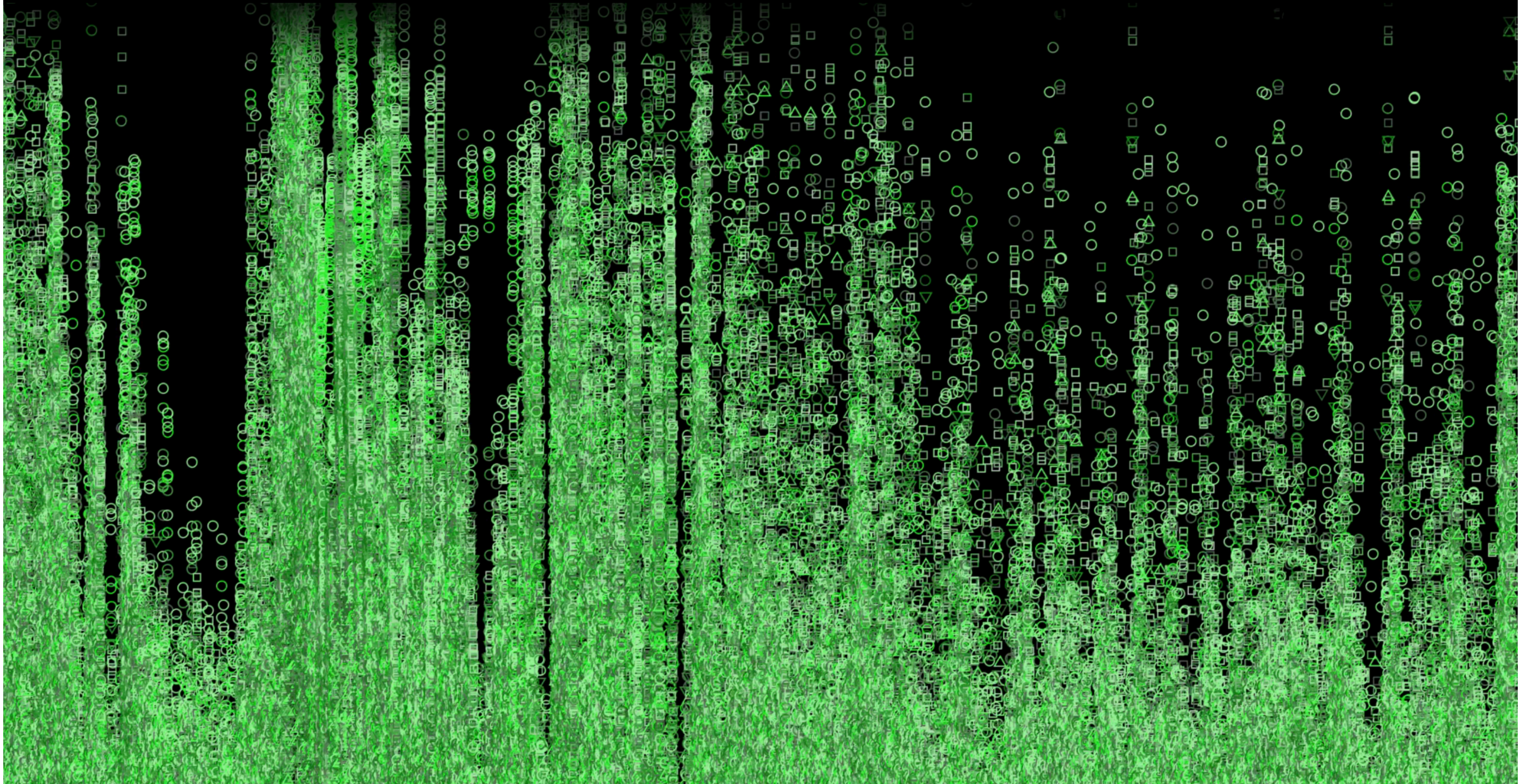
PANDORA UKBv1

Population Archive of Neuroimaging Data Organised for Rapid Analysis

FMRIB (OxCIN) Oxford



What is PANDORA UKBv1?

- **A massive archive of all UK Biobank brain imaging data, containing 82,000 subjects' standard space maps, from 99 sub-modalities, pre-processed to allow easy voxel-level regression against subject-wise variables such as genetics or lifestyle factors.**
- Hence for each sub-modality, (e.g., FA from dMRI), all subjects' standard space images are collected together into a huge but easy-to-use file containing one data matrix,
  e.g., of size (82K subjects x 2M voxels)
- The matrix has auto-QC applied: outlier voxels and subjects are removed
- This is much easier to work with than the traditional data store having separate packaging of files for each subject.
- We also provide a simple regression tool, allowing to regress the data matrix against any set of regressors of interest and confounds. It processes voxels in chunks in order to allow the full data matrix to be analysed without needing a compute node with huge RAM.
- We also provide a much more efficient **supervoxel** version of the full voxel matrix.

What are PANDORA supervoxels?

- In addition to the full data matrix (subjects x 2M voxels), we also provide a hugely reduced but almost lossless data matrix of (subjects x 1K supervoxels).
- The supervoxels representation is much smaller and faster to work with than the full voxels, but loses virtually no signal or spatial detail, while also achieving denoising.

What are the 3 spatial representations in PANDORA?

- **voxel** - the data from each subject is stored in full “raw” voxelwise form: 2M voxels for sub-modalities at 1mm resolution, 200K voxels for 2mm sub-modalities, and 90K greyordinates for fMRI sub-modalities.

- **ICA1K** - the voxelwise data is reduced to 1K supervoxels for each sub-modality, using 1000-dimensional spatial-ICA. This gives a huge reduction in the size of the subject-level representation, meaning that file sizes are much smaller, and RAM and compute time greatly reduced. In many scenarios, this compression loses almost no spatial detail, while the associated denoising results in improved statistical sensitivity when carrying out associations against other variables (compared with full voxelwise analysis). However, in some scenarios (depending on the sub-modality and the question of interest), some spatial detail can be lost, compared with analyses using voxel or ICA10K versions of PANDORA.

