia risk  
~80% of Autism Spectrum Disorder (ASD) risk

Complex traits/diseases  
(many genes, environmental factors, complex functional mechanism)

Genetic signals are non-spare and weak:  
Need large sample size to detect weak signals
Brain Imaging Genetics Paradigm

**Neuroimaging:** an important component to help understand the complex biological pathways of brain disorders

- **Genes** → **molecules, brain cells, structure/function**
  - Gene expression at RNA and protein levels
  - Changes in neuron structure and function

**Biological causes**
- Genomics
- Transcriptomics
- Proteomics
- Metabolomics
- Interactomics

**Environmental, social and psychological factors**
- Genetic
- Epigenomic modifications
- De novo mutations

**Molecular function and cell metabolism**
- RNA, proteins, metabolites

**Brain disorders**
- Structure, circuits, physiology
- Changes in neural interactions, altered brain structure/function

**Symptoms**
- Behavioral tests

**Uncover the profile of brain abnormalities in each clinical outcome to study how disorders develop**
Cardiovascular Disease & Brain Health

(Neuro)imaging: help understand the complex interplay between brain and other human organs and their underlying genetic overlaps

Possible causal factors of brain structure changes, resulting in brain disorders like stroke, dementia and cognitive impairment

Many diseases (e.g., microvascular disease, high blood pressure) are multisystem disorders
Long-term Challenges in Brain Imaging Genetics

- Traditionally, neuroimaging data are expensive and have very limited sample size ($n \sim 100$)
- On the other hand, genetic risk factors are typically dense and have small effect size, and thus need large sample size to detect
- Imaging batch effects/confounders (e.g., image acquisition, processing procedures, and software)
“Big Data” Brain Imaging Genetics Cohorts

“Big data” Brain imaging genetics datasets become available in recent few years
Systematically collect publicly available individual-level data for > 50k individuals
Build the largest database in this field

- **Aging Brain**
  - BCP (Age [0, 5])
  - PING (Age [3, 21])
  - ABCD (n ~ 10k, Age [9, 11])
  - PNC (Age [14, 29])
  - HCP (Age [22, 35])
  - UK Biobank (n ~ 100k [Ongoing], Age [40, 69])
  - RADC (Age > 65)
  - ADNI (Age [55, 92])

- **Brain Development**
  - PNC (Age [14, 29])
  - ABCD (Age [9, 11])
  - BCP (Age [0, 5])
  - PING (Age [3, 21])

BIG-KP | https://bigkp.org/
Brain Imaging Modality Examples

Harmonize tools/pipelines to consistently generate the full spectrum of neuroimaging features

Cortical and subcortical structures

White matter microstructure
(Structural connectivity, diffusion MRI)

Functional networks
(Functional connectivity, functional MRI)
Regional Brain Volumes and Shape

Generate regional brain volumes and shape representations for 98 pre-specified brain regions and total grey matter, white matter, and brain volumes.

Subcortical structures (deep within the brain)

Cortical structures (outer layer of the cerebrum)
**White Matter Microstructure**

5 white matter microstructure measures (DTI parameters) for 21 white matter tracts

- 21 white matter tracts from ENIGAMA-DTI pipeline
- Fractional anisotropy (FA)
- Mean diffusivity (MD)
- Axial diffusivity (AD)
- Radial diffusivity (RD)
- Mode of anisotropy (MO)

Sensitive to specific types of microstructural changes and have also been widely used in clinical research.
White Matter Microstructure

Tract-specific functional principal component analysis (FPCA) to capture major variations within each white matter tract.
Resting/task functional MRI (fMRI)

Independent component analysis (ICA)-based methods to form 76 functional regions and generate 1,701 functional connectivity traits.

