https://doi.org/10.1038/s42255-023-00806-x

Single-cell expression profiling of islets generated by the Human Pancreas Analysis Program

Check for updates

In this Correspondence, we introduce the Human Pancreas Analysis Program and discuss ongoing efforts with a specific focus on the continuous release of integrated single-cell RNA sequencing (scRNA-seq) measurements across pancreatic tissues. Although it has been over 50 years since Willy Gepts described the pancreatic pathology in type 1 diabetes (T1D), detailed understanding of the initial molecular perturbations that occur during disease pathogenesis is still incomplete. Two nontrivial constraints hamper insights into these processes: (1) the inability to safely biopsy the human pancreas of living donors; and (2) the substantial disease progression and beta cell destruction that has occurred by the time patients are clinically diagnosed with T1D. These limitations have meant that the majority of T1D studies have been performed on peripheral leukocytes from the blood, which is not the site of pathogenesis. A revolution in single-cell transcriptomic technologies during the past decade has influenced the molecular profiling of islets. Several primary single-cell pancreatic islet atlases for mouse and humans have become available[1–8]. However, the limited numbers of cells and relatively small numbers of human donors used for the abovementioned studies could not provide comprehensive coverage of the human pancreatic islet transcriptome. In 2016, the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) charged a group of investigators with forming the Human Pancreas Analysis Program (HPAP)[9]. The goal of this initiative is to use cutting-edge but robust and tested technologies in pancreatic tissues and to release high-quality datasets for the T1D community. A central mission of HPAP is to organize and share data resulting from the study of these tissues through an open-access resource database called PANC-DB. Owing to the early success of HPAP-T1D, in 2020 HPAP-T2D was also funded by NIDDK to define the molecular pathogenesis of islet dysfunction by studying human pancreatic tissue samples from organ donors with T2D[10].