- **ICA10K** - the voxelwise data is reduced to 10K supervoxels. The resulting subject-level representation is therefore 10x larger than with ICA1K, and compute times when using this are closer to working with voxel data. We have not yet seen any cases where any spatial detail is lost (compared with voxel data), and in general the denoising means that statistical sensitivity is slightly better.

TLDR recommendation:

- **For very fast analyses, use ICA1K**
- **For the best/final analyses, use ICA10K. However, it may be worth also trying ICA1K, as this might give increased statistical sensitivity without losing spatial detail.**
- **But beware p-hacking!**

PANDORA preprocessing for full voxelwise matrix

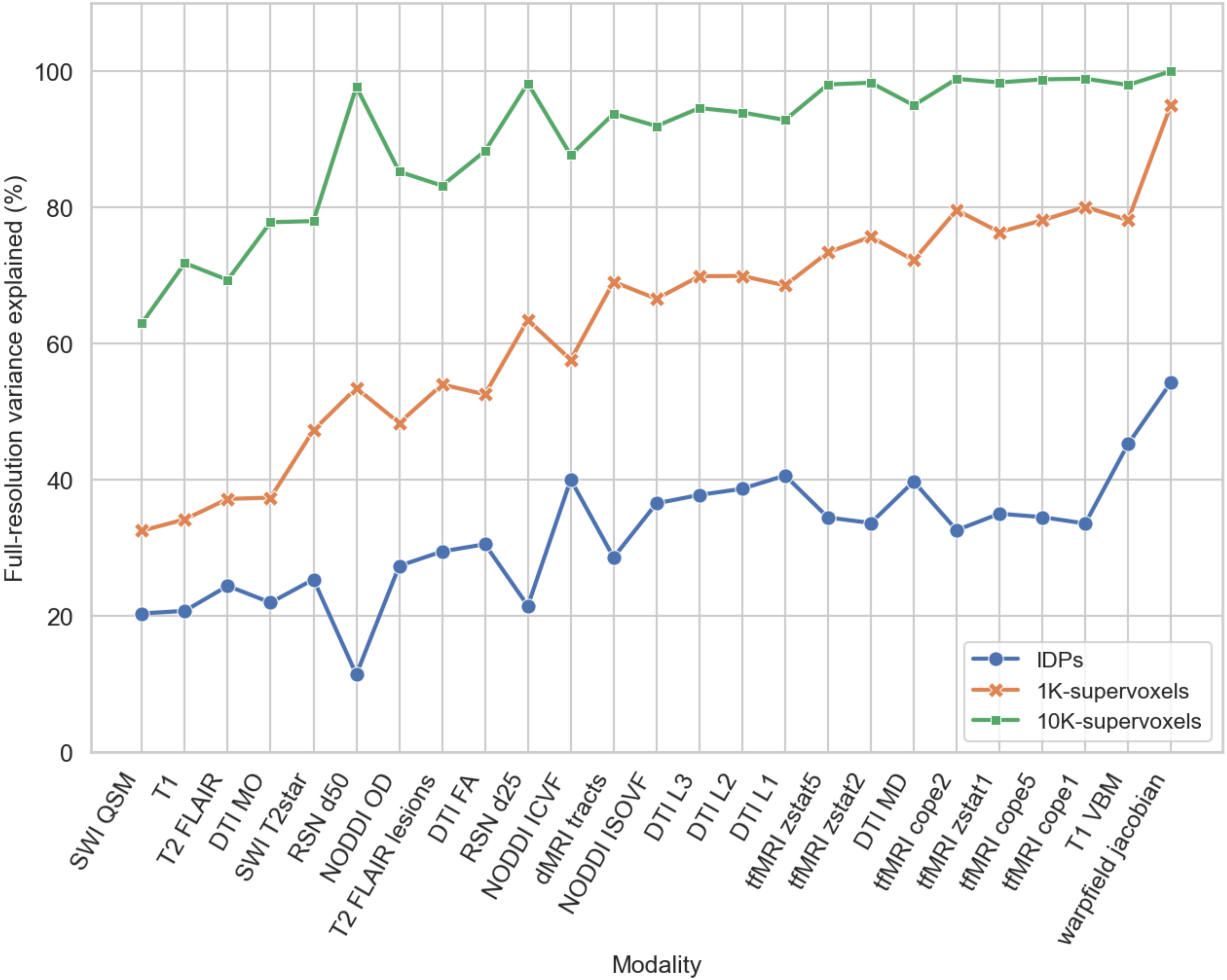
- For each sub-modality (raw T1, T1-VBM, T2*, DTI FA, DTI MO, etc.)
    - Take all subjects' maps in standard space and form the full data matrix, e.g., 80K subjects x 2M voxels
    - Each row is one subject's map (all voxels unwrapped into a long vector)
- Normalise each row if needed (most sub-modalities do not)
- Exclude columns (voxels) if non-brain, or too many outliers / missing data
- Exclude rows (subjects) if too many outliers / missing data

