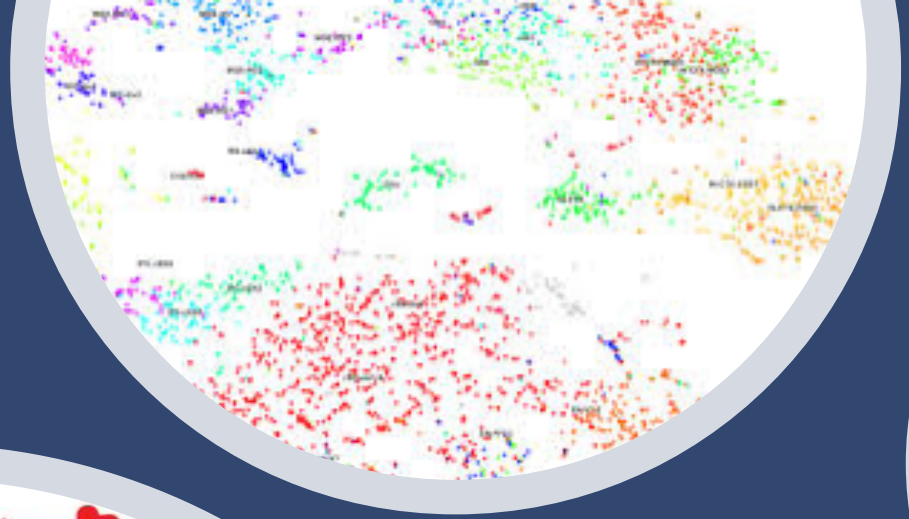
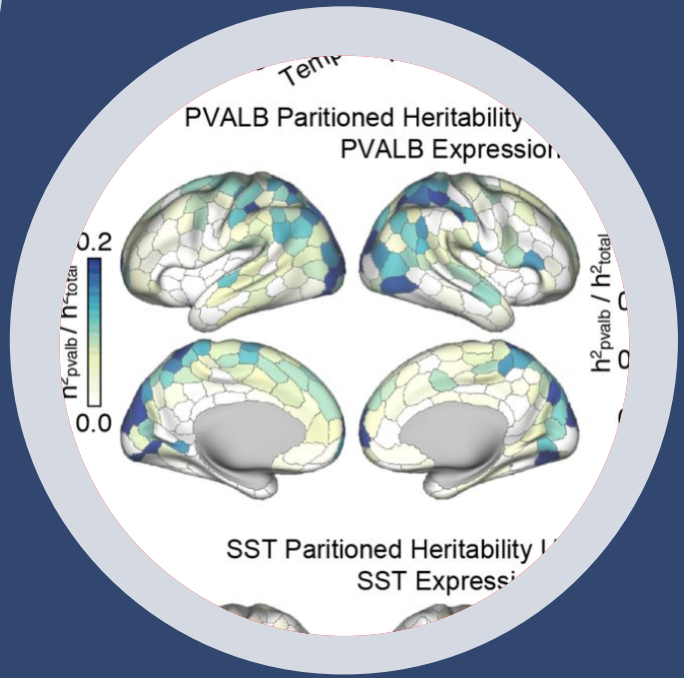
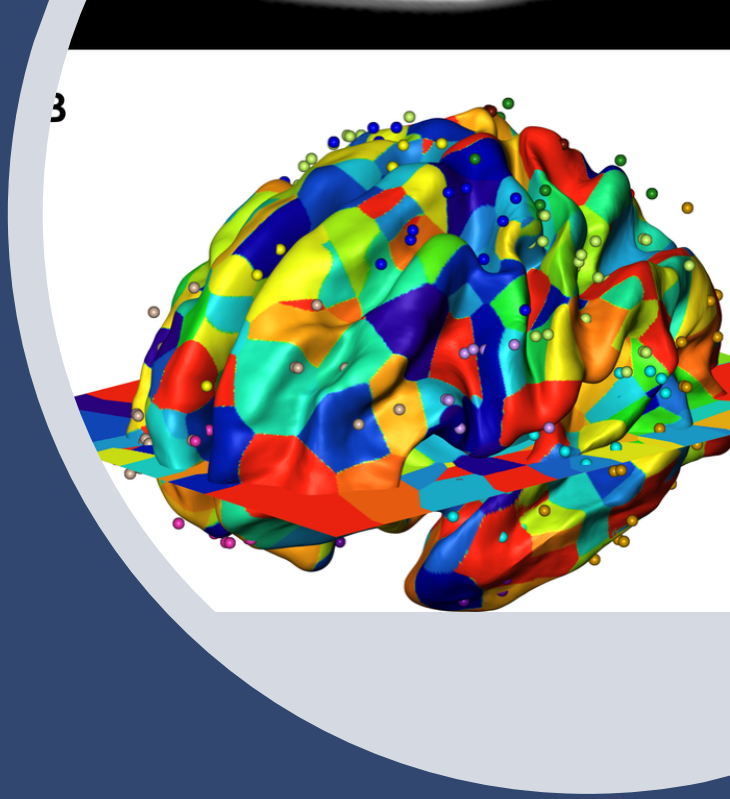
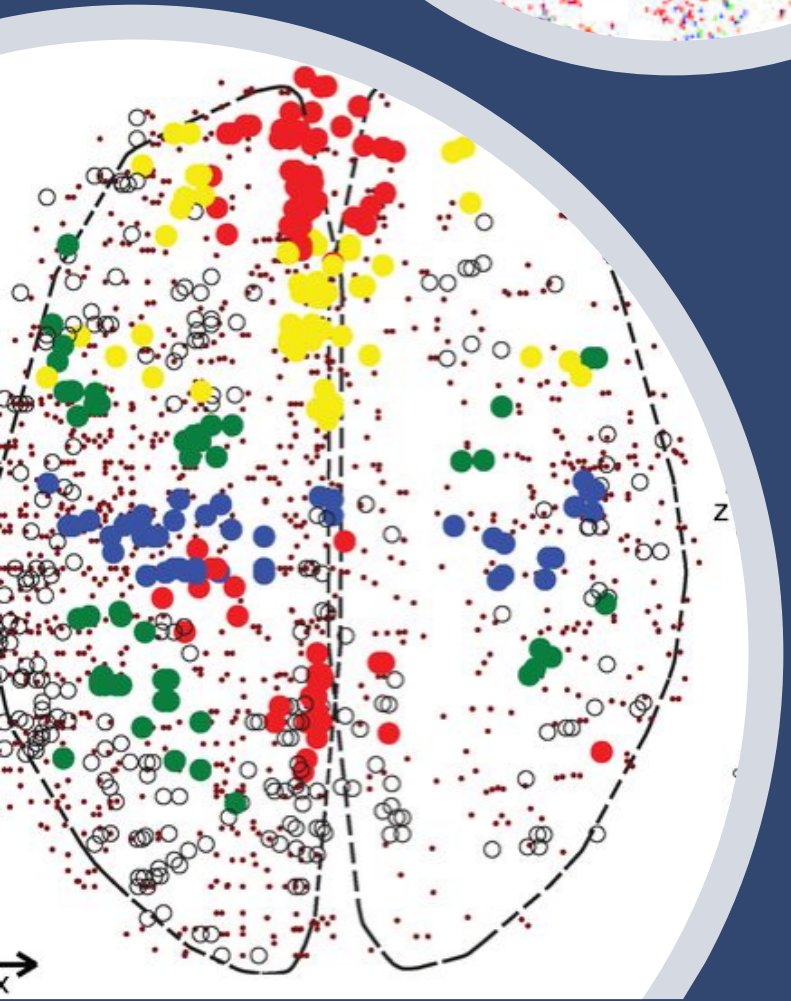


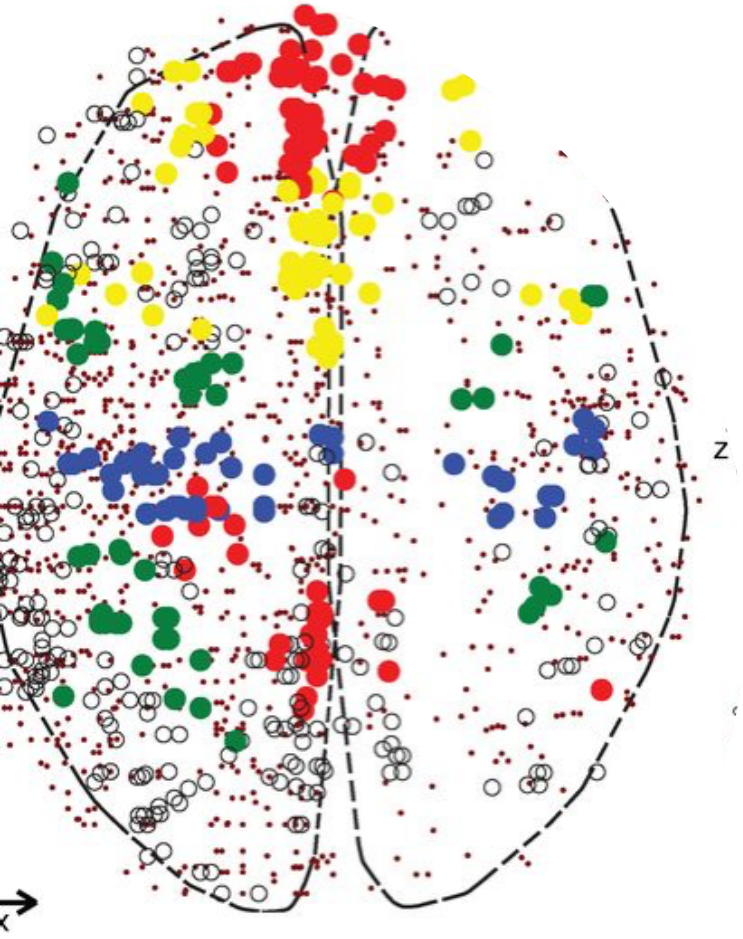






Integrating Neuroimaging Phenotypes and Post-Mortem Human Gene Expression Data

Integrating Neuroimaging Phenotypes and Post-Mortem Human Gene Expression Data



1. Gene Expression Quantification Techniques and Available Databases

2. AHBA Microarray Post-Processing

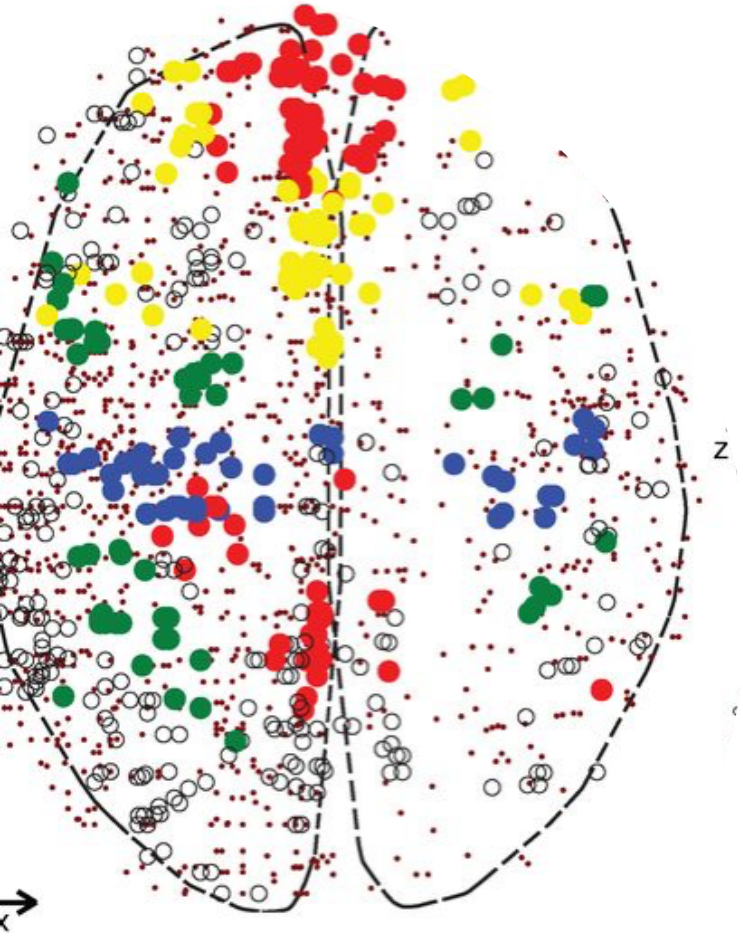
3. RNAseq Post-Processing

4. Integrating Gene Expression and Imaging Data
   - Assigning samples to brain regions
   - Removing donor-specific effects (individual variability)
   - Addressing spatially correlated gene expression

5. Gene Enrichment and Cell Type Analyses

6. Cool Papers

Integrating Neuroimaging Phenotypes and Post-Mortem Human Gene Expression Data



1. **Gene Expression Quantification Techniques and Available Databases**

2. AHBA Microarray Post-Processing

3. RNAseq Post-Processing

4. Integrating Gene Expression and Imaging Data
   - Assigning samples to brain regions
   - Removing donor-specific effects (individual variability)
   - Addressing spatially correlated gene expression

