Assessing the causal association of mtDNAcn with Alzheimer's disease

Dr. Shea Andrews

15 September, 2020

Contents

Abstract

Increasing evidence has implicated mitochondrial dysfunction in Alzheimer's Disease (AD). As AD features altered mitochondrial function, this suggests that therapeutics strategies aimed at preventing declines in mitochondrial function may modify the disease course in AD. However, it is unclear whether mitochondrial dysfunction causes, mediates, or is a by-product of AD pathogenesis. As mitochondria contain their own DNA outside of the nuclear genome, with every cell having between 100-10,000 copies of mitochondrial DNA, mitochondrial DNA copy number (mtDNA-CN) can be used as a surrogate measure of mitochondrial function. The overall objective of this research program is to evaluate whether mitochondrial dysfunction plays a causal role in AD pathogenesis. Our central hypothesis is that lower mtDNA-CN – indicative of mitochondrial dysfunction – will be associated with increased risk of AD. This study will disentangle the causal role of mitochondrial dysfunction in AD using traditional epidemiological approaches, polygenic risk scoring (PRS) and Mendelian randomization (MR). PRS are a measure of an individual's genetic propensity to a trait and can be used to evaluate the genetic overlap between two traits by testing whether the PRS of one trait predicts another trait, while MR uses genetic variants to estimate the causal effect of risk factors on disease outcomes. In the first aim, we will calculate in mtDNA-CN in AD cases and controls and evaluate the association between mtDNA-CN and AD. In the second aim, we will construct a PRS for mtDNA-CN and determine if genetically predicted mtDNA-CN is associated with AD outcomes. In the final aim, we will use MR to evaluate the causal effect of mtDNA-CN on AD outcomes and the causal effect of AD on mtDNA-CN. By establishing if mitochondrial dysfunction has a causal role in AD pathogenesis, this study will provide evidence regarding the utility of mitochondrial therapeutic strategies in AD.

Chapter 1

Introduction

1.1 Project Updates

Updates to project - just like a wet lab Lab Book

1.1.1 To Do

1.1.2 2020-09-09

- xCell: Got xCell working in R using MSBB bulk gene expression data.
  - Was using normalized (library size) gene expression data and adjusted for covariates. Need to use raw counts and adjust for gene length

1.1.3 2020-09-04

- Initial analysis of haplogroup assocations with ROSMAP pathology data. Some initial positive results between haplogroup K and tau pathology.

1.1.4 2020-09-02

- Ordinal Logistic Regression
  - Looking at running OLR for neuropathology outcomes in ROSMAP in particular. Cannot (well you technicaly can) use a linear regression as the assumption of LM will be violated, in particualr the constant marginal effect where the distances between successive points in the dependent variable are assumed to be identical - clearly not the case with a ordinal variable. Norris, C., et al. 2006
  - Proportional odds assumption: relationship between each pair of outcome groups is the same, such that the coefficients that describe the relationship between, say, the lowest versus all higher categories of

the response variable are the same as those that describe the relationship between the next lowest category and all higher categories, etc.
    * Can check assumption using a LRT or a Wald test (Yee pg 397), also graphically (Harrel; UCLA)
    * Assumption is frequently violated, however, the model can still be powerful and usefull (Harrell pg 313); the practical implications of violating this assumption are minimal. Assumption is sensetive to other misspeicifcations.
    * Can try using a different link function or introducing interactions; or partial proportional odds
- References
  - Abreu, M., Siqueira, A., Caiaffa, W. (2009). Regressão logística ordinal em estudos epidemiológicos. Revista de Saúde Pública 43(1), 183-194
  - Yee, T. (2015). Vector Generalized Linear and Additive Models, With an Implementation in R. Springer
  - Harrell, F. (2015). Regression Modeling Strategies, With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. Springer
  - UCLA statiscal consulting: https://stats.idre.ucla.edu/r/dae/ordinal-logistic-regression/
- R Packages
  - VGAM: Conducting OLR using `vlgm`
  - MASS: Conducting OLR using `polyr`
  - Brant: Checks proportional odds assumption when using `polyr` (also aproximate wald test)
  - QMSS: Checks proportional odds assumption when using `vlgm` (LR test)
- Sample size for logistic regression
  - Rule of thumb is a minimum sample size for logistic regression of ~10 events per predictor paramater Riley, R. et al. 2019
  - However, this is context specific, with the appropiate sample size depending not only on the number of events relative to the number of candidate predictor parameters but also on 1) the total number of participants, 2) the outcome proportion (incidence) in the study population and; 3) the expected predictive performance of the model

1.1.5 2020-08-22

- Initial visulizations of blood cell counts vs mtDNAcn do not show a strong relationship that would indicate a reason for the observed bimodal distribution in the samples with DNA isolated from blood.
- Maybe due to a batch effect? May be observable if we had date of visit to see if mtDNAcn distribution changes after a given date.

1.1.6 2020-08-22

- Succesfully used bc.bio to call conduct joint calling of AMP-AD + DIAN samples. At a macro haplogroup level, there is little difference in haplogroup assignments. For haplogroups, ~500 samples have a different haplogroup assignment.
- Obtained additional ROSMAP datasets including information on blood cell counts. Plan on evaluting assocation of mtDNAcn with platlets.
- mtDNAcn estimates from whole blood DNA can be confounded by cell type heterogeneity, and by the presence of platelets, as platelets do not have nDNA, but have mtDNA, which artificially inflates mtDNAcn. Han, L. et al 2019
- cell counts for samples with available RNA-sequencing data can be deconvoluted using gene expression measured in whole blood. Yang, S. et al 2020
  - Aran, D., Hu, Z., Butte, A. (2017). xCell: digitally portraying the tissue cellular heterogeneity landscape. Genome Biology 18(1), 220

1.1.7 2020-08-07

- trying to bcbio working
  - got it working on test set of five samples
  - used sed to rename chr lables in DIAN
- Not working on full sample list…

1.1.8 2020-08-06

- SampleIDs for DIAN
  - changed from WGS ID (file name) to sample ID
  - causes issues with fastmitocalc, which uses file name as prefix in output
- bcbio.variation.recall
  - Does not work when samples are mapped to differen builds. while MT positions are the same, chrm label is different which means the samples on other builds are not recalled.
- installed crossmap using conda install rather than pip
  - crossmap does not work due to chain files using chrM as label, does not lift over correctly
  - solution - use sed to rename all instances of chrM (or MT) to MT. Worked on example file from DIAN. allowed bcbio to work

1.1.9 2020-08-05

- Looking at conducting joint calling using bcbio
  - Issues with DIAN files, in particular it uses hg38, while AMP-AD uses 37

        * Not sure if this will be an issue, may just be able to use the same reference fasta
        * Tried using crossmap to liftover bam files, however mtDNA was unmapped, renames mtDNA as "M" rathern than "MT", also was rather dodgy to get working on minerva, ended up testing it locally
    - DIAN sampleIDs are different from file names, which must match read groups in bam files for bcbio
- Redoing sampleIDs for DIAN
    - encounted issue with broken symlinks - took a while to diagnoses and fix
    - pipeline is working again

1.2 Neuropathological Confirmed AD

There is consensus to disentangle the clinicopathologic term "Alzheimer's disease" from AD neuropathologic change. The former refers to clinical signs and symptoms of cognitive and behavioral changes that are typical for patients who have substantial AD neuropathologic change, and is the focus of recent NIA–AA-sponsored consensus reports on three defined stages in a clinical continuum that includes preclinical, mild cognitive impairment, anddementia. The latter refers to the presence and extent of neuropathologic changes of AD observed at autopsy, re-gardless of the clinical setting.

1.2.1 CERAD Criteria - 1991

Protocol provides neuropathologic definitions of such terms as "definite Alzheimer's disease" (AD), "probable AD," "possible AD," and "normal brain" to indicate levels of diagnostic certainty (**?**). The CERAD Neuritic Plaque score forms the basis of later neuropathological difinitions.

Sections are tacken from:

- middle frontal gyrus
- superior and middle temporal gyri
- inferior parietal lobule
- hippocampus and entorhinal cortex
- midbrain

And scored as a semiquantitative measurment:

- Absent
- Sparese
- Moderate
- Frequent

An age-related plaque score is then determined by combining the age of the patient at death and the semiquantitative measure of plaques in the *most severely*

affected region of the neocortex. This score is then intergrated with with clinical information the presence or absence of dementia.

1.2.2 NIA-Reagan Criteria - 1997

The modified NIA-Reagan diagnosis of Alzheimer's disease is based on consensus recommendations for postmortem diagnosis of Alzheimer's disease. The criteria rely on both neurofibrillary tangles (Braak) and neuritic plaques (CERAD). See NIA Working group consensus 1997 and corresponding editorial by Hyman et al 1997. Traditionaly, the criteria require a history of dementia, insofar as they were designed to help address the question of whether AD was the underlying cause of a patient's dementia.

- CERAD score is a semiquantitative measure of neuritic plaques
  - No neuritic plaques (C0)
  - Sparse/infrequent neuritic plaques (C1)
  - Moderate neuritic plaques (C2)
  - Frequent neuritic plaques (C3)
- Braak Stage is a semiquantitative measure of severity of neurofibrillary tangle (NFT) pathology.
  - no NFTs (B0)
  - stages I/II, with NFTs predominantly in en-torhinal cortex and closely related areas (B1)
  - stages III/IV, withNFTs more abundant in hippocampus and amygdala whileextending slightly into association cortex (B2)
  - stages V/VI,with NFTs widely distributed throughout the neocortex (B3)

CERAD / Braak	0	I/II	III/IV	V/VI
None	**Normal**	-	-	-
Sparse	-	**Low**	-	-
Moderate	-	-	**Intermediate**	-
Frequent	-	-	-	**High**

1.2.3 NIA-AA Criteria - 2012

The NIA-AA criteria updated and revised the 1997 NIA-Reagan criteria to recognize the pre-clinical stage of AD, enhance the assessment of AD to include amyloid accumulation as well as neurofibrillary change and neuritic plaques. Hyman et al 2012. The criteria relies on an 'ABC' score for AD neuropathologic change that incorporates histopathologic assessments of amyloid deposits (A - Thal phase), staging of neurofibrillary tangles (B - CERAD), and scoring of neuritic plaques (C - Braak Stage). See Hyman et al 2012 for guidlines and Montine et al 2012 for a practial guide.

