

# Polygenic Risk Score project

**Week 1**

2019-01-23

# Definition

## **Polygenic Risk Score (PRS) =**

- single value estimate of an individual's propensity to a phenotype
- calculated as a sum of their genome-wide genotypes weighted by corresponding genotype effect sizes from summary statistic GWAS data
- can be used to assess shared aetiology between phenotypes
- or to predict individuals at high risk for a disease

=> Question: computed at individual-level, how to switch to a cluster of individuals' score ? average of individual score inside a cluster?

=> Question: What's the difference between shared aetiology, shared heritability and shared genetic variance (genetic correlation)?

# State of the art

Polygenic Risk Score database => did not find any

Google search for “polygenic risk score database” => only one result:

Data - Cardiovascular Disease Genomics ( <http://www.broadcvdi.org/informational/data> ):

- make available lists of variants and weights comprising polygenic risk scores for five complex diseases :
  - Atrial fibrillation (297.3 MB) | Breast cancer (253 KB) | Coronary artery disease (292.9 MB) | Inflammatory bowel disease (305.1 MB) | Type 2 diabetes (305.6 MB)
- as described in:
  - Khera, Amit V., et al. "Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations." Nature genetics 50.9 (2018): 1219.

# State of the art

With no available database, we have to compute Polygenic Risk Scores (PRS) for a phenotype of interest with a software.

Identified 2 softwares which aim to compute PRS:

- PRSice (Euesden, Jack, Cathryn M. Lewis, and Paul F. O'reilly. "PRSice: polygenic risk score software." Bioinformatics 31.9 (2014): 1466-1468.), cited by 319
- LDpred (Vilhjálmsen, Bjarni J., et al. "Modeling linkage disequilibrium increases accuracy of polygenic risk scores." The American Journal of Human Genetics 97.4 (2015): 576-592.), cited by 189

# Bibliography

- Bralten, J., et al. "Autism spectrum disorders and autistic traits share genetics and biology." *Molecular psychiatry* (2017).
- Euesden, Jack, Cathryn M. Lewis, and Paul F. O'reilly. "PRSice: polygenic risk score software." *Bioinformatics* 31.9 (2014): 1466-1468.
- Khera, Amit V., et al. "Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations." *Nature genetics* 50.9 (2018): 1219.
- Choi, Shing Wan, Timothy Shin Heng Mak, and Paul O'Reilly. "A guide to performing Polygenic Risk Score analyses." *bioRxiv* (2018): 416545.

# Autism spectrum disorders and autistic traits share genetics and biology.

Method: “Shared genetic etiology analysis”:

- Points to the resource: large GWAS of ASDs by the Psychiatric Genomics Consortium (PGC) autism group (PGC-ASD GWAS, data for 5305 ASD cases and 5305 controls—these data are publicly available at: <http://www.med.unc.edu/pgc/downloads>)
- They used PGC-ASD GWAS as ‘base phenotype’ sample to generate Polygenic Risk Scores for ASDs with PRSice software
- They used summary statistics data from six independent GWAS of autistic traits as ‘target phenotype’ samples
- They applied PRS to compute shared etiology between ASD and autistic traits using PRSice
  - that is, the extent to which the combined SNPs from polygenic risk scores for ASD predict each of the six autistic trait phenotypes
- They observed genetic sharing between ASDs and the autistic traits ‘childhood behavior’, ‘rigidity’ and ‘attention to detail’

# PRSize: polygenic risk score software

- Can calculate PRS at any number of P-value thresholds (PT) and can thus identify the most predictive threshold.
- Automate PRS analyses
- Can test for shared aetiology between traits:
  - PRS on the base phenotype are calculated using GWAS results
  - PRS are then used as predictors of the target phenotype in a regression on individuals from an independent data set
- Input: GWAS results on a base phenotype and genotype data on a target phenotype
- Output: PRS for each individual and figures depicting the PRS model fit at a range of PT
- Option: can remove SNPs in LD

## Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations.

- Used Polygenic scores to identify individuals at high risk for a disease
- Results: developed and validated polygenic scores for 5 common diseases
- The approach identifies 8.0, 6.1, 3.5, 3.2, and 1.5% of the population at greater than threefold increased risk for coronary artery disease, atrial fibrillation, type 2 diabetes, inflammatory bowel disease, and breast cancer, respectively
- Method: PRS generation:
  - They used summary statistics from recent GWAS studies for 5 diseases as base phenotype
  - They used LDpred computational algorithm to generate seven candidate PRS for each disease
  - LDpred = “Bayesian approach that calculates a posterior mean effect size for each variant
- PRS calculation in a validation dataset of 120,280 participants from the UK Biobank phase 1 release
- Scores were generated by multiplying the genotype dosage of each risk allele for each variant by its respective weight, and then summing across all variants in the score using PLINK2 software
- For each of the five diseases, the score with the best discriminative capacity was determined based on the maximal AUC in a logistic regression model



# A guide to performing Polygenic Risk Score analyses

- From the same team that developed PRSice (currently reading)
- Provide guidelines and recommendation to obtain best PRS
- Quality control of Base and Target data
- Shrinkage of GWAS effect size estimates
- Important preprocessing: Controlling for Linkage Disequilibrium = clumping to select independent index SNPs for each linkage disequilibrium (LD) block in the genome

# Application

Base datasets: 2 meta-analyses

- PGC-ASD: GWAS summary statistics for Autism Spectrum Disorder
  - 5,305 ASD cases and 5,305 controls
- IGAP-LOAD: GWAS summary statistics for Late-Onset Alzheimer's Disease
  - 17,000 AD cases, 37,154 controls

# For next week:

- Continue reading PRSice documentation to understand how to compute PRS from base datasets
  - => read it, tried with a test dataset and with our data
- Identify which target dataset from UKBiobank to use: need raw genotype data, where is it stored?
  - => /neurospin/ukb/genetic/GENETIC\_DATA\_500k/CALL\_DATA/


# Polygenic Risk Score project

**Week 2**

2019-02-06


# PRSize: test dataset


- **Base dataset** : short GWAS summary statistics  
(91,062 variants with OR; binary phenotype: cases/controls)


  
TOY\_BASE\_GWAS.  
assoc

SNP	CHR	BP	A1	A2	P	OR
SNP_22857	4	103593179	1	2	0.2852	13.290
SNP_13879	2	237416793	1	2	0.8784	21.624
SNP_20771	4	16957461	1	2	0.1994	91.265
SNP_13787	2	235355721	1	2	0.7234	3.178
SNP_25383	4	189927377	1	2	0.3309	3.167

- **Target dataset** : raw genotype data (PLINK binary)  
(2,000 ind; 88,836 variants; binary phenotype: 1000 cases / 1000 controls)

  
TOY\_TARGET\_DATA.  
bed

  
TOY\_TARGET\_DATA.  
bim

  
TOY\_TARGET\_DATA.  
fam

- 88,836 variants in common

# PRSize: Compute PRS on test dataset

Command line:

```
Rscript PRSize.R --dir . \  
  --prsize ./PRSize \  
  --base ./data_test/TOY_BASE GWAS.assoc \  
  --target ./data_test/TOY_TARGET_DATA \  
  --thread 1 \  
  --stat OR \  
  --binary-target T \  
  --out ./output_test/test
```

# PRSize: Output of Results from test dataset

## Outputs:

FID 1  
CAS 1  
CAS 2  
CAS 3

test.best

- PRS for each individual at the best-fit PRS

FID	IID	In Regression	PRS
CAS_1	CAS_1	Yes	-0.00599501328
CAS_2	CAS_2	Yes	-0.00631017938
CAS_3	CAS_3	Yes	-0.00227495325
CAS_4	CAS_4	Yes	-0.00204360007
CAS_5	CAS_5	Yes	-0.000830676955
CAS_6	CAS_6	Yes	-0.00224943517