5. Gene Enrichment and Cell Type Analyses

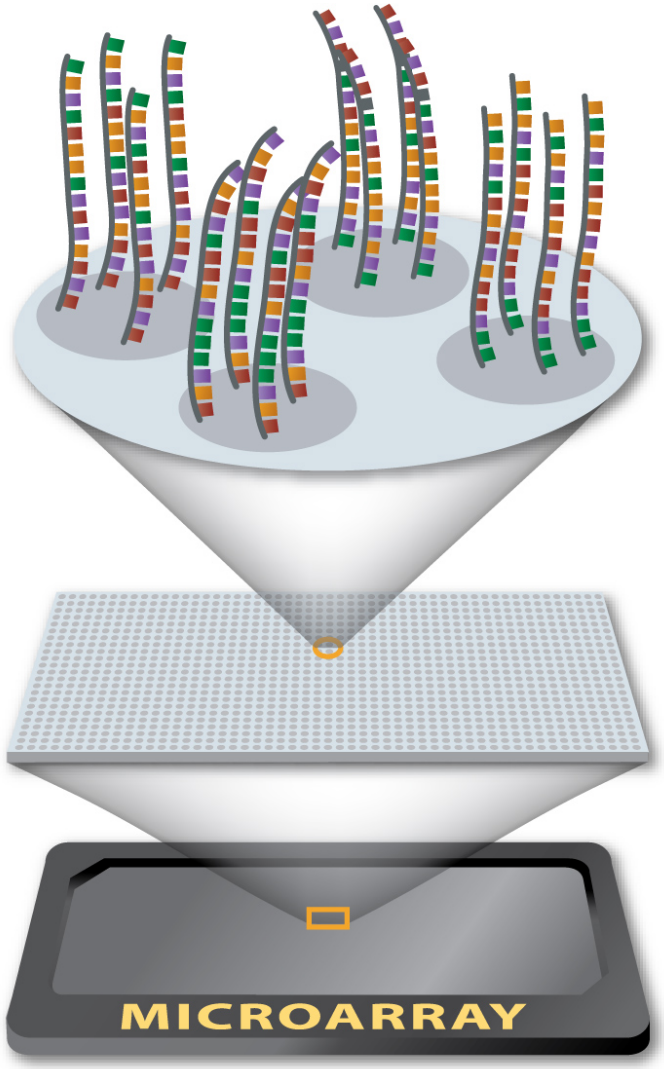
6. Cool Papers

Microarray

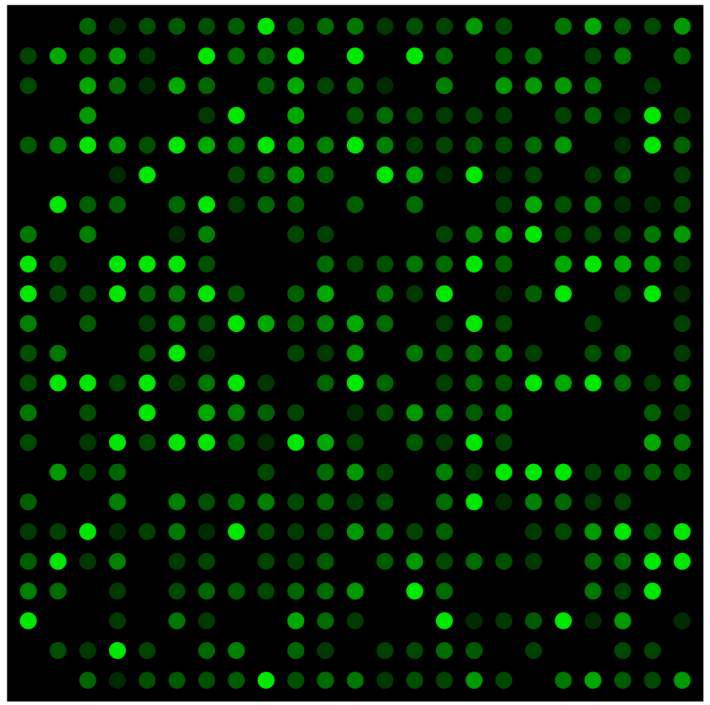
Microarray chips contain thousands of probe sequences (reporters) that are complimentary to known mRNA sequences. Tissue samples of interest are fluorescently labeled, and mRNA level is measured per sample via probe-specific fluorescence post hybridization.

Advantages:
- Cheap

Disadvantages:
- Laundry list



Microarray chip



Microarray

Microarray chips contain thousands of probe sequences (reporters) that are complimentary to known mRNA sequences. Tissue samples of interest are fluorescently labeled, and mRNA level is measured per sample via probe-specific fluorescence post hybridization.

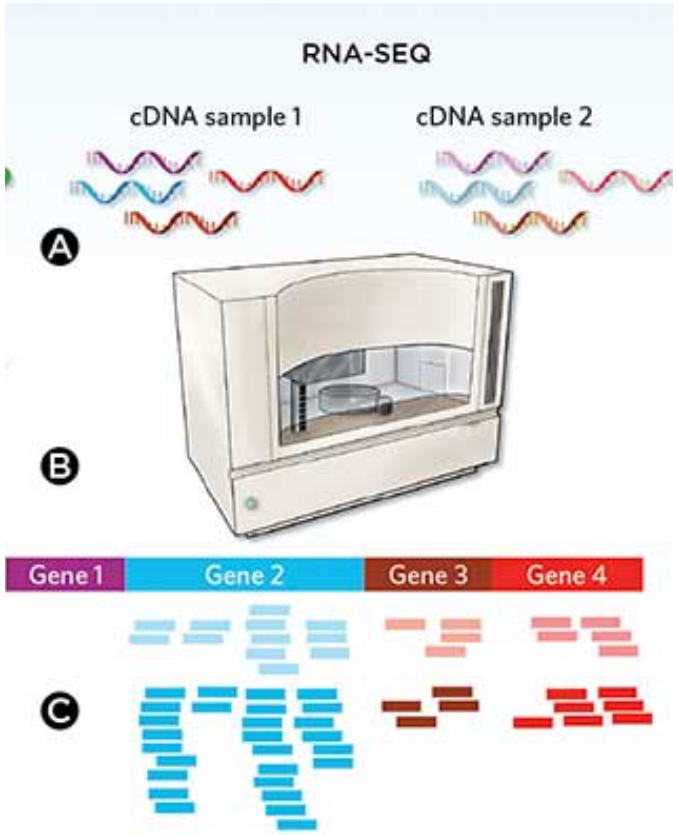
Advantages:
- Cheap

RNAseq

Tissue sample mRNA is isolated and reversed transcribed into cDNA. cDNA is then fragmented, sequenced, and mapped to a human reference genome. cDNA counts (table counts) allow for expression quantification.

Advantages:
- No background noise; no saturation
- No cross-hybridization or non-specific hybridization
- Higher sensitivity to genes expressed at low and high levels
- Lower technical variation

Disadvantages:
- cDNA not uniformly fragmented
- Length biases



Microarray

Microarray chips contain thousands of probe sequences (reporters) that are complimentary to known mRNA sequences. Tissue samples of interest are fluorescently labeled, and mRNA level is measured per sample via probe-specific fluorescence post hybridization.

Advantages:
- Cheap

RNAseq

Tissue sample mRNA is isolated and reversed transcribed into cDNA. cDNA is then fragmented, sequenced, and mapped to a human reference genome. cDNA counts (table counts) allow for expression quantification.

Advantages:
- No background noise; no saturation
- No cross-hybridization or non-specific hybridization
- Higher sensitivity to genes expressed at low and high levels
- Lower technical variation

Disadvantages:
- cDNA not uniformly fragmented
- Length biases

ISH

ISH uses labeled complementary sequences (probes) to identify the presence/location of mRNA of interest in tissue. Probes are radio-, fluorescent- or antigen- labeled and localized with fluorescence microscopy, autoradiography, or immunohistochemistry.

Advantages:
- Cell level resolution of gene expression

Disadvantages:
- Quantitation imperfect

Microarray

MIXED TISSUE SAMPLE

RNAseq

MIXED TISSUE SAMPLE

SINGLE CELL RNASEQ

ISH

TISSUE/LAYER/CELL SPECIFIC

Microarray

ALLEN (AHBA)

PRITZKER

STANLEY MEDICAL RESEARCH INSTITUTE

RNAseq

BRAINSPAN

GTEX

COMMON MIND

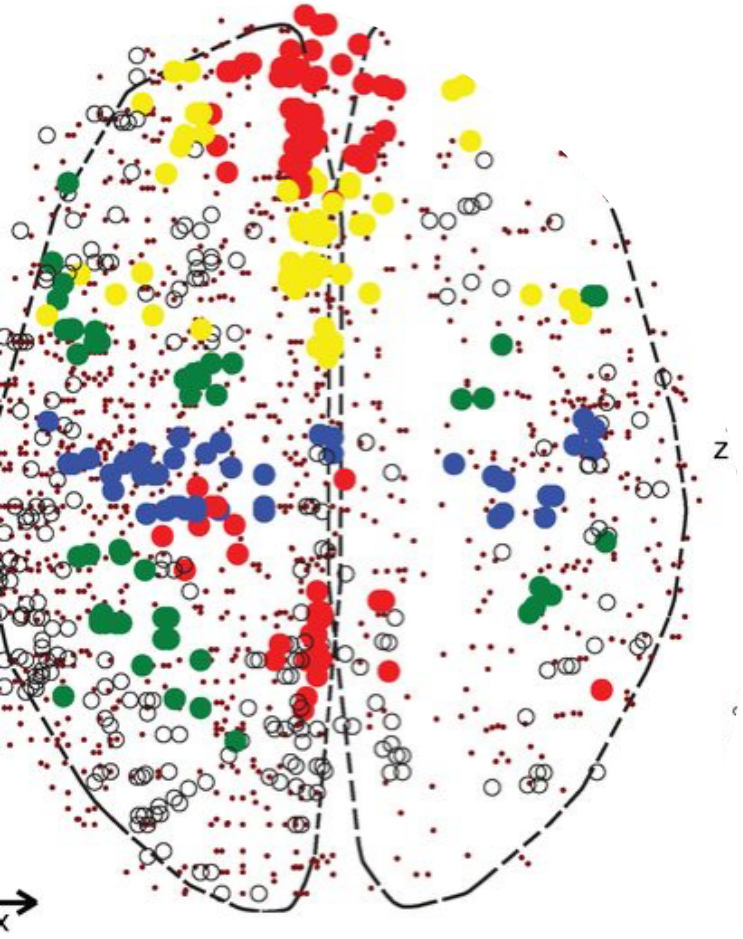
PSYCHENCODE

ISH

ALLEN (AHBA): NTs

DATABASE	GENE QUANT	BRAIN REGIONS SAMPLED	PARTICIPANT NUMBER	PARTICIPANT CHARACTERISTICS	OTHER
BrainSpan	RNAseq	From birth to 60+ years: DLPFC, VLPFC, medial PFC, orbitofrontal, M1, S1, A1, V1, posterior inferior parietal cortex, posterior superior temporal cortex, inferior temporal cortex, hippocampus, amygdala, striatum, thalamus MDN, cerebellar cortex.	57 total, throughout the lifespan (including prenatal) 21 post-natal 1340 tissue samples Not all regions in all participants	Healthy (excluded if had neurological, neurodegenerative or psychiatric disorder, prolonged agonal conditions, death by suicide or overdose, severe head injury)	Limited participant metadata ** three dimensional coordinates not available for these samples (i.e. no MRI space data)
Allen Brain Atlas	Microarray (120 RNAseq samples)	500 samples per hemisphere, most brain regions	6 adults	Healthy (no psychiatric or neuropathological history)	93% of genes represented with 2+ probes
Genotype Tissue Expression Project (GTEx)	RNAseq	Amygdala ACC Caudate Cerebellum Cortex BA9 Hippocampus Hypothalamus NAc Putamen Substantia nigra	129 147 194 175 205 175 165 170 202 170 114 Ages 20-79	Supposed to be “normal” but unclear if they screened for psychiatric disorders or have this information at all	
Pritzker	Microarray	DLPFC, nAcc, AnCg, HC, CB, Amygdala	80+ controls, 34 MDD	Healthy and Depressed Also data for schiz and bipolar?	Has age, sex, ethnicity, agonal factor scores, tissue pH, cause of death, TOD
CommonMind Consortium	RNAseq	DLPFC publically available, ACC and superior temporal gyrus collected	968 total	Healthy, schizophrenia, bipolar	
Stanley Medical Research Institute	Microarray	Broadman 6, 8/9, 10, 46 and cerebellum	All participants > 30	Healthy, schizophrenia, bipolar, MDD	Online Genomics Database comprised of derived data from 12 studies, 988 arrays across 6 platforms
BrainCloud GSE30272	Microarray	PFC, other regions coming?	1, 000 total? Developmental	No neuropathological or neuropsychiatric disorders	Has associated genetic and epigenetic information for people
PsychEncode	RNAseq	DLPFC, ACC, temporal cortex, hippocampus, amygdala, caudate, nucleus accumbens, cerebellum	1908 donors, 2996 brain samples across these studies: LIBD_szControl, CCMC_HBCC, CMC BrainSpan, BrainGVEX, BipSeq, Yale-ASD, UCLA-ASD, EpiGABA	Healthy, schizophrenia, BPD, ASD	https://dukespace.lib.duke.edu/dspace/bitstream/handle/10161/13706/psychENCODE_2015.pdf?sequence=1&isAllowed=y https://www.synapse.org/#!Synapse:syn4921369/wiki/390659

Integrating Neuroimaging Phenotypes and Post-Mortem Human Gene Expression Data



1. Gene Expression Quantification Techniques and Available Databases

2. **AHBA Microarray Post-Processing**

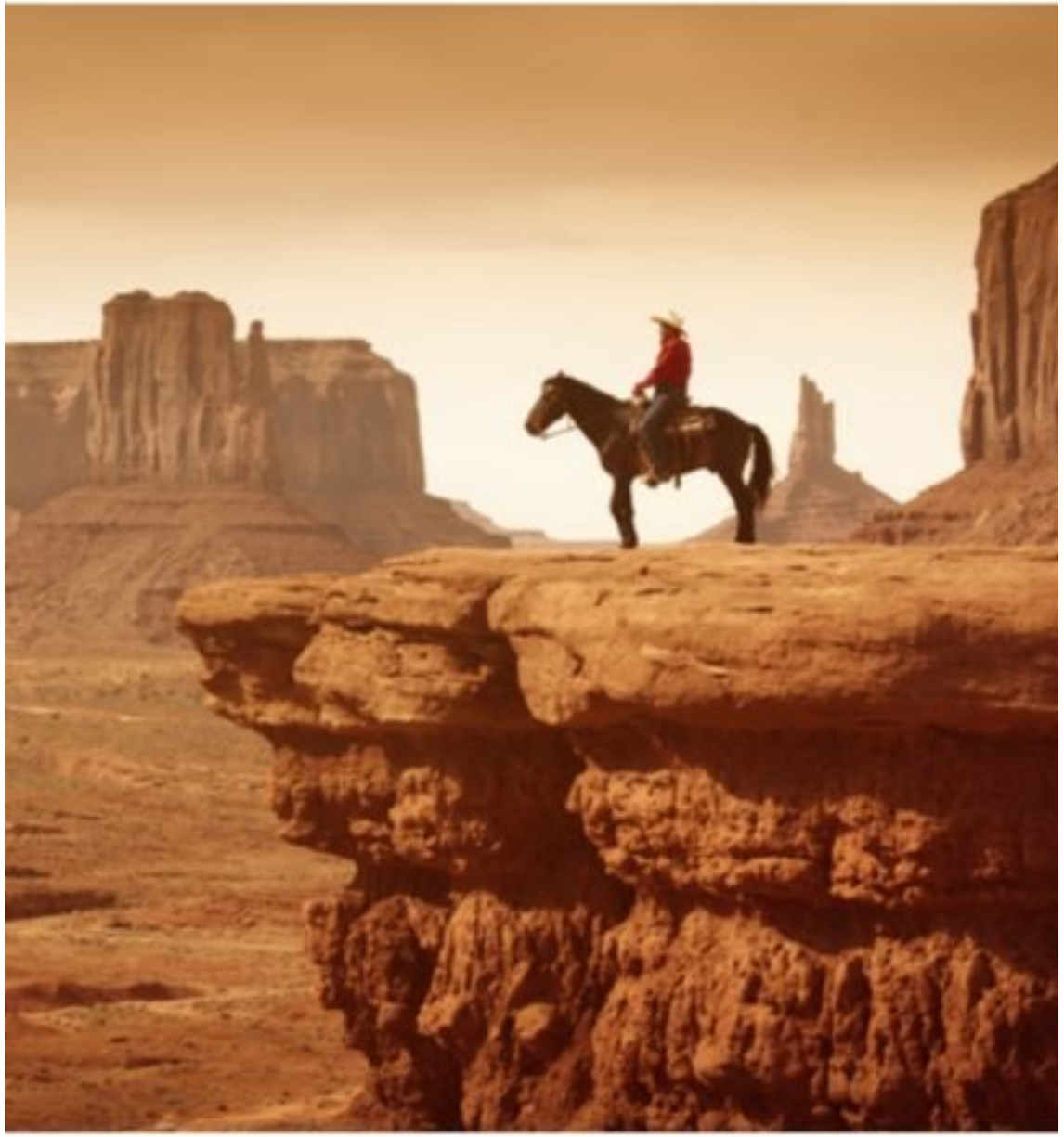
3. RNAseq Post-Processing

4. Integrating Gene Expression and Imaging Data
   - Assigning samples to brain regions
   - Removing donor-specific effects (individual variability)
   - Addressing spatially correlated gene expression