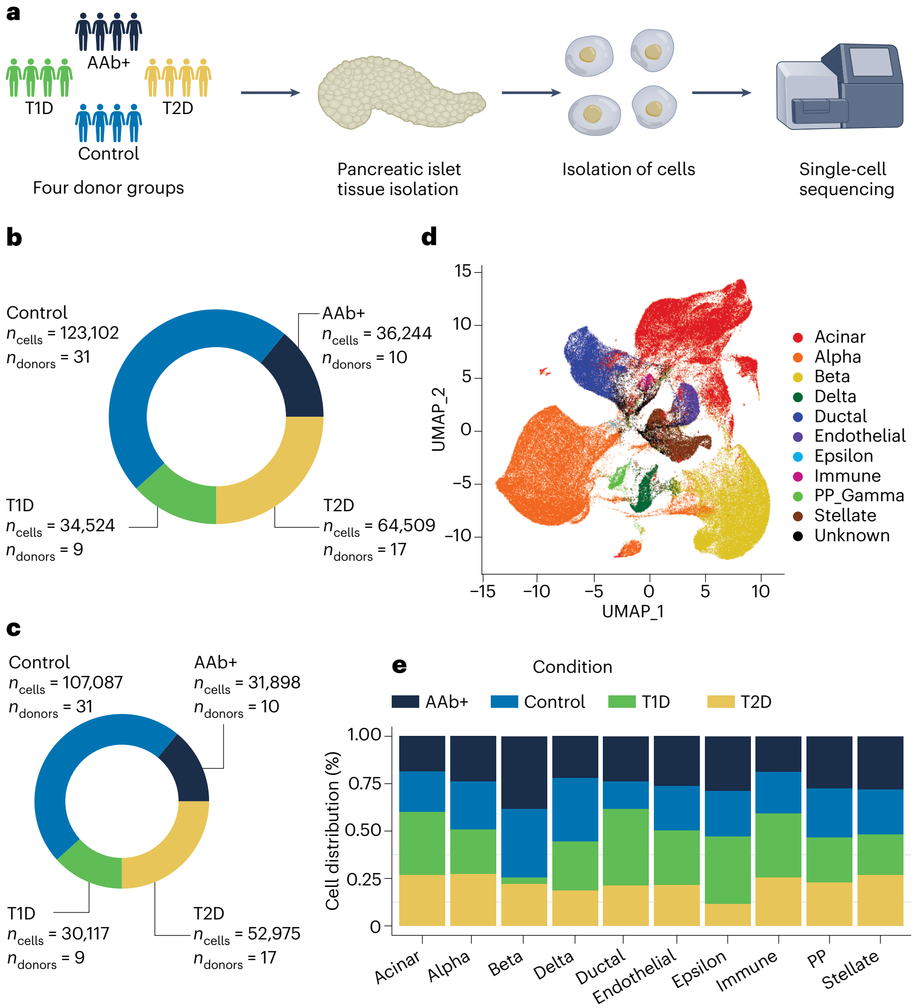

Fig. 1 | scRNA-seq reveals the cell populations of human pancreatic islets. a, Overview of the scRNA-seq workflow using human pancreatic islet tissue samples. **b**, Number of cells and donor distribution across different biological conditions before filtering. Control, non-diabetic donors. **c**, Number of cells and donor distribution across different biological conditions after filtering. **d**, UMAP plot showing the proper classification of islet cells. **e**, Stacked bar chart showing the percentage-wise distribution of cell types across different biological conditions.

Currently, the acceptance criteria for HPAP relates to donors who are younger than 40 years of age and have had T1D for 7 years or less, or donors without diabetes who are 40 years or younger with one or more T1D-associated autoantibody (AAb). These criteria have evolved over time from data demonstrating that residual insulin-positive islets and insulitis are more likely to be found using these acceptance criteria. Our goal is to process tissue from at least nine donors with recent-onset T1D per year, although the unpredictable nature of organ donor recovery does not allow us to provide a precise timeline for new donors. Multiple data modalities from molecular to functional measurements are being generated by HPAP, where raw measurements from individual donors become available on PANC-DB in real time. Among these data modalities, single-cell transcriptomic measurements are distinct, considering that the high-throughput nature of this technology and the maturity of computational techniques for these assays have been under active development for robust data integration across donors. Although the raw FASTQ files for scRNA-seq data are uploaded to PANC-DB once scRNA-seq libraries from each donor have been sequenced, HPAP periodically releases fully processed integrated scRNA-seq data on PANC-DB's 'Interactive Analysis' section once new batches of data from 10 donors have been sequenced and passed various quality control metrics (Supplementary Fig. 1). In this Correspondence, we highlight various aspects of the release of scRNA-seq data integration across HPAP donors.

The first set of scRNA-seq measurements collected by HPAP (about 80,000 cells from pancreatic islets of 24 human donors) consisted of three categories: healthy, AAb-positive but normoglycemic (AAb+), and diagnosed with T1D[11]. The latest scRNA-seq integration (February 2023) highlighted in this Correspondence includes around 260,000 high-quality human pancreatic islet cells from 67 donors consisting of four categories: non-diabetic (n = 31), AAb+ (n = 10), T1D (n = 9) and T2D (n = 17), forming, to our knowledge, the largest existing scRNA-seq dataset of human pancreatic tissues (Fig. 1b). After quality control (QC) and data pre-processing[12,13], including doublet removal[14,15] and normalization[16], the total number of cells was 222,077 from 67 donors (Fig. 1c). Next, based on the marker genes described in (Supplementary Information: Note 1), we used the scSorter method[17] and annotated cells into 10 clusters as shown in the uniform manifold approximation and projection (UMAP) analysis (Fig. 1d). The percentage of cells for each cell type across different conditions is shown in Fig. 1e. Each release of our integrative analysis using scRNA-seq measurements in pancreatic tissues has three major components, aiming to save many hours of computational work for diabetes investigators. (1) A fully documented R code relying on Seurat's pipeline used to process the latest release of data integration with detailed explanations of the pre-processing, normalization and scaling, cell annotation and other computational issues incorporated in the processed data. In our opinion, this code and accompanying explanations can provide training opportunities to individuals who are learning various aspects of scRNA-seq analysis. (2) Processed data based on Seurat's pipeline in two formats (rds and h5ad), enabling investigators to apply their own filtering, cell annotation and additional computational work to scRNA-seq data. (3) An interactive tool called CellxGene, developed by the Chan Zuckerberg initiative to navigate and visualize single cells from each release (Fig. 1). CellxGene[18] exploits modern web development techniques and enables interactive visualization of large numbers of cells, allowing hypotheses to be tested with functionalities including subsetting, differential expression and others, to help the community to explore the HPAP data. With every new batch of donors, these components are being updated and released on PANC-DB, providing a reliable reference for single-cell pancreatic islet biology studies, especially into diabetes-related conditions. Together, our expanding single-cell transcriptomic map of human pancreatic islet cells and the continuous release of processed scRNA-seq measurements across HPAP donors aim to democratize access to high-quality molecular profiling data in islets, facilitating the discovery of new cures for T1D and T2D.