- Thal Phase is a semiquantitiative measure of the distribution of AB

- phase 0 or no amyloid
- phase 1 or isocortical
- phase 2 or limbic
- phase 3 or basal ganglia
- phase 4 or basal forebrain and midbrain
- phase 5 or pons/medulla oblongata and cerebellum

Thal	CERAD	Braak: None or I/II (B0 or B1)	III/IV (B2)	V/VI (B3)	
0 (A0)	None (C0)	Other§	Other§	Other§	
1/2 (A1)	None - Sparse (C0 or C1)	Low	Low	Low¶	
	Modearte - Frequent C2 or C3)	Low†	Intermediate	Interm	
3 (A2)	Any C	Low†	Intermediate	Interm	
4/5 (A3)	None - Sparse (C0 or C1)	Low†	Intermediate	Interm	
	Modearte - Frequent C2 or C3)	Low†	Intermediate	High	

§Medial temporal lobe NFTs in the absence of significant Ab or neuritic plaques occur in older people and may be seen in individuals without cognitiveimpairment, with mild impairment, or with cognitive impairment from causes other than AD. Consider other diseases when clinically or pathologically indicated.

¶Widespread NFTs with some Ab/amyloid plaques or limited neuritic plaques are relatively infrequent, and when they occur, other diseases, particularly-tauopathies, should be considered. Such cases may not fit easily into a specific Braak stage, which is intended for categorization of AD-type NFTs.

†Higher levels of Ab or neuritic plaques with low Braak stage should prompt consideration of contribution by comorbidities such as vascular brain injury,LBD, or HS. Also, consider additional sections as well as repeat or additional protocols to demonstrate other non-AD lesions

For individuals **without cognitive impairmentat** the time tissue was obtained, it is possible that AD neuropathologic change may predate onset ofsymptoms by years. For individuals **with cognitive impairmentat** the time tissue was obtained, “Intermediate” or “High” level (Table 2) of AD neuropathologic change should be considered adequate explanation of cognitive impairment or dementia. When “Low” level of AD neuropathologic change is observed in the setting of cognitive impairment, it is likely that other diseases are present. In all cases with cognitive impairment, regardless of the extent of AD neuropathologicchange, it is essential to determine the presence or absence, as well as extent, of other disease(s) that might have contributed to the clinical deficits.

Possibility that Thal amyloid stages do not substantially contribute to predicting antemortem cognition compared to CERAD neuritic plaque scores and Braak NFT stages Serrano-Pozo et al 2016.

Chapter 2

Methods

This section describes the general methods used for calling mitochondrial haplogroups, estimating mtDNAcn and the cohorts used in the analysis.

2.1 Mitochondrial DNA

2.1.1 ADNI

Ridge, P., et al. (2018). Assembly of 809 whole mitochondrial genomes with clinical, imaging, and fluid biomarker phenotyping Alzheimer's & Dementia 14(4), 514-519

Called mitochondrial SNVs in ADNI using freebayes and assigned haplogroups using Phy-Mer

- ADNI was mapped to Hg19 which uses of the mitochondrial genome, represented as chrM, corresponding to NC_001807
- Since chrM and NC_012920 only differ by a few bases, we were able to extract only those reads that mapped to chrM (with SAMtools [46]), rather than all reads corresponding to the whole nuclear and mitochondrial genomes.
- Extracted reads were remapped to NC_012920 using BurrowsWheeler Aligner.
- performed local realignments around indels and base recalibration with Genome Analysis Toolkit to refine the new mappings.
- Used FreeBayes to joint-call variants
  - ploidy 1
  - min-alternate-fraction 0.6
  - removed variants with quality less than 20
- converted the resulting variant call format (VCF) file to fasta with vcf2fasta

- annotated mitochondrial haplotypes with PhyMer from fasta

2.1.2 AMP-AD

For futher details on alignement see: ROSMAP, Mayo and MSBB WGS germline analysis

- Whole Genome data are processed on NYGC automated pipeline.
- Paired-end 150bp reads were aligned to the GRCh37 human reference using the Burrows-Wheeler Aligner (BWA-MEM v0.7.08)
- processed using the GATK best-practices workflow that includes marking of duplicate reads by the use of Picard tools v1.83, local realignment around indels, and base quality score recalibration (BQSR) via Genome Analysis Toolkit (GATK v3.4.0).

2.1.3 DIAN

Unknown...

2.1.4 Calling Mitochondrial Variants in AMP-AD & DIAN

Mitochondrial Variants were called per sample using Freebayes

- Freebayes 1.3.2 and removed variants with a quality score less then 20
  - rCRS used as reference genome (either hg19 or hg38)
  - min-mapping-quality 30
  - min-base-quality 24
  - min-alternate-fraction 0.6
  - min-alternate-count 4
  - ploidy 1

bcbio variation recall was then used to squaring off multiple samples, called independently, by recalling at all identified genomic positions.

Alternative approaches

- gatk Mutect2: mitochondrial calling pipline based on Mutect2 from gatk. Very little documentation
- mity: Pipeline for calling mitochhondrial SNVs and INDELs, based on freebayes.
- Joint calling with freebayes: Run into memory issues trying to call all samples jointly. Possibly call in batchs (chort). Need to work out genome references between chorts (i.e. MT (hg19) vs chrM (hg38))

2.2 Haplogroup Assignment

2.2.1 Haplogrep

Weissensteiner, H. et al. (2016). HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. Nucleic acids research 44(W1), W58-63

- assigns haplogroups based on phylotree and uses a generic rule-based system for immediate quality control
- vcf input

Alternative approaches

- **Phy-Mer**: novel mitochondrial genome haplogroup-defining algorithm using a k-mer approach by decomposes a mitochondrial sequence into a set of all possible k-mers, which are then compared against each of the k-mer sets of all haplogroups. Uses NGS data (.bam, .cram).
- However, Resulted in low quality scores for AMP-AD, but not for DIAN
- Navarro-Gomez, D et al (2014). Bioinformatics (Oxford, England) 31(8), 1310-2.

2.3 Estimating mtDNAcn

Mitochondrial DNA Copy Number estimation

- mtDNA-CN can be estimated as the ratio of the average mitochodnrial DNA coverage by the average autosomal DNA coverage
  - mtDNA-CN = (mtDNA average coverage / autosomal DNA average coverage) * 2

2.3.1 Mosedepth

Pedersen, B., Quinlan, A. (2017). **Mosdepth: quick coverage calculation for genomes and exomes** Bioinformatics 34(5), 867-868.

- Mosdepth uses a simple algorithm that is computationally efficient enableing it to quickly calculating genome-wide sequencing coverage. Not specifically designed for estimating mtDNA-CN, but provides coverage estimates of the autosome and mitochondrial genome.

Alternative approaches

- **fastMitoCalc**: uses a randomly selected small subset (0.1%) of the nuclear genome to estimate autosomal DNA coverage accurately for estimation of the mtDNA-CN.
  - However, a ceiling effect was observed in samples with DNA isolated from brain tissue.
  - Qian, Y., et al. (2017). Bioinformatics 33(9), 1399-1401.

2.4 Cell Type Enrichment

- mtDNAcn estimates from whole blood DNA can be confounded by cell type heterogeneity, and by the presence of platelets, as platelets do not have nDNA, but have mtDNA, which artificially inflates mtDNAcn. Han, L. et al 2019
- Estimates from brain tissue with clinical disease are also potentially confounded due to cell loss
- cell counts for samples with available RNA-sequencing data can be deconvoluted using gene expression measured in whole blood. Yang, S. et al 2020

2.4.1 xCell

Aran, D., Hu, Z., Butte, A. (2017). **xCell: digitally portraying the tissue cellular heterogeneity landscape**. Genome Biology 18(1), 220

- Git Repo
- Webtool
- Uses an adapation of ssGSEA to calculated enrichment scores for 64 cell types to identify particular pathways or gene sets that are differentialy expressed in tissue and represent distinct cell types.
  - Cell types span multiple adaptive and innate immunity cell, hematopoietic progenitors, epithelial cells, extracellular matrix cells. Includes neurons and astrocytes.
  - Uses 489 (three for each cell type, from each data source) signatures learned from six sources to estimate enrichment of cell types
  - Raw scores are the average single-sample GSEA of all signatures corresponding to a cell type, which are transformed to linear scores, allowing for comparison of scores across cell types and across samples
  - Spillover compenstation correction to account for corrleated scores between closely related cell types
- Calculating scores for a mixture
  - Input is a n x m matrix with, rows corresponding to genes (gene symbols) and columns samples
  - Recomended to use data sets containing the majority of the 10,808 genes used by xCell for scoring
  - Missing values are treated as missing genes
  - Use as many samples as possible, with highly expected variation in cell type fractions
  - xCell uses the expression levels ranking and not the actual values, thus normalization does not have an effect, however normalizing to gene length (RPKM/FPKM/TPM/RSEM) is required
    - Gene Expression Units explained
    - Counts to TPM
    - Misuse of RPKM or TPM normalization. Zhao et al 2020. RNA

        * RNA-seq Analysis.
        * RNA-seq resources
    - Produces enrichment scores, not percentages, which means that the main usage is for comparing across samples, not across cell types
- Has several advantages over deconvolution approaches
    - Suitiable for cross-platform transcriptomic measurement
    - Agnostic to normalization methods, batch effects
    - No decline in performance with increase cell types
    - Simple and easy to adjust

Alternative approaches

- **CIBERSORTx**: uses a deconvulution approach to estimate cell-type frequences from bulk gene expression data
    - Requires a signature matraix, possibly from a reference scRNA-seq dataset. May have one to use.
    - Newman, A., et al (2019). Nature Biotechnology 37(7), 773-782
    - Webtool
- Further reading
    - Liu, C., et al (2019). Computational approaches for characterizing the tumor immune microenvironment. Immunology 158(2), 70-84
    - Bortolomeazzi, M., et al (2019). Identification of non-cancer cells from cancer transcriptomic data. Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms 1863(6), 194445

2.5 Cohorts

Accelerating Medicine Partnership in Alzheimer's Disease (AMP-AD)

Whole genome sequencing data was obtained from three cohorts using AMP-AD knowledge portal.

2.5.1 ROSMAP

- Study Detailes
    - Whole Genome Sequencing
    - Bulk Brain RNA seq
        * "un-normalized" files indicate the pre-normalization expression data, which is FPKM calls from RSEM
    - Bulk Blood RNA seq

2.5.2 Mayo

2.5.3 MSBB

- Study Detailes

  - Whole Genome Sequencing
  - RNA seq
    * The gene level read counts data were normalized as counts per million (CPM) using the trimmed mean of M-values normalization (TMM) method to adjust for sequencing library size difference.