5. Gene Enrichment and Cell Type Analyses

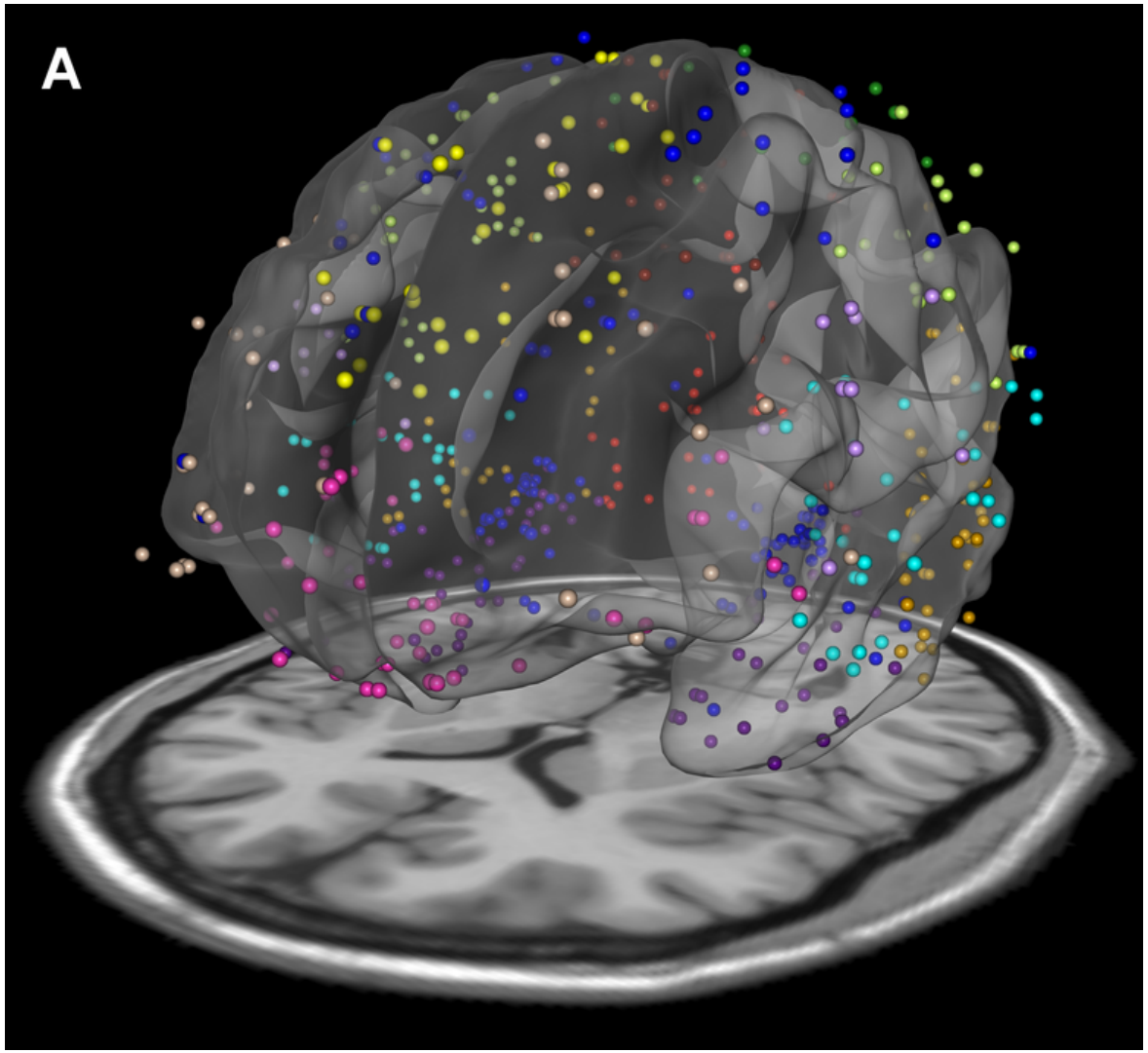
6. Cool Papers

Diffusion MRI



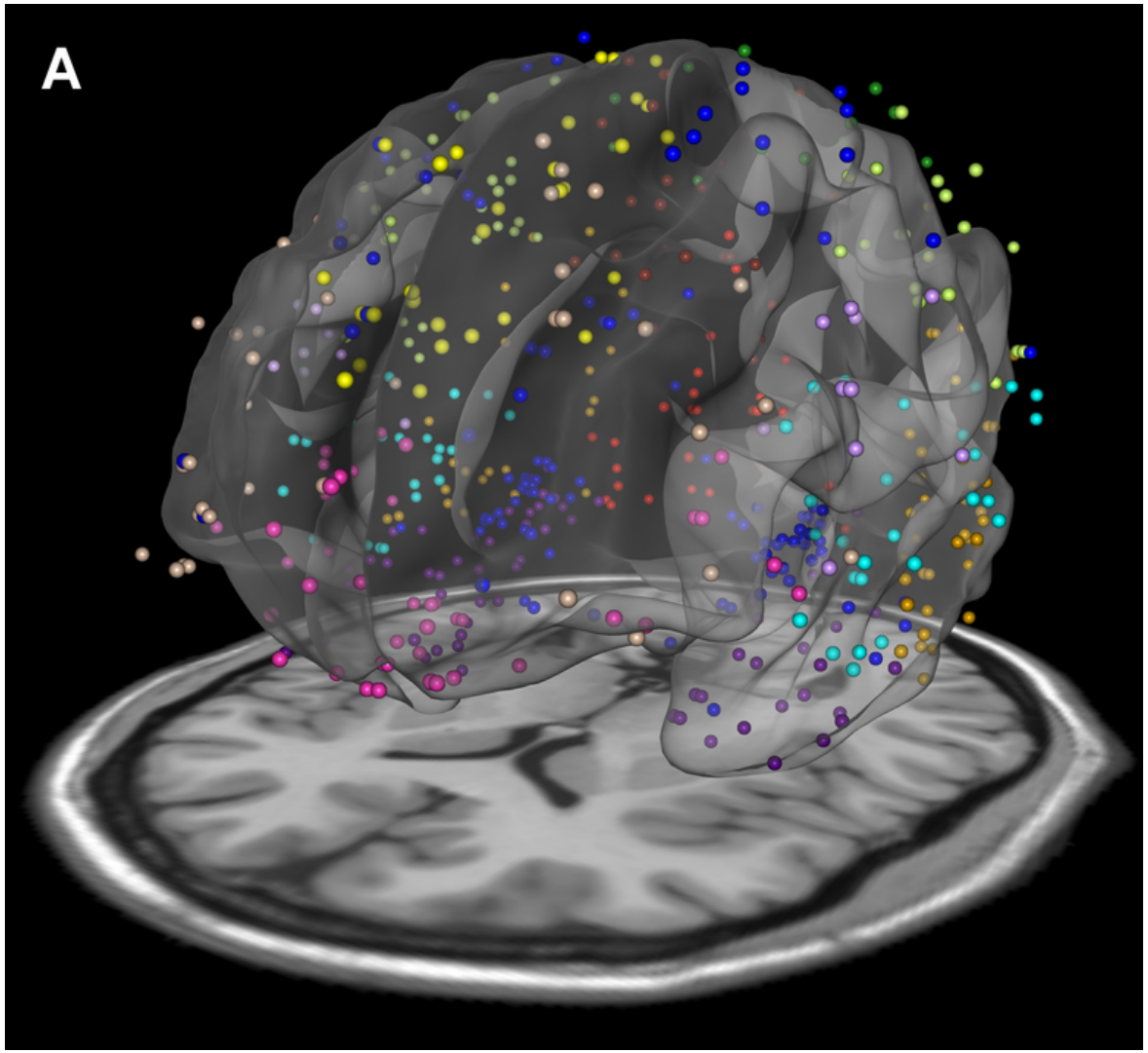
Gene Expression

AHBA Microarray Post-Processing



6 individuals (4 M 2 F)
6 with LH, 2 with RH
~500 samples per hemisphere
4,000 unique anatomical samples
20,000 + genes
Cortical and subcortical
MNI coordinates of probes

AHBA Microarray Post-Processing



Probe to gene annotation

Background noise filtering

Probe selection

Within sample normalization

AHBA Microarray Post-Processing

Probe to gene annotation

Background noise filtering
Probe selection
Within sample normalization

Probe to gene mapping gets updated over time as sequencing databases get updated. If you want to use most up-to-date information, need to Re-annotate probes, rather than relying on annotations provided by AHBA.

Outdated annotation of probes becomes an increasing problem in publicly available gene expression catalogues such as the ALLEN brain atlas as researchers tend to use the provided expression data as is, that is, without further validity checks and quality control. Arloth et al., 2015, PLOS One

AHBA Microarray Post-Processing

Probe to gene annotation

Background noise filtering
Probe selection
Within sample normalization

Probe to gene mapping gets updated over time as sequencing databases get updated. If you want to use most up-to-date information, need to Re-annotate probes, rather than relying on annotations provided by AHBA.

Outdated annotation of probes becomes an increasing problem in publicly available gene expression catalogues such as the ALLEN brain atlas as researchers tend to use the provided expression data as is, that is, without further validity checks and quality control. Arloth et al., 2015, PLOS One

Re-annotation software: **Re-Annotator** (command line, maps to human mRNA reference first, then to human reference genome using **ANNOVAR**)

- RefSeq database created by concatenated exon isoforms to generate one mRNA for matching in silico

AHBA Microarray Post-Processing

Probe to gene annotation

Background noise filtering

Probe selection

Within sample normalization

Microarrays have background noise (background level of fluorescence) due to non-specific binding of cDNA to non-complimentary probes. 30% of AHBA probes do not exceed background in at least 50% of samples (13,844 of 45,812).

AHBA Microarray Post-Processing

Probe to gene annotation

Background noise filtering

Probe selection

Within sample normalization

Microarrays have background noise (background level of fluorescence) due to non-specific binding of cDNA to non-complimentary probes. 30% of AHBA probes do not exceed background in at least 50% of samples (13,844 of 45,812).

AHBA has a binary indicator as to whether a given probe has an expression level above background signal (from negative controls) in a specific sample. Can elect to disregard probes where expression level is less than background in given % of samples (e.g. 50%). Above background=

1. mean signal of probe's expression is significantly greater than background (t-test)
2. difference between background signal and background subtracted probe signal is > 2.6

AHBA Microarray Post-Processing

Probe to gene annotation
Background noise filtering
Probe selection
Within sample normalization

Multiple probes are used to measure expression level of the same gene (probes target different exons). Probe-specific expression levels vary for same gene due to probe-specific differences in hybridization specificity, splice variants, probe immobilization during manufacturing, inaccurate probe to gene annotation

AHBA Microarray Post-Processing

Probe to gene annotation
Background noise filtering
Probe selection
Within sample normalization

Multiple probes are used to measure expression level of the same gene (probes target different exons). Probe-specific expression levels vary for same gene due to probe-specific differences in hybridization specificity, splice variants, probe immobilization during manufacturing, inaccurate probe to gene annotation

Approaches to probe selection:

- mean of all probes
- probe with highest expression in sample
- probe with highest variances across brain regions
- probe with highest loading onto PC1 from PCA of all probes
- probe exceeding background noise in highest proportion of samples
- probe with highest differential stability

AHBA Microarray Post-Processing

Probe to gene annotation
Background noise filtering
Probe selection
Within sample normalization

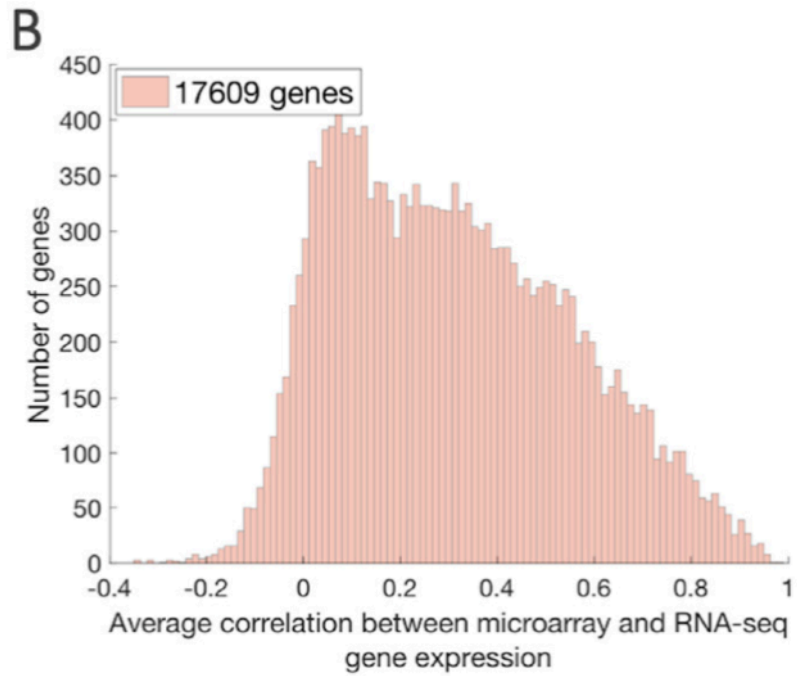
Multiple probes are used to measure expression level of the same gene (probes target different exons). Probe-specific expression levels vary for same gene due to probe-specific differences in hybridization specificity, splice variants, probe immobilization during manufacturing, inaccurate probe to gene annotation

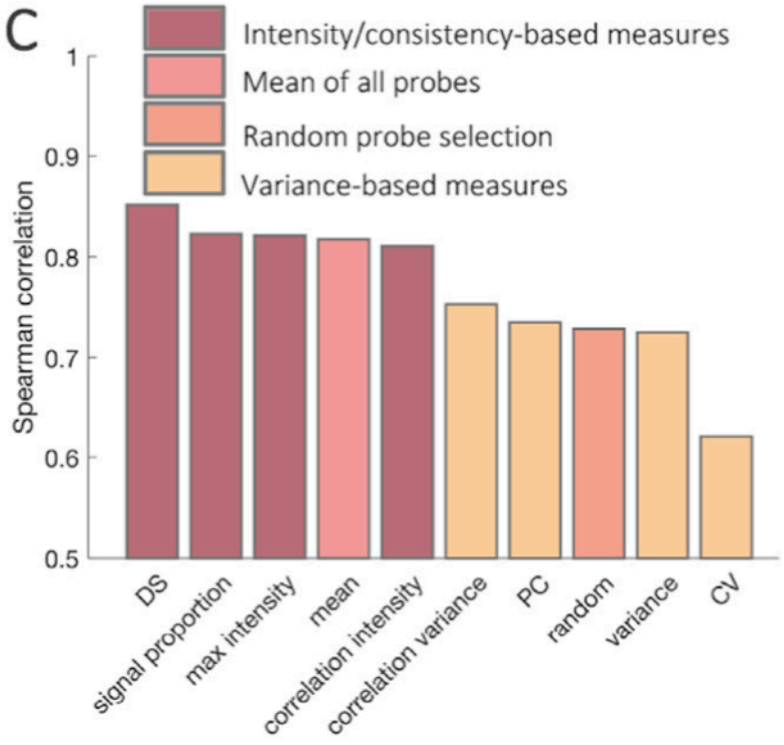
Approaches to probe selection: **Probe with highest correlation to RNAseq or Differential Stability**

* can also set RNAseq correlation minimum thresholds and DS thresholds to remove probes/genes

AHBA Microarray Post-Processing

Probe to gene annotation
Background noise filtering
Probe selection
Within sample normalization





Differential Stability: assesses how consistent regional variation in expression is across all 6 AHBA donors for a given probe

Arnatkevičiūtė et al., 2019, NeuroImage

AHBA Microarray Post-Processing

Probe to gene annotation
Background noise filtering
Probe selection
Within sample normalization

Tissue samples in close proximity can show significantly (and artefactually) different expression profiles across all genes (i.e. all genes have higher expression in sample X compared to Y), confounding analysis of relative gene expression across regions.