PRSize  
https  
(C) 2

test.log

SetTh  
Base0  
Base0  
Base0

test.prsize

- PRS model fit R2 across thresholds

Set	Threshold	R2	P	Coefficient	Std.Error	Num_SNP
Base	0.00025	0.0133696	8.43169e-06	-0.197266	0.0442903	2
Base	0.0003	0.00824473	0.000456434	-0.225204	0.0642503	3
Base	0.0004	0.0089725	0.000256089	-0.350267	0.0958035	5
Base	0.00045	0.0101339	0.000102845	-0.445497	0.114707	6
Base	0.00065	0.00532975	0.004775	-0.402003	0.142462	8

Pheno  
-Base


test.summary

- Information of the best model fit of each phenotype and gene set

Phenotype	Set	Threshold	PRS.R2	Full.R2	Null.R2	Prevalence	Coefficient	P	Num_SNP
- Base	0.4463	0.0520082	0.0520082	0	-	86.288	9.96331	4.69368e-18	36759

# PRSize: UKB call data

- **Base dataset** : PGC ASD GWAS summary statistics  
(9,499,589 variants with OR, Binary Phenotype: 5,305 ASD-diagnosed cases / 5,305 controls)

  
daner\_pgc\_asd\_  
euro\_all\_  
25Mar2015

CHR	SNP	BP	A1	A2	FRQ_A_5305	FRQ_U_5305	INFO	OR	SE	P	ngt
1	rs3107975	55326	T	C	0.990765	0.990765	0.469	1.072080	0.1342	0.60400	0
1	rs147589465	240789	T	G	0.98681	0.98681	0.436	0.891990	0.1496	0.44500	0
1	rs60791385	526970	T	C	0.994723	0.994723	0.420	0.824400	0.1711	0.25930	0
1	chr1_532258_D	532258	D	I	0.01583	0.01583	0.437	1.260990	0.1527	0.12880	0
1	chr1_540540_D	540540	I	D	0.98549	0.98549	0.438	0.850870	0.1836	0.37880	0

- **Target dataset** : UKB genetic data 500k call data (non imputed), chr9 ( arbitrary chr for testing)  
(488,377 ind; 31,544 variants; Unknown Phenotype)

```
tr256641@is220756:/neurospin/ukb/genetic/GENETIC_DATA_500k/CALL_DATA$ ls -lh chr9*  
-rwxrwxrwx 1 yl247234 yl247234 4,0G juil. 12 2017 chr9_v2.bed  
-rwxrwxrwx 1 yl247234 yl247234 949K juil. 13 2017 chr9_v2.bim  
-rwxrwxrwx 1 yl247234 yl247234 16M juil. 20 2017 chr9_v2.fam
```

- 26,441 variants in common => 10,001 after clumping



# PRSize: Compute PRS on UKB call data

Command line:

```
Rscript PRSize.R --dir . \  
  --prsize ./PRSize \  
  --base /home/tr256641/Documents/polygenic_risc_score_project/PRS_base_dataset/  
  daner_pgc_asd_euro_all_25Mar2015 \  
  --target /neurospin/ukb/genetic/GENETIC_DATA_500k/CALL_DATA/chr9_v2 \  
  --thread 1 \  
  --stat OR \  
  --binary-target T \  
  --out ./output_ukb_call_data/chr9
```

=> `488377 sample(s) with invalid phenotype`  
`Error: All sample has invalid phenotypes!`

=> It's because UKB individuals have unknown phenotype, we want to predict it

- In that case PRSize author recommends to use the “--no-regress option”

# PRSize: Compute PRS on UKB call data

Command line:

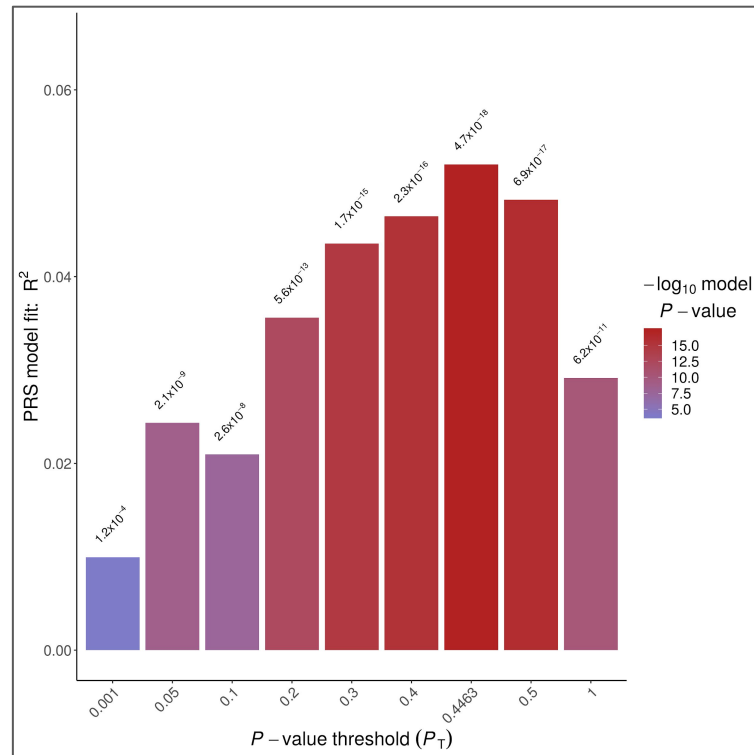
```
Rscript PRSize.R --dir . \  
  --prsize ./PRSize \  
  --base /home/tr256641/Documents/polygenic_risc_score_project/PRS_base_dataset/  
  daner_pgc_asd_euro_all_25Mar2015 \  
  --target /neurospin/ukb/genetic/GENETIC_DATA_500k/CALL_DATA/chr9_v2 \  
  --thread 1 \  
  --stat OR \  
  --binary-target T \  
  --out ./output_ukb_call_data/chr9 \  
  --no-regress
```

=> it solved the problem, PRSize ran without error

# PRSize: Compute PRS on UKB call data

=> It worked because :

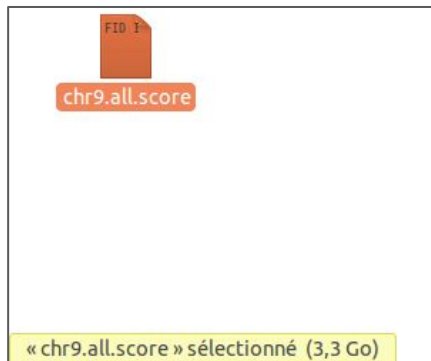
- Usually PRSize does a regression after computing PRS to evaluate the association between the PRS and the target phenotype (for example to compute shared aetiology between 2 known traits)
- It does the regression for each score at each p-value threshold
- It allows to identify the best p-val threshold that gives the best score to predict a known phenotype



# PRSize: PRS result file on UKB call data


=> But in our case :

- the target phenotype of UKB is unknown, so we cannot do the regression and thus we cannot identify the best p-value threshold for our score
- => after PRSize computation, we obtain a very large file: 488,377 ind x 5299 PRS/threshold (thresholds : 0.0001 to 0.5 with 0.00005 increments = 5299 thresholds)



# PRSize: UKB imputed data

- **Base dataset** : PGC ASD GWAS summary statistics  
(9,499,589 variants with OR, Binary Phenotype: 5,305 ASD-diagnosed cases / 5,305 controls)

  
daner\_pgc\_asd\_  
euro\_all\_  
25Mar2015

CHR	SNP	BP	A1	A2	FRQ_A_5305	FRQ_U_5305	INFO	OR	SE	P	ngt
1	rs3107975	55326	T	C	0.990765	0.990765	0.469	1.072080	0.1342	0.60400	0
1	rs147589465	240789	T	G	0.98681	0.98681	0.436	0.891990	0.1496	0.44500	0
1	rs60791385	526970	T	C	0.994723	0.994723	0.420	0.824400	0.1711	0.25930	0
1	chr1_532258_D	532258	D	I	0.01583	0.01583	0.437	1.260990	0.1527	0.12880	0
1	chr1_540540_D	540540	I	D	0.98549	0.98549	0.438	0.850870	0.1836	0.37880	0