2.5.4 AMP-AD Cross-Study RNAseq Harmonization

AMP-AD consortium efforts to harmonize RNAseq data + Study Details - RNA Seq - Differential Expression Analysis

Chapter 3

ROSMAP

The samples that we have profiled come from two prospective studies of aging-The Religious order Study (ROS) and the Memory and Aging Project (MAP)-that recruit older individuals without known dementia and include (1) detailed cognitive, neuroimaging and other ante-mortem phenotyping and (2) an autopsy at the time of death that includes a structured neuropathologic examination. A subset of the ROSMAP samples (n=1200 for 1179 unique deceased participants) underwent whole genome sequencing, with DNA coming from brain tissue (n=806), whole blood (n=389) or lymphocytes transformed with EBV virus (n=5) (**?**).

Data Dictionaries for ROSMAP can be found at:

- AMP-AD
- RADC

```
## Phenotypic data
rosmap.pheno_bl <- readxl::read_xlsx("data/AMPAD_extra/rosmap/dataset_899_basic_08-22-2020.xlsx")
  mutate(projid = as.numeric(projid))
rosmap.pheno_long.raw <- readxl::read_xlsx("data/AMPAD_extra/rosmap/dataset_899_long_08-22-2020.x
  mutate(projid = as.numeric(projid))

# wgs_clinical.raw <- read_csv('data/AMPAD_extra/rosmap/ROSMAP_Clinical_2019-05_v3.csv')
## Mitochondrial
rosmap.wgsqc <- read_csv("data/AMPAD_extra/rosmap/WGS_sample_QC_info.csv", guess_max = 10000)
mosdepth <- read_tsv("data/mosdepth/mosdepth_mtDNAcn_All.txt")
haplogrep <- read_tsv("data/haplogrep/haplogrep_jointAll.txt")

apoe <- read_tsv("data/AMPAD_extra/rosmap/wgs_apoe.tsv")

xcell <- read_csv("data/xcell/ampad_xCell.csv") %>%
  filter(study == "ROSMAP") %>%
```

```
  select(-SampleID, -study) %>%
  rename(rna_seq_tissue = Tissue, rna_seq_batch = batch) %>%
  mutate(ID = as.numeric(ID))

# Neuroimaging
tot_vol.raw <- bind_rows(
  read_csv("data/AMPAD_extra/rosmap/mri_total_volumes_share_ROS.csv"),
  read_csv("data/AMPAD_extra/rosmap/mri_total_volumes_share_MAP.csv")
) %>%
  select(-X9) %>%
  mutate(projid = as.numeric(projid),
         visit = as.numeric(visit))
subcort_v6 <- read_csv("data/AMPAD_extra/rosmap/mri_subcortical_v6_share.csv")  %>%
  select(-X69)
```

```
## extract baseline and last vist values and spread
rosmap.pheno_long <- rosmap.pheno_long.raw %>%
  group_by(projid) %>%
  slice(c(which.min(fu_year), which.max(fu_year))) %>%
  mutate(visit = case_when(fu_year == min(fu_year) ~ "bl",
                           fu_year == max(fu_year) ~ "lv")) %>%
  distinct(., fu_year, .keep_all = TRUE) %>%
  ungroup() %>%
  pivot_wider(names_from = visit, values_from = c(-projid, -study, -scaled_to))

## Merge WGS files
rosmap_wgs <- rosmap.wgsqc %>%
  filter(QC == "Pass") %>%
  select(projid, SampleID = WGS_id, Source.Tissue.Type, QC) %>%
  left_join(mosdepth, by = "SampleID") %>%
  left_join(haplogrep, by = "SampleID") %>%
  mutate(macro = case_when(
    str_detect(Haplogroup, "^L|^HV|^JT") ~ substr(Haplogroup, start = 1, stop = 2),
    TRUE ~ substr(Haplogroup, start = 1, stop = 1)
  ))

## extract last MRI measurment
tot_vol <- tot_vol.raw %>%
  group_by(projid) %>%
  arrange(visit) %>%
  slice(which.max(visit)) %>%
  ungroup()

### 32 duplicate IDS
# get_dupes(rosmap_wgs, projid)  %>%
```

```
#   print(n = Inf)

rosmap <- rosmap.pheno_bl %>%
  left_join(apoe, by = "projid") %>%
  left_join(rosmap.pheno_long, by = c("projid", "study", "scaled_to")) %>%
  left_join(tot_vol, by = "projid") %>%
  left_join(xcell, by = c("projid" = "ID")) %>%
  left_join(rosmap_wgs, by = "projid") %>%
    mutate(race7 = as_factor(race7),
        race7 = fct_recode(race7,'White' = '1', 'Black' = '2', "AMR" = "3", "Oceania" = "4",
                           "Asian" = "5", "Other" = "6", "Other" = "7"),
        z_mtdnacn = scale(mtcn_avg, center = TRUE, scale = TRUE)[,1],
        spanish = as_factor(spanish),
        spanish = fct_recode(spanish,'Yes' = '1', 'No' = '2'),
        organ = recode(Source.Tissue.Type, 'Blood' = 'blood', 'Blood-PBMC' = 'blood', 'Whole Bloo
                       'Blood-Cerebellum' = 'brain',  'Brain-Anterior Caudate' = 'brain',
                       'Brain-Cerebellum' = 'brain', 'Brain-DLPFC' = 'brain',
                       'Brain-Frontal Cortex (BA unknown)' = 'brain',
                       'Brain-Frontal Pole (BA10-12,32)' = 'brain',
                       'Brain-Occipital Association Cortex (BA18,19)' = 'brain',
                       'Brain-PCC' = 'brain', 'Brain-Posterior Cingulate Cortex' = 'brain',
                       'Brain-region unknown' = 'brain',
                       'lymphocytes _transformed _with EBV virus'= 'lymphocytes'),
        organ = as_factor(organ),
        apoe_genotype = ifelse(is.na(apoe_genotype), apoe, apoe_genotype),
        apoe4 = recode(apoe_genotype, '22' = 'e4-', '23' = 'e4-', '33' = 'e4-', '24' = 'e4+', '34
        tomm40_hap = as_factor(tomm40_hap),
        tomm40_hap = fct_recode(tomm40_hap, 'S/S' = '1', 'S/L' = '2', "S/VL" = "3", "L/L" = "4",
        aod_cat = cut(age_death, c(50, 60, 70, 80, 90, Inf), c('50-59', '60-69', '70-79', '80-89'
        aod_cat = ordered(aod_cat, levels = c('50-59', '60-69', '70-79', '80-89', '90+')),
        msex = as.factor(msex),
        msex = fct_recode(msex, 'M' = '1', 'F' = '0'),
        cogdx = factor(cogdx),
        r_pd_bl = factor(r_pd_bl),
        r_pd_lv = factor(r_pd_lv),
        r_stroke_bl = factor(r_stroke_bl),
        r_stroke_lv = factor(r_stroke_lv),
        dcfdx_bl = factor(dcfdx_bl),
        dcfdx_lv = factor(dcfdx_lv),
        apoe_genotype = as.factor(apoe_genotype),
        apoe4 = as.factor(apoe4),
        study = as.factor(study),
        braaksc = ordered(braaksc, levels = c('0', '1', '2', '3', '4', '5', '6')),
        ceradsc = ordered(ceradsc, levels = c('4', '3', '2', '1')),
        dlbdx = as.factor(dlbdx),
```

```
        ci_num2_mct = as.factor(ci_num2_mct),
        ci_num2_gct = as.factor(ci_num2_gct),
        cvda_4gp2 = as.factor(cvda_4gp2),
        caa_4gp = as.factor(caa_4gp),
        arteriol_scler = as.factor(arteriol_scler),
         hspath_typ = as.factor(hspath_typ),
         tdp_st4 = as.factor(tdp_st4),
        niareagansc = ordered(niareagansc, levels = c('4', '3', '2', '1')),
        CDR = cut(cts_mmse30_lv, breaks = c(-Inf, 11, 21, 26, 30, Inf), labels = c(3, 2
  filter(!is.na(study))

write_rds(rosmap, 'output/rosmap.rds')

df <- rosmap %>%
    select(study, age_bl, msex, educ, apoe_genotype, cogdx, age_first_ad_dx, Source.Tis
```

3.1 Demographics

Demographic variables avaliable in ROSMAP are shown in Table **??**.

Data Summary

variables

definitions

types

missing_percent

unique_count

study

Study

factor

0.00

2

race7

Racial group

factor

0.00

6

spanish

Spanish ethnicity

factor

0.00

2

msex

Sex

factor

0.00

2

educ

Education

numeric

0.03

31

age_bl

Age at baseline

numeric

0.00

3087

age_death

Age at death

numeric

45.56

1750

Descriptive statistics of numerica varibles are presented in Table **??**.

Variable type: Numeric

col_name

min

q1

median

mean

q3

max

sd

pcnt_na

educ

0.00

14.00

16.00

16.32

18.00

30.00

3.77

0.03

age_bl

36.50

72.90

79.10

78.40

84.00

102.15

7.91

0.00

age_at_visit_lv

54.62

81.54

87.05

86.27

91.61

108.14

7.56

7.52

age_death

65.91

84.97

89.67

89.18

93.69

108.28

6.65

45.56

fu_year_bl

0.00

0.00

0.00

0.00

0.00

0.00

0.00

0.03

fu_year_lv

1.00

3.00

7.00

7.97

11.00

26.00

5.71

7.52

Frequency and proportions of categorical varibles are presented in Table **??**.

