DE LA RECHERCHE À L'INDUSTRIE

IMAGERIE GÉNÉTIQUE : DÉFIS COMPUTATIONNELS ET IMPLÉMENTATION GROS GRAINS SUR CLUSTER

Forum TERATEC | Vincent Frouin

Déluge de données

S Laguitton (CATI-CEA)
B daMota (INRIA-CEA)

28 JUIN 2012

NeuroSpin

www.cea.fr

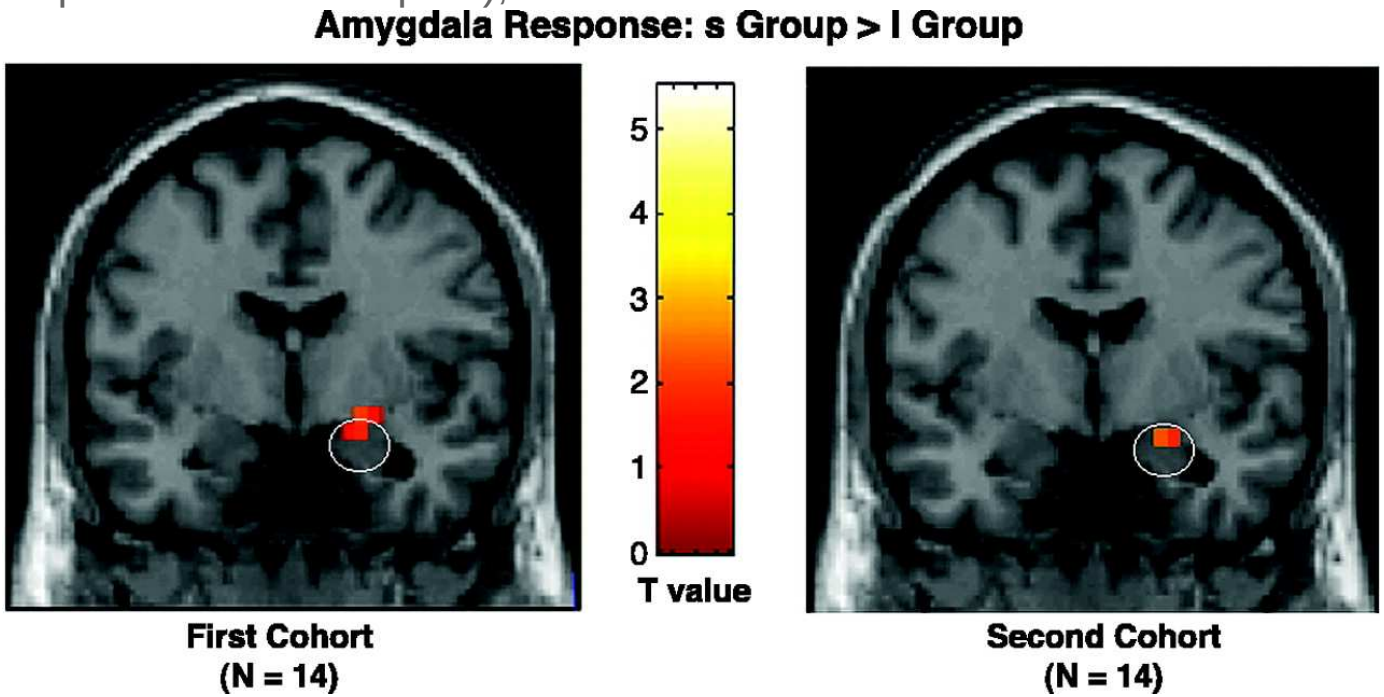
NEUROIMAGERIE-GÉNÉTIQUE :

CONTEXTE

NEUROIMAGERIE-GÉNÉTIQUE : MOTIVATIONS

Etude de groupe, basée sur le génotype

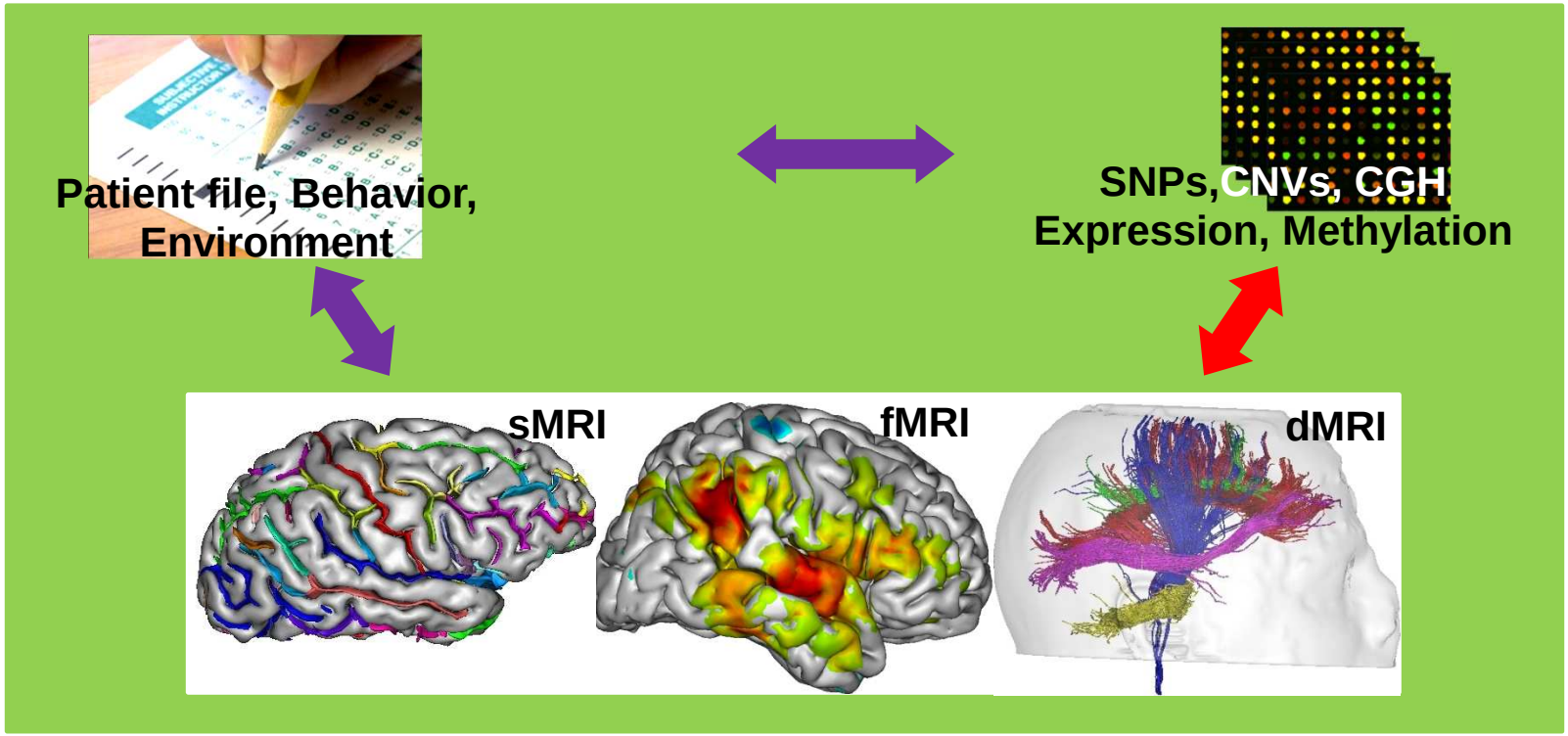
- L'étude inclut ~30 sujets
- 2 groupes : 5-HTT muté *vs* 5-HTT
- Activité plus grande dans l'amygdale à droite (structure impliquée dans le comportement lié à la peur),