Characterize major functional brain regions and their connectivity.
Resting/task functional MRI (fMRI)

Map 76 ICA-nodes onto automated anatomical labeling (AAL) parcellation and pre-defined functional networks

Seven networks in Yeo et al., 2011

Eight networks in Finn et al., 2015
Brain Imaging Genetics: Learning Problems

(a) Learning Problems in Brain Imaging Genomics
- Heritability estimation of brain imaging phenotypes
- Imaging genomics associations
  - Fundamentals
  - Meta-analysis
  - Multivariate regression
  - Bi-multivariate correlation
- Integrating imaging and genomics for outcome prediction

(b) Biomedical Application Considerations
- Brain imaging
  - Voxels, ROIs, spatial correlation
  - Multimodal, longitudinal studies
  - Prior knowledge, brain network
- Clinical outcome
  - Diagnosis
  - Progression
  - Impairment score

(c) Statistical & Machine Learning Considerations
- Increasing power
  - Quantitative traits
  - Multiple comparison
  - Meta/mega analysis
  - Multivariate models
- Overfitting control
  - Dimensionality reduction
  - Regularization
  - Knowledge-guided learning
  - Outcome-guided learning
- Other topics
  - Biological interpretation
  - Scalability
  - Biased sampling
  - Interaction

Heritability estimation of brain imaging phenotypes
Integrating imaging and genomics for outcome prediction

Genomics
- SNPs, LD blocks
- Genes, pathways, networks
- Polygenic risk scores (PRSs)

Genetic variation

Brain Imaging: MRI, AV45 PET, FDG PET

Clinical outcome: Cognitive Condition, Diagnosis

(Shen & Thompson Proc of the IEEE 2020)
Methodological Challenges

- Database/Tool/Theory for Brain Imaging Genetics
  - Multiple Biobanks Integration (e.g., Heterogeneity in global populations)
  - Omics Data Integration (e.g., new tech, biological pathway)

- New Computational Tools (e.g., challenge of dense signal in biobank-scale database)

- Advanced Methods for Dense Signals (e.g., deep learning)

 UNC Biostatistics

Multiple Biobanks Integration
(e.g., Heterogeneity in global populations)

New Computational Tools
(e.g., challenge of dense signal in biobank-scale database)

Advanced Methods for Dense Signals
(e.g., deep learning)

https://bigkp.org/
Brain Imaging Genetics: Learning Problems


Brain Imaging Genetics Data Analysis

**Association tests**
- Identify and replicate novel genetic factors associated with brain structure and function

**Causal inference/Mediation analysis**
- Analyze the genetic links among brain structure, brain function, cognition, and major brain disorders.

**Data integration**
- Integrate external genetics/genomics data (e.g., the GTEx, Hi-C chromatin interactions) to uncover new biological insights

**Predictive model**
- Perform out-of-sample risk prediction for brain disorders using genetics, genomics, and imaging data

1) Output high-quality novel clinical findings
2) Identify, model, and address important statistical problems
3) Share our summary-level data/results to the research community
Part III

Novel Clinical findings
Brain Imaging Genetics Knowledge Portal (BIG-KP)

Genetics Discoveries in Human Brain by Big Data Integration

Aim to build the best knowledge database of neuroimaging genetics
Searchable database for 1593+ neuroimaging traits across four imaging modalities: (grey matter volume, white matter microstructure, resting-state and task functional activity/connectivity)
GWAS Locus Browser

Amplitude Trait (node activity)
(Precuneus)
(Default mode, Central executive)
The full set of GWAS summary statistics have been made freely available to the research community. Resources with the largest sample size (>4,000 page views since Sep 2019)
GWAS of White Matter Tracts

Overview of the ENIGMA-DTI pipeline and the multiple-stage design in GWAS

Apply the same pipeline in different datasets (UKB, ABCD, PING, PNC, HCP)
Genetic Architecture of White Matter

We observed 109 novel genomic regions (151 in total, $P < 2.3 \times 10^{-10}$, $5 \times 8/215$) associated with white matter microstructure.