Variable type: Factor

col_name

level

prop
cnt
msex
F
0.73
2664
msex
M
0.27
1008
race7
White
0.93
3412
race7
Black
0.06
212
race7
Other
0.01
20
race7
AMR
0.00
13
race7
Asian
0.00
13
race7

Oceania

0.00

2

spanish

No

0.95

3487

spanish

Yes

0.05

185

study

MAP

0.60

2191

study

ROS

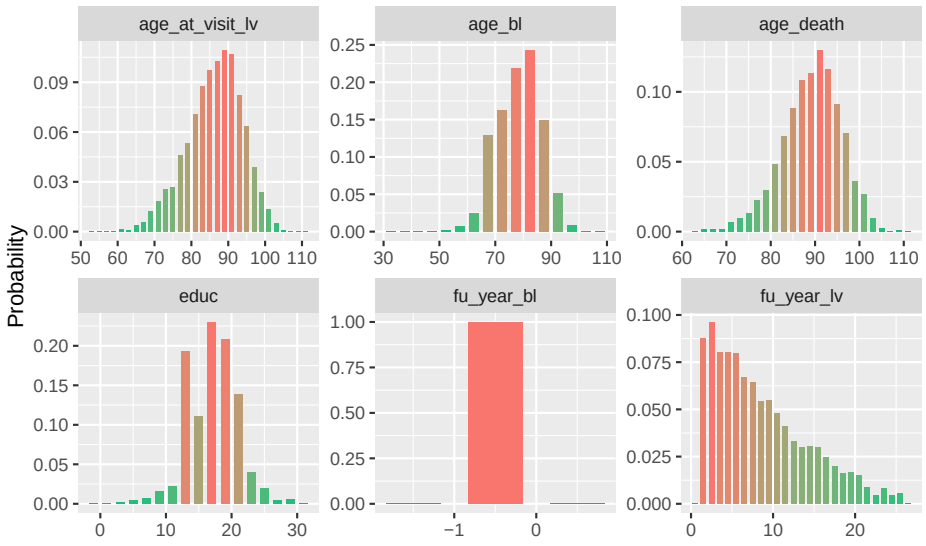
0.40

1481

3.1.1 Plots

```
demo_n %>% show_plot()
```

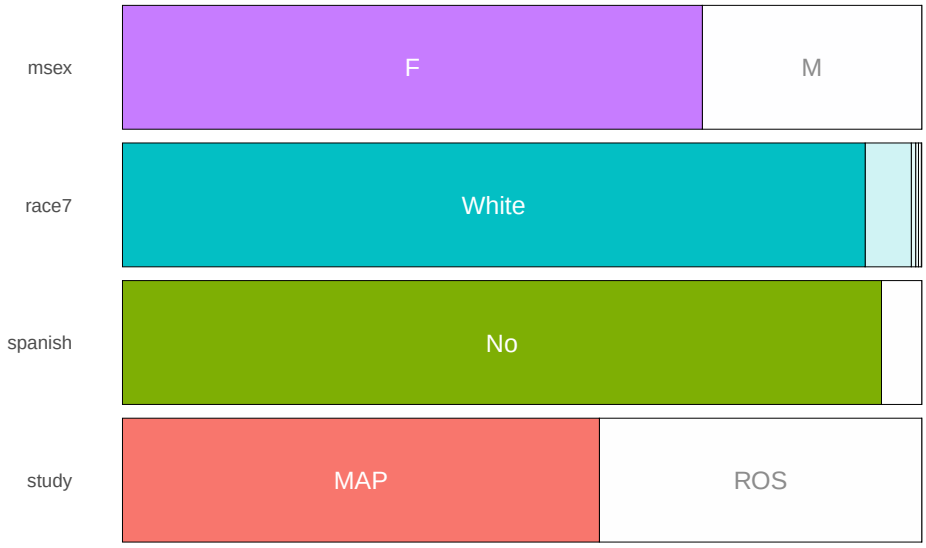
Histograms of numeric columns in df::rosmap




```
demo_c %>% show_plot(high_cardinality = 5)
```

Frequency of categorical levels in df::rosmap

Gray segments are missing values



3.2 Genetics

Data Summary

variables

definitions

types

missing_percent

unique_count

Source.Tissue.Type

Source tissue for DNA

character

67.89

15

organ

collapsed source tissue into organ

factor

67.89

4

apoe_genotype

APOE genotypes

factor

14.30

7

apoe4

APOE e4 carriers

factor

14.30

3

tomm40_hap

TOMM40 genotype

factor

36.19

7

Variable type: Factor

col_name
level
prop
cnt
apoe_genotype
33
0.53
1947
apoe_genotype
34
0.18
665
apoe_genotype
NA
0.14
525
apoe_genotype
23
0.11
390
apoe_genotype
24
0.02
72
apoe_genotype
44
0.02
56
apoe_genotype
22
0.00

17
apoe4
e4-
0.64
2354
apoe4
e4+
0.22
793
apoe4
NA
0.14
525
organ
NA
0.68
2493
organ
brain
0.22
796
organ
blood
0.10
378
organ
lymphocytes
0.00
5
Source.Tissue.Type
NA

0.68

2493

Source.Tissue.Type

Brain-DLPFC

0.13

460

Source.Tissue.Type

Whole Blood

0.10

355

Source.Tissue.Type

Brain-Cerebellum

0.07

256

Source.Tissue.Type

Brain-Posterior Cingulate Cortex

0.02

67

Source.Tissue.Type

Other

0.01

41

tomm40_hap

NA

0.36

1329

tomm40_hap

S/VL

0.23

833

tomm40_hap

S/S

0.13

485

tomm40_hap

VL/VL

0.13

469

tomm40_hap

S/L

0.07

269

tomm40_hap

L/VL

0.07

250

tomm40_hap

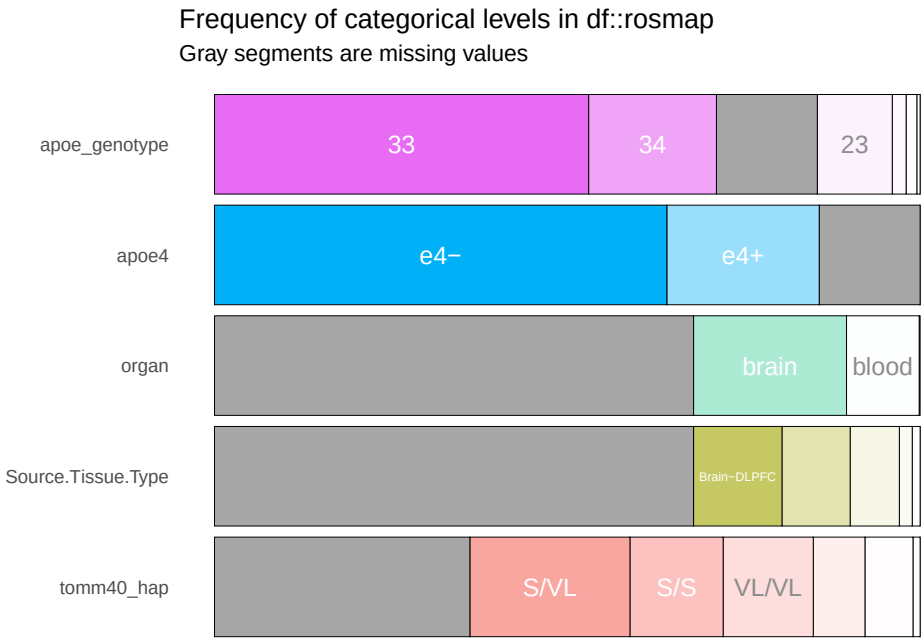
L/L

0.01

37

3.2.1 Plots

```
genetic_c %>% show_plot(high_cardinality = 5)
```



3.3 Mitochondria

Variable type: Numeric

col_name

min

q1

median

mean

q3

max

sd

pcnt_na

autosomal_coverage

26.90

33.74

36.41

36.59

39.18

60.26

4.45

67.89

mt_coverage

580.79

10127.78

25800.50

31538.67

53264.42

88911.89

23864.72

67.89

mtcn_avg

41.37

564.44

1418.11

1735.36

2945.84

4988.99

1306.14

67.89

Quality

0.50

0.90

0.93

0.93

0.96

1.00

0.05

67.89

Variable type: Factor

col_name

level

prop

cnt

Haplogroup

NA

0.68

2493

Haplogroup

Other

0.27

987

Haplogroup

V+@72

0.01

35

Haplogroup

H1

0.01

24

Haplogroup

H

0.01

22

Haplogroup

HV

0.01

19

Haplogroup

T2b

0.01

19

Haplogroup

T1a1

0.00

17

Haplogroup

H1a

0.00

16

Haplogroup

U5a1

0.00

16

Haplogroup

H1c

0.00

12

Haplogroup

J1c

0.00

12

macro

NA

0.68

2493

macro

H

0.14

507

macro

U
0.05
187
macro
T
0.03
111
macro
J
0.03
105
macro
K
0.02
78
macro
V
0.01
54
macro
I
0.01
37
macro
HV
0.01
35
macro
Other
0.01
26