- **Target dataset** : UKB genetic data 25k imputed data, chr9 ( arbitrary chr for testing)  
(24,277 ind; 3,396,128 variant; Unknown Phenotype)

```
tr256641@is220756:/neurospin/ukb/genetic/GENETIC_DATA_500k/CALL_DATA$ ls -lh chr9*  
-rwxrwxrwx 1 yl247234 yl247234 4,0G juil. 12 2017 chr9_v2.bed  
-rwxrwxrwx 1 yl247234 yl247234 949K juil. 13 2017 chr9_v2.bim  
-rwxrwxrwx 1 yl247234 yl247234 16M juil. 20 2017 chr9_v2.fam
```

- 291,475 variants in common => 20,432 after clumping

# PRSize: Compute PRS on UKB imputed data

Same problem as with the call data :

- Generates a large file with 24,277 ind x 5299 PRS/threshold

# Perspective

- How to evaluate the best p-value threshold for our score?

=> the answer is to find a phenotype for our target dataset and run the software for shared aetiology between base and target phenotypes

# For next week:

- Find in the literature, a relevant phenotype for autism
- e.g. structural connectivity measures for an autism-relevant fibre tract
- Try to run PRSice on:
  - Base dataset: PGC ASD GWAS summary statistics
  - Base phenotype: Binary trait : ASD and control
  - Target dataset: UKB imputed genetic data with corresponding dMRI measures
  - Target phenotype: Quantitative trait: structural connectivity measures

=> PRSice should be able to identify the best p-value threshold that gives the best polygenic score to maximize the target phenotype (structural connectivity measures), to maximize the shared aetiology between the two traits



# Polygenic Risk Score project

**Week 3**

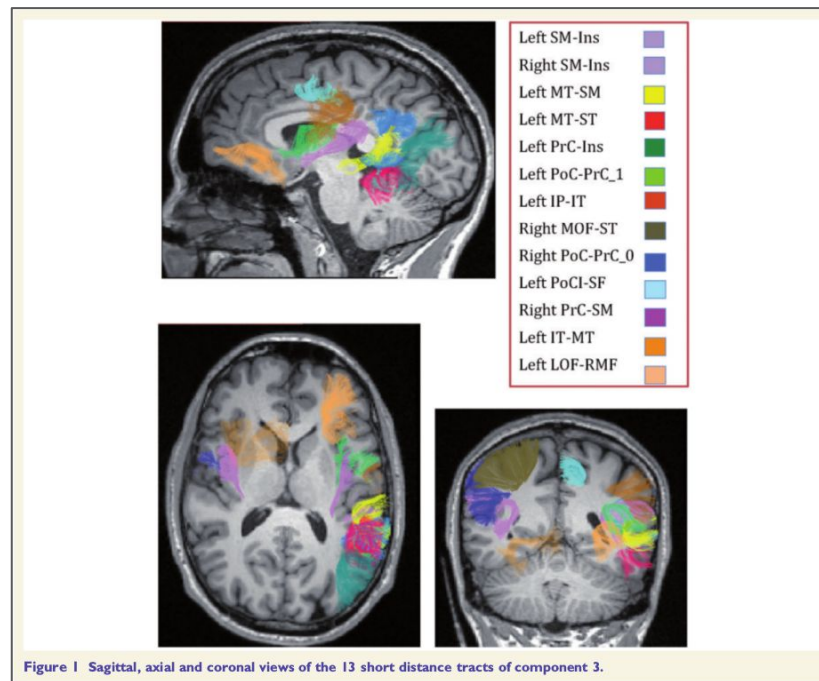
2019-02-13

# Search for a phenotype associated with autism



- Suggested phenotype: Superficial white matter composed of U-shaped association fibres

# Search for a phenotype associated with autism



- Results : decreased local structural connectivity in **13 short tracts** in ASD subjects compare to controls
- Specific regions : short-tract atypicalities in **temporal lobe** and **insula** significantly associated with clinical manifestation of ASD

# Objectives

- Without available data for short distance tracts
- Testing of PRSice with available data: FA and ICVF for long distance tracts:
  - Middle\_cerebellar\_peduncle
  - Pontine\_crossing\_tract\_(a\_part\_of\_MCP)
  - Genu\_of\_corpus\_callosum
  - Body\_of\_corpus\_callosum
  - Splenium\_of\_corpus\_callosum
  - Fornix(column\_and\_body\_of\_fornix)
  - Corticospinal\_tract\_L/R
  - Medial\_lemniscus\_L/R
  - Inferior\_cerebellar\_peduncle\_L/R
  - Superior\_cerebellar\_peduncle\_L/R
  - Cerebral\_peduncle\_L/R
  - Anterior\_limb\_of\_internal\_capsule\_L/R
  - Posterior\_limb\_of\_internal\_capsule\_L/R
  - Retrolenticular\_part\_of\_internal\_capsule\_L/R
  - Anterior\_corona\_radiata\_L/R
  - Superior\_corona\_radiata\_L/R
  - Posterior\_corona\_radiata\_L/R
  - Posterior\_thalamic\_radiation\_(include\_optic\_radiation)\_L/R
  - Sagittal\_stratum\_(include\_inferior\_longitudinal\_fasciculus\_and\_inferior\_fronto-occipital\_fasciculus)\_L/R
  - External\_capsule\_L/R
  - Cingulum\_(cingulate\_gyrus)\_L/R
  - Cingulum\_(hippocampus)\_L/R
  - Fornix\_(cres)/Stria\_terminalis\_(can\_not\_be\_resolved\_with\_current\_resolution)\_L/R
  - Superior\_longitudinal\_fasciculus\_LR
  - Superior\_fronto-occipital\_fasciculus\_(could\_be\_a\_part\_of\_anterior\_internal\_capsule)\_L/R
  - Uncinate\_fasciculus\_L/R
  - Tapetum\_L/R

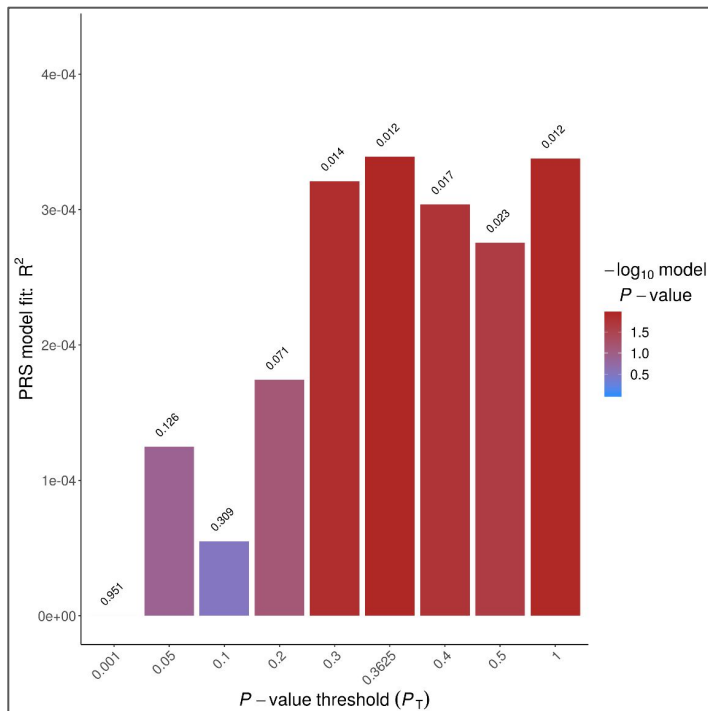
# PRSize: shared aetiology between ASD and a long distance tract

Phenotype tested: “FA\_Middle\_cerebellar\_peduncle”

run PRSize on:

- Base dataset: PGC ASD GWAS summary statistics
- Base phenotype: Binary trait : ASD and control
- Target dataset: UKB genetic call data with FA measures
- Target phenotype: Quantitative trait: FA\_Middle\_cerebellar\_peduncle

# PRSize: shared aetiology between ASD and a long distance tract



Giving a target phenotype to PRSize as input, PRSize can identify the best p-value threshold that gives the best scores to maximize the target phenotype ('FA\_Middle\_cerebellar\_peduncle')

=> PRSize identified the best p-value threshold at 0.3625 which includes 4894 SNPs.