A R Hariri et al. Science 2002;297:400-403

NEUROIMAGERIE-GÉNÉTIQUE : MOTIVATIONS (2)

- Un outil pour trouver de nouveaux biomarqueurs (via l'endo-phénotype d'imagerie)
- Un outil pour étudier des questions de recherche appliquée ou fondamentale
  - Integration des mesures obtenues à des échelles différentes
  - Révélateur des réseaux moléculaires ou fonctionnels à l'oeuvre



SOMMAIRE

Neuroimagerie Génétique

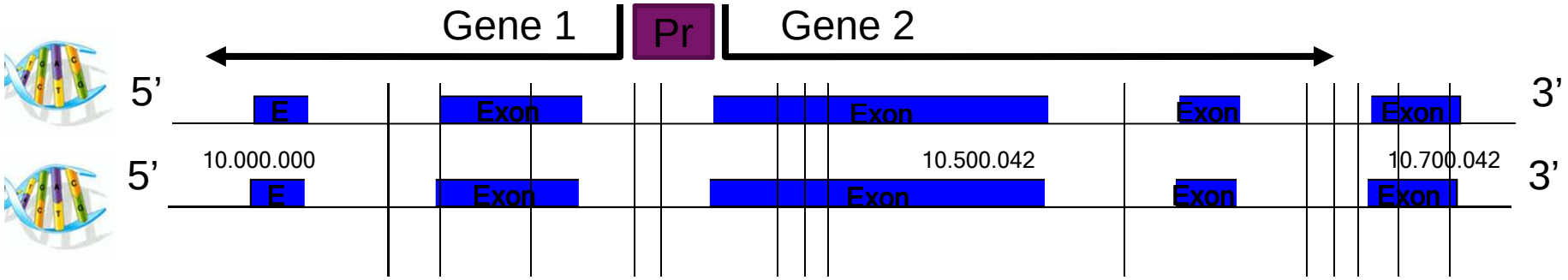
- Le contexte, les mesures,
- L'integration des données produites par des outils de neuroimagerie et de génomique matures.

Implémentation "gros-grain" du modèle de dosage allélique

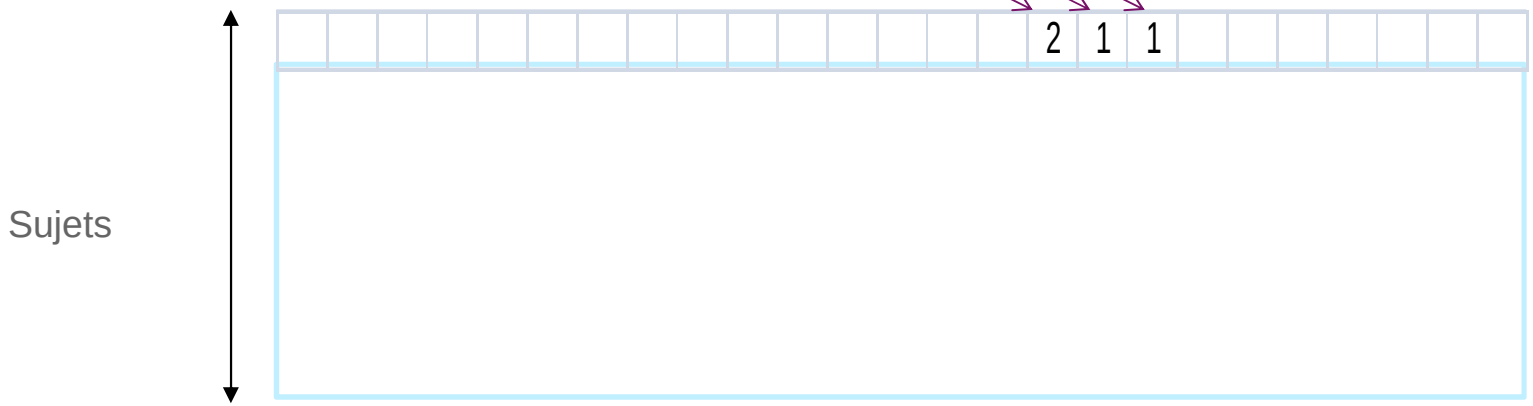
- Besoins de run multiples d'un code données (permutation, validation-croisée)
- MapReduce pattern, Soma-Workflow
- Cluster avec 240 cores

NEUROIMAGERIE-GÉNÉTIQUE : SNP/GENOTYPE

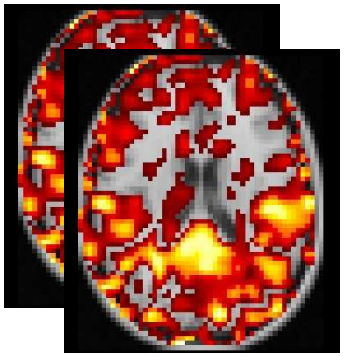
Single Nucleotide Polymorphism : 90% de la variabilité de l'ADN ~10M loci)

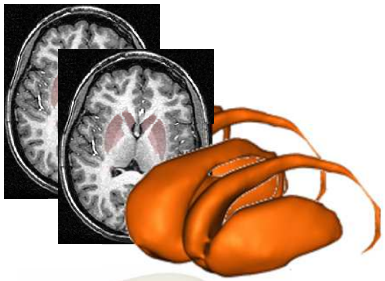


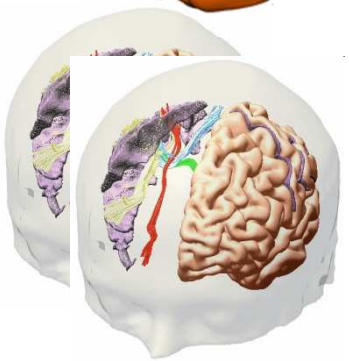
Configuration Hom-M, hom -m ou het(0,2,1)