macro

X

0.01

21

macro

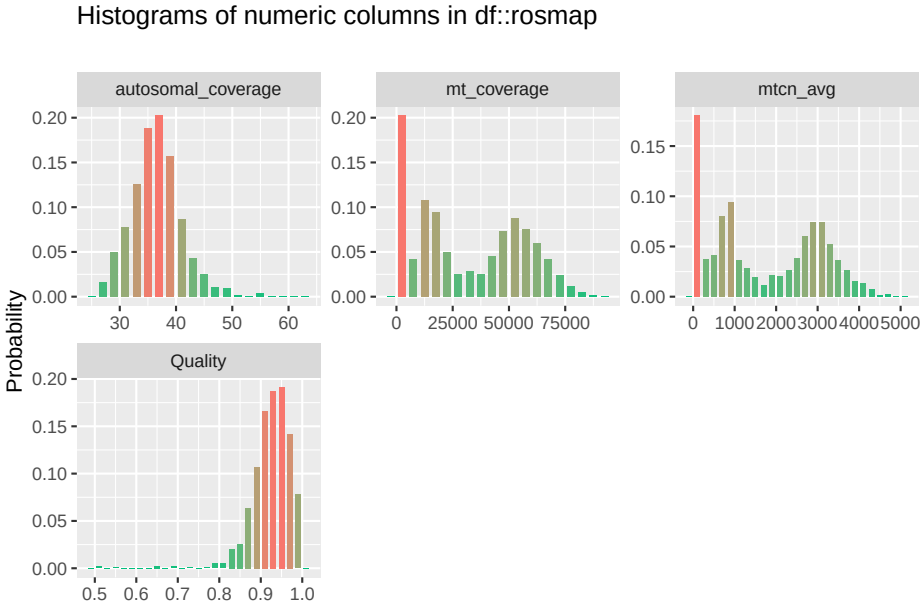
W

0.00

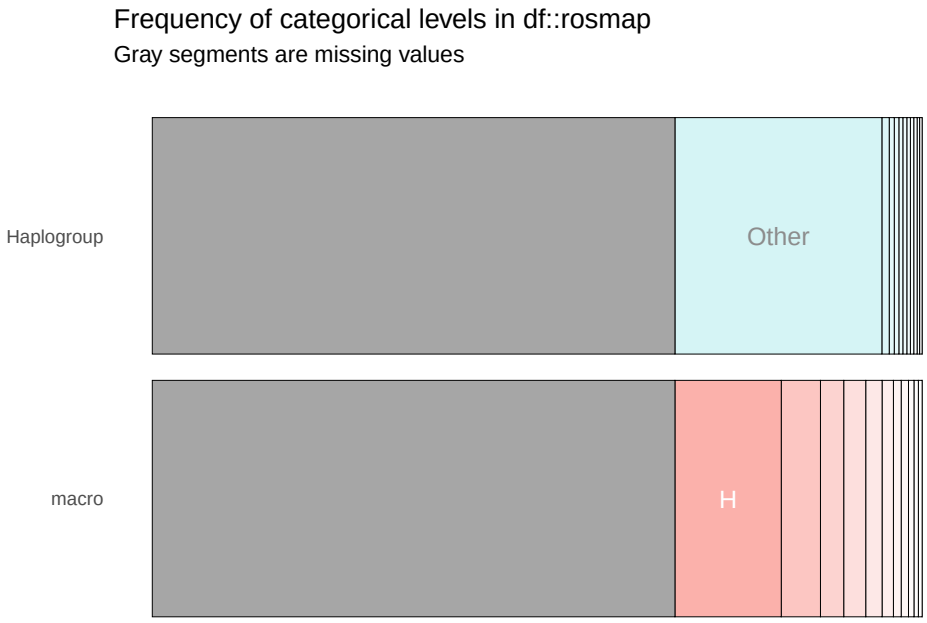
18

3.3.1 Plots

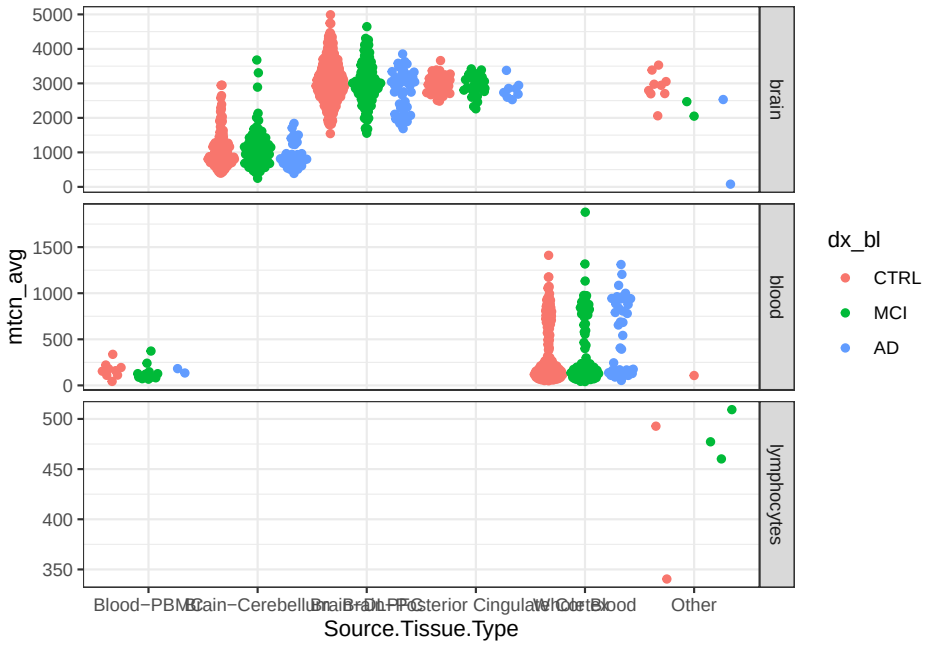
```
mt_n %>% show_plot()
```



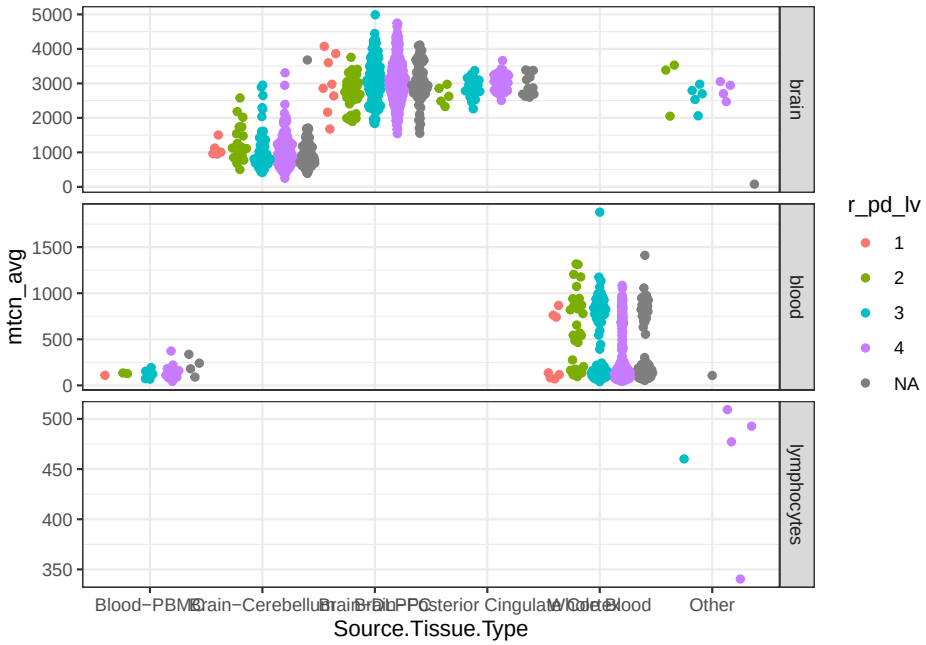

```
mt_c %>% show_plot(high_cardinality = 5)
```




```
mtcn_dx <- rosmap %>%
    filter(!is.na(mtcn_avg)) %>%
    mutate(Source.Tissue.Type = fct_lump_n(Source.Tissue.Type, 5),
      dx_lv = fct_recode(dcfdx_lv, "CTRL" = "1",
                         "MCI" = "2", "MCI" = "3",
                         "AD" = "4", "AD" = "5", "AD" = "6"),
        dx_bl = fct_recode(dcfdx_bl, "CTRL" = "1",
                         "MCI" = "2", "MCI" = "3",
                         "AD" = "4", "AD" = "5", "AD" = "6"))

mtcn_dx %>%
  ggplot(., aes(x = Source.Tissue.Type, y = mtcn_avg, colour = dx_bl)) +
    ggbeeswarm::geom_quasirandom(dodge.width=1) +
    facet_grid(organ ~ ., scales = "free") +
  theme_bw()
```




```
mtcn_dx %>%
  ggplot(., aes(x = Source.Tissue.Type, y = mtcn_avg, colour = r_pd_lv)) +
    ggbeeswarm::geom_quasirandom(dodge.width=1) +
    facet_grid(organ ~ ., scales = "free") +
  theme_bw()
```