Maths behind supervoxels

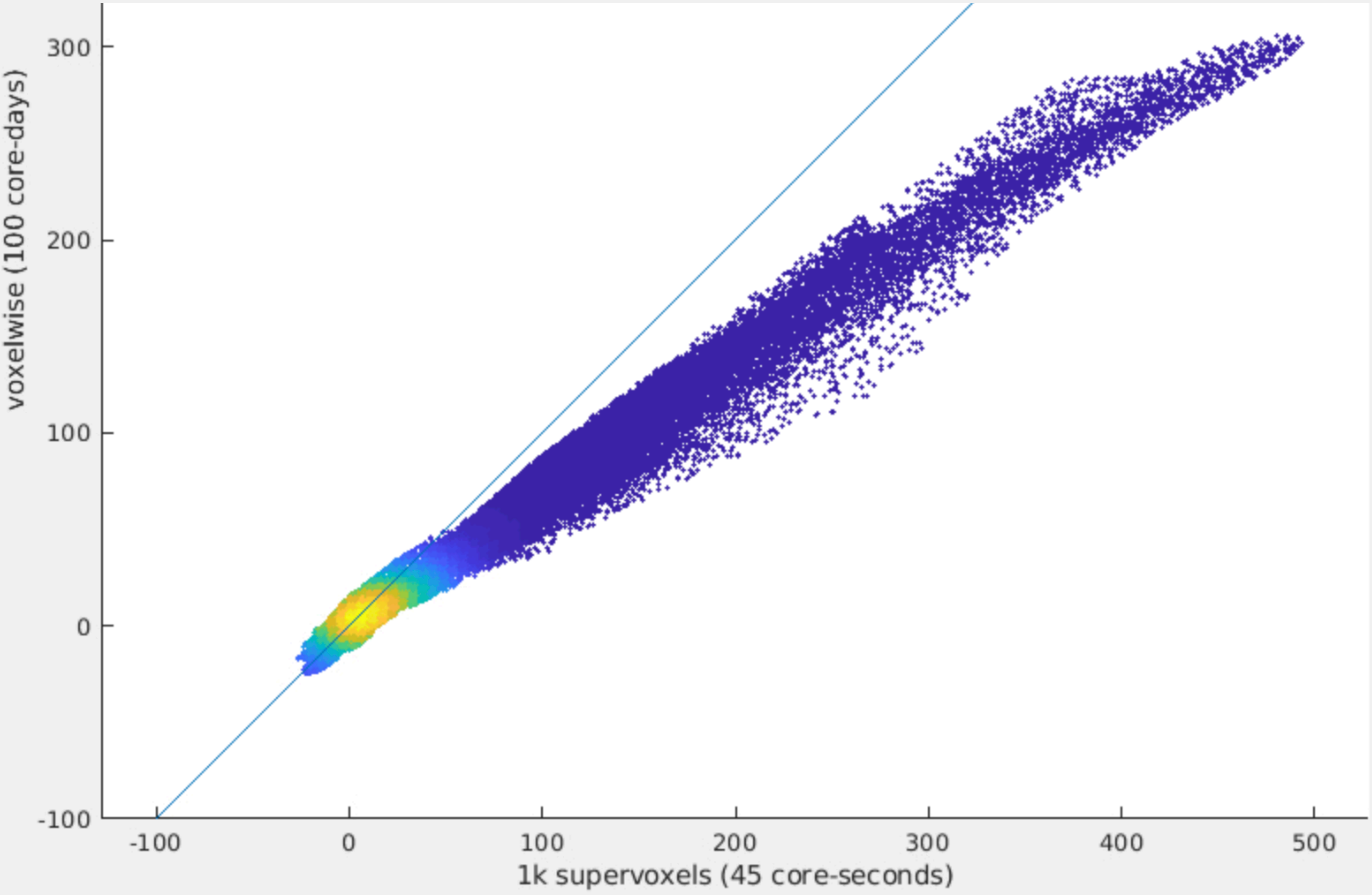
- The original full data matrix is for example Y = 82K subjects x 2M voxels
- We demean each row and reduce the matrix using spatial-ICA, to give a close approximation:
- Y = A (82K subjects x 1K supervoxels) x S (1K supervoxels x 2M voxels) + noise
- A is the matrix of subject weights used in regressions against other variables, e.g., age.
- S is the matrix of spatial maps - one map per supervoxel.
- Because of the reduction of full data to supervoxels, there is some denoising, which gives a boost in statistical sensitivity during the regression, while the very high dimensionality (e.g., 1K) means that almost no signal is lost.
- The regression primarily uses just A (hence it is very fast and low-RAM), but then takes S into account, giving an identical result to using the huge AxS matrix for the regressions:

  $Y = AS = X\beta + \epsilon \qquad \beta = (pinv(X)\ A)\ S$ etc.
- This gives a voxelwise regression map without ever forming massive matrices (e.g., Y*)*.

PANDORA supervoxels outperform IDPs



Scatterplot (one point per voxel) of association Z-statistic between T2* and age, showing increased sensitivity in supervoxels (N=50K)