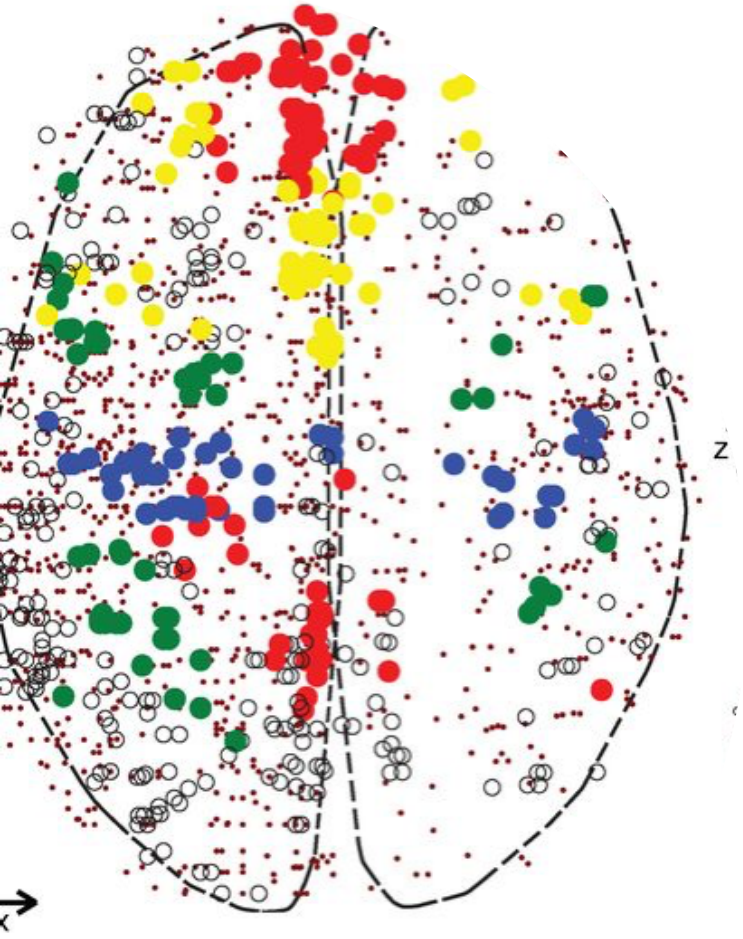
AHBA Microarray Post-Processing

Probe to gene annotation
Background noise filtering
Probe selection
Within sample normalization

Tissue samples in close proximity can show significantly (and artefactually) different expression profiles across all genes (i.e. all genes have higher expression in sample X compared to Y), confounding analysis of relative gene expression across regions.

Can be addressed via within-sample, across genes normalization to capture relative gene expression

Integrating Neuroimaging Phenotypes and Post-Mortem Human Gene Expression Data



1. Gene Expression Quantification Techniques and Available Databases

2. AHBA Microarray Post-Processing

3. **RNAseq Post-Processing**

4. Integrating Gene Expression and Imaging Data
   - Assigning samples to brain regions
   - Removing donor-specific effects (individual variability)
   - Addressing spatially correlated gene expression

5. Gene Enrichment and Cell Type Analyses

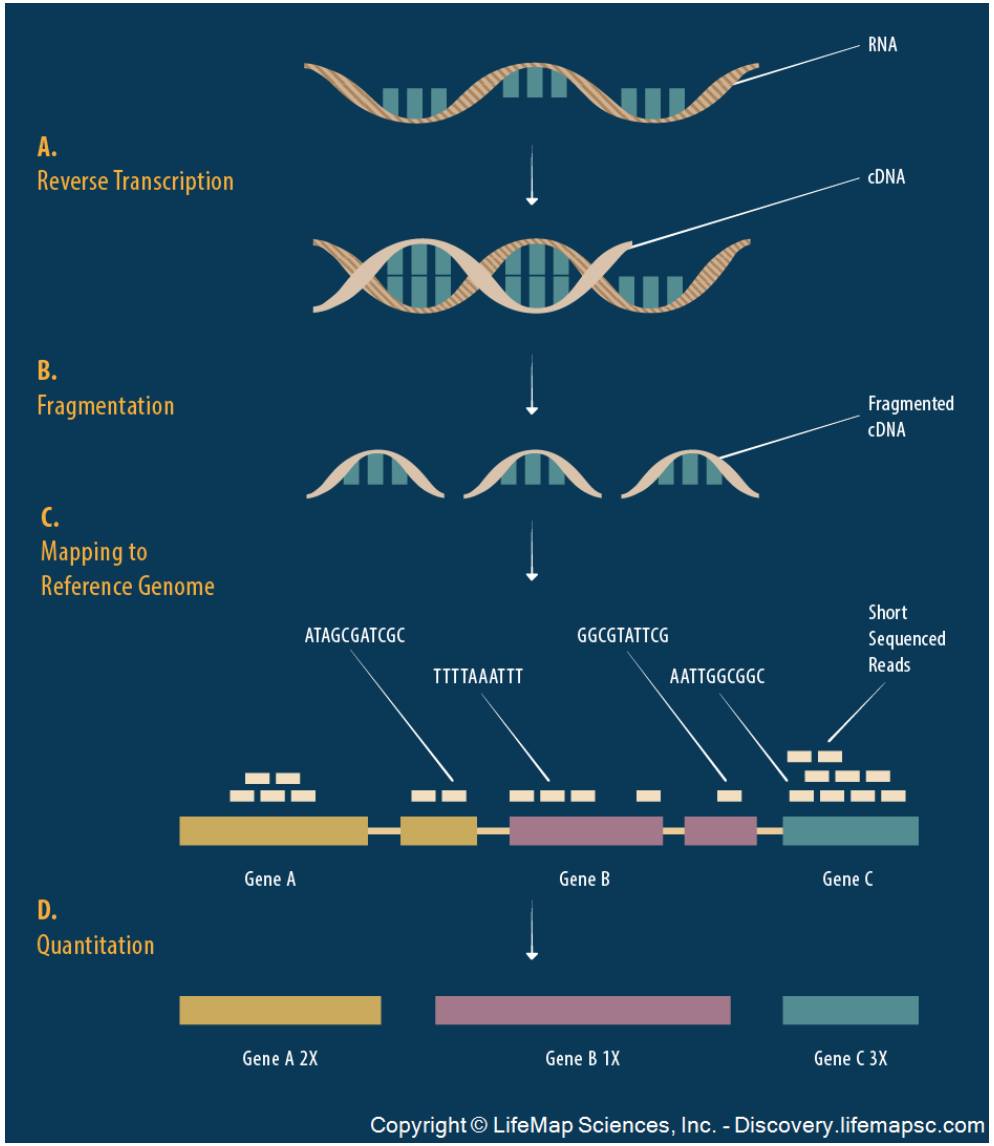
6. Cool Papers

RNASeq Post-Processing

Gene length and library size normalization

Remove genes with low counts, log transform

Expression value normalization/rescaling



RNASeq Post-Processing

Gene length and library size normalization

Remove genes with low counts, log transform

Expression value normalization/rescaling

The number of cDNA fragments that map to a given gene in the human reference genome will be influenced by:

1. **The length of the gene (longer gene = longer mRNA = more fragments = more counts)**
2. **The “library size”, i.e., the number of cDNA fragments made per experiment**

Hence, gene size and library size will greatly bias gene expression estimates

RNASeq Post-Processing

Gene length and library size normalization

Remove genes with low counts, log transform
Expression value normalization/rescaling

The number of cDNA fragments that map to a given gene in the human reference genome will be influenced by:

1. **The length of the gene (longer gene = longer mRNA = more fragments = more counts)**
2. **The "library size", i.e., the number of cDNA fragments made per experiment**

Hence, gene size and library size will greatly bias gene expression estimates

Gene length normalization: divide raw counts by length of gene in kilobases (reads per kilobase)
Library size normalization: divide library size (total sum of RPK) by 1 million

RNASeq Post-Processing

Gene length and library size normalization

Remove genes with low counts, log transform

Expression value normalization/rescaling

RNASeq Post-Processing

Gene length and library size normalization
Remove genes with low counts, log transform

Expression value normalization/rescaling

TPM and RPKM/FPKM values are still biased by the overall RNA repertoire of a sample (i.e. overall expression properties of entire sample). Need to scale TPM or RPKM/FPKM values for a given sample such that we can compare cross samples.

RNASeq Post-Processing

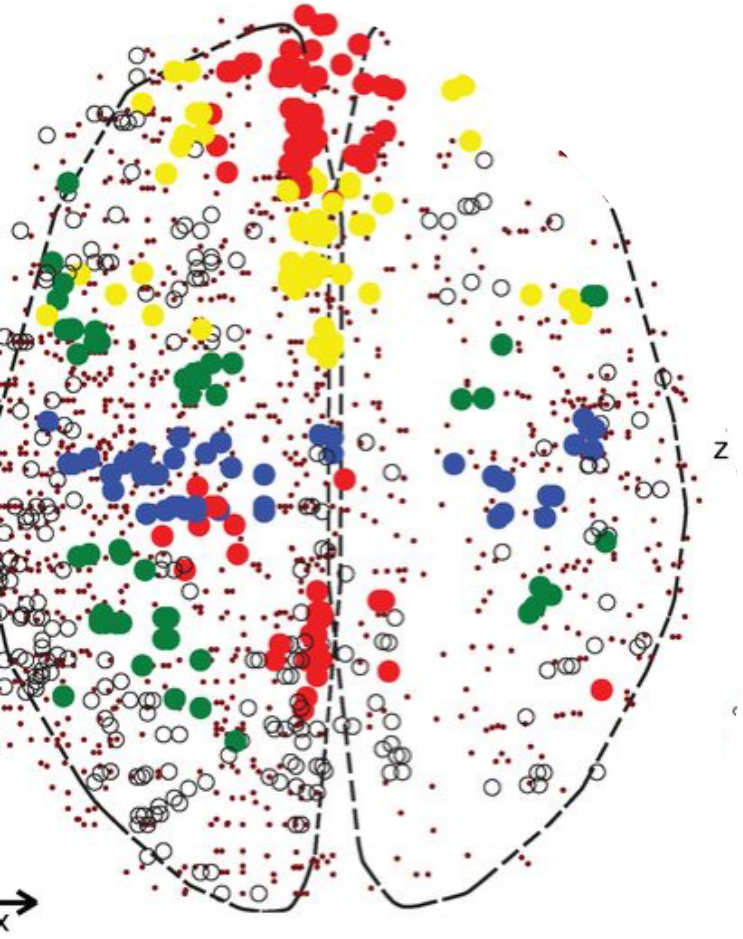
Gene length and library size normalization
Remove genes with low counts, log transform

Expression value normalization/rescaling

TPM and RPKM/FPKM values are still biased by the overall RNA repertoire of a sample (i.e. overall expression properties of entire sample). Need to scale TPM or RPKM/FPKM values for a given sample such that we can compare across samples.

Trimmed mean of M values (TMM): ignore highest expressed and lowest expression genes in normalization procedure, scale rest of genes such that “total” RNA expression across all non-differentially expressed genes is now equal across samples/regions → rescaling/correction factors

Integrating Neuroimaging Phenotypes and Post-Mortem Human Gene Expression Data



1. Gene Expression Quantification Techniques and Available Databases

2. AHBA Microarray Post-Processing

3. RNAseq Post-Processing

4. **Integrating Gene Expression and Imaging Data**
   - **Assigning samples to brain regions**
   - **Removing donor-specific effects (individual variability)**
   - **Addressing spatially correlated gene expression**

5. Gene Enrichment and Cell Type Analyses

6. Cool Papers

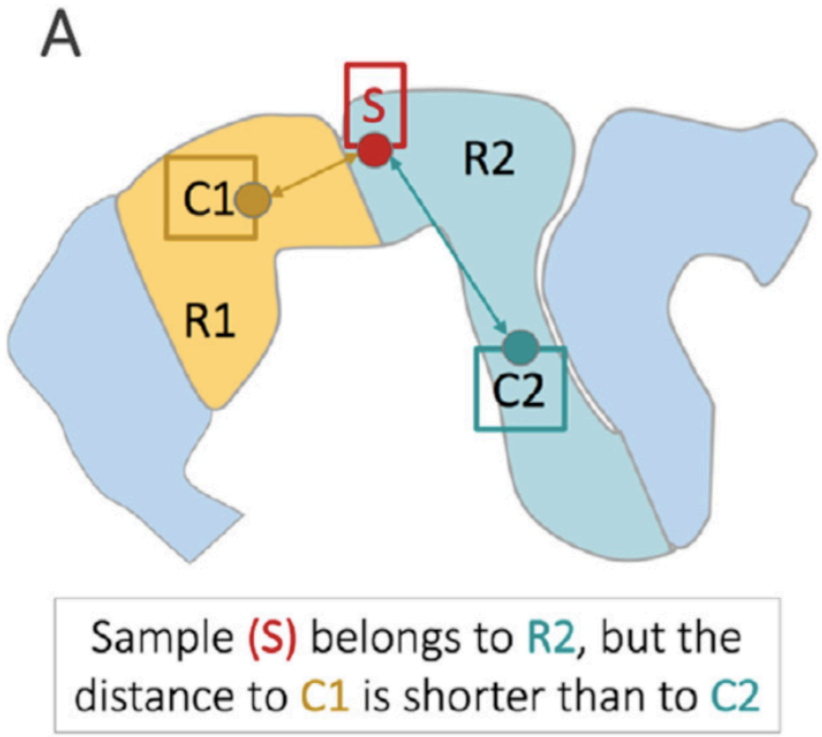
Integrating Gene Expression and Neuroimaging Data: Assigning Samples to Regions