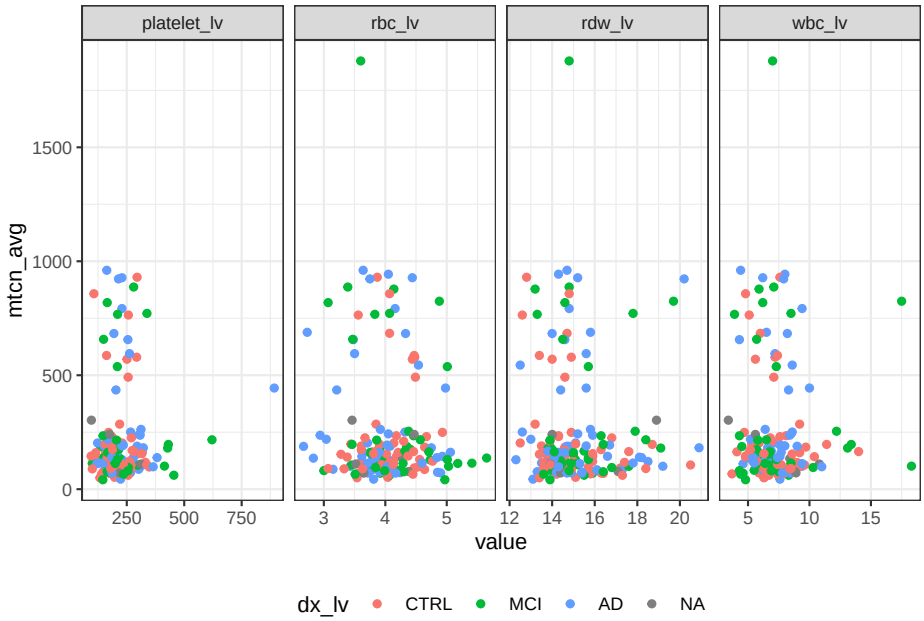


3.3.3 Blood cell counts and mtDNAcn

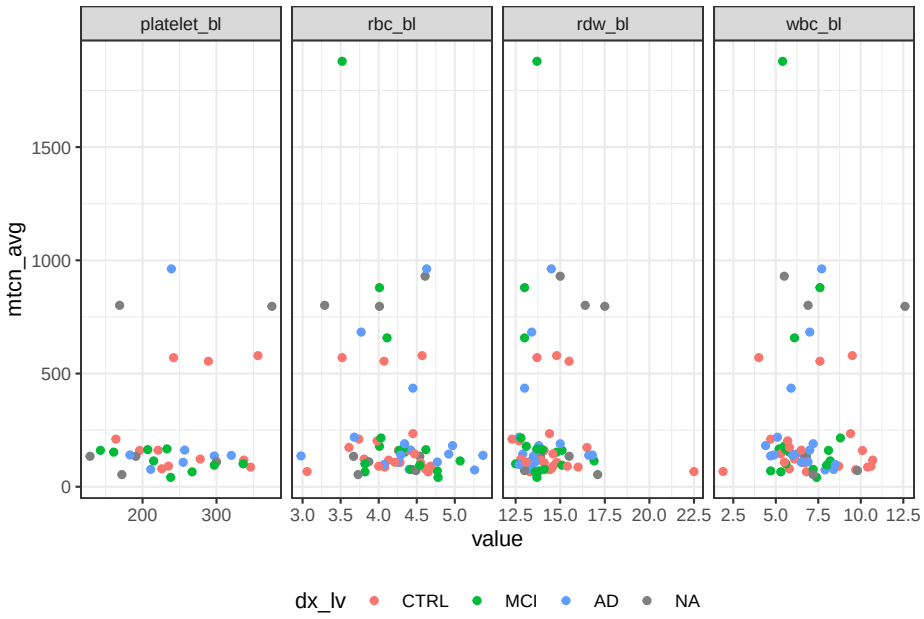
- mtDNAcn estimates from whole blood DNA can be confounded by cell type heterogeneity, and by the presence of platelets, as platelets do not have nDNA, but have mtDNA, which artificially inflates mtDNAcn. Han, L. et al 2019
- A subset of patients in ROSMAP have longitudinal measures of blood cell counts - can test if bimodal distribution of mtDNAcn in whole blood is due to cell type heterogeneity

```
blood_mt <- rosmap %>%
  mutate(dx_lv = fct_recode(dcfdx_lv, "CTRL" = "1",
                            "MCI" = "2", "MCI" = "3",
                            "AD" = "4", "AD" = "5", "AD" = "6"),
         dx_bl = fct_recode(dcfdx_bl, "CTRL" = "1",
                            "MCI" = "2", "MCI" = "3",
                            "AD" = "4", "AD" = "5", "AD" = "6")) %>%
  select(projid, study, Source.Tissue.Type,
         age_at_visit_bl, age_at_visit_lv,
         platelet_bl, rbc_bl, rdw_bl, wbc_bl,
         platelet_lv, rbc_lv, rdw_lv, wbc_lv,
         dx_bl, dx_lv,
         mtcn_avg) %>%
  pivot_longer(c(-study, -Source.Tissue.Type, -projid, -mtcn_avg, -dx_lv, -dx_bl, -age_
               names_to = "cells", values_to = "value")
```

```
blood_mt %>%
  filter(Source.Tissue.Type == "Whole Blood") %>%
  filter(str_detect(cells, "platelet_lv|rbc_lv|rdw_lv|wbc_lv")) %>%
  ggplot(., aes(x = value, y = mtcn_avg, colour = dx_lv)) +
    facet_grid(. ~ cells, scales = "free") +
    geom_point() +
    theme_bw() +
    theme(legend.position = "bottom")
```




```
blood_mt %>%
  filter(Source.Tissue.Type == "Whole Blood") %>%
  filter(str_detect(cells, "platelet_bl|rbc_bl|rdw_bl|wbc_bl")) %>%
  ggplot(., aes(x = value, y = mtcn_avg, colour = dx_lv)) +
    facet_grid(. ~ cells, scales = "free") +
    geom_point() +
    theme_bw() +
    theme(legend.position = "bottom")
```



3.4 RNA-seq

- Neuronal and astrocyte cell type enrichment are estimated using xCell to adjust for potential neuronal cell loss in dementia
- Only avaliable in a subset of participants

Variable type: Numeric

col_name

min

q1

median

mean

q3

max

sd

pcnt_na

Neurons

0.00

0.05

0.06

0.06

0.07

0.09

0.02

82.41

Astrocytes

0.01

0.02

0.03

0.03

0.04

0.09

0.01

82.41

RIN

5.00

6.30

7.20

7.06

7.90

9.90

1.00

82.43

Characteristic

DLPFC

Unknown

Total

Source.Tissue.Type

Blood

0 (0%)

1 (<0.1%)

1 (<0.1%)

Blood-Cerebellum

0 (0%)

1 (<0.1%)

1 (<0.1%)

Blood-PBMC

0 (0%)

22 (0.6%)

22 (0.6%)

Brain-Anterior Caudate

2 (<0.1%)

2 (<0.1%)

4 (0.1%)

Brain-Cerebellum

111 (3.0%)

145 (3.9%)

256 (7.0%)

Brain-DLPFC

217 (5.9%)

243 (6.6%)

460 (13%)

Brain-Frontal Cortex (BA unknown)

0 (0%)

1 (<0.1%)

1 (<0.1%)

Brain-Frontal Pole (BA10-12,32)

0 (0%)

1 (<0.1%)

1 (<0.1%)

Brain-Occipital Association Cortex (BA18,19)

2 (<0.1%)

2 (<0.1%)

4 (0.1%)

Brain-PCC

0 (0%)

1 (<0.1%)

1 (<0.1%)

Brain-Posterior Cingulate Cortex

62 (1.7%)

5 (0.1%)

67 (1.8%)

Brain-region unknown

0 (0%)

1 (<0.1%)

1 (<0.1%)

lymphocytes _transformed _with EBV virus

2 (<0.1%)

3 (<0.1%)

5 (0.1%)

Whole Blood

188 (5.1%)

167 (4.5%)

355 (9.7%)

Unknown

62 (1.7%)

2,431 (66%)

2,493 (68%)

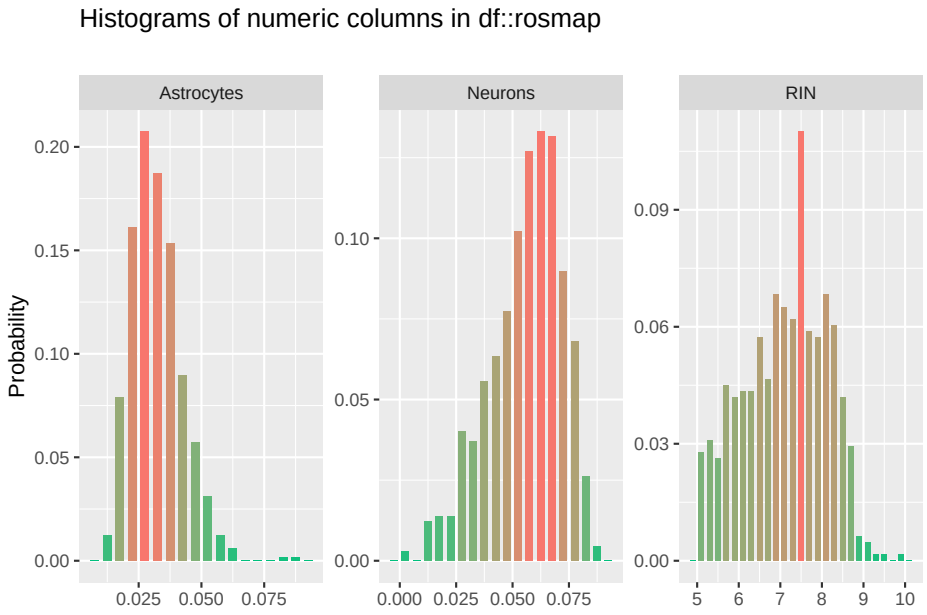
Total

646 (18%)

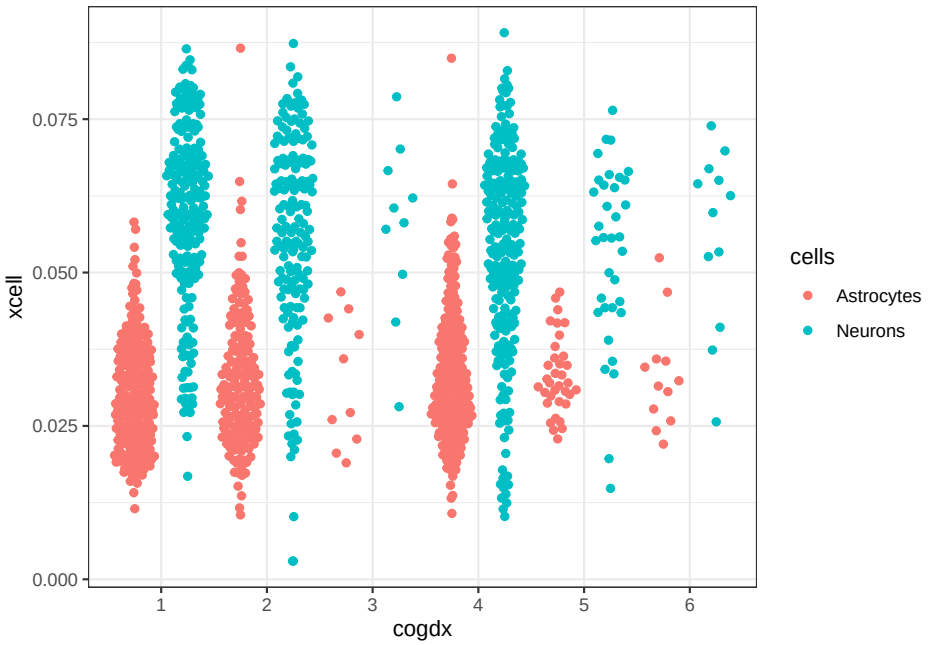
3,026 (82%)

3,672 (100%)