**Heritability $h^2 \sim 45\%$**

Sample size is essential for gene discovery of traits with highly polygenic genetic architecture.
For the 25 known genomic risk regions of Glioma/GBM, 11 are associated with white matter microstructure.
Colocalization with Stroke

Genetic colocalizations among vascular risk factors (e.g., obesity, diabetes, high blood pressure), white matter microstructure, and stroke

C

chr10, Region: 10q24.33

D

rs2295786 (10q24.33), MD map

Superior corona radiata
Anterior corona radiata
Uncinate fasciculus
Posterior limb of internal capsule

rs7859727 (9p21.3), MD map

Superior corona radiata
Anterior corona radiata
Uncinate fasciculus
Posterior limb of internal capsule

Superior fronto-occipital fasciculus
Uncinate fasciculus
Retrolenticular part of internal capsule

Posterior limb of internal capsule

Genu of corpus callosum
Splenium of corpus callosum
Anterior limb of internal capsule

External capsule

rs2295786
rs188186531

SNP in GWAS panels

GTI index SNP

co-localized GWAS index SNP

$\rho > 0.8$

$0.8 > \rho > 0.6$

$0.6 > \rho > 0.4$

$0.4 > \rho > 0.2$

$0.2 > \rho > 0.1$

$\rho < 0.1$

$r > 0.6$

$r < 0.6$

GWAS Catalog Category

Glioma/Glioblastoma
Cognitive Traits
Neurological Disorders
Psychiatric Disorders
Psychological Traits
Sleep
Smoking/Drinking
Anthropometric measurements
Bone Mineral Density
Alzheimer's disease Biomarkers
Educational Attainment
Stroke

SFO AD
105.5 mb
105.6 mb
105.7 mb
105.8 mb
0
2
4
6
8
10

15
7

-\log(10)(P)
Genetic Correlations with Brain Disorders

Strong genetic correlation between white matter microstructure and small vessel stroke subtype
Spatial Overlap with Brain Structures

Strong genetic correlation between white matter microstructure and the grey matter volume of neighboring regions
White matter is largely composed of glial cell types (oligo, microglia, astrocyte)

Gross cell types (neuron, non-neuron [glia, including oligo, microglia, astrocyte])

Heritability Enrichment in Brain Cells
Identify brain cell types where genetic variation leads to changes in white matter connectivity

Oligo annotation accounted for 10.4% heritability while only composed 0.3% of the genetic variants
Glial cell enrichment was widely observed in white matter tracts and was most significant in posterior corona radiata (PCR), posterior limb of internal capsule (PLIC), and genu of corpus callosum (GCC).
White matter is largely composed of glial cell types (oligo, microglia, astrocyte). Heritability of 49 complex traits was significantly enriched in genetic regions influencing white matter microstructure, such as stroke, schizophrenia, ADHD, bipolar Alzheimer's Disease, T2D, high blood pressure, and coronary artery disease. DTI annotation enrichment analysis using DTI annotations (defined by significant genes of DTI parameters of white matter microstructure).
The salience network (SN) plays a crucial role in dynamic switching between the central executive (CE) and default mode (DM) networks.

Three core functional networks that support efficient cognition:

- Related to major brain disorders, such as Alzheimer’s disease (AD), Parkinson’s disease (PD), and major depressive disorder (MDD).
The level of genetic control is higher in the triple networks, which closely control multiple cognitive functions and affect major brain disorders.
Genetics of Functional Brain

Ideogram of the loci influencing rsfMRI traits of intrinsic brain activity at the significance level 2.8e-11 (5e-8/1777)
Colocalization with AD and SCZ

Colocalization between brain function in the default mode (DM) and central executive (CE) networks with Alzheimer's disease (AD) and Schizophrenia (SCZ)

Alzheimer's disease (APOE)

Schizophrenia
Colocalization at \textit{APOE}

\textit{APOE} gene has stronger genetic relationships with brain function than brain structures.
Colocalization at 17q21.31 regions

Neurological disorders (e.g., Parkinson’s disease, Alzheimer’s disease, corticobasal degeneration)

Psychiatric disorders (e.g., autism spectrum disorder, depression)

Education, cognitive ability

Psychological traits (e.g., neuroticism)

Alcohol use disorder
Colocalization with Sleep and Cognition

C

chr2, Region: 2q14.1

chr2, Region: 2q24.2

Cognitive traits/education
Genetic Overlap with Brain Structures

Shared genetic influences between functional connectivity of default mode and central executive networks and insula volume

Location of the right insula and its neighboring brain regions whose functional connectivity strengths were genetically correlated with the right insula volume

Spatial colocalizations between regional brain volumes and their genetically correlated functional connectivity traits

left pericalcarine volume was genetically correlated with the connectivity strengths among its neighboring regions