- Generate a surface reconstruction of each individual donor's brain, and warp atlas of interest to surface reconstruction. Then generate volumetric parcellation (subcortical mapping must all be volumetric). Use subject-space sample coordinates

Integrating Gene Expression and Neuroimaging Data: Assigning Samples to Regions

- Generate a surface reconstruction of each individual donor's brain, and warp atlas of interest to surface reconstruction. Then generate volumetric parcellation (subcortical mapping must all be volumetric). Use subject-space sample coordinates

- Assign samples to regions based on Euclidean distance in 3D space (find minimum distance between a sample and any voxel in a region) using a maximum distance of 2mm

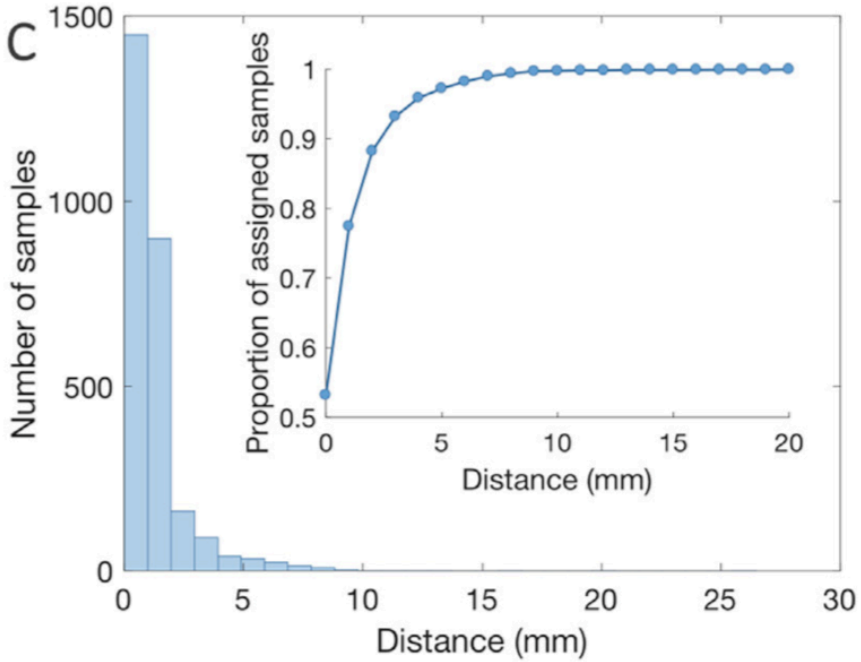


Arnatkevičiūtė et al., 2019, NeuroImage

Integrating Gene Expression and Neuroimaging Data: Assigning Samples to Regions

- Generate a surface reconstruction of each individual donor's brain, and warp atlas of interest to surface reconstruction. Then generate volumetric parcellation (subcortical mapping must all be volumetric). Use subject-space sample coordinates

- Assign samples to regions based on Euclidean distance in 3D space (find minimum distance between a sample and any voxel in a region) using a maximum distance of 2mm



Arnatkevičiūtė et al., 2019, NeuroImage

Integrating Gene Expression and Neuroimaging Data: Assigning Samples to Regions

- Generate a surface reconstruction of each individual donor's brain, and warp atlas of interest to surface reconstruction. Then generate volumetric parcellation (subcortical mapping must all be volumetric). Use subject-space sample coordinates

- Assign samples to regions based on Euclidean distance in 3D space (find minimum distance between a sample and any voxel in a region) using a maximum distance of 2mm

- Perform sample-to-region distance-based mapping separately for LH and RH cortex and for cerebellum

Integrating Gene Expression and Neuroimaging Data: Assigning Samples to Regions

- Generate a surface reconstruction of each individual donor's brain, and warp atlas of interest to surface reconstruction. Then generate volumetric parcellation (subcortical mapping must all be volumetric). Use subject-space sample coordinates

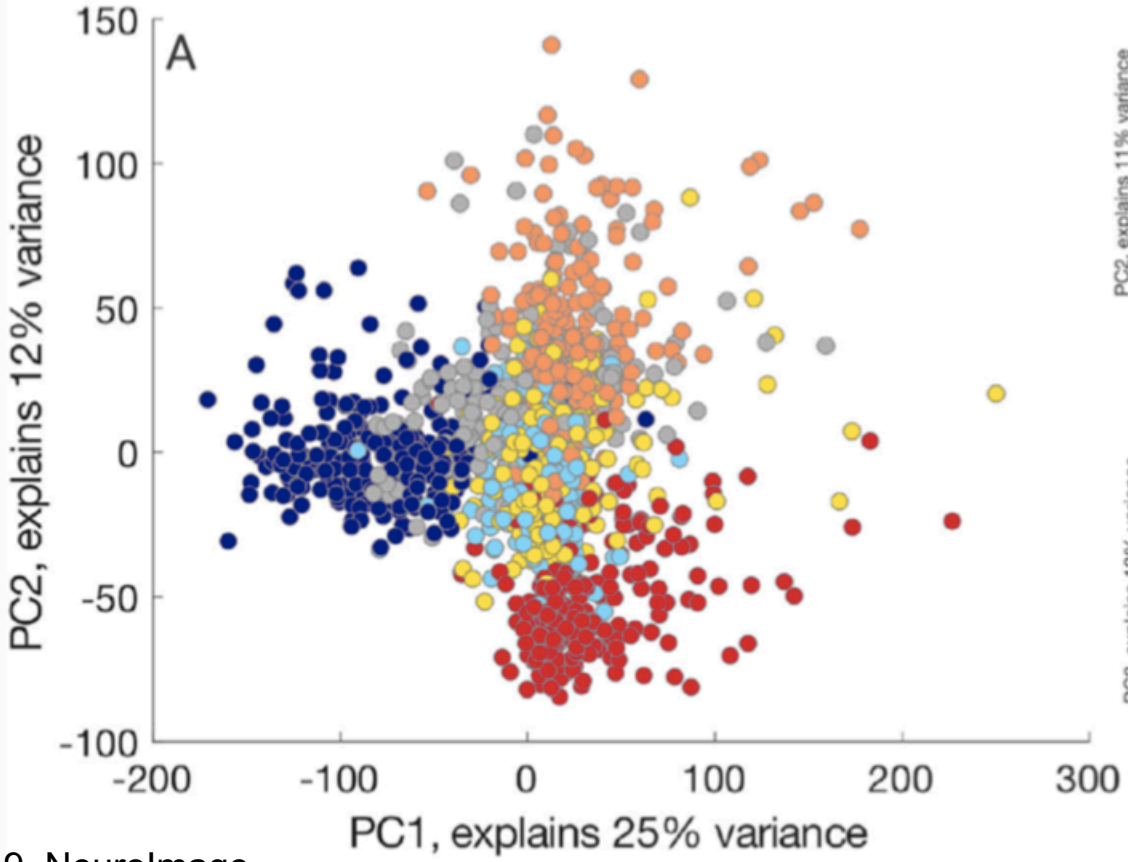
- Assign samples to regions based on Euclidean distance in 3D space (find minimum distance between a sample and any voxel in a region) using a maximum distance of 2mm

- Perform sample-to-region distance-based mapping separately for LH and RH cortex and for cerebellum

- Impute missing voxel gene expression data based on weighted linear combination of nearest sample values, using a Gaussian process regression model

Integrating Gene Expression and Neuroimaging Data: Remove Donor-Specific Effects

We want to remove differences in gene expression that arise due to donor-specific variability in age, sex, race, medical history, cause of death, post-mortem interval, etc.



Arnatkevičiūtė et al., 2019, NeuroImage

Integrating Gene Expression and Neuroimaging Data: Remove Donor-Specific Effects

We want to remove differences in gene expression that arise due to donor-specific variability in age, sex, race, medical history, cause of death, post-mortem interval, etc.

Z-score

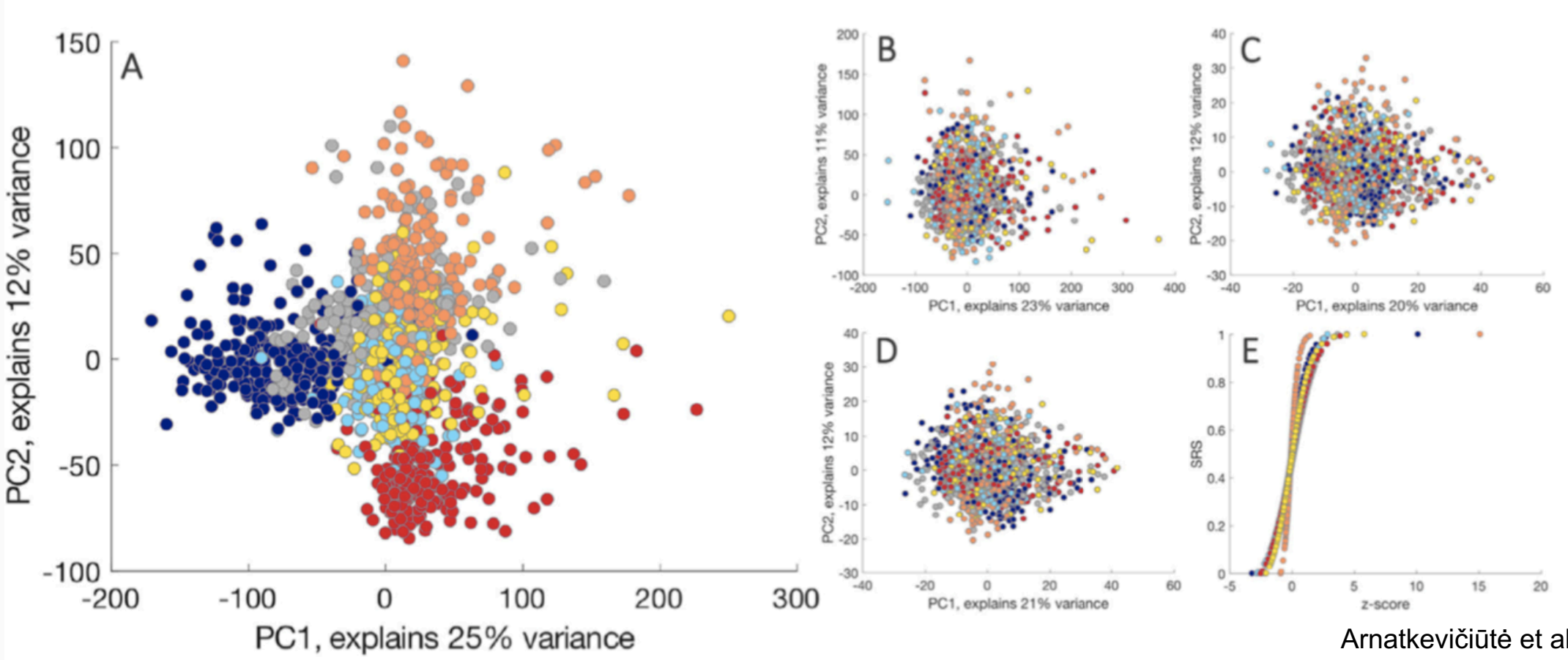
Scaled robust sigmoid normalization (SRS)

limma (linear model as batch effects) + SRS

Combat

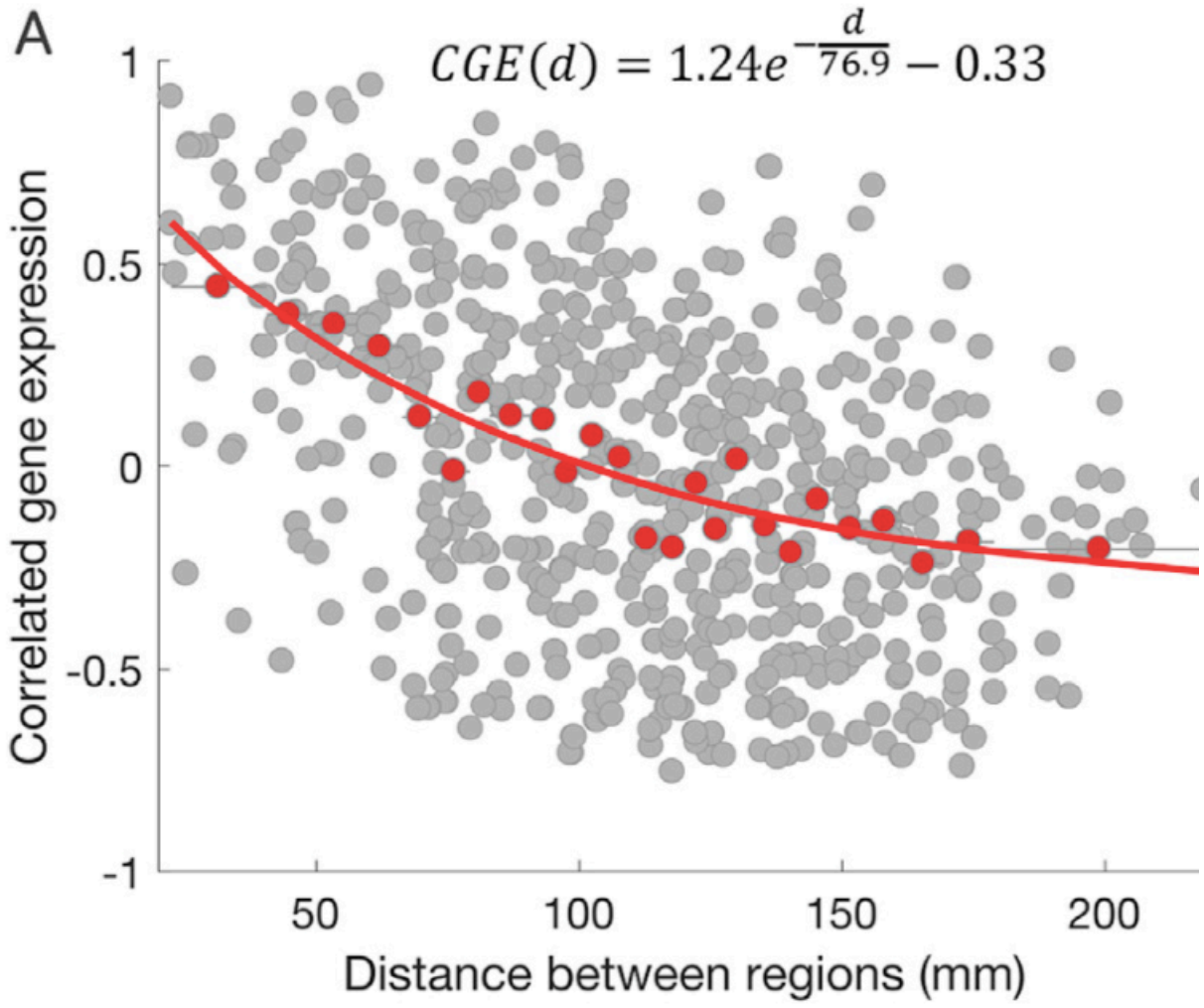
Integrating Gene Expression and Neuroimaging Data: Remove Donor-Specific Effects

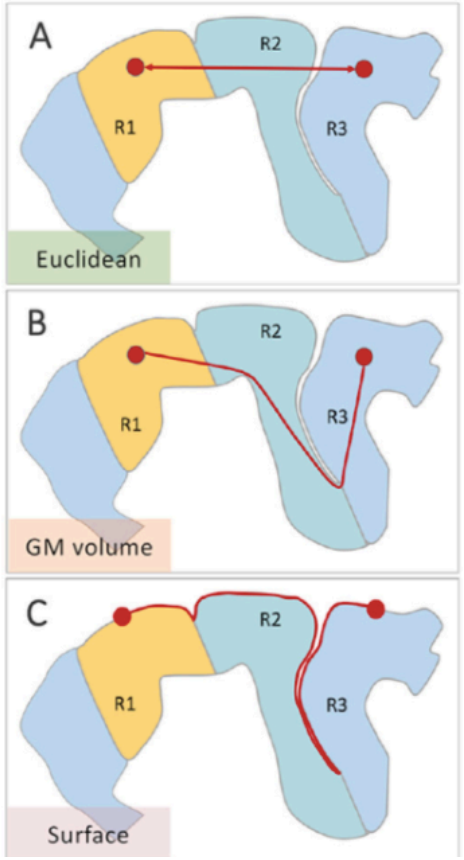
We want to remove differences in gene expression that arise due to donor-specific variability in age, sex, race, medical history, cause of death, post-mortem interval, etc.