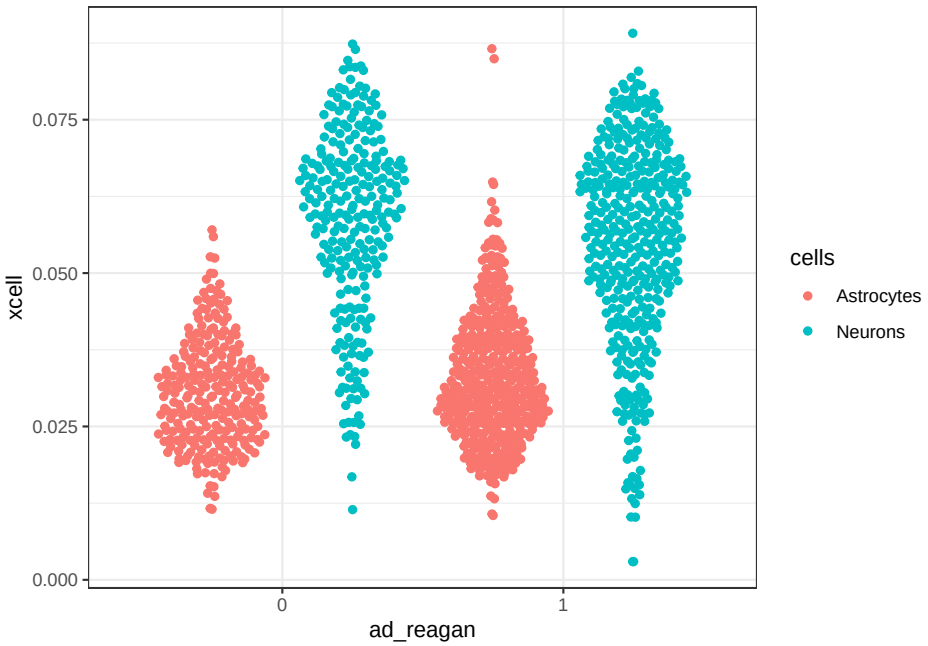
3.4.1 Plots



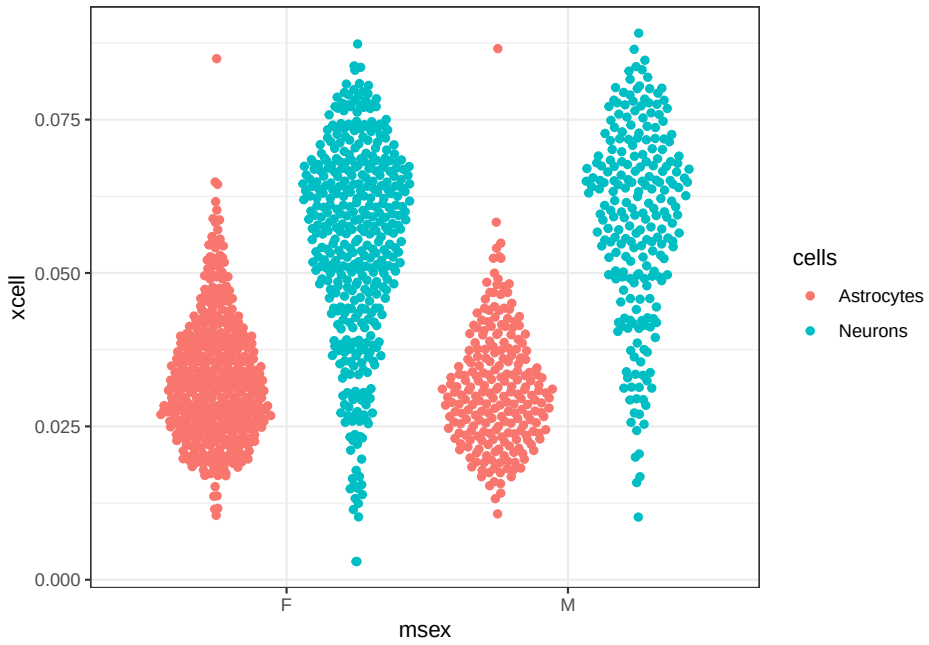

```
rosmap_xcell <- select(rosmap, Neurons, Astrocytes, rna_seq_tissue, Source.Tissue.Type
                       age_death, cogdx, ad_reagan, mtcn_avg, msex
                       ) %>%
  mutate(ad_reagan = as_factor(ad_reagan)) %>%
  filter(!is.na(Neurons)) %>%
  pivot_longer(c(Neurons, Astrocytes), names_to = "cells", values_to = "xcell")

ggplot(rosmap_xcell, aes(x = cogdx, y = xcell, colour = cells)) +
  ggbeeswarm::geom_quasirandom(dodge.width=1) +
  theme_bw()
```



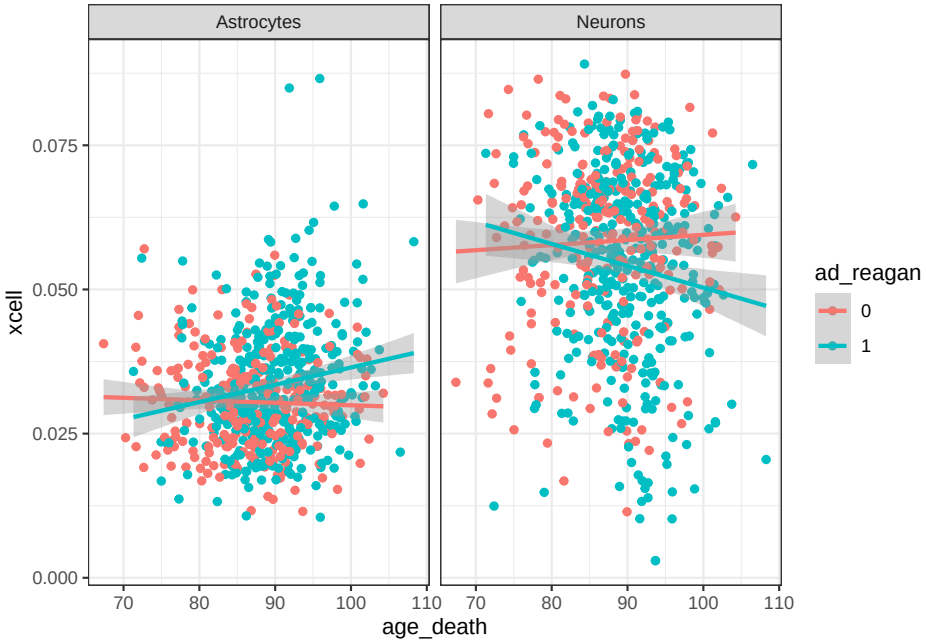

```
ggplot(rosmap_xcell, aes(x = ad_reagan, y = xcell, colour = cells)) +
  # facet_grid(. ~ cells) +
  ggbeeswarm::geom_quasirandom(dodge.width=1) +
  theme_bw()
```



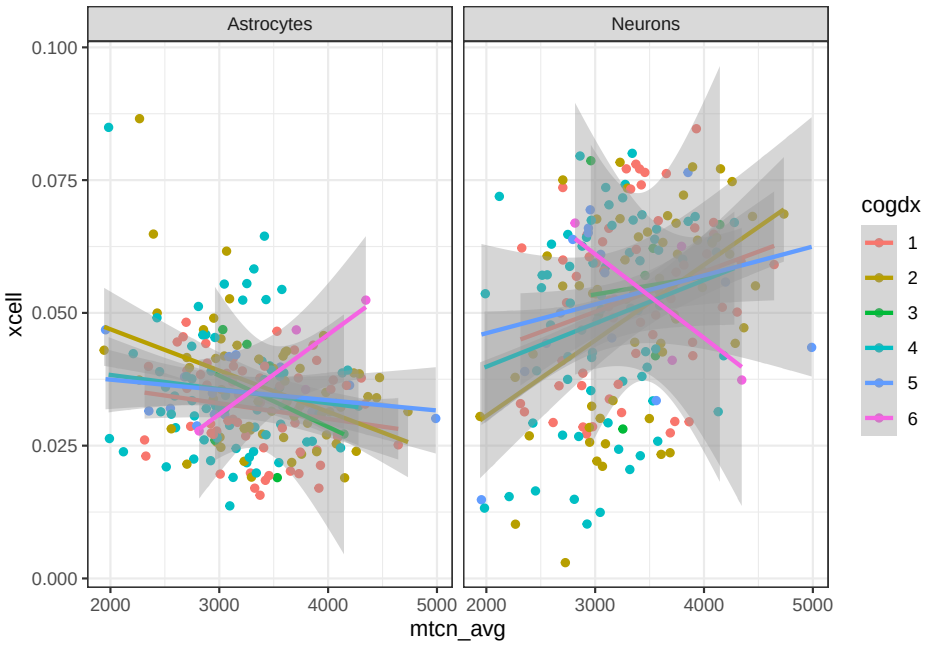
```
ggplot(rosmap_xcell, aes(x = msex, y = xcell, colour = cells)) +
  # facet_grid(. ~ cells) +
  ggbeeswarm::geom_quasirandom(dodge.width=1) +
  theme_bw()
```




```
ggplot(rosmap_xcell, aes(x = age_death, y = xcell, colour = ad_reagan)) +
  facet_grid(. ~ cells) +
  geom_point() +
  geom_smooth(method = lm) +
  theme_bw()
```




```
rosmap_xcell %>%
  filter(Source.Tissue.Type == "Brain-DLPFC") %>%
ggplot(., aes(x = mtcn_avg, y = xcell, colour = cogdx)) +
  facet_grid(. ~ cells) +
  geom_point() +
  geom_smooth(method = lm) +
  theme_bw()
```



3.5 Clinical Diagnosis

- Clinical cognitive diagnosis summary: `dcfdx` Physician's overall cognitive diagnostic category
  - 1 = NCI: No cognitive impairment (No impaired domains)
  - 2 = MCI: Mild cognitive impairment (One impaired domain) and NO other cause of CI
  - 3 = MCI: Mild cognitive impairment (One impaired domain) AND another cause of CI
  - 4 = AD: Alzheimer's dementia and NO other cause of CI (NINCDS PROB AD)
  - 5 = AD: Alzheimer's dementia AND another cause of CI (NINCDS POSS AD)
  - 6 = Other dementia: Other primary cause of dementia
- Age at first Alzheimer's dementia dx: `age_first_ad_dx` Age at cycle where first Alzheimer's dementia diagnosis was given
- Final consensus cognitive diagnosis: `cogdx` Clinical consensus diagnosis of cognitive status at time of death - same coding as `dcfdx`
- Clinical Parkinson's disease: `r_pd` is made by a clinician through review of self report questions, neurological exam (when available), cognitive testing, and interview of participant.
  - 1 = Highly Probable
  - 2 = Probable
  - 3 = Possible

  - 4 = Not Present
- Stroke diagnosis: `r_stroke` s made by a clinician through review of self report questions, neurological exam (when available), cognitive testing, and interview of participant.
  - 1 = Highly Probable
  - 2 = Probable
  - 3 = Possible
  - 4 = Not Present

Data Summary

variables

definitions

types

missing_percent

unique_count

cogdx

Clinical consensus diagnosis of cognitive status at time of death

factor

46.00

7

age_first_ad_dx

Age at cycle where first Alzheimer's dementia diagnosis was given

numeric

75.95

838

dcfdx_bl

Physician's overall cognitive diagnostic category - baseline

factor

0.03

7

dcfdx_lv

Physician's overall cognitive diagnostic category - last visit

factor

17.10

7

r_pd_bl

Diagnosis of Parkinson's disease - baseline

factor

1.39

6

r_pd_lv

Diagnosis of Parkinson's disease - last visit

factor

38.26

6

r_stroke_bl

Diagnosis of stroke made by clinician - baseline

factor

6.92

5

r_stroke_lv

Diagnosis of stroke made by clinician - last visit

factor

28.70

5

Variable type: Numeric

col_name

min

q1

median

mean

q3

max

sd

pcnt_na

age_first_ad_dx

64.06

83.4

87.93

87.56

92.16

107.23

6.58

75.95

Variable type: Factor

col_name

level

prop

cnt

cogdx

NA

0.46

1689

4

0.20

732

1

0.17

641

2

0.12

440

5

0.03

103

3

0.01
36
6
0.01
31
dcfdx_bl
1
0.69
2545
2
0.24
896
4
0.05
196
6
0.00
12
3
0.00
11
5
0.00
11
NA
0.00
1
dcfdx_lv
1
0.39
1433

4
0.21
787
2
0.18
654
NA
0.17
628
5
0.03
105
6
0.01
44
3
0.01
21
r_pd_bl
4
0.83
3056
3
0.13
463
2
0.02
80
NA
0.01
51

1
0.01
21
9
0.00
1
r_pd_lv
4
0.44
1633
NA
0.38
1405
3
0.12
436
2
0.04
165
1
0.01
31
8
0.00
2
r_stroke_bl
4
0.80
2949
NA
0.07