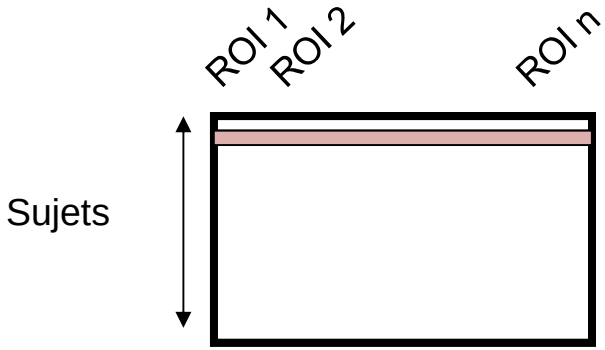
NEUROIMAGERIE-GÉNÉTIQUE : PHENOTYPE







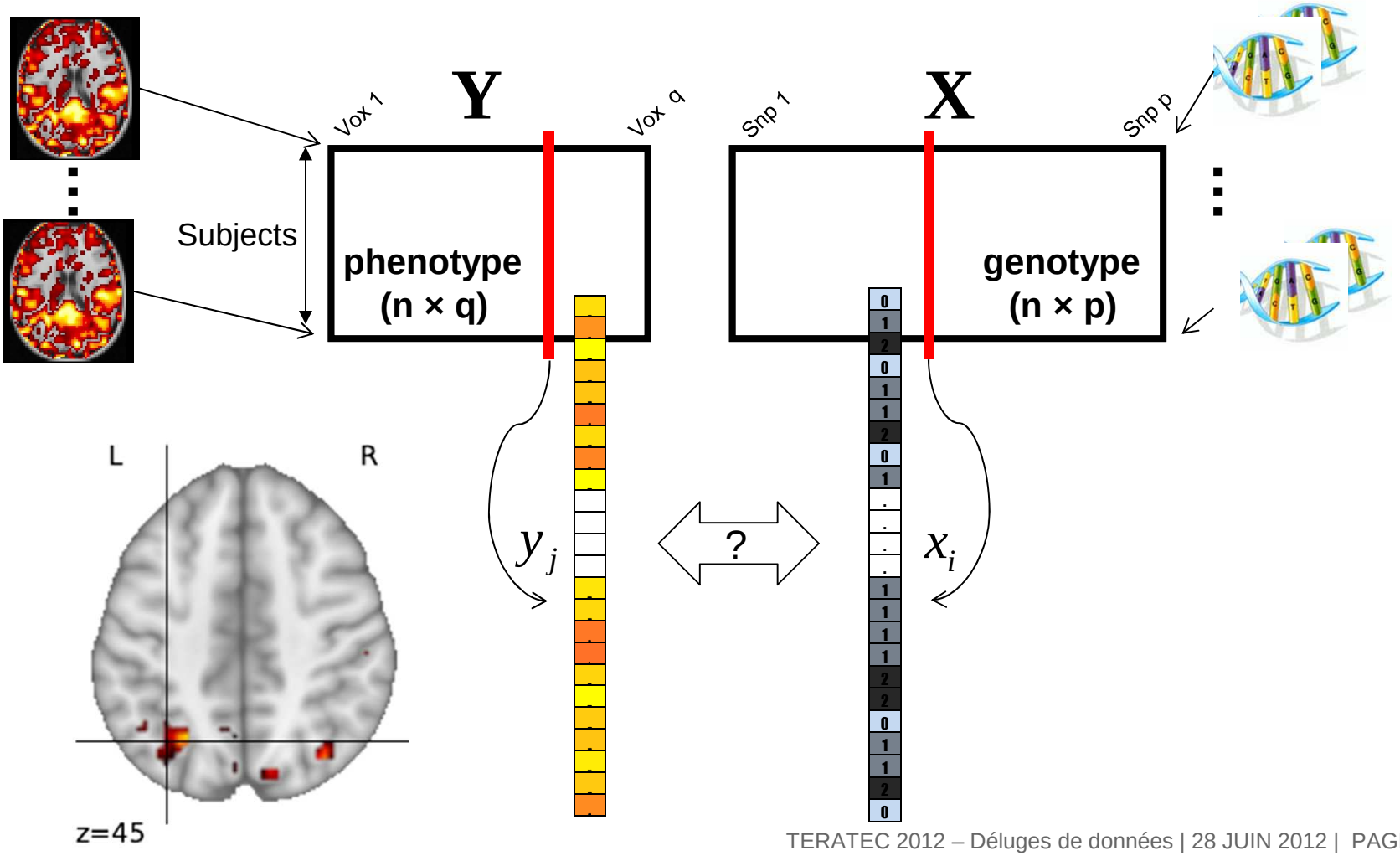
+ ROIs



Cartes statistiques des associations avec des SNPs

- 2000 sujets normaux, acquisition fMRI et tâche « Impulsivité »,
- genotype 1 M SNPs

Schuman *et al.* 2010 ***Mol. Psych.***

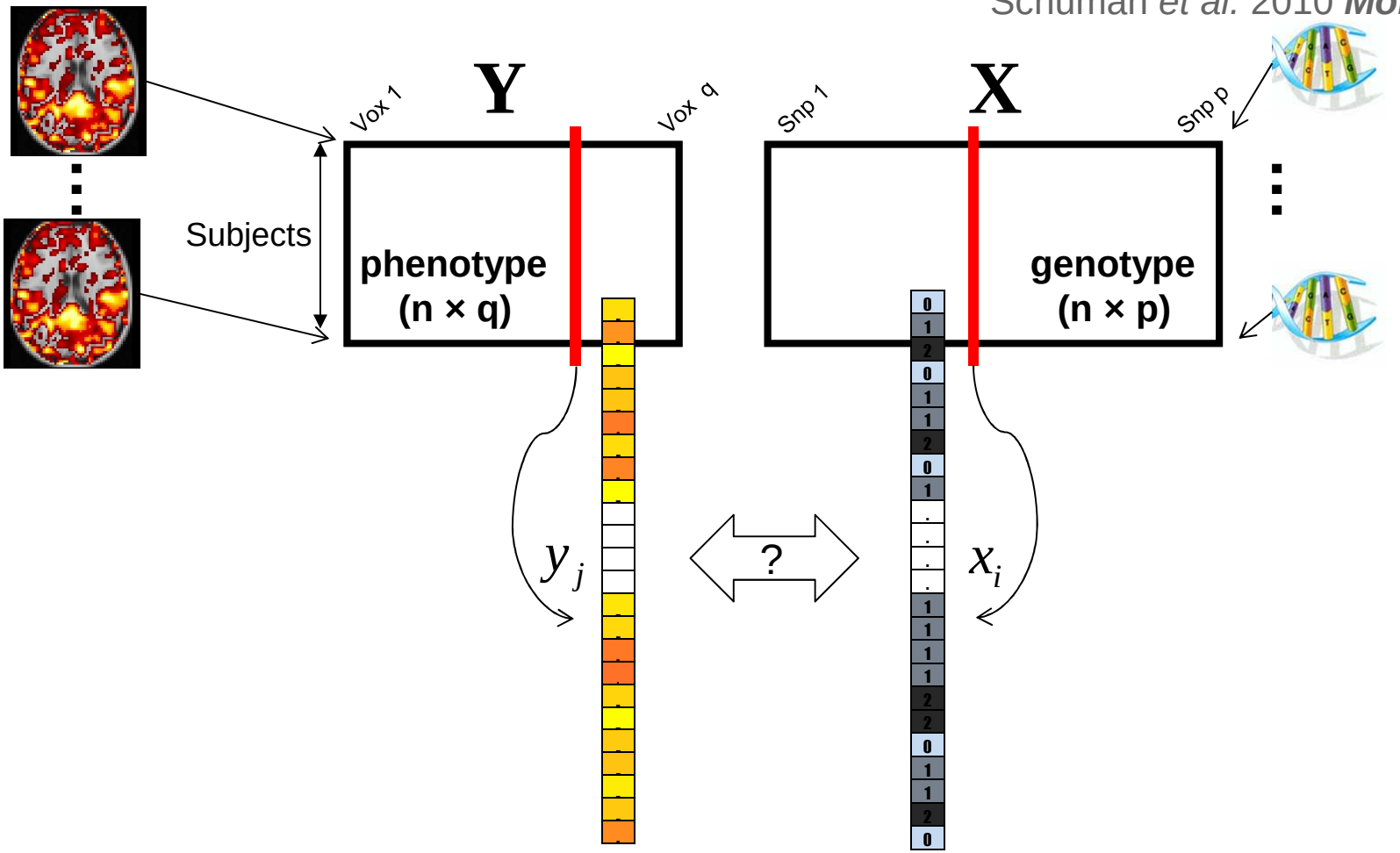


MODÈLE LINÉAIRE MASSIVEMENT UNIVARIÉ

Cartes statistiques des associations avec des SNPs

- 2000 sujets normaux, acquisition fMRI et tâche « Impulsivité »,
- genotype 1 M SNPs