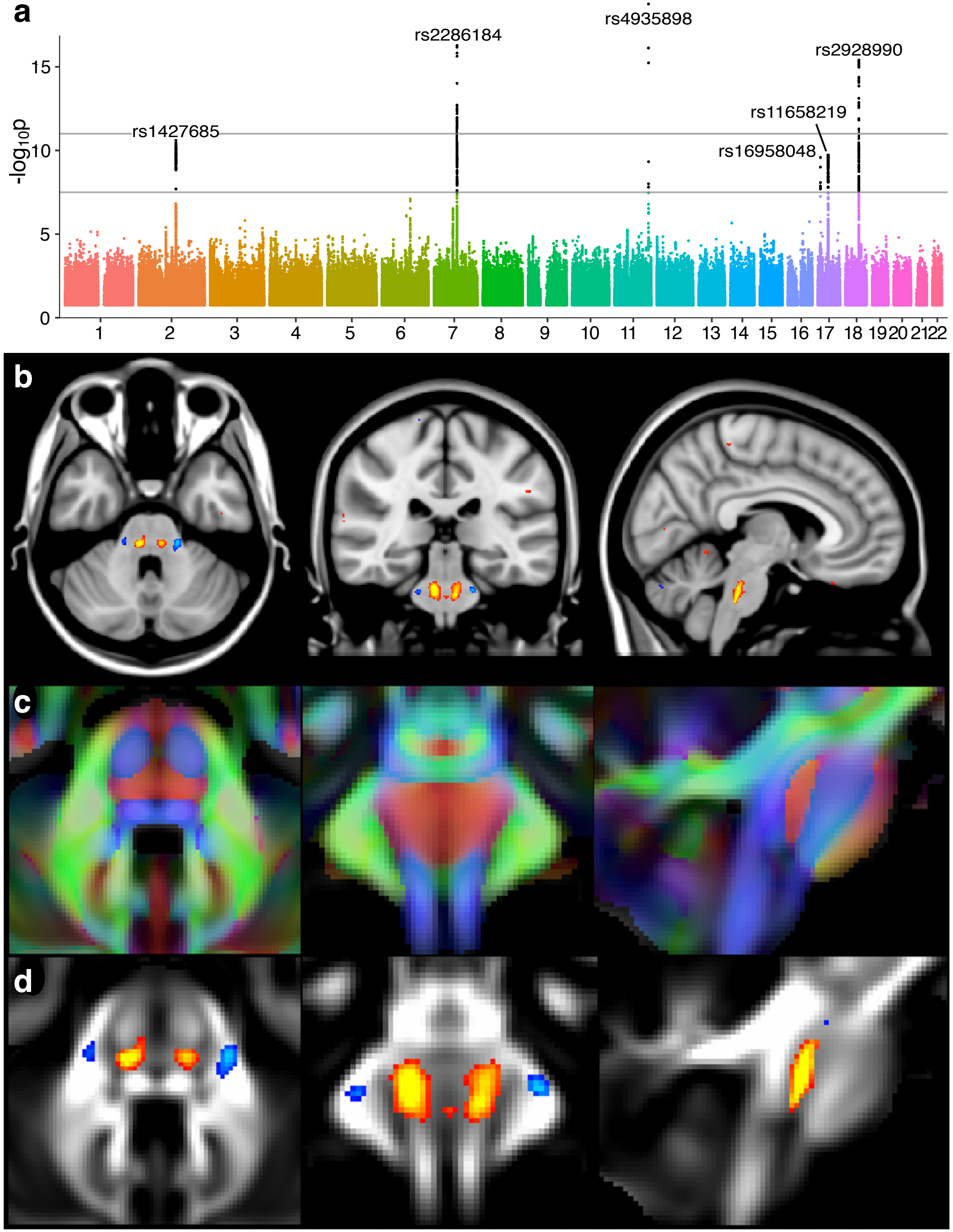


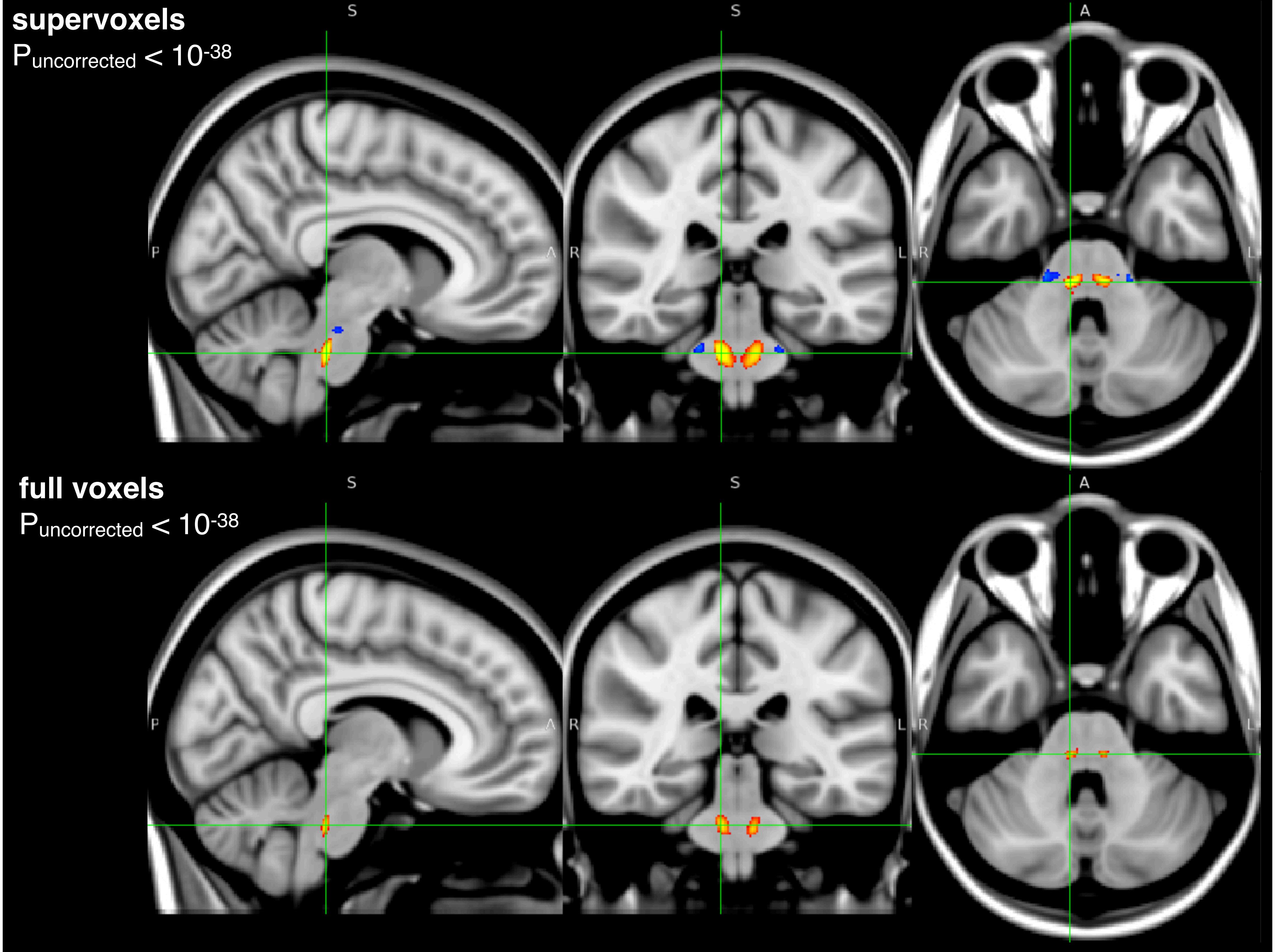
Original GWAS in Elliott 2018