254

2

0.06

221

3

0.05

184

1

0.02

64

r_stroke_lv

4

0.67

2443

NA

0.29

1054

3

0.02

73

2

0.02

72

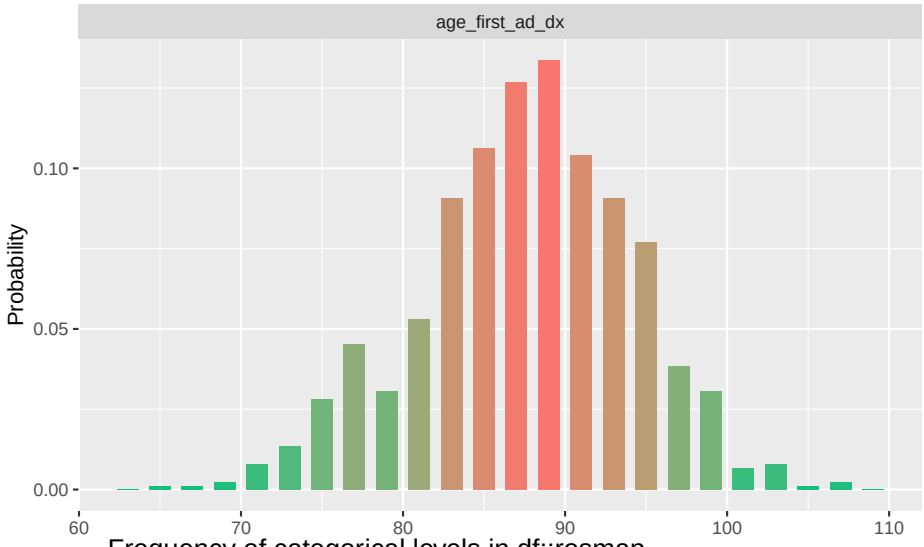
1

0.01

30

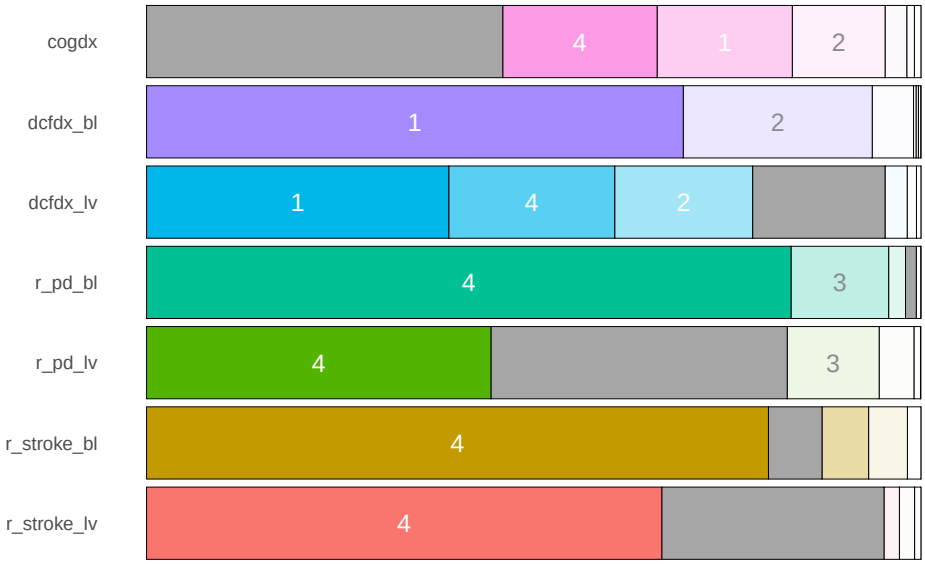
3.5.1 Plots

Histograms of numeric columns in df::rosmap



Frequency of categorical levels in df::rosmap

Gray segments are missing values



3.5.2 Cross-tabs

Characteristic

1

2

3

4

5

6

Unknown

Total

cogdx

1

535 (15%)

22 (0.6%)

0 (0%)

1 (<0.1%)

0 (0%)

2 (<0.1%)

81 (2.2%)

641 (17%)

2

19 (0.5%)

362 (9.9%)

5 (0.1%)

1 (<0.1%)

0 (0%)

0 (0%)

53 (1.4%)

440 (12%)

3

6 (0.2%)

18 (0.5%)

8 (0.2%)

1 (<0.1%)

0 (0%)
0 (0%)
3 (<0.1%)
36 (1.0%)
4
1 (<0.1%)
20 (0.5%)
0 (0%)
593 (16%)
29 (0.8%)
9 (0.2%)
80 (2.2%)
732 (20%)
5
0 (0%)
0 (0%)
0 (0%)
31 (0.8%)
51 (1.4%)
14 (0.4%)
7 (0.2%)
103 (2.8%)
6
1 (<0.1%)
0 (0%)
1 (<0.1%)
9 (0.2%)
6 (0.2%)
12 (0.3%)
2 (<0.1%)
31 (0.8%)

Unknown

871 (24%)

232 (6.3%)

7 (0.2%)

151 (4.1%)

19 (0.5%)

7 (0.2%)

402 (11%)

1,689 (46%)

Total

1,433 (39%)

654 (18%)

21 (0.6%)

787 (21%)

105 (2.9%)

44 (1.2%)

628 (17%)

3,672 (100%)

3.6 Pathology

Pathology: post-mortem neuropathologic evaluation

- Alzheimer's disease
  - NIA-Reagan diagnosis of AD: `niareaganscc` modified NIA-Reagan diagnosis of Alzheimer's disease is based on consensus recommendations for postmortem diagnosis of Alzheimer's disease. The criteria rely on both neurofibrillary tangles (Braak) and neuritic plaques (CERAD).
    - 1 = High; 2 = Intermediate; 3 = Low; 4 = No AD
  - Dichomtomized NIA-Reagan: `ad_reagan`
  - CERAD score: `ceradsc` CERAD score is a semiquantitative measure of neuritic plaques. A CERAD neuropathologic diagnosis of AD required moderate (probable AD) or frequent neuritic plaques (definite AD) in one or more neocortical regions.
    - 1 = Definite -> frequent (C3); 2 = Probable -> moderate (C2); 3 = Possible -> Sparse (C1); 4 = No AD -> None (C0)

  - Braak stage: `braaksc` Braak Stage is a semiquantitative measure of severity of neurofibrillary tangle (NFT) pathology.
    * 0 = 0; 1 = I (entorhinal); 2 = II (entorhinal); 3 = III (limbic); 4 = IV (limbic); 5 = V (neocortical); 6 = VI (neocortical)
  - Global AD pathology burden: `gpath` Global AD pathology burden is a quantitative summary of AD pathology derived from counts of three AD pathologies: neuritic plaques (n), diffuse plaques (d), and neurofibrillary tangles (nft)
- Beta-Amyloid
  - amyloid: `amyloid` Overall amyloid level - Mean of 8 brain regions
  - plaq_d: `plaq_d` Diffuse plaque summary based on 5 regions
  - plaq_n: `plaq_n` Neuritic plaque summary based on 5 regions
- PHF tau Tangles
  - Tangle: `tangles` Tangle density - Mean of 8 brain regions
  - NFT burden: `nft` Neurofibrillary tangle summary based on 5 regions
- Lewy Body disease: `dlbdx` Pathologic diagnosis of Lewy body diseases - 4 stages
  - 0 = Not present; 1 = nigral-predominant; 2 = limbic-type; 3 = neocortical-type
- Vascular
  - gross infarcts: `ci_num_gct` Cerebral Infarctions - Binary - Gross-Chronic-Any Location
  - micro infarcts: `ci_num2_mct` Cerebral Infarctions - Binary - Micro-Chronic-Any Location
  - Cerebral atherosclerosis: `cvda_4gp2` Cerebral Atherosclerosis Rating
    * 0 = None; 1 = Mild; 2 = Moderate; 3 = Severe
  - Cerebral amyloid angiopathy: `caa_4gp` Cerebral amyloid angiopathy
    * 0 = None; 1 = Mild; 2 = Moderate; 3 = Severe
  - Arteriolosclerosis: `arteriol_scler` Arteriolosclerosis
    * 0 = None; 1 = Mild; 2 = Moderate; 3 = Severe
- Hippocampal sclerosis (Typical): `hspath_typ` Definite presence of typical hippocampal sclerosis
- TDP-43 stage: `tdp_st4` TDP-43 pathology from 8 regions
  - 0 = None; 1 = Amygdala; 2 = Amygdala + Limbic; 3 = Amygdala + Limbic + Neocortical
  - dichotomized version: 0 = No TDP-43 pathology or TDP-43 pathology in amygdala only (Stages 0 and 1); 1 = TDP-43 pathology extending beyond amygdala (Stages 2 and 3)
- Microglia: Immunohistochemistry for microglia
  - mglia123_caud_vm
  - mglia123_it
  - mglia123_mf
  - mglia123_put_p
  - mglia23_caud_vm
  - mglia23_it
  - mglia23_mf

- mglia23_put_p
- mglia3_caud_vm
- mglia3_it
- mglia3_mf
- mglia3_put_p

Data Summary

variables

definitions

types

missing_percent

unique_count

niareagansc

NA

ordered

53.76

5

ceradsc

NA

ordered

53.76

5

braaksc

NA

ordered

53.76

8

gpath

NA

numeric

53.95

1605

amyloid

NA

numeric

55.47

1374

plaq_d

NA

numeric

53.95

1308

plaq_n

NA

numeric

53.95

1264

tangles

NA

numeric

55.28

1622

nft

NA

numeric

53.95

1368

dlbdx

NA

factor

55.23

5

ci_num2_gct

NA

factor

53.79

3

ci_num2_mct

NA

factor

53.79

3

cvda_4gp2

NA

factor

53.40

5

caa_4gp

NA

factor

54.90

5

arteriol_scler

NA

factor

54.08

5

hspath_typ

NA

factor

56.05

3

tdp_st4

NA

factor

56.86

5

Variable type: Numeric

col_name

min

q1

median

mean

q3

max

sd

pcnt_na

gpath

0

0.19

0.64

0.75

1.14

3.20

0.63

53.95

amyloid

0

0.53

2.83

3.93

6.24

22.94

3.95

55.47

plaq_d

0

0.09

0.53

0.73

1.10

4.93

0.77

53.95

plaq_n

0

0.06

0.71

0.86

1.34

5.36

0.84

53.95

tangles

0

1.59

4.23

7.28

9.46

78.52

8.83

55.28

nft

0

0.14

0.37

0.65