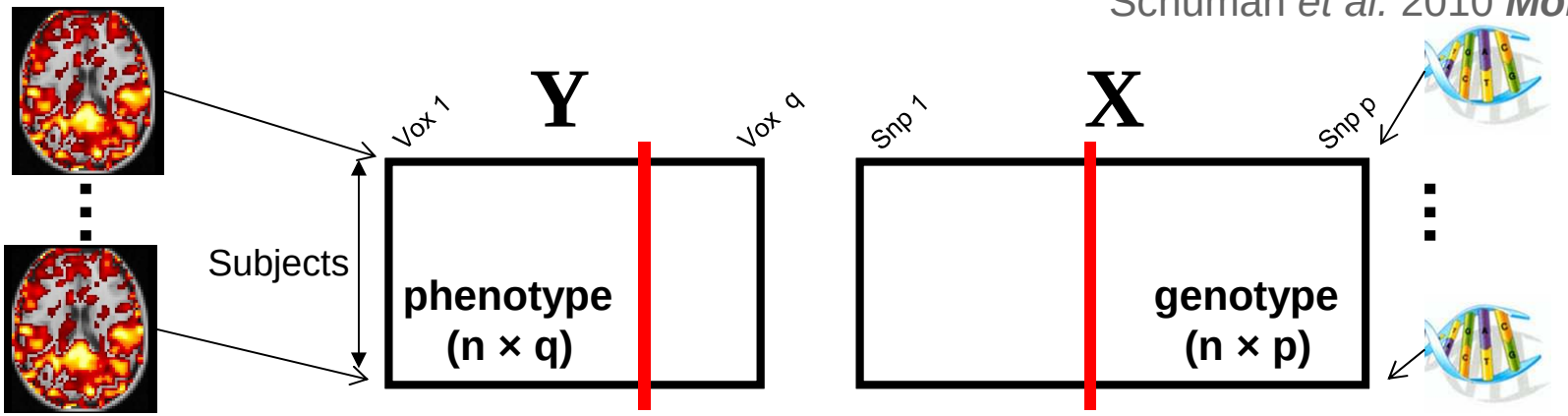
Schuman *et al.* 2010 ***Mol. Psych.***



Cartes statistiques des associations avec des SNPs

- 2000 sujets normaux, acquisition fMRI et tâche « Impulsivité »,
- genotype 1 M SNPs

Schuman *et al.* 2010 ***Mol. Psych.***



Modèle de dosage allèlique :

Ajuster un modèle pour chaque (SNP, voxel)

Cartographier la statistique (eg. *F-score*)

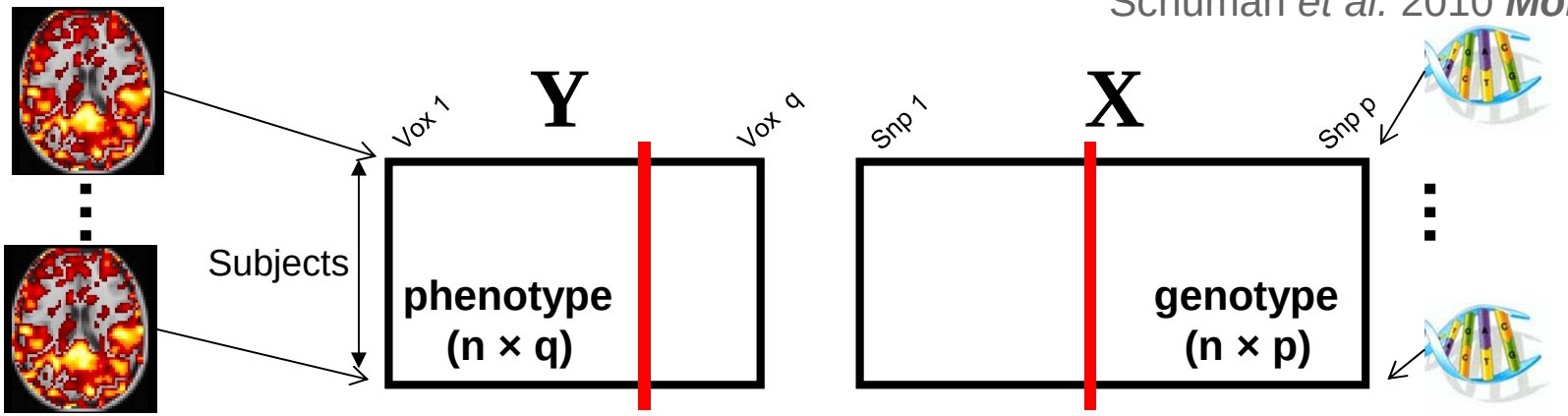
$$y_j = \beta_{jk} x_k + \epsilon$$

$$j = 1, \cdots, q \qquad k=1, \cdots, p$$

Cartes statistiques des associations avec des SNPs

- 2000 sujets normaux, acquisition fMRI et tâche « Impulsivité »,
- genotype 1 M SNPs

Schuman *et al.* 2010 ***Mol. Psych.***



Modèle de dosage allèlique :

Ajuster un modèle pour chaque (SNP, voxel)

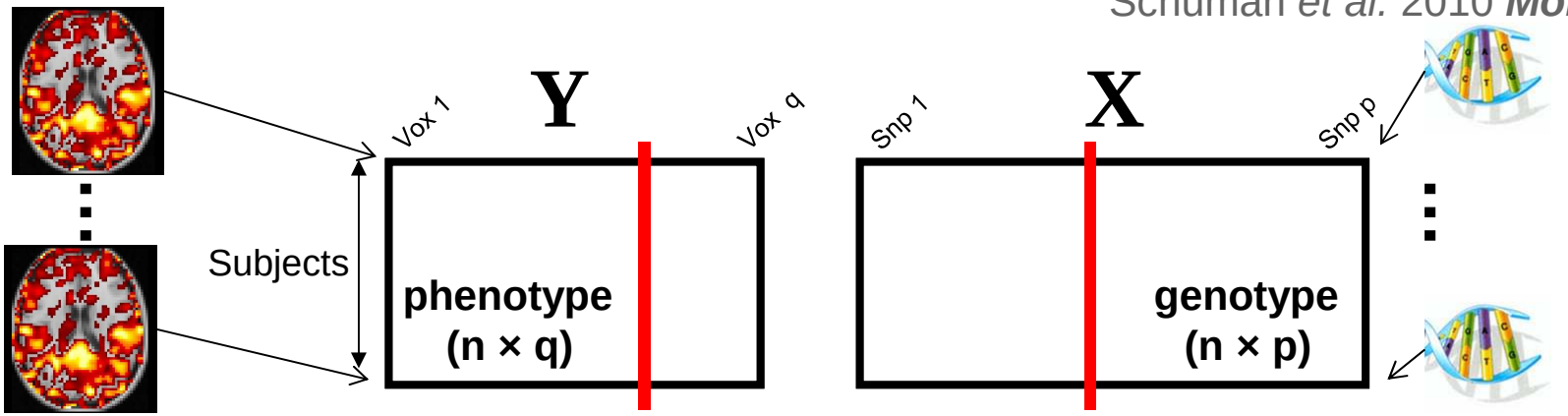
Cartographier la statistique (eg. *F-score*)

$$X^T Y \rightarrow \textit{F-score pour chaque association}$$

Cartes statistiques des associations avec des SNPs

- 2000 sujets normaux, acquisition fMRI et tâche « Impulsivité »,
- genotype 1 M SNPs

Schuman *et al.* 2010 ***Mol. Psych.***



Modèle de dosage allèlique :

Ajuster un modèle pour chaque (SNP, voxel)

Cartographier la statistique (eg. *F-score*)