Highlights:
- Genetic Overlap with Brain Structures
- Shared genetic influences between functional connectivity of default mode and central executive networks and insula volume
- Location of the right insula and its neighboring brain regions whose functional connectivity strengths were genetically correlated with the right insula volume
- Spatial colocalizations between regional brain volumes and their genetically correlated functional connectivity traits

Note: The images show various brain structures including Precentral, Middle cingulate, Superior frontal, Middle frontal, Inferior frontal, Middle temporal, Insula (right), Precuneus, Cuneus, Superior occipital, Calcarine, Lingual, and Pericalcarine (left). The insula is associated with multiple functions, including emotion, addiction, and cognition, through extensive connections to neocortex, the limbic system, and amygdala.
Spatial Overlap of Genetics Effects

Shared genetic influences between brain functional connectivity and structural connectivity

Location of the SLF and its neighboring brain regions whose functional connectivity strengths were genetically correlated with the structural connectivity of SLF

Genetic evidence on how distributed functional networks communicate across large distances
Regions whose functional connectivity genetically related to brain disorders and intelligence

- **ADHD** (Attention-Deficit / Hyperactivity Disorder)
- **SCZ** (Schizophrenia)
- **MDD** (Major Depression Disorder)

**Brain Regions:**
- Superior frontal
- Precentral
- Postcentral
- Anterior cingulate
- Middle frontal
- Insula
- Superior temporal
- Inferior temporal
- Cerebellum
- Angular
- Precuneus
- Superior parietal
- Middle temporal
- Middle cingulate
- Putamen
- Caudate
- Inferior frontal
- Superior frontal

**Disorders:**
- Attention-Deficit / Hyperactivity Disorder
- Schizophrenia
- Major Depression Disorder

**Intelligence**
Gene expression-informed gene-level PRS + GWAS PRS has higher prediction accuracy

Construct gene-level PRS (polygenic risk scores) by leveraging gene expression reference panels (e.g., GTEx) in TWAS
It's just a beginning

Hundreds of associated genetic variants for 1593+ neuroimaging traits across three modalities:
(grey matter volume, white matter microstructure, resting-state functional connectivity + rfMRI, task fMRI, shape, heart)

Publications (2018+)
Genetic influences on the intrinsic and extrinsic functional organizations of the cerebral cortex (2021). medRxiv, 21261187. LINK
Transcriptome-wide association analysis of brain structures yields insights into pleiotropy and complex neuropsychiatric traits (2021). Nature Communications, 842872. LINK
Genome-wide association analysis of 19,629 individuals identifies variants influencing regional brain volumes and refines their genetic co-architecture with cognitive and mental health traits (2019). Nature Genetics, 51(11), 1637-1644. LINK nature genetics [Cover Feature]
Large-scale GWAS reveals genetic architecture of brain white matter microstructure and genetic overlap with cognitive and mental health traits (n= 17,706) (2019). Molecular Psychiatry, in press. LINK
Heritability of regional brain volumes in large-scale neuroimaging and genetic studies (2018). Cerebral Cortex, 29(7), 2904-2914. LINK

Genetics discovery in human brain by big data integration
Heart Imaging Genetics Knowledge Portal (Heart-KP)

Integrating knowledge and data for human heart research

Welcome to Heart-KP!

This knowledge portal serves as a platform for accelerating research into human hearts.
Ongoing/Future Directions

1. Causal relationships among disease, brain structures, and brain functionalities (e.g., the genetic pathway among vascular risk factors, white matter, and stroke)

2. Build optimal models for complex traits and diseases prediction using imaging and genetics data (e.g., deep learning)

3. Compare and identify the best practical strategy and pipelines to process different neuroimaging modalities (e.g., ICA for fMRI)

4. Model brain changes and genetics effects across the life span

5. Align and integrate different neuroimaging modalities
Acknowledgement

Funding: U.S. NIH Grants MH086633 and MH116527

Pictures: Copyrights belong to their own authors and/or holders.

Data: We thank Bingxin Zhao, Tengfei Li and other members of the UNC BIG-S2 lab (https://med.unc.edu/bigs2/) for processing the neuroimaging data.

UK Biobank resource application number: 22783.