# PRSize: shared aetiology between ASD and 4 long distance tracts

Option for testing multiple phenotypes:

```
--pheno-file ./phenotype/all_FA_skeletonised.tsv \  
--pheno-col Middle_cerebellar_peduncle,Genu_of_corpus_callosum,  
Body_of_corpus_callosum,Corticospinal_tract_R
```

# PRSize: shared aetiology between ASD and 4 long distance tracts

chr9.Body\_of\_corpus\_callosum.best x example\_co

1	FID	IID	In Regression	PRS
2	4767298	4767298	No	0.0102627637
3	2605573	2605573	No	0.00734067787
4	2449175	2449175	No	0.00459266603
5	3658048	3658048	No	0.00103840145
6	4910718	4910718	No	0.00156362229
7	1865602	1865602	No	-0.00134547055
8	5811216	5811216	No	0.00380187952
9	2633756	2633756	No	0.00202764788
10	1577784	1577784	No	0.00669873248

chr9.Genu\_of\_corpus\_callosum.best x example\_co

1	FID	IID	In Regression	PRS
2	4767298	4767298	No	0.00371388886
3	2605573	2605573	No	0.00265676067
4	2449175	2449175	No	-0.0124698361
5	3658048	3658048	No	-0.0182165141
6	4910718	4910718	No	-0.00671672881
7	1865602	1865602	No	-0.0437474907
8	5811216	5811216	No	-0.0209545969
9	2633756	2633756	No	-0.0260745605
10	1577784	1577784	No	0.0114493173

chr9.Corticospinal\_tract\_R.best x example\_command

1	FID	IID	In Regression	PRS
2	4767298	4767298	No	0.00163835959
3	2605573	2605573	No	0.000923451143
4	2449175	2449175	No	0.00113334457
5	3658048	3658048	No	0.000589316755
6	4910718	4910718	No	-0.000245026972
7	1865602	1865602	No	0.000213269808
8	5811216	5811216	No	0.000935911664
9	2633756	2633756	No	0.00108811393
10	1577784	1577784	No	0.00084882605

chr9.Middle\_cerebellar\_peduncle.best x example\_co

1	FID	IID	In Regression	PRS
2	4767298	4767298	No	0.00105581844
3	2605573	2605573	No	0.000622464356
4	2449175	2449175	No	0.000647592359
5	3658048	3658048	No	0.000347149649
6	4910718	4910718	No	0.000190573612
7	1865602	1865602	No	9.25228617e-05
8	5811216	5811216	No	0.000447883302
9	2633756	2633756	No	0.000852313503
10	1577784	1577784	No	0.000584202549

Phenotype	Set Threshold	PRS.R2	Full.R2	Null.R2	Prevalence	Coefficient	Standard.Error	P	Num_SNP
Middle_cerebellar_peduncle Base	0.3625	0.000338951	0.000338951	0	-	-1.07808	0.427778	0.0117373	4894
Genu_of_corpus_callosum Base	0.0004	0.000135473	0.000135473	0	-	0.0314441	0.0197374	0.11115	13
Body_of_corpus_callosum Base	0.00635	0.000148793	0.000148793	0	-	0.091195	0.0546205	0.0950127	130
Corticospinal_tract_R Base	0.1832	0.000648249	0.000648249	0	-	-1.58075	0.453482	0.00049177	2747



# Perspective

- Currently testing to run PRSice for 22 chromosomes and one phenotype “FA\_Middle\_cerebellar\_peduncle” on call data
  - => around 5 hours to run instead of 15min ( $15\text{min} \times 22\text{chr}$ )
- Upscale to 50 phenotypes
- Identify the phenotype with the highest polygenic score
- Upscale to imputed data

# Polygenic Risk Score project

**Week 4**

2019-02-20

# Singularity container for PRSice

- We built a singularity container containing PRSice and its dependencies
- Aim = run the container in Alambic cluster because imputed genetic data are big (558 Go) and PRSice requires as much memory to save it before computing PRS

# Split clumping and PRS computation

- We split “clumping step” and “PRS computation step” because the clumping step was limiting (take much longer to process than actually computing the scores)
- The idea is to clump the target imputed genetic data once and then run PRSice on clumped data
- We worked on a modular script where the user can choose to compute clumping with PLINK or compute PRS with PRSice or both
- The clumping is done for HCP and UKB genetic data.
- It outputs a csv with the most significant SNPs for ASD without the SNPs in LD with them.

# Compute PRS for UKB

- We have not run it yet because imputed genetic data are huge.
- We first testing it on HCP data (much smaller: imputed genetic data for all autosomes = 7 Go).

# Compute PRS for HCP

- Over 1000 subjects: 950 have genetic and imaging data
- Test run for one phenotype: one connexion between two areas
- Added covariates data (age, sex, 4th first dimension of PCA),
  - need to find covariates for ukb: do you know where it is?
- It is currently running on alambic cluster and it starts creating files, it seems to work

# Perspective

- If test run works for one phenotype for HCP => upscale to 1000 phenotypes (connexions between different areas + metrics of the whole network), phenotype files are separated to run PRS processes simultaneously
- Then upscale to ukb imputed data for 1 phenotype
- and then for 250 phenotypes (tbss-long distance tracts)

# Polygenic Risk Score project

**Week 5**

2019-03-06



# Compute PRS for UKB (imputed)

UKB	Disease	
	Alzheimer (IGAP)	Autism (PGC)
Phenotype	TBSS ✓	TBSS ✓
	Opening ✓	Opening ✓

# Compute PRS for HCP (non-imputed)

HCP	Disease	
	Alzheimer (IGAP)	Autism (PGC)
Phenotype	TBSS ✓	TBSS ✓
	Connectivity ✓	Connectivity ✓
	Network ✓	Network ✓
	Opening ✓	Opening ✓

# Perspective

- Recompute imputation for HCP and rerun PRSice
- Run test on non-imputed data for “height” phenotype to check if PRSice explain correctly the variability ( $> 0.2$ )
  - If it's correct, compute imputation

# Compute heritability ( $h^2$ )

- Solar code cleaning is finished and Solar is ready to run
- Currently running to compute  $h^2$  for all the phenotypes in HCP
- Perspective: compute heritability of the same phenotypes for UKB with GCTA
- 2nd perspective : compute shared  $h^2$  between phenotypes with GCTA or BOLT-REML or LD score regression (LDSC)
  - Idea: plan to do a thinklab on computing on genetic correlation (include shared heritability and PRS)