Avec

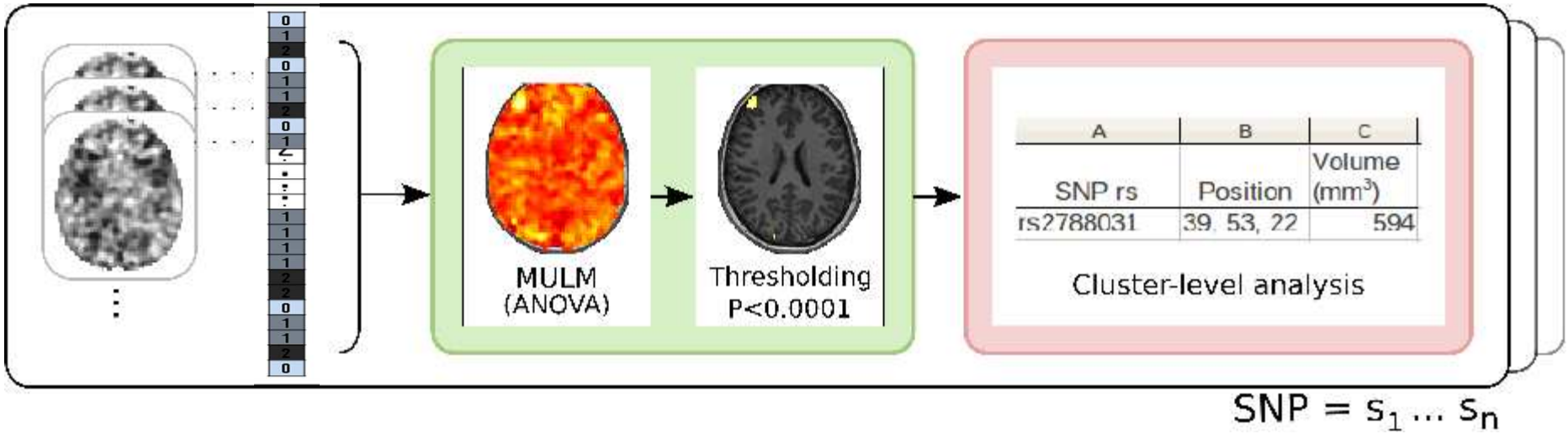
- $p=10^6$
- $q=10^5$
- SNP en LD
- Voxels sont corrélés

Procédure

- Calibration des statistiques
- Correction pour les tests multiples

PERMUTATIONS, MAX(T)

Cluster-based Analysis



Algorithme complet :

- BOUCLE: Données « observées » et « (sujet)-permutées »
  - BOUCLE : Pour chaque SNP
    - Calculer la carte statistique
    - Seuiller la carte et enregistrer les clusters
- De la distribution empirique sous H0, déterminer les clusters significatifs

MULM
CLUSTER –BASED ANALYSIS :

MAP-REDUCE PATTERN ET SOMA-WORKFLOW

ACCÉLÉRATION PAR LA PARALLÉLISATION

Données: 2000 subj., ~50k voxels MRI, ~500K SNP

Unité: 1 GWAS : dosage allélique 500k SNP *vs 1 phenotype*

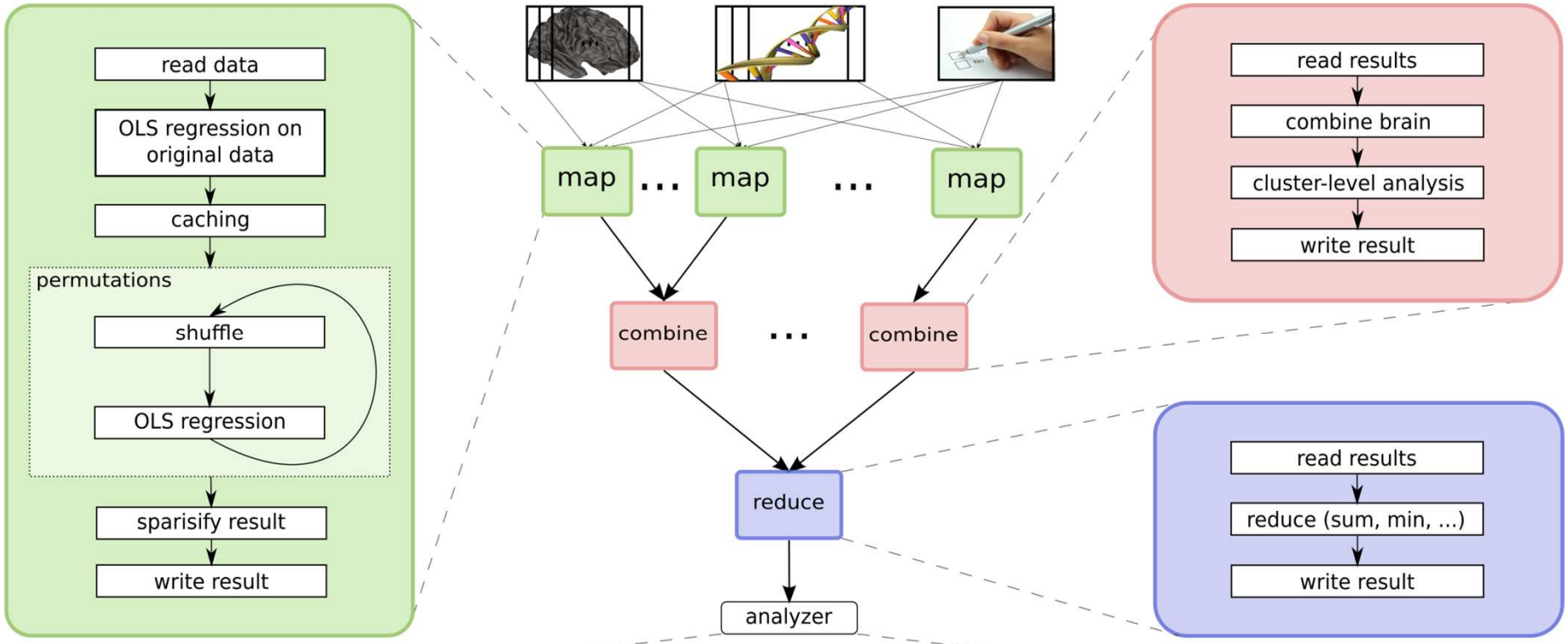
- Code de référence PLINK séquentiel sur un core : 0.5 GWAS/s/c
  - 50k voxels, 1k perm : 4,7 années
  - : 400 Tb de données à partager.

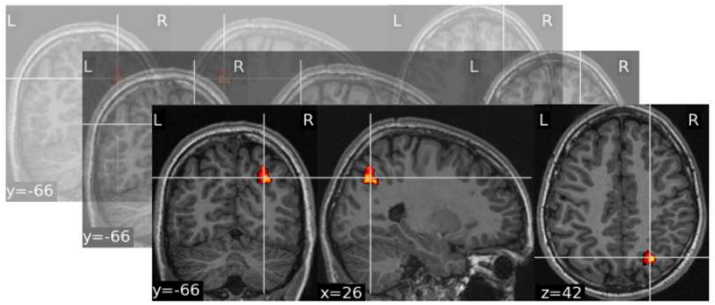
- L'accélération peut être obtenue par
  - le profilage du code unitaire
  - la distribution du code sur les nœuds d'un cluster
  - par l'optimisation-parallélisation conjointe