Reporting Summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All pancreatic islet sequencing data have been uploaded to PANC-DB (https://hpap.pmacs.upenn.edu/) and the interactive session can be accessed on the cellxgene database hosted on PANC-DB (https://faryabi16.pmacs.upenn.edu/view/T1D_T2D_public.h5ad/). The raw islet sequencing data in the form of FASTQ files for each HPAP sample can be obtained from the PANC-DB database. The processed data can be downloaded in two formats, RDS or h5ad, which can be loaded for analysis using Seurat[12] and Scanpy[19] pipelines.

Code availability

The scripts used for data processing and analysis of the scRNA-seq data are available on GitHub (https://github.com/faryabiLab/HPAP-scRNA-seq-Workflow-2022). The cellxgene interactive tool with HPAP data can be accessed via the cellxgene database (https://faryabi16.pmacs.upenn.edu/view/T1D_T2D_public.h5ad/)

Abhijeet R. Patil [1,2,3,4], **Jonathan Schug** [1,4], **Ali Naji** [2,5], **Klaus H. Kaestner** [1,3,4], **Robert B. Faryabi** [2,3,5,6,7] & **Golnaz Vahedi** [1,2,3,4,7] ✉

[1]Department of Genetics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA. [2]Institute for Immunology, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA. [3]Epigenetics Institute, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA. [4]Institute for Diabetes, Obesity and Metabolism, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA. [5]Department of Surgery, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA. [6]Department of Pathology and Laboratory Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA. [7]Abramson Family Cancer Research Institute, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA.

✉e-mail: vahedi@pennmedicine.upenn.edu

Published online: 15 May 2023

References

1. Basile, G. et al. *Genome Med.* **13**, 1–17 (2021).
2. Xin, Y. et al. *Cell Metab.* **24**, 608–615 (2016).
3. Segerstolpe, Å. et al. *Cell Metab.* **24**, 593–607 (2016).
4. Grün, D. et al. *Cell Stem Cell* **19**, 266–277 (2016).
5. Muraro, M. J. et al. *Cell Syst.* **3**, 385–394.e3 (2016).
6. Baron, M. et al. *Cell Syst.* **3**, 346–360.e344 (2016).
7. Fang, Z. et al. *Cell Rep.* **26**, 3132–3144.e7 (2019).
8. Lawlor, N. et al. *Genome Res.* **27**, 208–222 (2017).
9. Kaestner, K. H., Powers, A. C., Naji, A. & Atkinson, M. A. *Diabetes* **68**, 1394–1402 (2019).
10. Shapira, S. N., Naji, A., Atkinson, M. A., Powers, A. C. & Kaestner, K. H. *Cell Metab.* **34**, 1906–1913 (2022).
11. Fasolino, M. et al. *Nat. Metab.* **4**, 284–299 (2022).
12. Hao, Y. et al. *Cell* **184**, 3573–3587.e29 (2021).
13. Amezquita, R. A. et al. *Nat. Methods* **17**, 137–145 (2020).
14. Germain, P.-L., Lun, A., Macnair, W. & Robinson, M. D. *F1000 Res.* **10**, 979 (2021).
15. Zheng, G. X. Y. et al. *Nat. Commun.* **8**, 14049 (2017).
16. Hafemeister, C. & Satija, R. *Genome Biol.* **20**, 296 (2019).
17. Guo, H. & Li, J. *Genome Biol.* **22**, 69 (2021).
18. Megill, C. et al cellxgene: a performant, scalable exploration platform for high dimensional sparse matrices. Preprint at *bioRxiv* https://doi.org/10.1101/2021.04.05.438318 (2021).
19. Wolf, F. A., Angerer, P. & Theis, F. J. *Genome Biol.* **19**, 15 (2018).