# Polygenic Risk Score project

**Week 6**

2019-03-27

# Results

- Sort phenotypes by pval
- Keep 10 most significant phenotypes

# Results ASD - UKB - 10 best

Measure	Phenotype	Threshold	PRS.R2	Full.R2	Null.R2	Coefficient	Standard.Error	P	Num_SNP	-log(P)
all_OD_skeletonised_British_filtered	Posterior_limb_of_internal_capsule_L	0.00035	0.00114441	0.0109551	0.00981069	-0.606922	0.138792	1.23369E-05	802	4.9087939555
all_OD_skeletonised_British_filtered	Posterior_limb_of_internal_capsule_R	0.00035	0.000992739	0.0113942	0.0104015	-0.562633	0.138113	4.64862E-05	802	4.3326759536
all_OD_skeletonised_British_filtered	Cingulum-hippocampus-L	0.0008	0.000806346	0.0298918	0.0290854	-1.56622	0.422588	0.000211038	1538	3.6756393376
all_OD_skeletonised_British_filtered	Posterior_corona_radiata_R	0.00315	0.000747479	0.0806698	0.0799224	-1.77283	0.483638	0.00024751	4458	3.6064072498
all_OD_skeletonised_British_filtered	Superior_corona_radiata_L	0.0043	0.000688915	0.0282442	0.0275552	-2.09237	0.611295	0.000621181	5635	3.2067818365
all_FA_skeletonised_British_filtered	Posterior_corona_radiata_R	0.00255	0.000683714	0.0186061	0.0179224	2.22744	0.656457	0.000692631	3766	3.1594980747
all_OD_skeletonised_British_filtered	Superior_corona_radiata_R	0.0059	0.000677137	0.0175278	0.0168507	-2.40302	0.712026	0.000740096	7238	3.130711943
all_OPENING_scaled	FPO_left	0.0003	0.000749042	0.173958	0.173209	14.7599	4.46645	0.000953877	717	3.0205076228
all_OPENING_scaled	FCLa_left	0.00015	0.000794561	0.110717	0.109923	14.9474	4.55864	0.00104501	426	2.9808795536
all_MO_skeletonised_British_filtered	Superior_fronto-occipital_fasciculus-part_of_anterior_internal_capsule-L	0.0068	0.000622903	0.0119846	0.0113617	-11.2664	3.49037	0.00124971	8103	2.903190755

# Summary file fields

1. Phenotype - Name of Phenotype
2. Threshold - Best P-value Threshold
3. PRS.R2 - Variance explained by the PRS. If prevalence is provided, this will be adjusted for ascertainment
4. Full.R2 - Variance explained by the full model (including the covariates). If prevalence is provided, this will be adjusted for ascertainment
5. Null.R2 - Variance explained by the covariates. If prevalence is provided, this will be adjusted for ascertainment
6. Coefficient - Regression coefficient of the model. Can provide insight of the direction of effect.
7. P - P value of the model fit
8. Num\_SNP - Number of SNPs included in the model



# Results ASD - HCP - 10 best

Measure	Phenotype	Threshold	PRS.R2	Full.R2	Null.R2	Coefficient	Standard.Error	P	Num_SNP	-log(P)
connectivity	ctx-lh-inferiorparietal_ctx-lh-insula	0.00945	0.0342342	0.0865838	0.0523495	-270616	44789.7	2.15839E-09	4898	8.6658700798
connectivity	ctx-lh-lateraloccipital_ctx-lh-insula	0.007	0.023263	0.0527861	0.0295231	-239228	48912.9	1.17215E-06	3839	5.9310168081
connectivity	ctx-lh-superiorfrontal_ctx-rh-precuneus	0.00475	0.0219392	0.0381945	0.0162553	157813	33480.8	2.7856E-06	2852	5.5550812462
connectivity	ctx-lh-lingual_ctx-lh-insula	0.00725	0.0201929	0.0604258	0.0402328	-149236	32618.2	5.36453E-06	3939	5.2704683217
connectivity	ctx-rh-superiorparietal_ctx-rh-inferiorparietal	0.0015	0.0191118	0.0339563	0.0148445	3631020	827174	1.25805E-05	1179	4.9003020979
connectivity	ctx-lh-insula_ctx-rh-lateraloccipital	0.00755	0.0175908	0.103188	0.0855976	-10040	2297	1.36894E-05	4082	4.8636155864
connectivity	ctx-rostralmiddlefrontal_ctx-lh-transversetemporal	0.00025	0.0179471	0.0807235	0.0627764	2871.52	658.504	1.43329E-05	275	4.8436659292
connectivity	ctx-lh-fusiform_ctx-lh-insula	0.00715	0.0174557	0.0399539	0.0224981	-77336.2	18377.2	0.00002809	3897	4.5514482608
connectivity	ctx-lh-inferiortemporal_ctx-lh-insula	0.00725	0.0167913	0.0469514	0.03016	-67611.4	16321.3	3.73057E-05	3939	4.4282248065
connectivity	ctx-lh-insula_ctx-rh-inferiorparietal	0.02355	0.0143875	0.166016	0.151629	-40202.7	9807.58	4.48894E-05	10077	4.3478561994

# Results ASD - HCP tbbs

Measure	Phenotype	Threshold	PRS.R2	Full.R2	Null.R2	Coefficient	Standard.Error	P	Num_SNP	-log(P)
tbss	Superior_corona_radiata_L	0.00055	0.0100976	0.0688645	0.0587669	3.006	0.913737	0.00103738	514	2.9840621292
tbss	Inferior_cerebellar_peduncle_L	0.0235	0.00918049	0.112728	0.103548	27.5728	8.58047	0.00135321	10062	2.8686348015
tbss	Fornix_column_and_body_of_fornix	0.0065	0.00687537	0.223359	0.216483	18.392	6.18766	0.00302539	3618	2.5192186329
tbss	Fornix_cres_Stria_terminalis	0.00645	0.00569909	0.293819	0.28812	11.5079	4.05497	0.0046311	3606	2.3343158411
tbss	Inferior_cerebellar_peduncle_R	0.0235	0.00637028	0.096646	0.0902757	25.3177	9.54352	0.00810669	10062	2.0911564341
tbss	Anterior_corona_radiata_L	0.0007	0.00667817	0.0477165	0.0410384	2.73229	1.0328	0.00828367	630	2.0817772106
tbss	Superior_corona_radiata_R	0.0007	0.00615416	0.0850117	0.0788575	2.61507	1.00935	0.00971241	630	2.0126729926

# Comparison ASD between ukb and hcp

Measure	Phenotype	Threshold	PRS.R2	Full.R2	Null.R2	Coefficient	Standard.Error	P	Num_SNP	-log(P)
all_OD_skeletonised_British_filtered	Posterior_limb_of_internal_capsule_L	0.00035	0.00114441	0.0109551	0.00981069	-0.606922	0.138792	1.23369E-05	802	4.9087939555
all_OD_skeletonised_British_filtered	Posterior_limb_of_internal_capsule_R	0.00035	0.000992739	0.0113942	0.0104015	-0.562633	0.138113	4.64862E-05	802	4.3326759536
all_OD_skeletonised_British_filtered	Cingulum-hippocampus-L	0.0008	0.000806346	0.0298918	0.0290854	-1.56622	0.422588	0.000211038	1538	3.6756393376
all_OD_skeletonised_British_filtered	Posterior_corona_radiata_R	0.00315	0.000747479	0.0806698	0.0799224	-1.77283	0.483638	0.00024751	4458	3.6064072498
all_OD_skeletonised_British_filtered	Superior_corona_radiata_L	0.0043	0.000688915	0.0282442	0.0275552	-2.09237	0.611295	0.000621181	5635	3.2067818365
all_FA_skeletonised_British_filtered	Posterior_corona_radiata_R	0.00255	0.000683714	0.0186061	0.0179224	2.22744	0.656457	0.000692631	3766	3.1594980747
all_OD_skeletonised_British_filtered	Superior_corona_radiata_R	0.0059	0.000677137	0.0175278	0.0168507	-2.40302	0.712026	0.000740096	7238	3.130711943
all_OPENING_scaled	FPO_left	0.0003	0.000749042	0.173958	0.173209	14.7599	4.46645	0.000953877	717	3.0205076228
all_OPENING_scaled	FCLa_left	0.00015	0.000794561	0.110717	0.109923	14.9474	4.55864	0.00104501	426	2.9808795536
all_MO_skeletonised_British_filtered	Superior_fronto-occipital_fasciculus-part_of_anterior_internal_capsule-L	0.0068	0.000622903	0.0119846	0.0113617	-11.2664	3.49037	0.00124971	8103	2.903190755