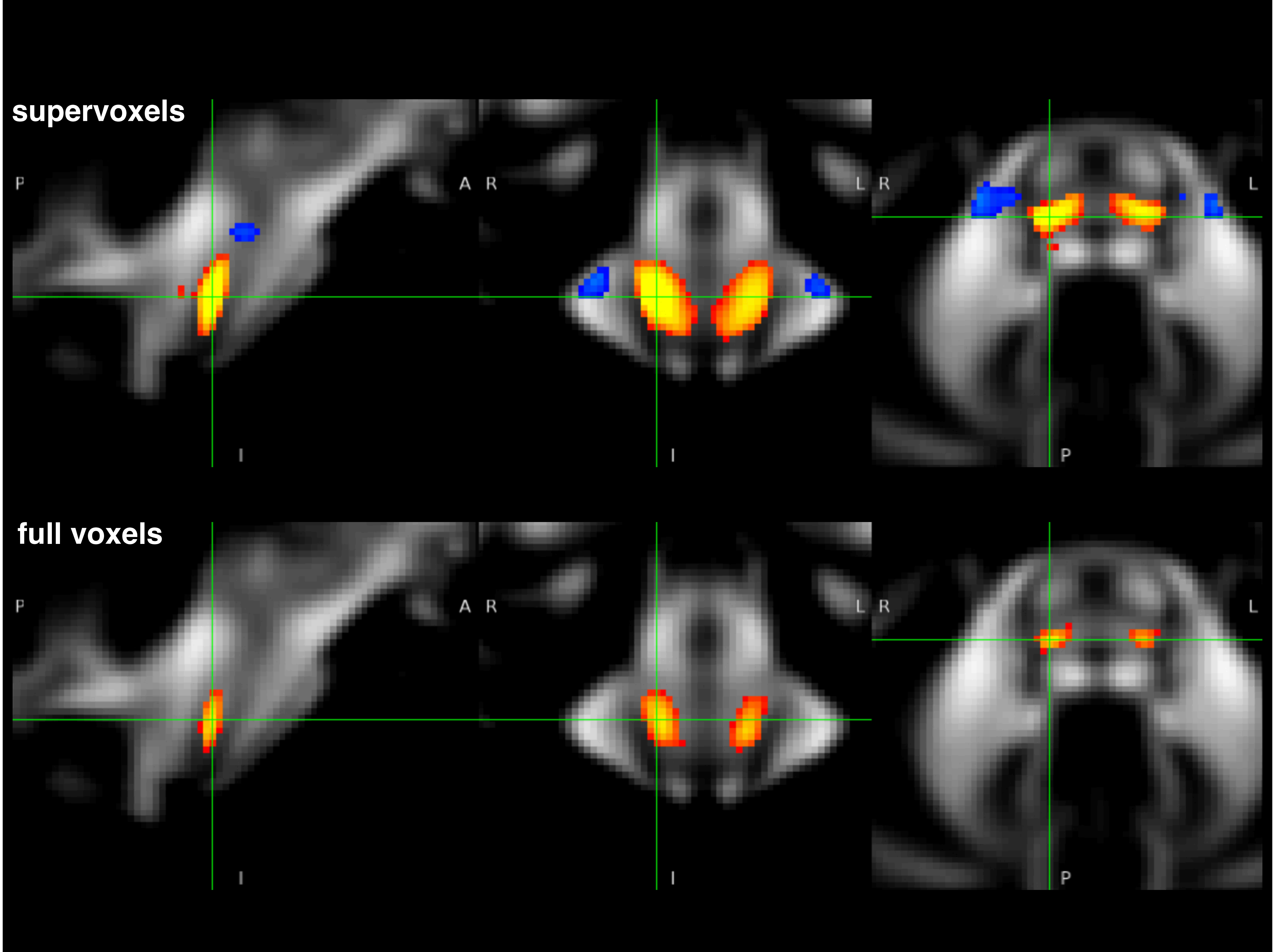
GWAS of dMRI tensor mode in crossing pontine tract