PARALLÉLISATION

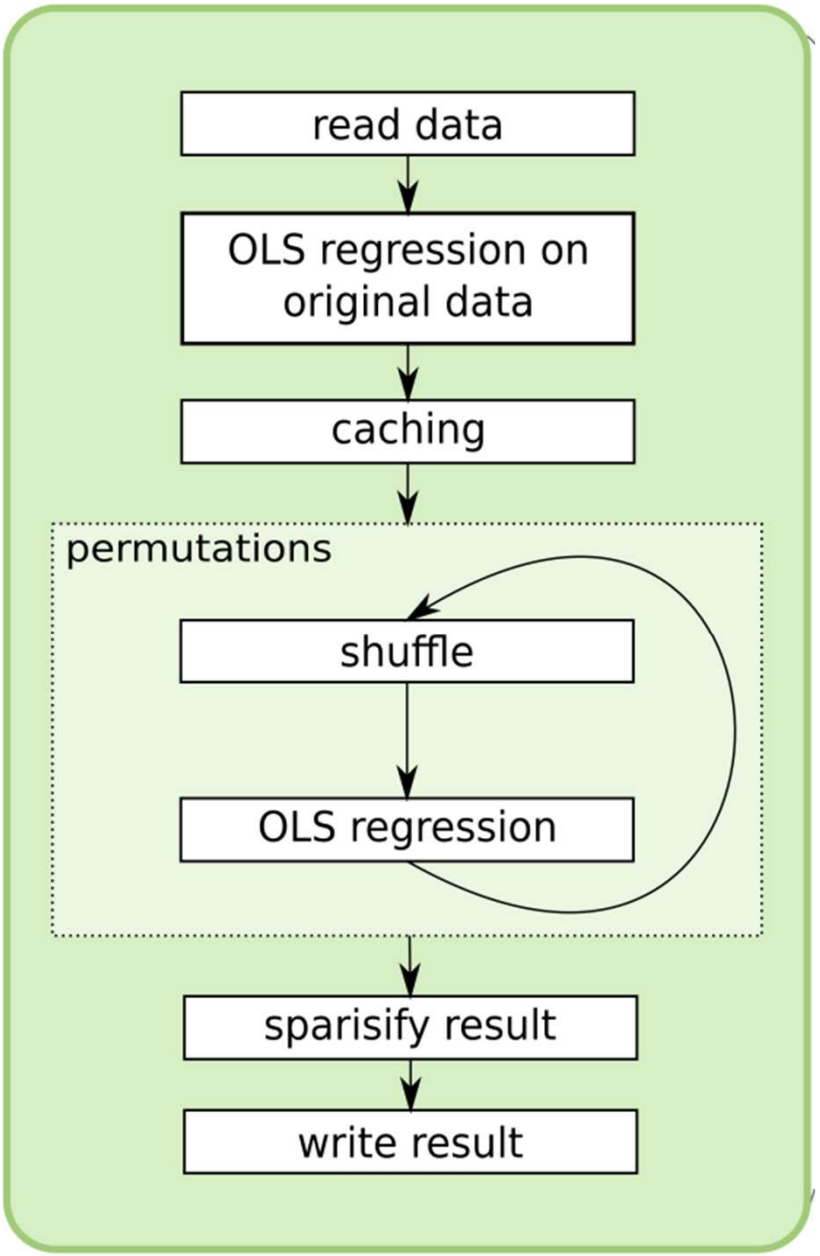
- Repenser l'algorithme pour minimiser les E/S :
  - Optimiser le remplissage de la mémoire cache du processeur
  - Ne charger que les données utiles
- Approche « gros grain » est facilitée car le problème est naturellement parallèle en :
  - SNP, Voxels
  - Permutations
- Utilisation d'un code facile à lire et compatible avec différentes architectures.