Arnatkevičiūtė et al., 2019, NeuroImage

Integrating Gene Expression and Neuroimaging Data: Addressing Spatially Correlated Expression





Imaging metrics and gene expression show high spatial autocorrelation (regions closer to each other show highly correlated phenotypes/gene expression). Thus, associations observed between an imaging-derived phenotype and gene expression values may arise due to spatial effects. Need to ensure that an associations are stronger than those expected due to low-order spatial gradients.

Arnatkevičiūtė et al., 2019, NeuroImage

Integrating Gene Expression and Neuroimaging Data: Addressing Spatially Correlated Expression

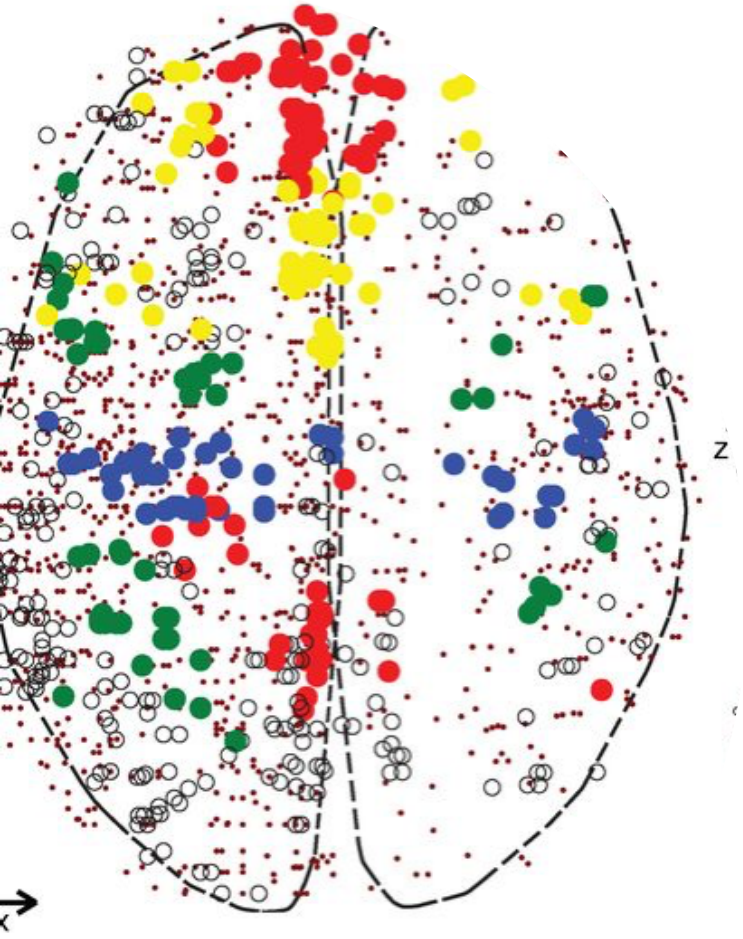
Spin testing

Spatially-constrained permutation

- **block permutation**
- **“distance aware” permutation**

Fit exponential model to data and use residuals in analysis

Integrating Neuroimaging Phenotypes and Post-Mortem Human Gene Expression Data



1. Gene Expression Quantification Techniques and Available Databases

2. AHBA Microarray Post-Processing

3. RNAseq Post-Processing

4. Integrating Gene Expression and Imaging Data
   - Assigning samples to brain regions
   - Removing donor-specific effects (individual variability)
   - Addressing spatially correlated gene expression

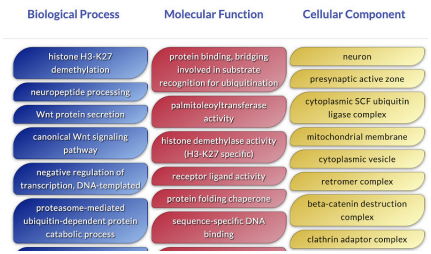
5. **Gene Enrichment and Cell Type Analyses**

6. Cool Papers

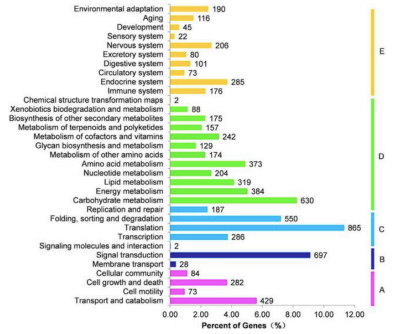
Characterizing Differentially Expressed Genes

Gene Ontology (GO) Terms

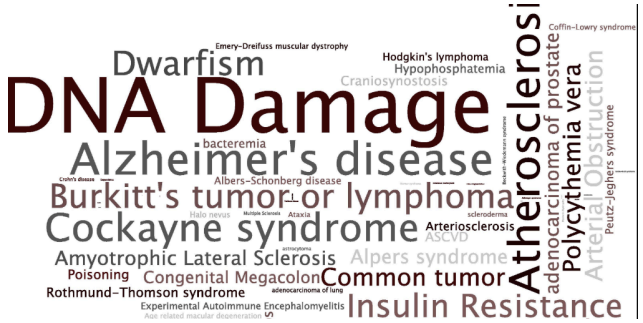
Molecular Function
Cellular Component
Biological Process



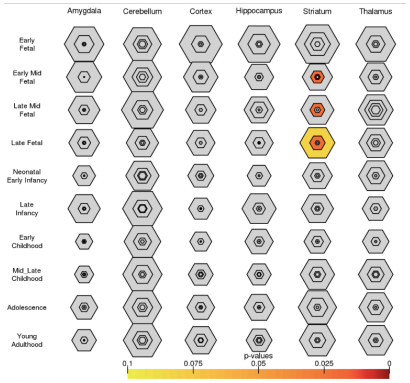
Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathways



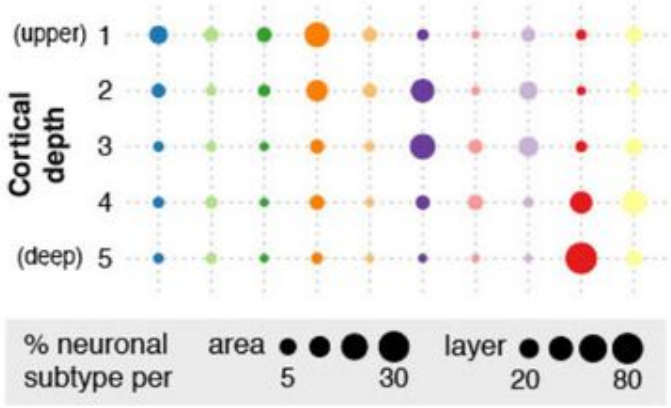
Disease Ontology Semantic and Enrichment (DOSE)



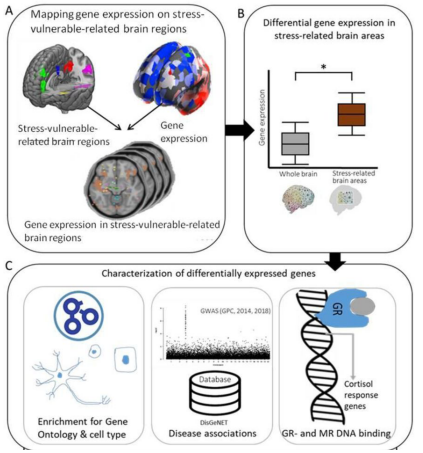
Specific Expression Analysis across Development (BrainSpan)



Layer Enrichment Analysis

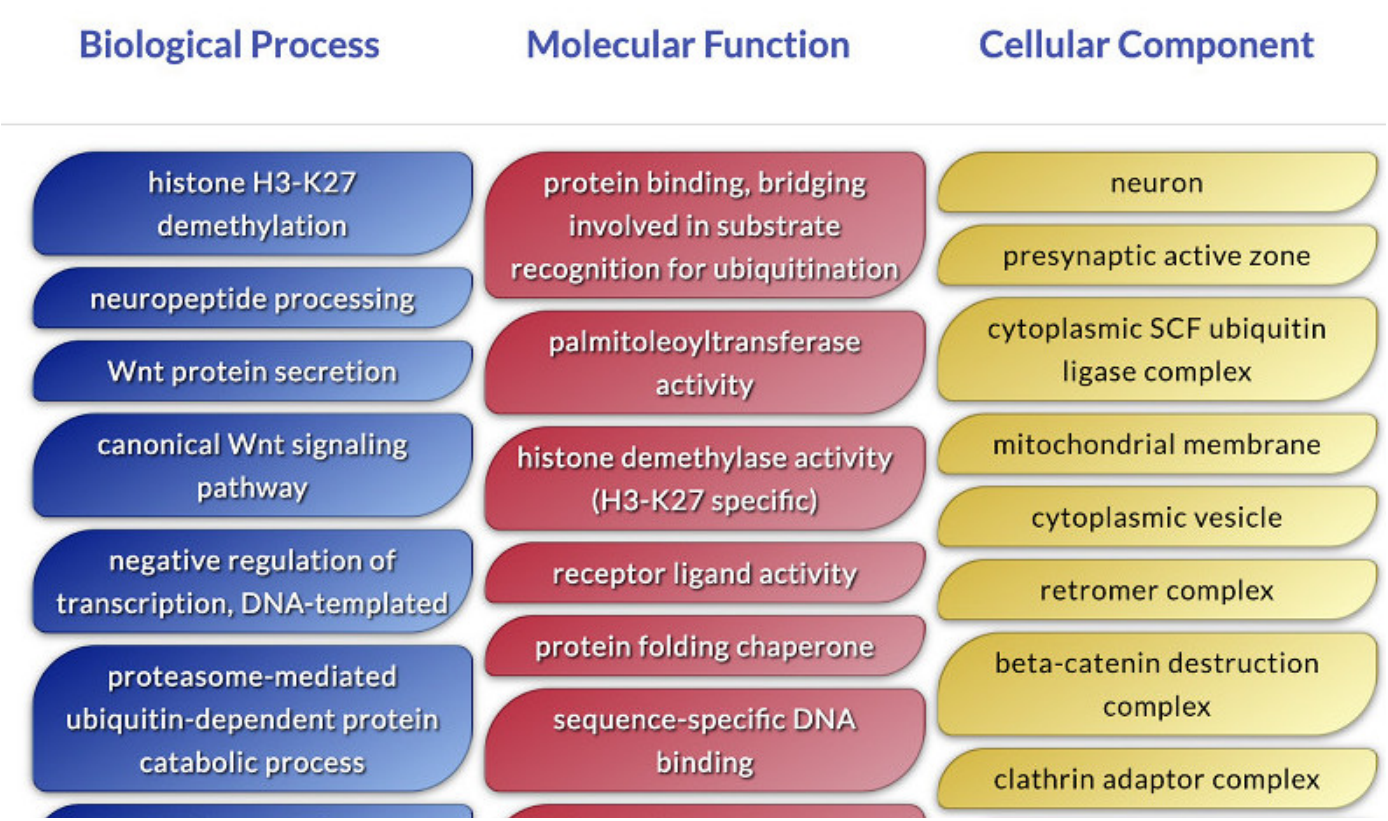


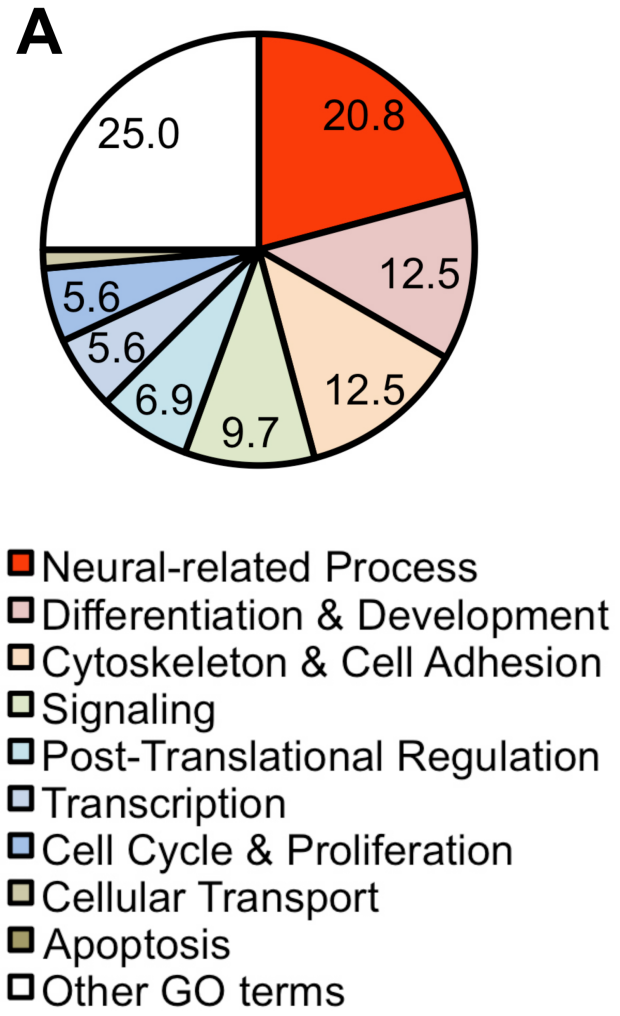
GWAS Integration (MAGMA, DisGeNET)