Measure	Phenotype	Threshold	PRS.R2	Full.R2	Null.R2	Coefficient	Standard.Error	P	Num_SNP	-log(P)
tbss	Superior_corona_radiata_L	0.00055	0.0100976	0.0688645	0.0587669	3.006	0.913737	0.00103738	514	2.9840621292
tbss	Inferior_cerebellar_peduncle_L	0.0235	0.00918049	0.112728	0.103548	27.5728	8.58047	0.00135321	10062	2.8686348015
tbss	Fornix_column_and_body_of_fornix	0.0065	0.00687537	0.223359	0.216483	18.392	6.18766	0.00302539	3618	2.5192186329
tbss	Fornix_cres_Stria_terminalis	0.00645	0.00569909	0.293819	0.28812	11.5079	4.05497	0.0046311	3606	2.3343158411
tbss	Inferior_cerebellar_peduncle_R	0.0235	0.00637028	0.096646	0.0902757	25.3177	9.54352	0.00810669	10062	2.0911564341
tbss	Anterior_corona_radiata_L	0.0007	0.00667817	0.0477165	0.0410384	2.73229	1.0328	0.00828367	630	2.0817772106
tbss	Superior_corona_radiata_R	0.0007	0.00615416	0.0850117	0.0788575	2.61507	1.00935	0.00971241	630	2.0126729926

# Results ALZ - UKB

Measure	Phenotype	Threshold	PRS.R2	Full.R2	Null.R2	Coefficient	Standard.Error	P	Num_SNP	-log(P)
all_FA_skeletonised_British_filtered	Posterior_thalamic_radiation-include_optic_radiation-R	0.0001	0.000741209	0.114477	0.113736	-0.36497	0.0981302	0.000200483	296	3.6979224476
all_FA_skeletonised_British_filtered	Middle_cerebellar_peduncle	0.0001	0.000637054	0.0638557	0.0632187	-0.231716	0.0690965	0.000799705	296	3.0970701886
all_ISOVF_skeletonised_British_filtered	Cingulum-cingulate_gyrus-L	0.0001	0.000677765	0.00450687	0.00382911	0.185022	0.0551827	0.000801495	296	3.0960991826
all_FA_skeletonised_British_filtered	Tapetum_L	0.0001	0.000631319	0.0336804	0.0330491	-0.654907	0.199311	0.00101886	296	2.9918854876
all_MD_skeletonised_British_filtered	Posterior_thalamic_radiation-include_optic_radiation-R	0.0001	0.000601853	0.0787606	0.0781588	0.00046536	0.000141627	0.001019	296	2.991825816
all_OPENING_scaled	SOTlatpost_right	0.00785	0.000775043	0.102207	0.101432	-23.8982	7.47411	0.00139006	8042	2.8569664536
all_OD_skeletonised_British_filtered	Genu_of_corpus_callosum	0.0123	0.000576706	0.0115436	0.0109668	1.20249	0.387258	0.00190504	11371	2.720095901
all_ISOVF_skeletonised_British_filtered	Cingulum-cingulate_gyrus-R	0.0001	0.000579274	0.00537445	0.00479517	0.165904	0.0534988	0.00193145	296	2.7141165301
all_FA_skeletonised_British_filtered	Posterior_thalamic_radiation-include_optic_radiation-L	0.0001	0.000512487	0.102163	0.10165	-0.311422	0.101396	0.00213452	296	2.6706997716
all_OD_skeletonised_British_filtered	Superior_fronto-occipital_fasciculus-part_of_anterior_internal_capsule-R	0.00015	0.00053954	0.00736447	0.00682493	-0.361406	0.120585	0.00272969	394	2.5638866713



# Results ALZ - HCP

Measure	Phenotype	Threshold	PRS.R2	Full.R2	Null.R2	Coefficient	Standard.Error	P	Num_SNP	-log(P)
connectivity	ctx-lh-lateralorbitofrontal_ctx-rh-postcentral	0.0012	0.021627	0.0255439	0.00391685	-363.599	78.2035	3.78294E-06	840	5.4221705469
tbss	Middle_cerebellar_peduncle	0.00035	0.0185839	0.123885	0.105301	-2.13267	0.463521	4.74023E-06	314	5.3242005855
connectivity	ctx-lh-insula_Left-Pallidum	0.00225	0.0182191	0.107413	0.0891941	6780.73	1520.75	9.19152E-06	1391	5.0366126635
connectivity	ctx-lh-superiorfrontal_Right-Hippocampus	0.0042	0.017488	0.113397	0.0959089	5321.89	1214.18	1.29568E-05	2298	4.8875022449
connectivity	ctx-lh-superiorfrontal_ctx-lh-transversetemporal	0.0001	0.0179883	0.0620248	0.0440365	918.391	212.495	1.70389E-05	157	4.7685584459
connectivity	ctx-rh-rostralmiddlefrontal_Right-Putamen	0.01965	0.0168002	0.0857456	0.0689454	797263	188451	2.54833E-05	7978	4.593744333
morphologist	surface_native	0.0045	0.0162071	0.0451997	0.0289926	-50382.9	12032	3.05991E-05	2421	4.5142913471
connectivity	ctx-lh-parsorbitalis_ctx-rh-lateraloccipital	0.02005	0.016278	0.0779492	0.0616712	400.934	96.6874	0.000036643	8108	4.4360089774
connectivity	ctx-rh-superiortemporal_Right-Putamen	0.01105	0.0162274	0.0700721	0.0538448	-146194	35460.9	4.06114E-05	4983	4.3913520388
tbss	Anterior_corona_radiata_L	0.00485	0.0156967	0.056735	0.0410384	6.8374	1.6778	4.95863E-05	2563	4.3046382964

# Comparison ALZ between ukb and hcp

Measure	Phenotype	Threshold	PRS.R2	Full.R2	Null.R2	Coefficient	Standard.Error	P	Num_SNP	-log(P)
all_FA_skeletonised_British_filtered	Posterior_thalamic_radiation-include_optic_radiation-R	0.0001	0.000741209	0.114477	0.113736	-0.36497	0.0981302	0.000200483	296	3.6979224476
all_FA_skeletonised_British_filtered	Middle_cerebellar_peduncle	0.0001	0.000637054	0.0638557	0.0632187	-0.231716	0.0690965	0.000799705	296	3.0970701886
all_ISOVF_skeletonised_British_filtered	Cingulum-cingulate_gyrus-L	0.0001	0.000677765	0.00450687	0.00382911	0.185022	0.0551827	0.000801495	296	3.0960991826
all_FA_skeletonised_British_filtered	Tapetum_L	0.0001	0.000631319	0.0336804	0.0330491	-0.654907	0.199311	0.00101886	296	2.9918854876
all_MD_skeletonised_British_filtered	Posterior_thalamic_radiation-include_optic_radiation-R	0.0001	0.000601853	0.0787606	0.0781588	0.00046536	0.000141627	0.001019	296	2.991825816
all_OPENING_scaled	SOTlatpost_right	0.00785	0.000775043	0.102207	0.101432	-23.8982	7.47411	0.00139006	8042	2.8569664536
all_OD_skeletonised_British_filtered	Genu_of_corpus_callosum	0.0123	0.000576706	0.0115436	0.0109668	1.20249	0.387258	0.00190504	11371	2.720095901
all_ISOVF_skeletonised_British_filtered	Cingulum-cingulate_gyrus-R	0.0001	0.000579274	0.00537445	0.00479517	0.165904	0.0534988	0.00193145	296	2.7141165301
all_FA_skeletonised_British_filtered	Posterior_thalamic_radiation-include_optic_radiation-L	0.0001	0.000512487	0.102163	0.10165	-0.311422	0.101396	0.00213452	296	2.6706997716
all_OD_skeletonised_British_filtered	Superior_fronto-occipital_fasciculus-part_of_anterior_internal_capsule-R	0.00015	0.00053954	0.00736447	0.00682493	-0.361406	0.120585	0.00272969	394	2.5638866713