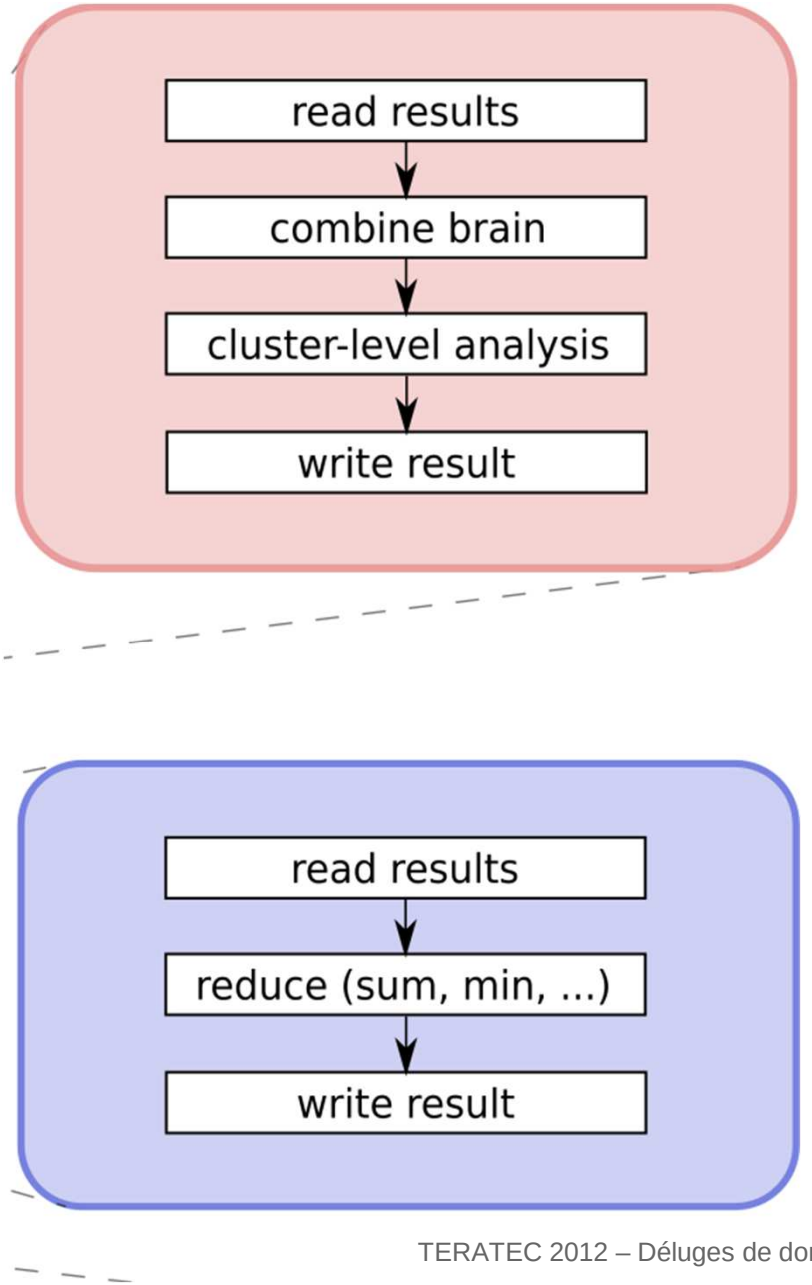
Implémentation Python

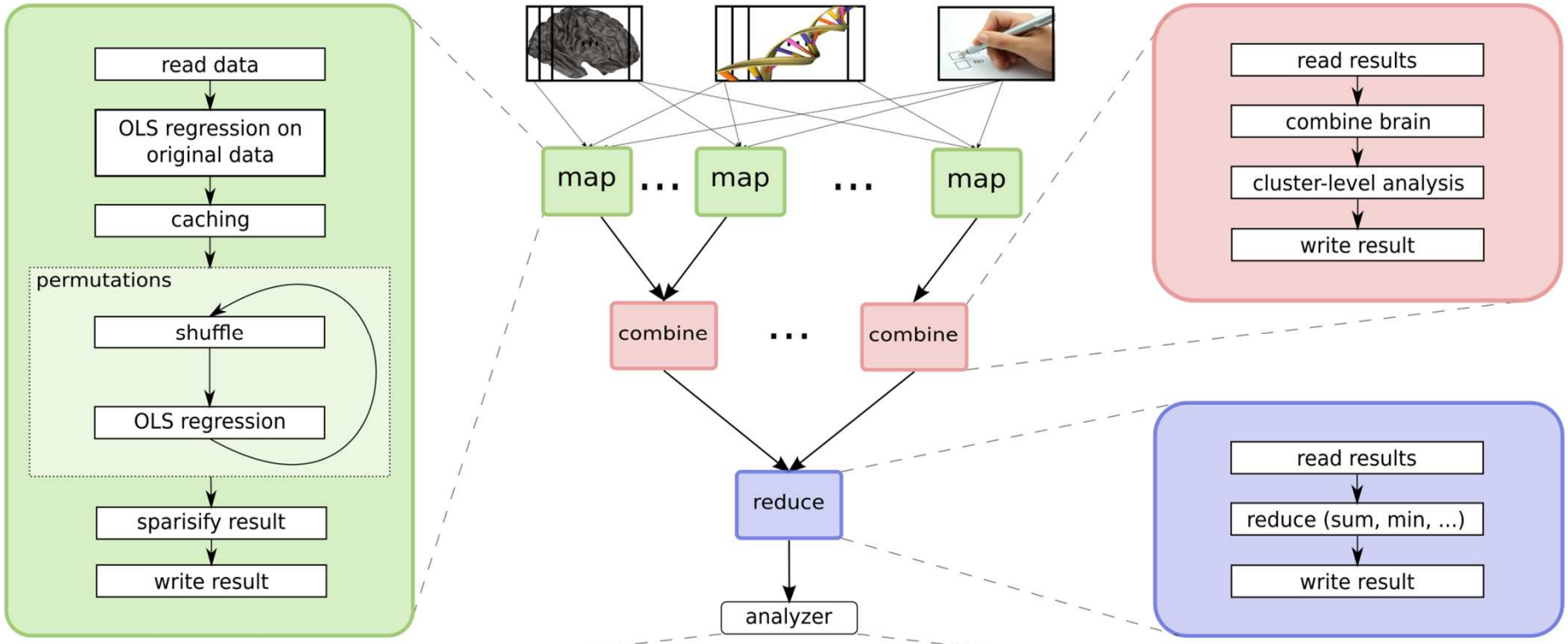


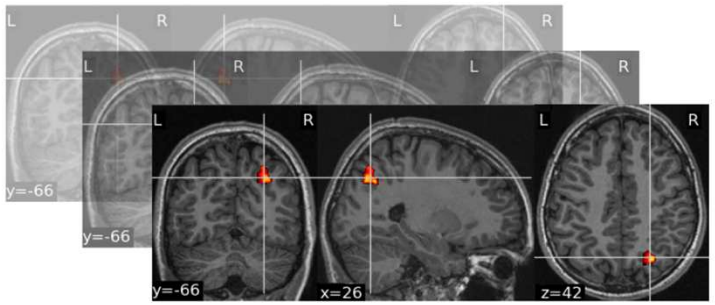


A	B	C	D	E	F	G	H
SNP rs	Chr	Volume (mm^3)	Corr. P-value	Cluster Volume (mm^3)	Corr. P-value	Minimum P-value	Corr. P-value
rs2788031	1	810	1	810	1	1,43E-006	1
rs2185077	1	648	1	540	1	3,18E-005	1
rs10888855	1	1053	1	1026	1	2,51E-005	1
rs10888857	1	2241	1	2241	0,9995	3,76E-007	1
rs3753009	1	2592	0,9888	2565	0,9651	1,90E-007	1
rs7551981	1	540	1	540	1	5,98E-005	1
rs4433442	1	783	1	783	1	5,69E-005	1
rs12755760	1	594	1	594	1	1,94E-005	1


read data
OLS regression on original data
caching
permutations
shuffle
OLS regression
sparisify result
write result

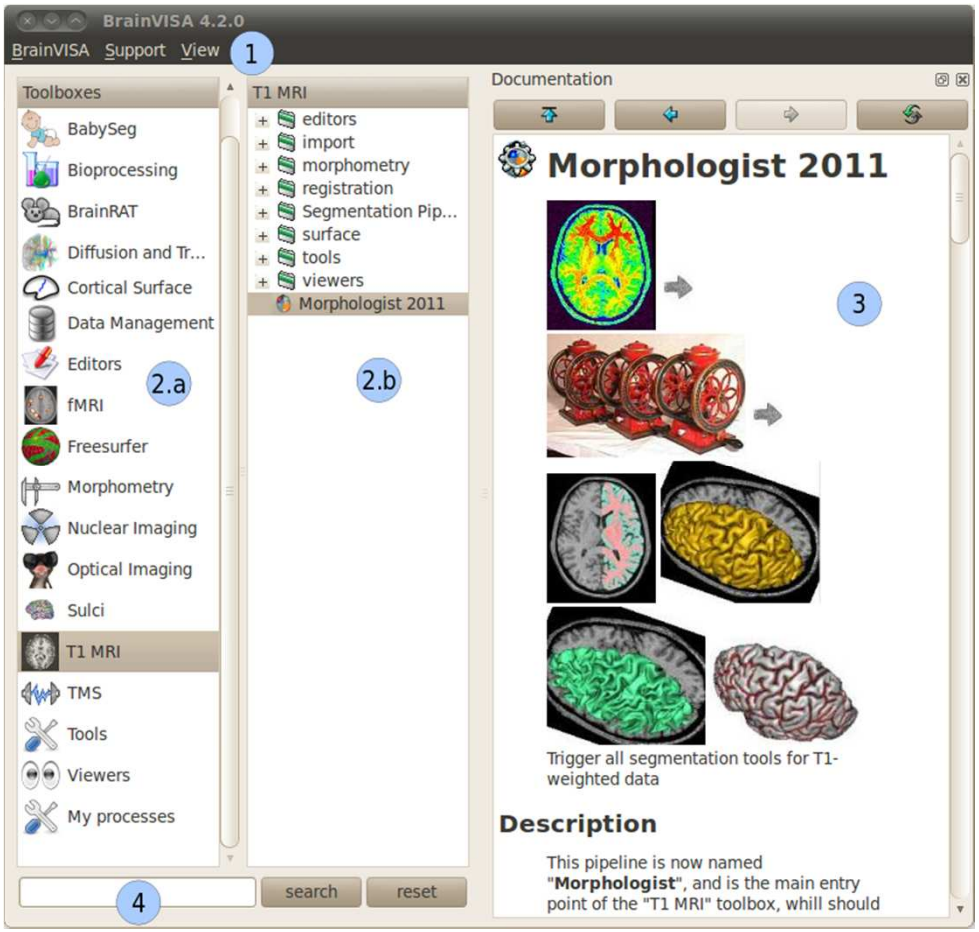

read results
combine brain
cluster-level analysis
write result
read results
reduce (sum, min, ...)
write result