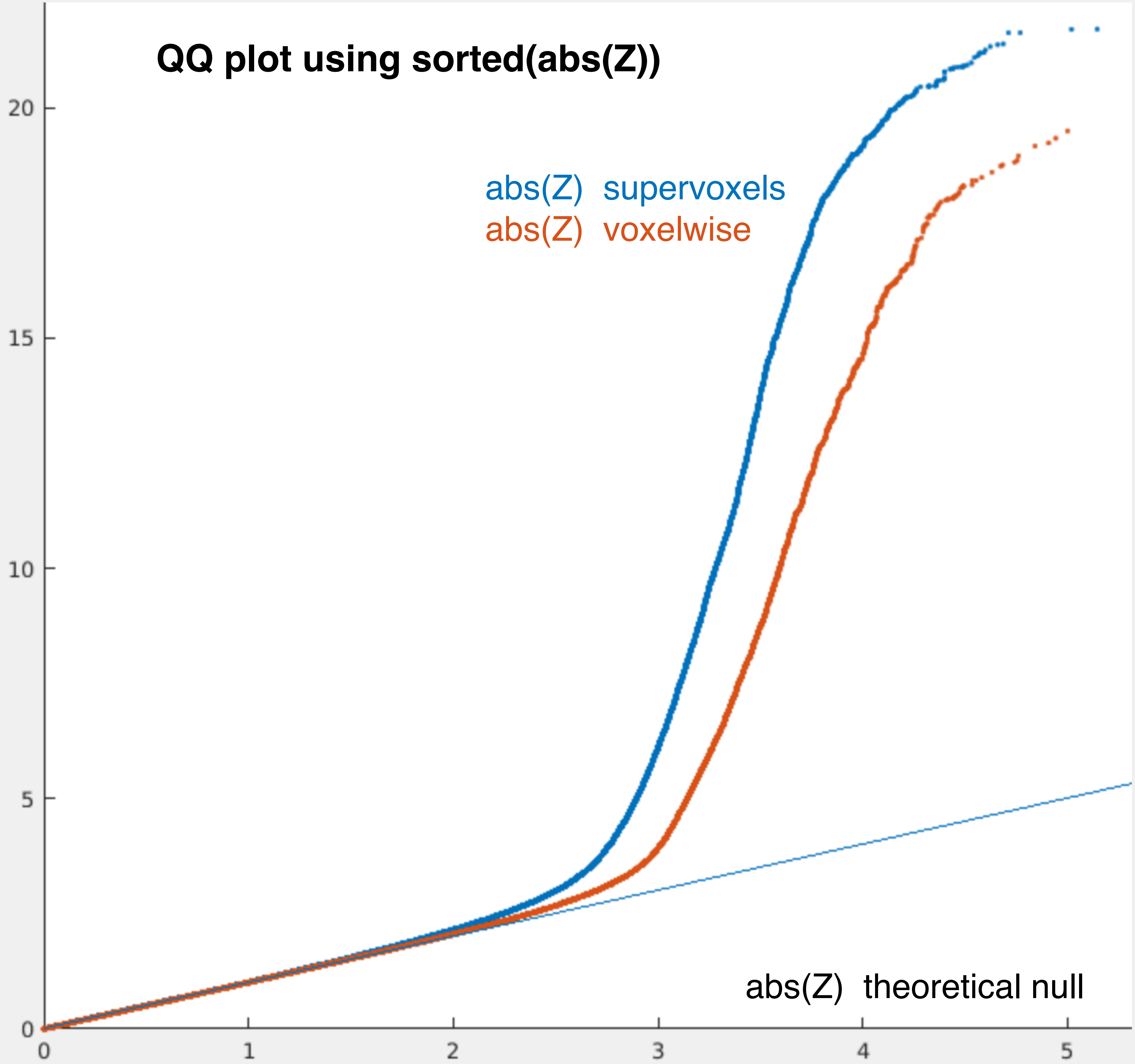
Voxelwise maps of MO associated with rs4935898

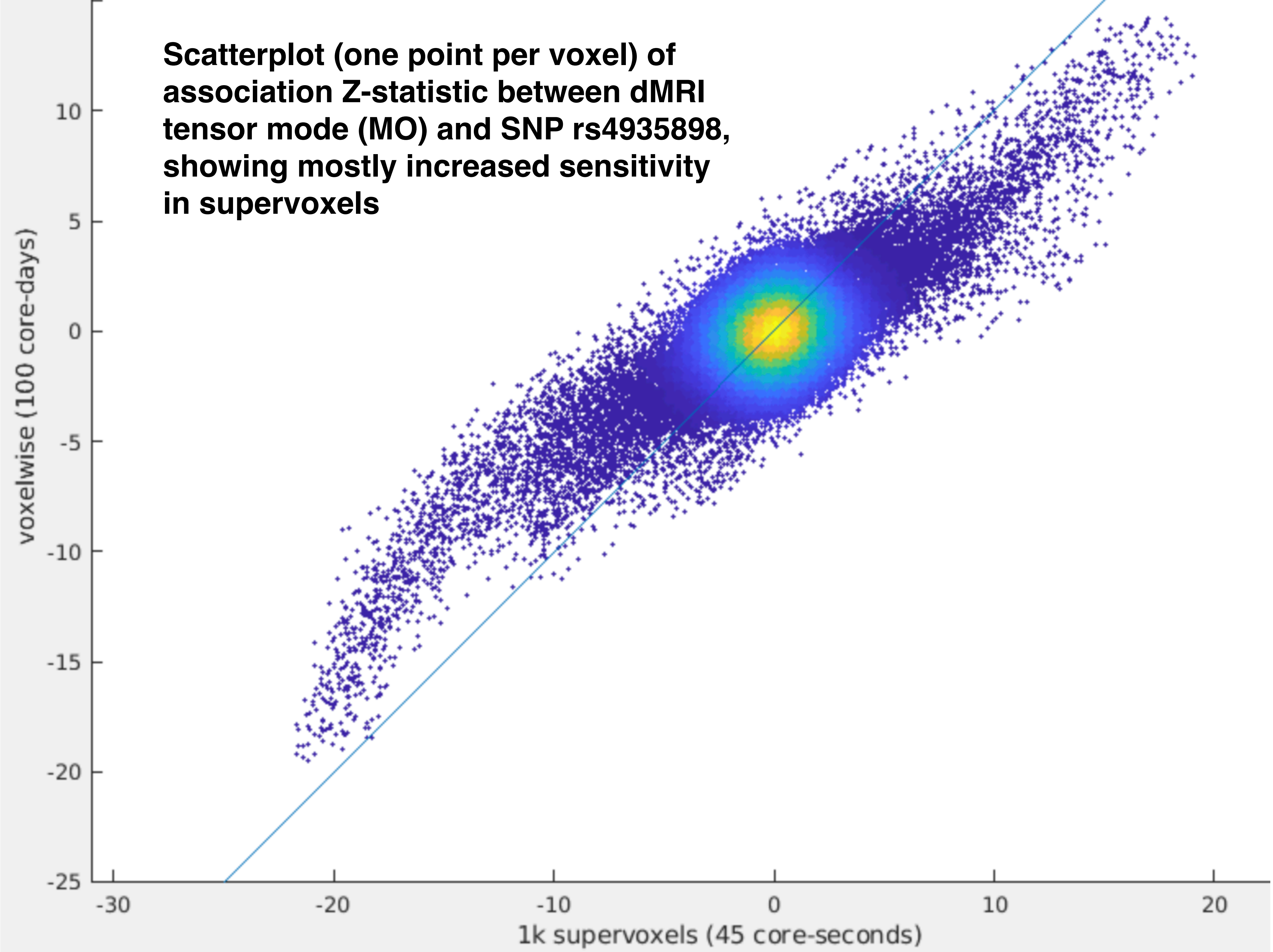
Genes SEMA3D & ROBO3 regulate axon development and are associated with horizontal gaze palsy, a disorder of these crossing axons