Characterizing Differentially Expressed Genes

Gene Ontology (GO) Terms
Molecular Function
Cellular Component
Biological Process

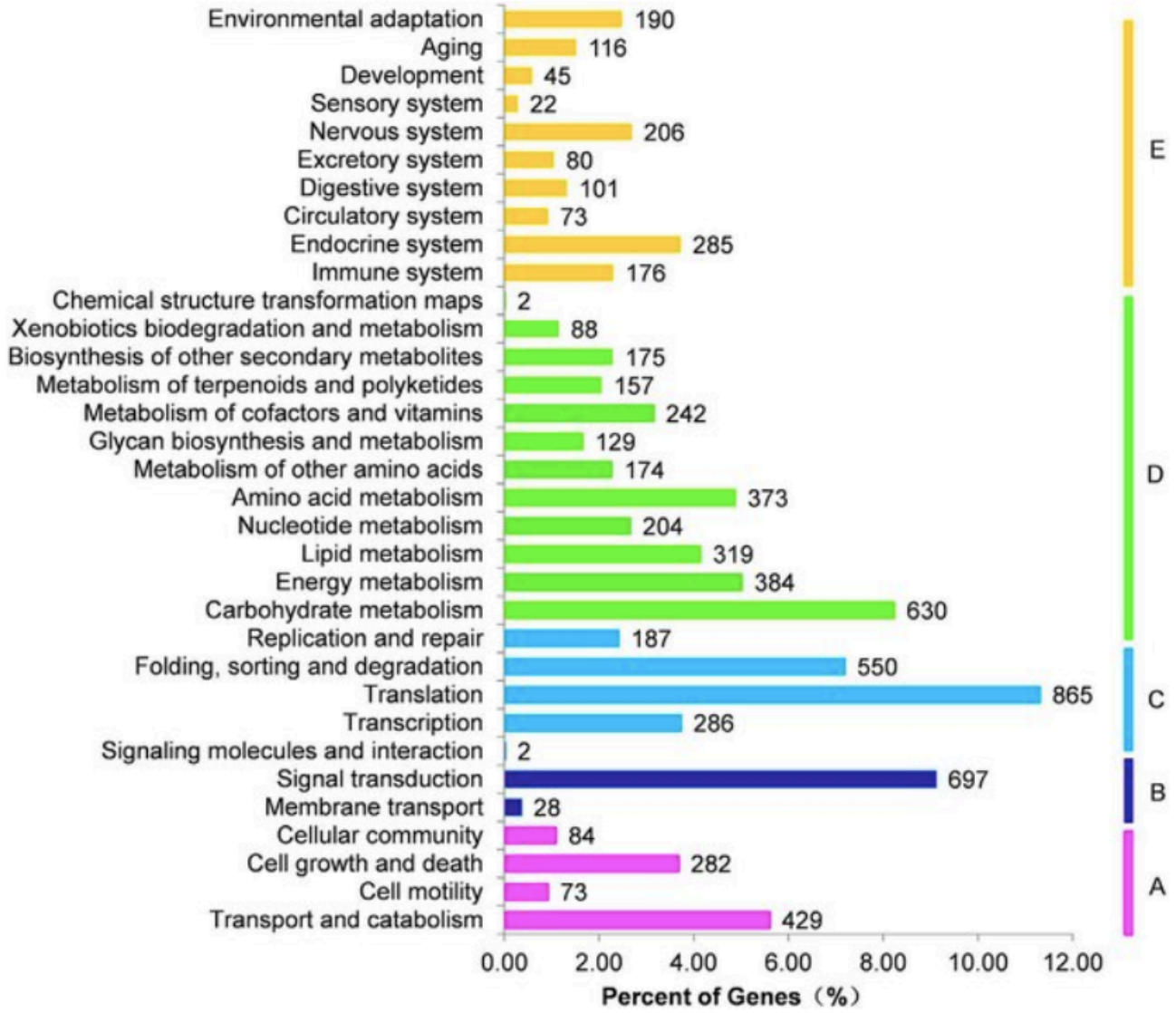




Characterizing Differentially Expressed Genes

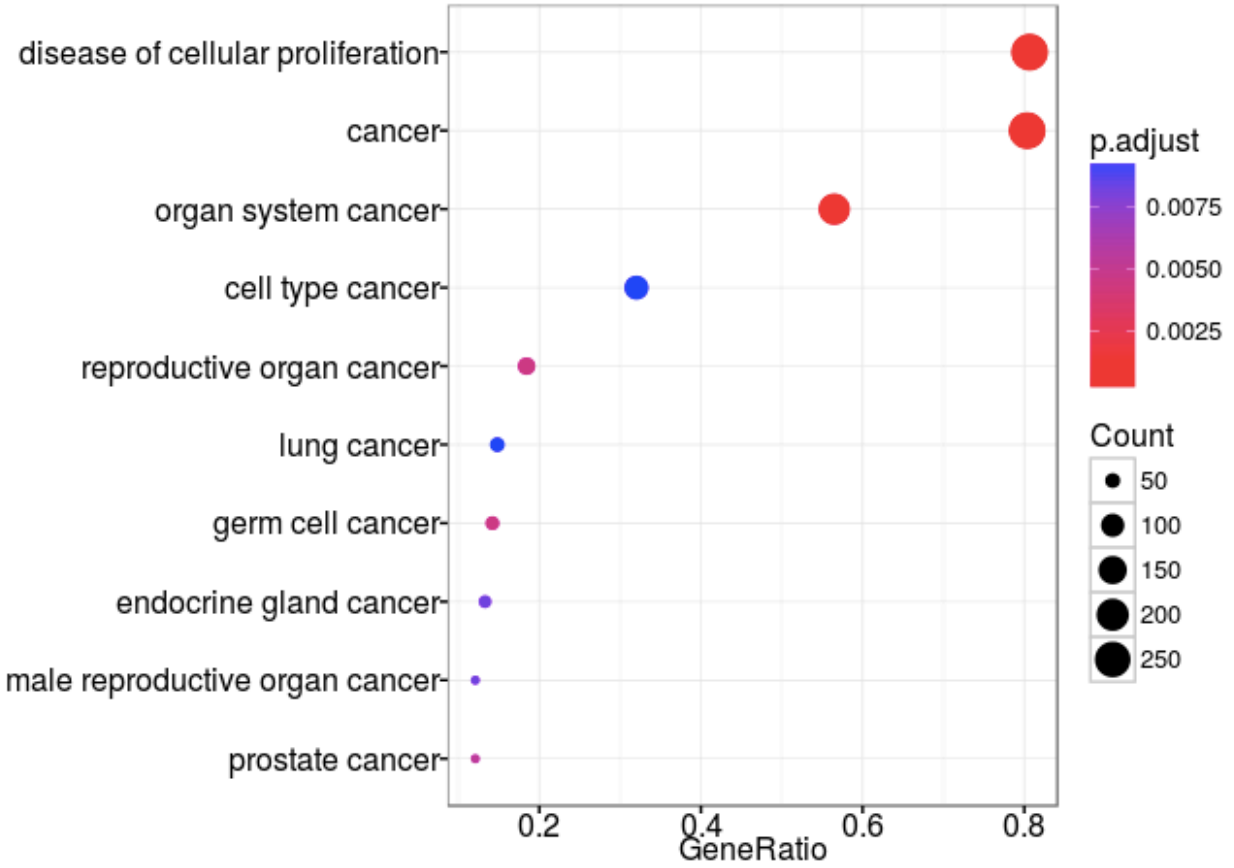
Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathways

Metabolism
Genetic Information Processing
Environmental Processing
Cellular Processes
Organismal Systems
Human Diseases
Drug Development



Characterizing Differentially Expressed Genes

Disease Ontology Semantic and Enrichment (DOSE)

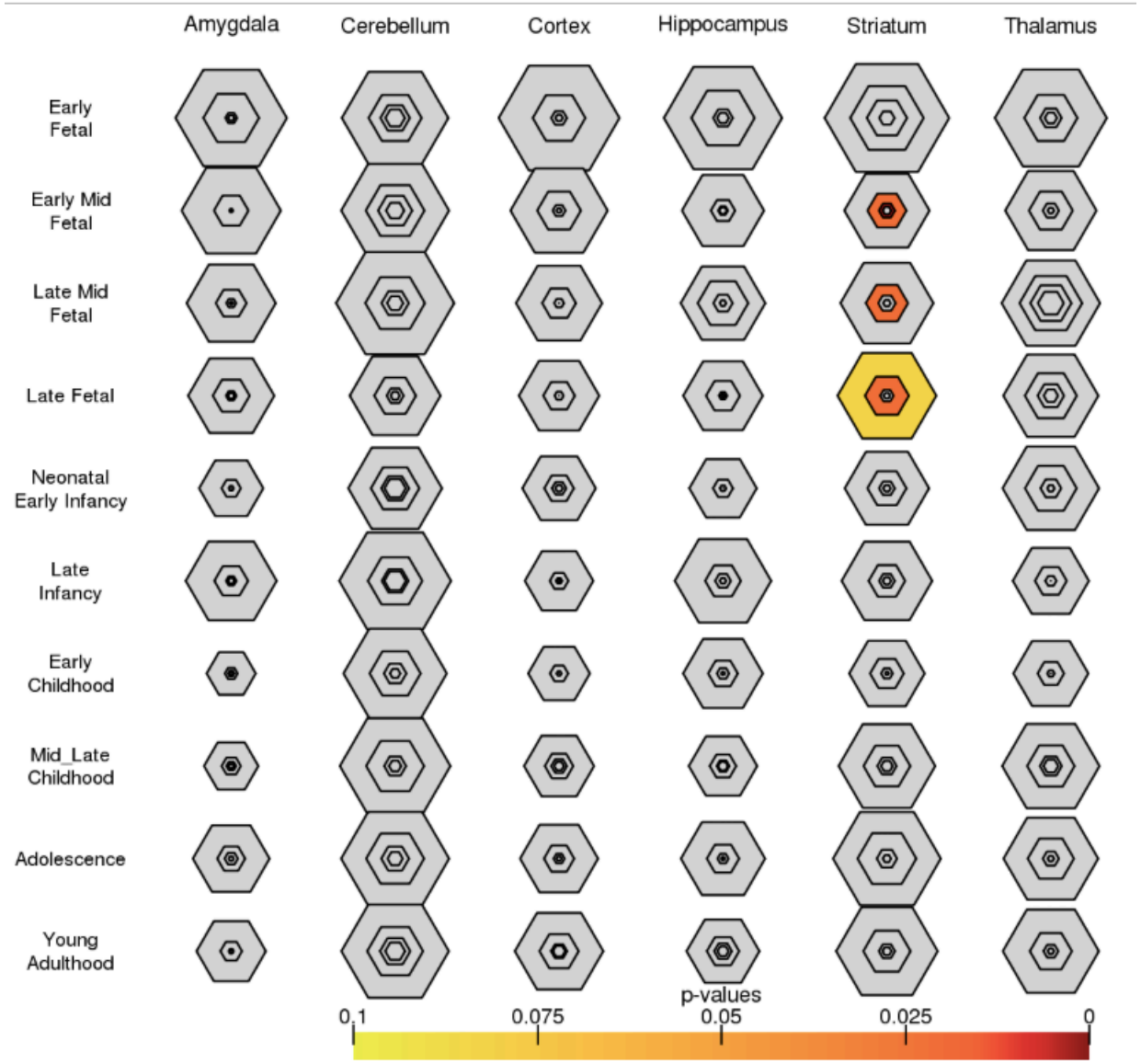


Characterizing Differentially Expressed Genes

Specific Expression Analysis across Development (BrainSpan)

Tool:
http://genetics.wustl.edu/jdlab/csea-tool-2/

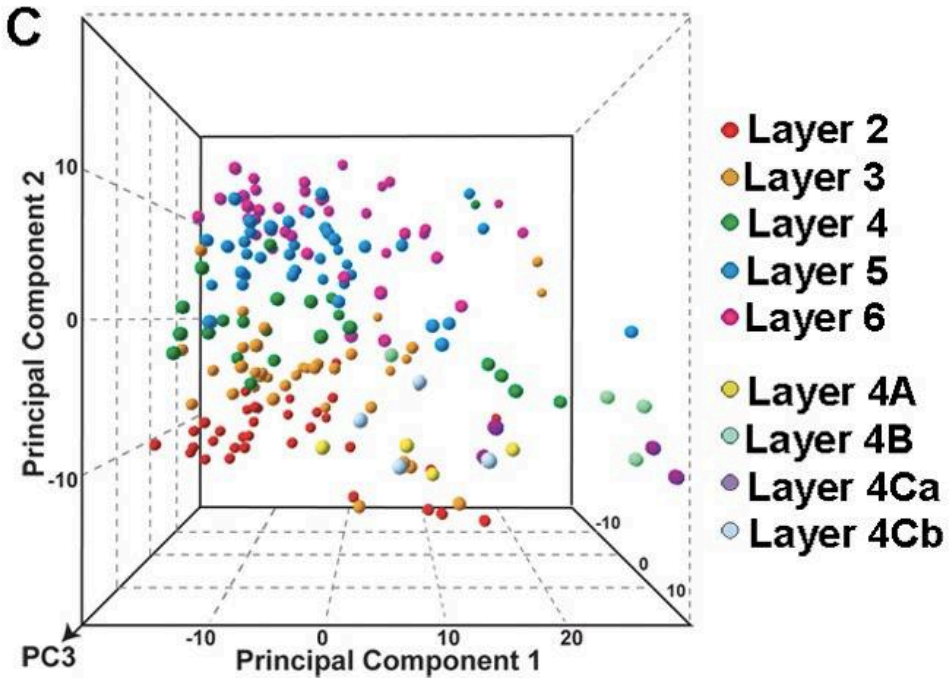
Paper:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3898298/



Characterizing Differentially Expressed Genes

Layer Enrichment Analysis

“The most striking features were the robust molecular signatures associated with different cortical layers…Gene set analysis suggests these layer-associated clusters are associated with neuronal function, including neuronal activity, LTP/LTD, calcium, glutamate and GABA signaling”