Measure	Phenotype	Threshold	PRS.R2	Full.R2	Null.R2	Coefficient	Standard.Error	P	Num_SNP	-log(P)
connectivity	ctx-lh-lateralorbitofrontal_ctx-rh-postcentral	0.0012	0.021627	0.0255439	0.00391685	-363.599	78.2035	3.78294E-06	840	5.4221705469
tbss	Middle_cerebellar_peduncle	0.00035	0.0185839	0.123885	0.105301	-2.13267	0.463521	4.74023E-06	314	5.3242005855
connectivity	ctx-lh-insula_Left-Pallidum	0.00225	0.0182191	0.107413	0.0891941	6780.73	1520.75	9.19152E-06	1391	5.0366126635
connectivity	ctx-lh-superiorfrontal_Right-Hippocampus	0.0042	0.017488	0.113397	0.0959089	5321.89	1214.18	1.29568E-05	2298	4.8875022449
connectivity	ctx-lh-superiorfrontal_ctx-lh-transversetemporal	0.0001	0.0179883	0.0620248	0.0440365	918.391	212.495	1.70389E-05	157	4.7685584459
connectivity	ctx-rh-rostralmiddlefrontal_Right-Putamen	0.01965	0.0168002	0.0857456	0.0689454	797263	188451	2.54833E-05	7978	4.593744333
morphologist	surface_native	0.0045	0.0162071	0.0451997	0.0289926	-50382.9	12032	3.05991E-05	2421	4.5142913471
connectivity	ctx-lh-parsorbitalis_ctx-rh-lateraloccipital	0.02005	0.016278	0.0779492	0.0616712	400.934	96.6874	0.000036643	8108	4.4360089774
connectivity	ctx-rh-superiortemporal_Right-Putamen	0.01105	0.0162274	0.0700721	0.0538448	-146194	35460.9	4.06114E-05	4983	4.3913520388
tbss	Anterior_corona_radiata_L	0.00485	0.0156967	0.056735	0.0410384	6.8374	1.6778	4.95863E-05	2563	4.3046382964

# Problem

How to account for multiple testing?

One solution proposed by PRSice is permutation.

# Summary file fields

1. Phenotype - Name of Phenotype
2. Set - Name of Gene Set
3. Threshold - Best P-value Threshold
4. PRS.R2 - Variance explained by the PRS. If prevalence is provided, this will be adjusted for ascertainment
5. Full.R2 - Variance explained by the full model (including the covariates). If prevalence is provided, this will be adjusted for ascertainment
6. Null.R2 - Variance explained by the covariates. If prevalence is provided, this will be adjusted for ascertainment
7. Prevalence - Population prevalence as indicated by the user. "-" if not provided.
8. Coefficient - Regression coefficient of the model. Can provide insight of the direction of effect.
9. P - P value of the model fit
10. Num\_SNP - Number of SNPs included in the model
11. Empirical-P - Only provided if permutation is performed. This is the empirical p-value and should account for multiple testing and over-fitting



# Permutation option

- To account for multiple testing and over-fitting

- `--perm`

Number of permutation to perform. This will generate the empirical p-value. Recommend to use value larger than or equal to 10,000

## **Note**

When permutation is required, PRSice will perform the following operation

1. Perform normal PRSice across all thresholds and obtain p-value of the most significant threshold
2. Repeat PRSice analysis  $N$  times with permuted phenotype. Count the number of time where the p-value of the most significant threshold for the permuted

# Correct for multiple testing

Other options:

- Bonferroni: divide by number of phenotypes
  - 48 tracts \* 6 measures (FA,ICVF,ISOVF,MD,MO,OD) = 288 TBBS,
  - 129 opening
  - 6972 connectivity
  - 70 network measures
  - 130 morphologist
- FDR

# Visualization of significant tracts

- We will work on FSL visualization with Antoine
- But problem of triplanar representation: they won't possibly be on the same plan

# JOBIM abstract

## Outline:

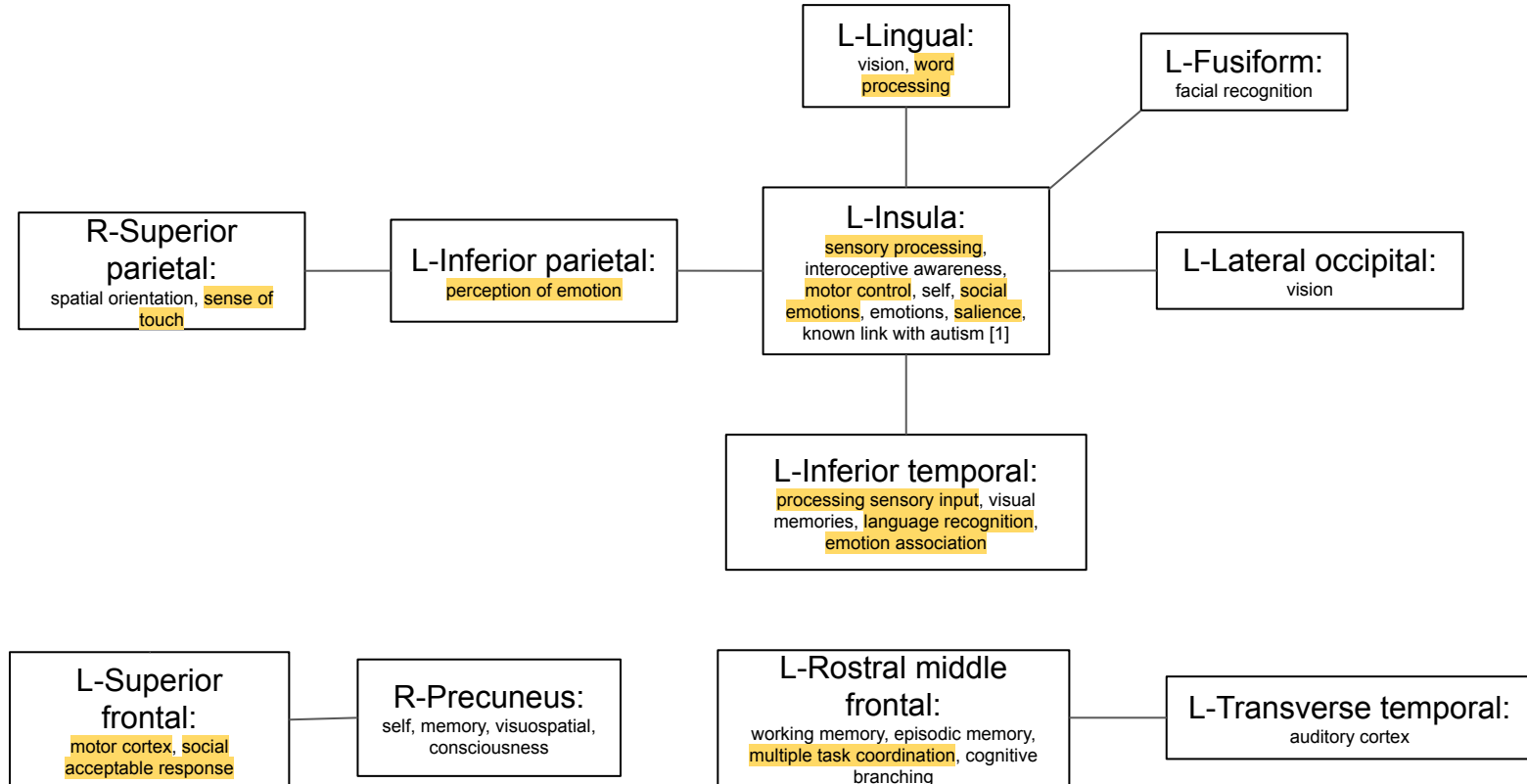
- Present PRS method
- Focus on ASD (ALZ in perspective)
- Figures:
  - a table with 10 most significant phenotypes associated with ASD-PRS (UKB and HCP)
  - A triplanar image of the most significant tract in ASD (corona radiata, only one common between UKB and HCP)
- Discussion: the role of this/these tract(s) in literature, how it make sense for autism