supervoxels
P_uncorrected < 10^-38
S
S
A
P
A
R
L
R
L
full voxels
P_uncorrected < 10^-38
S
S
A
P
A
R
L
R
L


supervoxels
P
A
R
L
R
L
I
I
P
full voxels
P
A
R
L
R
L
I
I
P

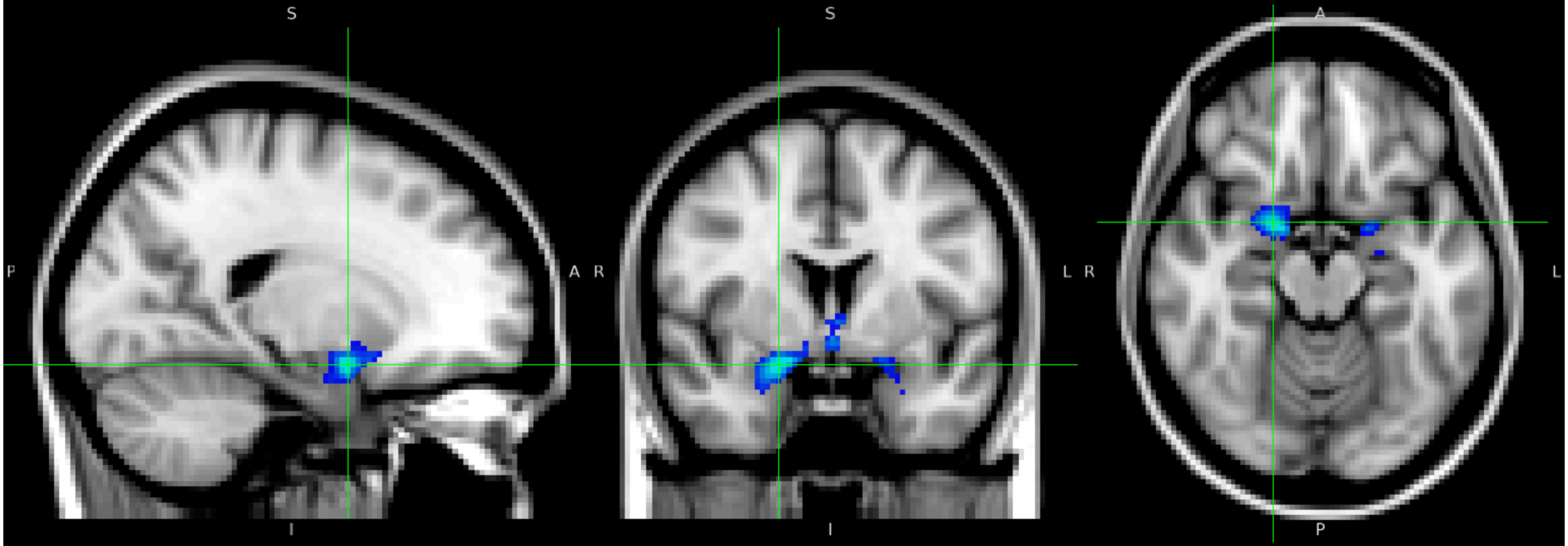

QQ plot using sorted(abs(Z))
abs(Z) supervoxels
abs(Z) voxelwise
abs(Z) theoretical null
20
15
10
5
0
0
1
2
3
4
5