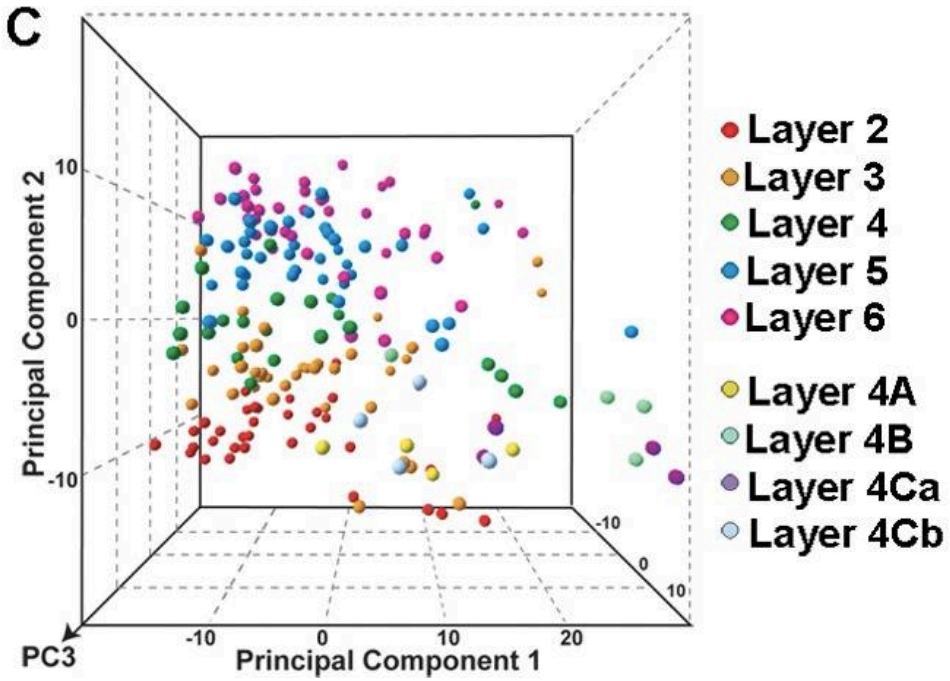


Bernard et al., 2013, Neuron

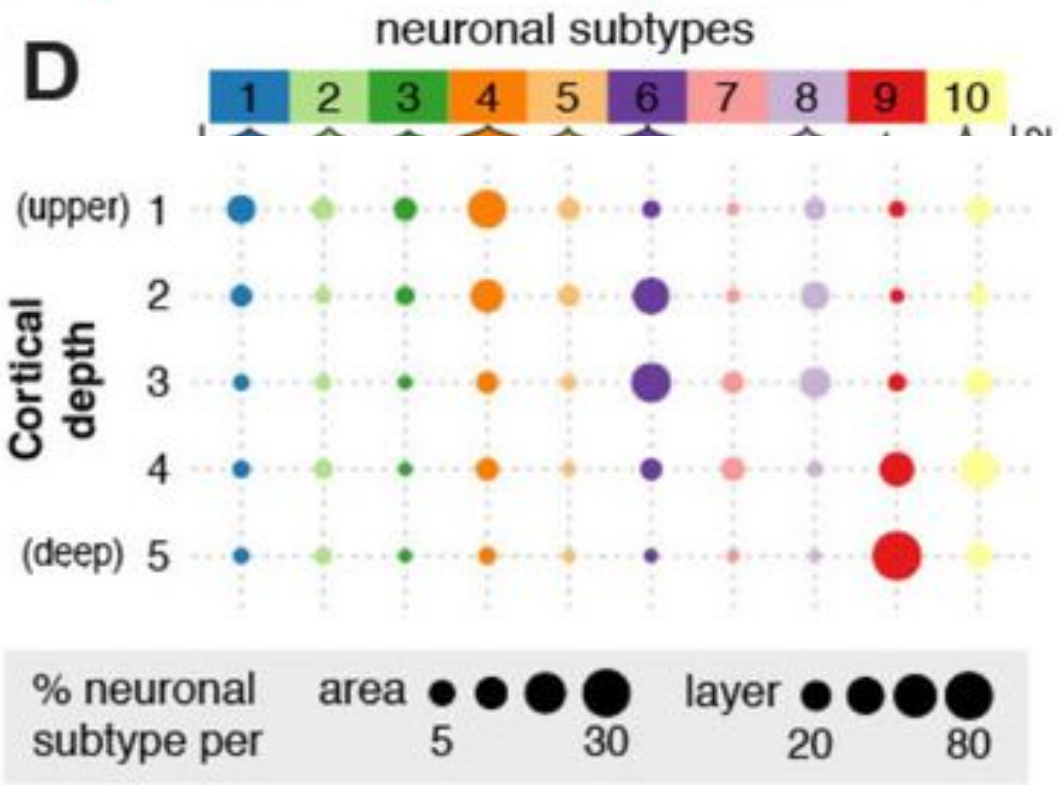
Characterizing Differentially Expressed Genes

Layer Enrichment Analysis

“The most striking features were the robust molecular signatures associated with different cortical layers…Gene set analysis suggests these layer-associated clusters are associated with neuronal function, including neuronal activity, LTP/LTD, calcium, glutamate and GABA signaling”



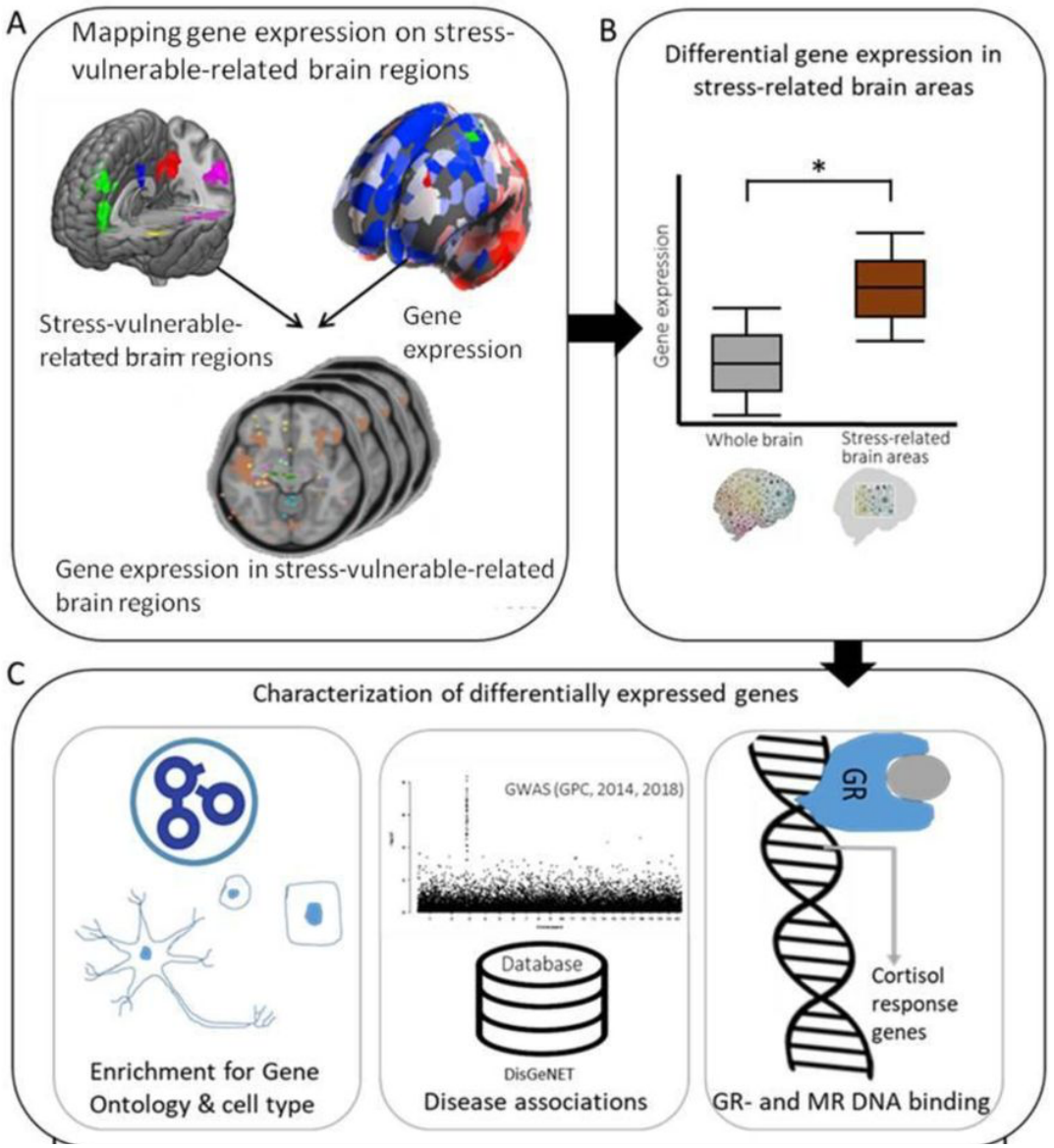
Bernard et al., 2013, Neuron



Bayraktar et al., 2018, BioRxiv

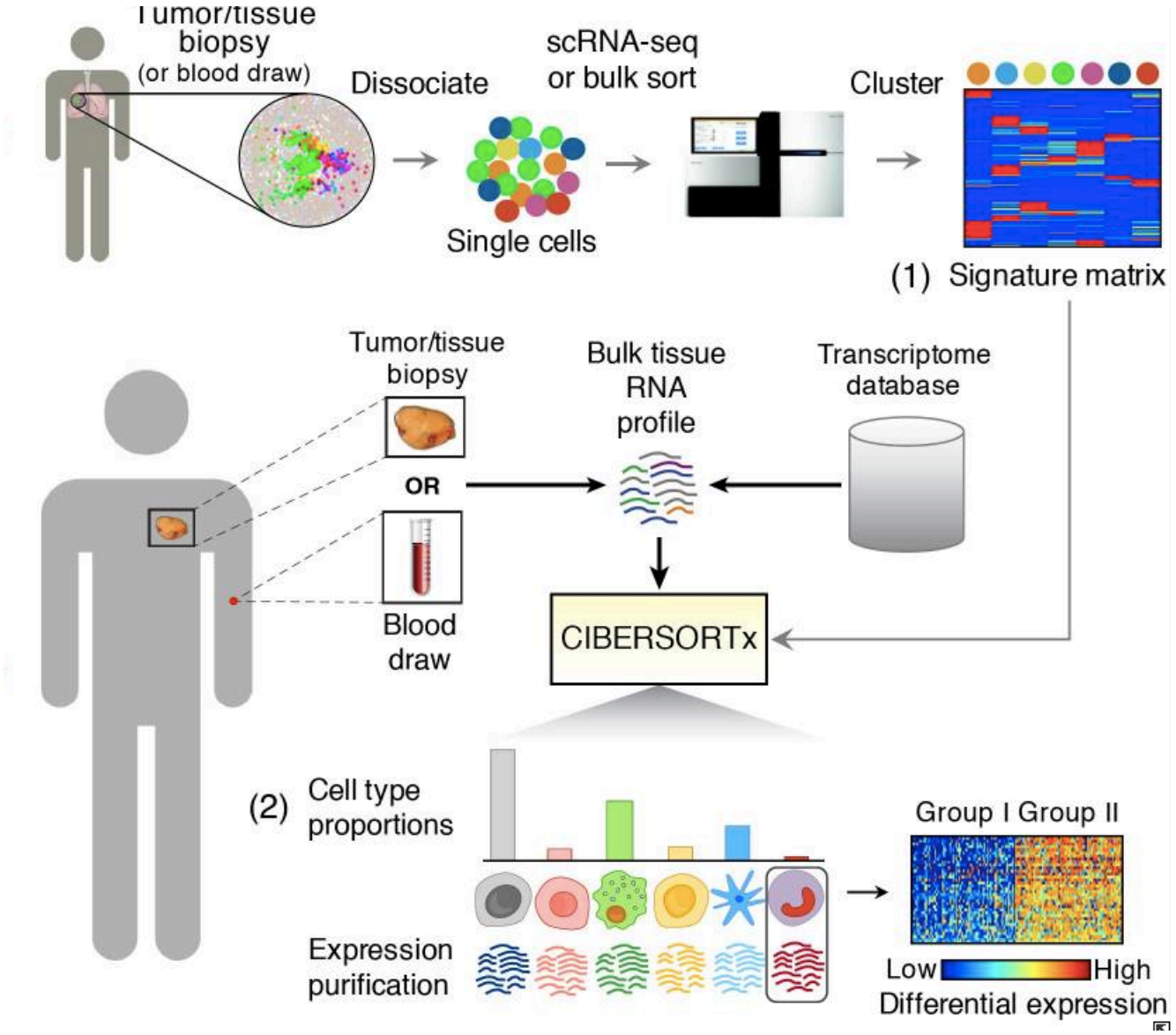
Characterizing Differentially Expressed Genes

GWAS Integration (MAGMA, DisGeNET)



Meijer et al., 2019, BioRxiv

Cell-Type Deconvolution



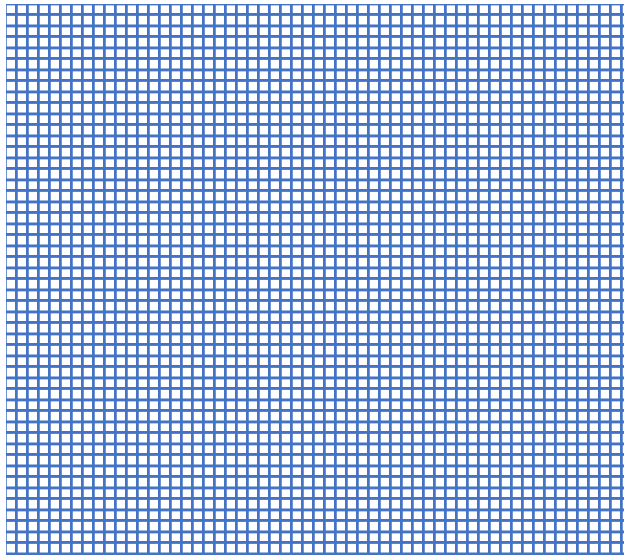
Newman et al., 2019, Nature Biotechnology

Cell-Type Deconvolution

AHBA/GTEX/BRAINSPAN

K tissue samples

n genes



=

n cell-type specific genes

SIGNATURE MATRIX
(single cell RNAseq)

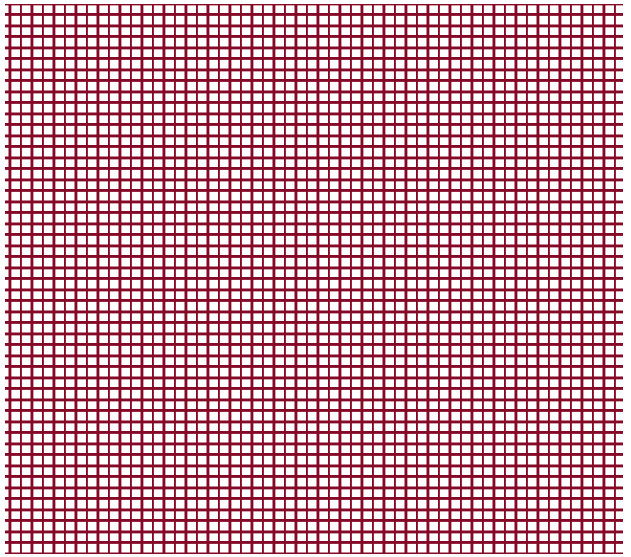
C cell types

X

C cell types

CELL TYPE PROPORTION MATRIX

K fractional abundance (proportion)



Newman et al., 2019, Nature Biotechnology