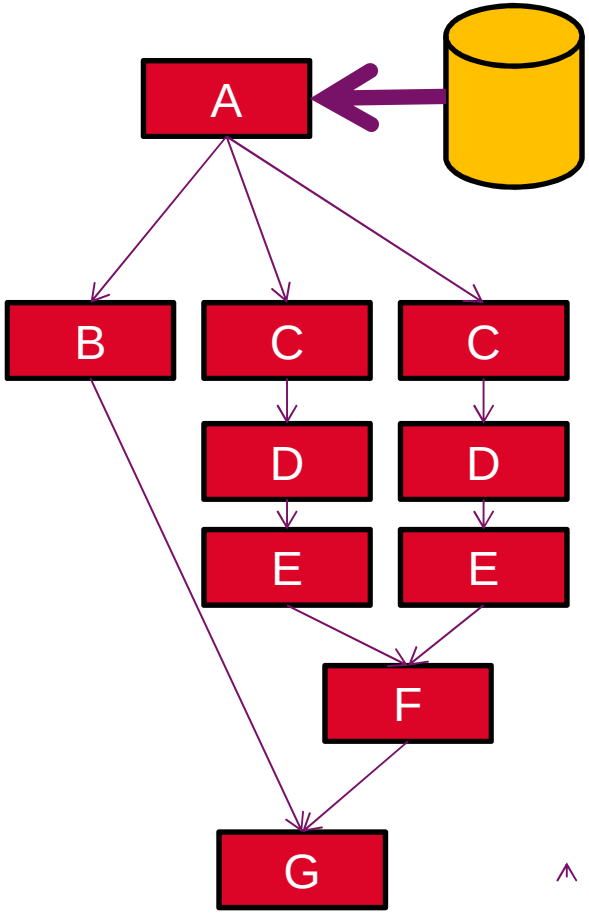




A	B	C	D	E	F	G	H
SNP rs	Chr	Volume (mm³)	Corr. P-value	Cluster Volume (mm³)	Corr. P-value	Minimum P-value	Corr. P-value
rs2788031	1	810	1	810	1	1,43E-006	1
rs2185077	1	648	1	540	1	3,18E-005	1
rs10888855	1	1053	1	1026	1	2,51E-005	1
rs10888857	1	2241	1	2241	0,9995	3,76E-007	1
rs3753009	1	2592	0,9888	2565	0,9651	1,90E-007	1
rs7551981	1	540	1	540	1	5,98E-005	1
rs4433442	1	783	1	783	1	5,69E-005	1
rs12755760	1	594	1	594	1	1,94E-005	1

SOMA-WORKFLOW

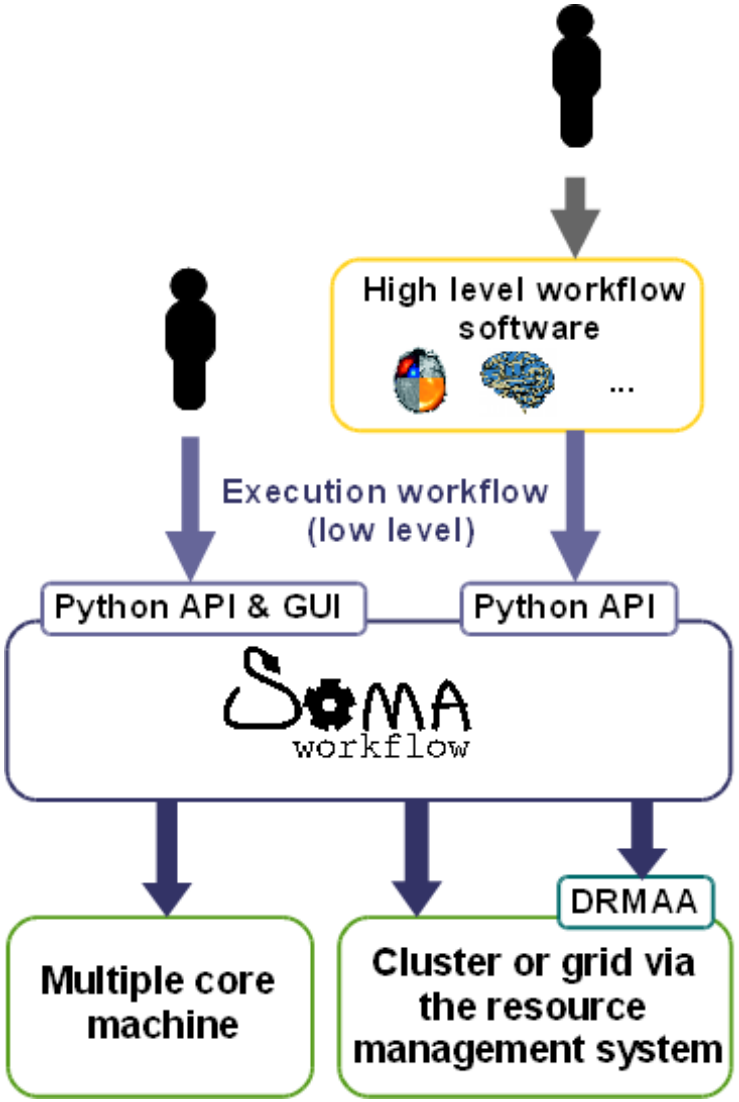




SOMA-WORKFLOW (2)

- Permet de décrire un ensemble tâches indépendantes qui vont s'exécuter suivant le graphe

- Permet de contrôler l'exécution du graphe en soumettant des jobs à un système de file d'attente classique sur un cluster de calcul.



PERFORMANCES

Données: 2000 subj., ~50k voxels MRI, ~500K SNP

Unité: 1 GWAS : dosage allèlique 500k SNP *vs 1 phénotype*

- Code de référence PLINK séquentiel sur un core : 0.5 GWAS/s/c
  - 50k voxels, 1k perm : 4,7 années

- LONI cluster: 300 nodes → (2400 @2.4GHz cores) ? (Stein et al. Neuroimage'10)
  - Code Natif Plink comme code unitaire
  - Implémentation Naïve parallèle de PLINK code
  - 0.002 GWAS/s/c

- DSVcluster: 20 nodes → 240 cores @2.6GHz (B daMota et al. Compstat'12)
  - Notre implémentation optimisée (corrélation, cache size)
  - Map-reduce pattern et soma-workflow
  - 12.0 GWAS/s/c

DISCUSSION - CONCLUSION

BESOINS RÉCURRENTS CONCEPTS RÉUTILISABLES

Besoins : définitivement la parallélisation à “gros grain”

- Codes pour étudier la variabilité (ex GWAS 1 phénotype vs full génotype) par des approches statistiques classiques ou apprentissage automatique
  - **CHARGE** : PERMUTATIONS - CROSS VALIDATION, BOOTSTRAP

- Codes pour le traitement d’image dédié à l’analyse d’un sujet (efficace, débuggé)
  - **CHARGE** : BOUCLE SUR LES IMAGES de la DB

Concepts

- **Map Reduce pattern**
  - Efficient pour un problème « embarrassingly parallel » comme les GW-BW association analysis
  - Plusieurs environnements supportent ce pattern : Hadoop /Apache, SomaWorkflow

- **Optimisation/Parallélisation conjointe** en neuroimagerie-génétique:
  - Permutations (ou CV) gagnent à être intégrées dans le code unitaire

DEVELOPEMENTS FUTURS

Evolutions Soma-workflow

- Adapter Soma-Workflow à des clusters opérés classiquement :
  - En utilisant **openMP**

Optimisation

- Nouvelles implémentations du mapper (Brainomics : NeuroSpin and AS+)
  - Nœuds avec des cartes GPU

BIBLIOGRAPHIE

- A R Hariri et al. Serotonin transporter genetic variation and the response of the human amygdala. Science 2002;297:400-403

- Stein et al 2010. Voxelwise genome-wide association study (vGWAS) Neuroimage, 53 (2010), pp. 1160–1174

- daMota et al 2012. A fast computational framework for genome-wide association studies with neuroimaging data. COMPSTAT conference, Limassol, Cyprus.

NeuroSpin

S. Laguitton
E. Duchesnay
JB. Poline
JF. Mangin

Collaboration :

Inria *PARIETAL*

B DaMota
B Thirion

AS+

Brainomics

M Cadennes
V Ducrot
S Monot

Commissariat à l'énergie atomique et aux énergies alternatives
Centre de Saclay | 91191 Gif-sur-Yvette Cedex
T. +33 (0)1 69 08 79 74 | F. +33 (0)1

DSV
I2BM
NeuroSpin