Acknowledgements

We thank the Vahedi, Faryabi and Kaestner lab members for discussions. This work was supported by National Institute of Health grants UC4 DK112217, U01DK112217 (A.N., K.H.K., R.B.F. and G.V.), R01HL145754, U01DK127768 and U01DA052715 and awards from the Burroughs Wellcome Fund, the Chan Zuckerberg Initiative, W. W. Smith Charitable Trust, the Penn Epigenetics Institute and the Sloan Foundation (G.V.).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42255-023-00806-x.

Peer review information *Nature Metabolism* thanks Yan Li and Benoit Gauthier for their contribution to the peer review of this work.

nature portfolio

Corresponding author(s): Golnaz Vahedi

Last updated by author(s): Feb 23, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	□	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	□	A description of all covariates tested
☒	□	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	□	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	□	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	□	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	scRNA-seq For samples prepared using ‘The Single Cell 3’ Reagent Kit v2’, the following chemistry was performed on an Illumina HiSeq4000: Read 1: 26 cycles, i7 Index: 8 cycles, i5 index: 0 cycles, and Read 2: 98 cycles. For samples prepared using ‘The Single Cell 3’ Reagent Kit v3 and v3.1’, the following chemistry was performed on an Illumina HiSeq 4000: Read 1: 28 cycles, i7 Index: 8 cycles, i5 index: 0 cycles, and Read 2: 91 cycles. Cell Ranger (10x Genomics; v3.0.1) was used for bcl2fastq conversion, aligning (using the hg38 reference genome), filtering, counting, cell calling, and aggregating (--normalize=none)
Data analysis	scRNA-seq data processing using Seurat package. The merged filtered feature-barcode matrix included counts from all the HPAP samples adding up to around 260,000 cells. We built the pipeline for downstream analysis using standard scRNA-seq packages such as Seurat v4.1.0, scDblFinder v1.8.0, SingleCellExperiment v1.16.0, sctransform v0.3.3, and scSorter v0.0.2 for building our pipeline carried out the following filtering steps. Initial Pre-processing. In the first step, we considered only those cells that contained expression values >0 for at least 200 genes using the min.features parameter. The min.cells parameter indicates to include only features detected in at least three cells. Doublet finding using scDblFinder. Doublets are commonly found in scRNA-seq data generated from droplet-based protocols. They arise due to cell capture or sorting mistakes, especially when thousands of cells are involved. Doublets are undesirable in scRNA-seq data as they lead to artifactual findings. We employed the scDblFinder method to identify and remove potential doublets. The scDblFinder uses an object of class SingleCellExperiment that contains at least the counts. It trains an iterative classifier on the neighborhood of real cells and artificial doublets. We measured the doublets for each donor sample separately by passing a vector of sample ids given as a parameter to the scDblFinder function. The program was run on multiple threads using the BPPARAM parameter to achieve faster computation. The scDblFinder

automatically chooses the doublet rate for 10x data; therefore, we left this parameter empty. The output includes two key columns, scDblFinder.score and scDblFinder.class. While the former represents the doublet score assigned for each cell, the latter shows whether the cell was assigned as a doublet or singlet.

Second-stage pre-processing. We further filtered the data based on nFeature_RNA and nCount_RNA, representing the number of genes detected in each cell and the total number of molecules (UMIs) detected within a cell. While the low nFeature_RNA in a cell may be dead/ dying or represent an empty droplet, the high nCount_RNA and nFeature_RNA indicate that the cell may be a doublet or multiplet. This filtering, along with the filtering of mitochondrial reads (Percent.mt), is a crucial pre-processing step because removing such outliers from these groups will also remove some doublets or dead/empty droplets. In addition, we removed the cells marked as doublets from the scDblFinder in the previous step.

Normalizing and scaling data using single cell transform (SCT). Seurat's SCTranform (SCT) function was used to normalize the counts that measure the differences in sequencing depth per cell for each sample. The SCT method is built on the regularized negative binomial model to perform normalization and variance stabilization of single-cell RNA-seq data. It removes the variation due to sequencing depth (nUMIs). The vars.to.regress parameter allowed us to regress the variation from other sources, such as the percentage of mitochondrial reads. The output of the SCT model is the normalized and scaled expression levels for all the transcripts. Lastly, the variable.features.n parameter is used to select the variable features and is set as 3,000.