# Polygenic Risk Score project

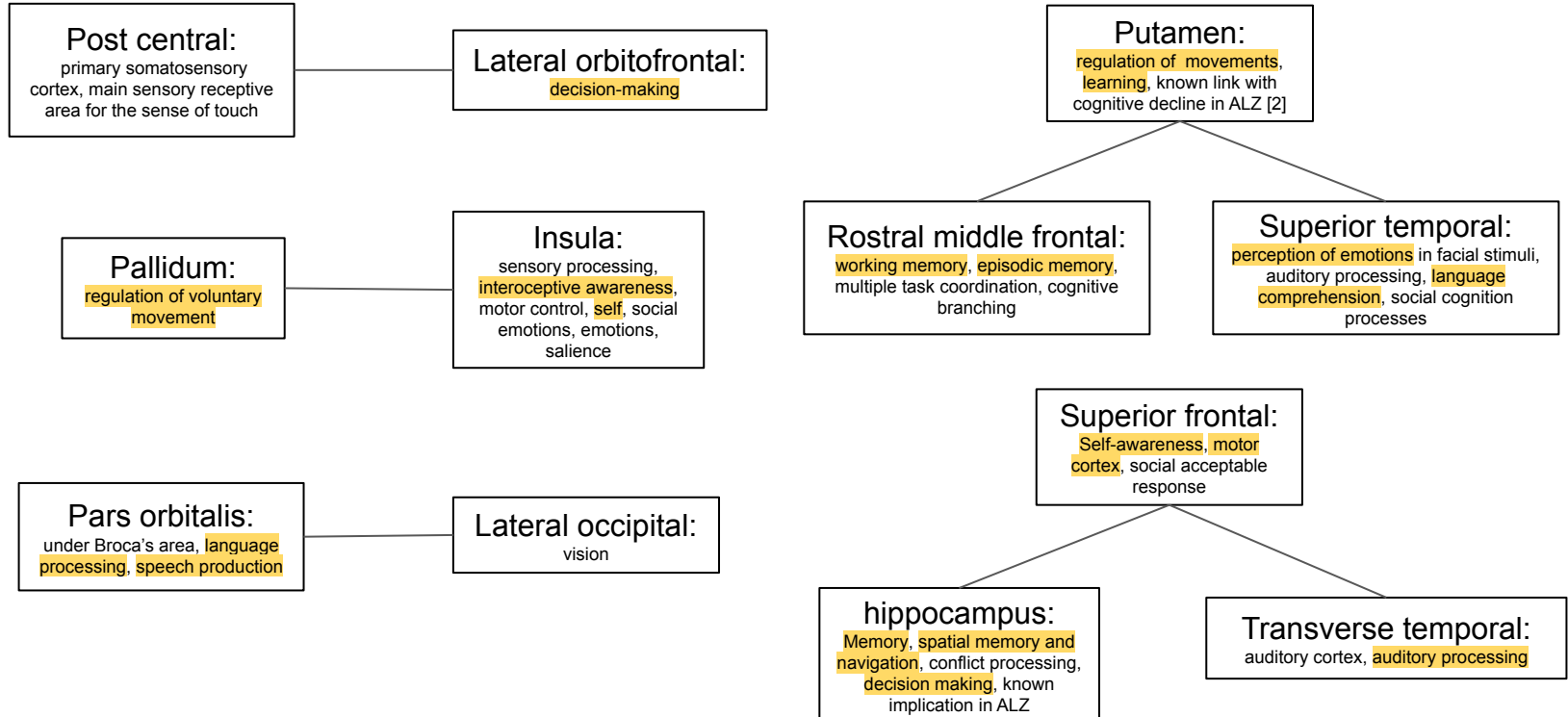
**Week 7**

2019-04-17

# Results Cortical Connectivity Autism



# Results Cortical Connectivity Alzheimer



# Things to do

- Retrieve results for SULCUS OPENING in HCP  
(only phenotype in common between HCP and UKB besides TBBS)
- Run imputation on HCP genotype data
- Compute  $h^2$  for TBBS and OPENING phenotypes in HCP with SOLAR
- and  $h^2$  in UKB with GCTA
- Compute shared  $h^2$  between TBBS tracts
- and shared  $h^2$  between TBBS tracts and Connectivity



# Polygenic Risk Score project

**Week 8**

2019-05-07

# PRSize data accessibility

For UKB PRSize data in :

/neurospin/ukb/derivatives/brainomics\_prs



For HCP PRSize data in:

/neurospin/ukb/analysis/brainomocs\_prs/

# PRSice data accessibility

For UKB PRSice data in :

1. 'GWAS' dir contains summary statistics for 'ALZHEIMER' and 'AUTISM' and clumped list of snps (SNPs in LD removed)
2. 'PHENOTYPES' dir contains 'OPENING' and 'TBSS' phenotypes formatted for PRSice
3. 'PRS' dir contains PRSice analysis results for > 'ALZHEIMER' and 'AUTISM' > source data > phenotype >
  - a. ".best" = all PRS/subject for one phenotype (tract or sulcus) ;
  - b. ".summary" = best PRS for all phenotypes (all tracts/sulcus in one summary file)

Same organisation for HCP in /neurospin/HCP

# PRSize script accessibility

- Need of GitLab repository to share PRSize script

# PRSice results: UKB opening

Phenotype	P	P_Bonferroni	P_FDR
SPaint_right	2.23E-05	2.75E-03	2.75E-03
SCLPC_right	4.36E-04	5.37E-02	2.16E-02
SPeCmedian_right	5.26E-04	6.47E-02	2.16E-02
SPoCsup_right	1.28E-03	1.58E-01	3.42E-02
SOTlatpost_right	1.39E-03	1.71E-01	3.42E-02
SForbitaire_left	2.01E-03	2.48E-01	3.95E-02
SPasup_right	2.32E-03	2.85E-01	3.95E-02
SPaint_left	3.08E-03	3.79E-01	3.95E-02
SPoCsup_left	3.34E-03	4.10E-01	3.95E-02
SC_left	3.49E-03	4.29E-01	3.95E-02
SLipost_right	3.54E-03	4.35E-01	3.95E-02
STipost_left	3.88E-03	4.77E-01	3.97E-02
FCLrretroCtr_left	4.20E-03	5.16E-01	3.97E-02
SPat_left	4.72E-03	5.80E-01	4.14E-02
FCMpost_right	5.15E-03	6.34E-01	4.22E-02

- Going to work on visualization
- `import statsmodels.stats.multitest as multi`
- `df['P_FDR'] = multi.multipletests(df.P, alpha=0.05, method='fdr_bh')[1]`

# PRSize results: HCP opening

- Need permission for /neurospin/hcp/ to access PRSize results