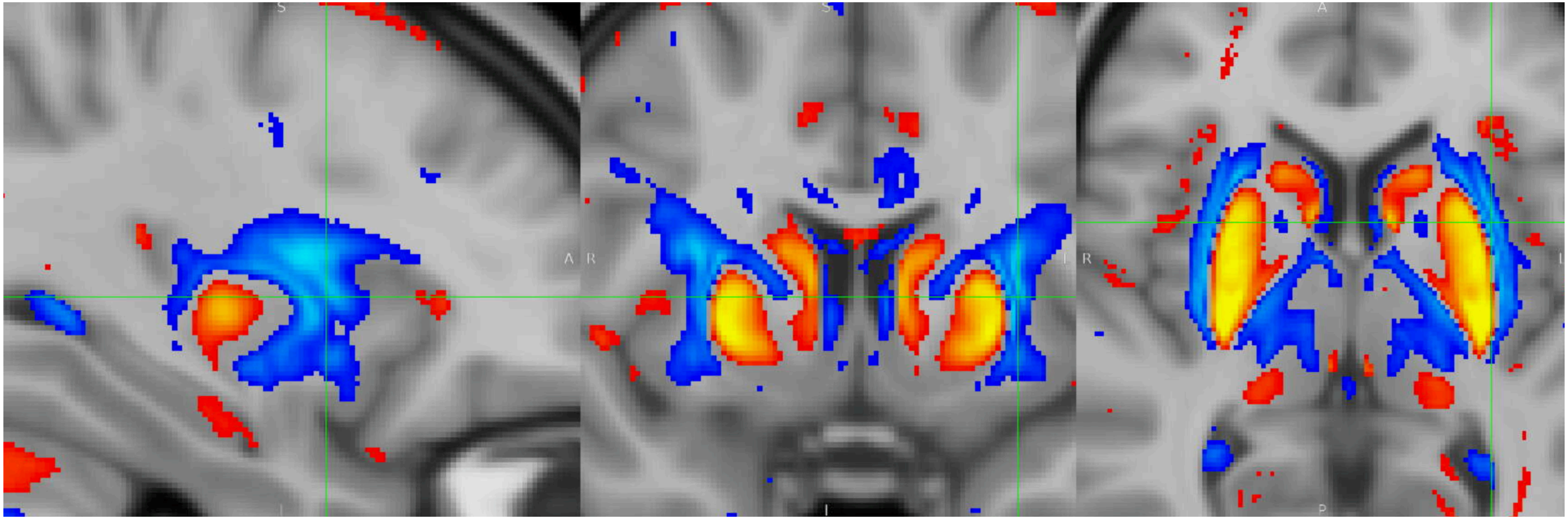
Scatterplot (one point per voxel) of association Z-statistic between dMRI tensor mode (MO) and SNP rs4935898, showing mostly increased sensitivity in supervoxels

T1_VBM associated with PRS for AD

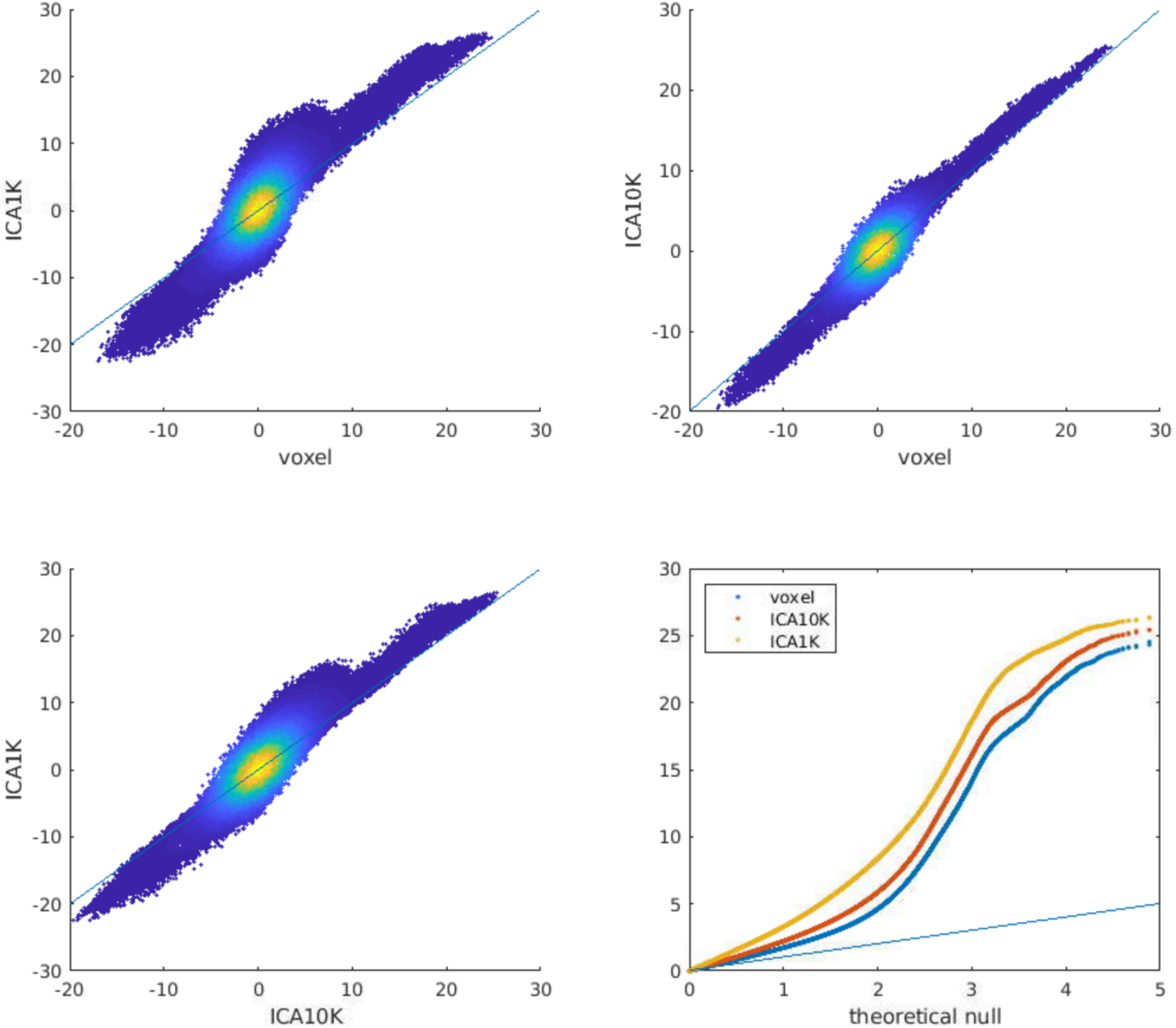
supervoxels $P_{uncorrected} < 10^{-12}$



Associating QSM with smoking history (N=70K)



Associating QSM with smoking history (N=70K)


ICA1K
voxel
ICA10K
voxel
ICA1K
ICA10K
voxel
ICA10K
ICA1K
theoretical null