PC calculation and clustering. We performed principal-component analysis (PCA) based on the top 3,000 most variable genes obtained from the SCT transformed data. We first calculated 50 PCs, and then the dimensionality was determined using an elbow plot showing each PC's standard deviation. Based on the elbow plot, we chose the top 20 PCs.
The FindNeighbors function in Seurat was used to compute the k-nearest neighbors in the dataset. The calculations were made on the reduced dimensions of the 20 PCs determined based on the elbow plot in the previous step. After calculating k-nearest neighbors, the FindClusters function. built on the concept of shared nearest neighbors (SNN), clustering algorithm was used to cluster the cells at the resolution of 0.8. Finally, the uniform manifold approximation and projection (UMAP) dimensional reduction technique was run on the selected 20 PCs.

Automatic cell annotation using the single-cell sorter method (scSorter). We employed the marker-based scSorter cell annotation method to assign cells to known cell types based on the marker genes. The cell type annotations were made based on the marker genes. The scSorter method is based on a semi-supervised learning algorithm. It does not require all genes but only a set of marker genes and highly variable genes from the expression data to perform the analysis. The scSorter function expects the pre-processed data as input. Therefore, we used the previously normalized, scaled, and transformed expression matrix generated by the SCT method as input. The top 3,000 most variable genes selected by the SCT method were also given as input for the scSorter method. Additionally, the scSorter function accepts "weight" for the marker genes. Therefore, we assigned equal weights to all markers through the default_weight option.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

https://hpap.pmacs.upenn.edu/analysis

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	*Use the terms sex (biological attribute) and gender (shaped by social and cultural circumstances) carefully in order to avoid confusing both terms. Indicate if findings apply to only one sex or gender; describe whether sex and gender were considered in study design whether sex and/or gender was determined based on self-reporting or assigned and methods used. Provide in the source data disaggregated sex and gender data where this information has been collected, and consent has been obtained for sharing of individual-level data; provide overall numbers in this Reporting Summary. Please state if this information has not been collected. Report sex- and gender-based analyses where performed, justify reasons for lack of sex- and gender-based analysis.*
Population characteristics	All donors were screened for autoantibodies prior to organ harvest, and AAb positivity was confirmed again post tissue processing and islet isolation. The contribution of covariates on our findings was assessed. Covariates were NOT found to affect cell clustering, and therefore, not affect our findings: autoantibody type, age, amylase levels, BMI, cold ischemic time, collection period, c-peptide levels, culture days, CyTOF purity, DCD or DBD, group (T1D, AAB +, or control), HbA1c%, individual, lipase levels, ancestry, sex, viability, or warm ischemic time.
Recruitment	Pancreatic islets were procured by the HPAP consortium (RRID:SCR_016202; https://hpap.pmacs.upenn.edu), part of the Human Islet Research Network (https://hirnetwork.org/), with approval from the University of Florida Institutional Review Board (IRB # 201600029) and the United Network for Organ Sharing (UNOS). A legal representative for each donor provided informed consent prior to organ retrieval. For T1D diagnosis, medical charts were reviewed and C-peptide levels were

measured in accordance with the American Diabetes Association guidelines (American Diabetes Association 2009). All donors were screened for autoantibodies prior to organ harvest, and AAb positivity was confirmed again post tissue processing and islet isolation.

Ethics oversight

Pancreatic islets were procured by the HPAP consortium under the Human Islet Research Network (https://hirnetwork.org/) with approval from the University of Florida Institutional Review Board (IRB # 201600029) and the United Network for Organ Sharing (UNOS).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

Describe how sample size was determined, detailing any statistical methods used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.

Data exclusions

Describe any data exclusions. If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.

Replication

Describe the measures taken to verify the reproducibility of the experimental findings. If all attempts at replication were successful, confirm this OR if there are any findings that were not replicated or cannot be reproduced, note this and describe why.

Randomization

Describe how samples/organisms/participants were allocated into experimental groups. If allocation was not random, describe how covariates were controlled OR if this is not relevant to your study, explain why.

Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☒	☐	Antibodies
☒	☐	Eukaryotic cell lines
☒	☐	Palaeontology and archaeology
☒	☐	Animals and other organisms
☒	☐	Clinical data
☒	☐	Dual use research of concern

Methods

n/a	Involved in the study	
☒	☐	ChIP-seq
☒	☐	Flow cytometry
☒	☐	MRI-based neuroimaging