DE LA RECHERCHE À L'INDUSTRIE

cea

NeuroSpin

UNATI/BrainOmics

www.cea.fr

ANALYSE D'ASSOCIATION GÉNOME/CERVEAU ENTIER

MISE EN ŒUVRE SUR CLUSTER GPU CURIE

Séminaire ASPROM.
Calculs intensifs : modélisation, simulation

Vincent FROUIN, Benoit da MOTA

30 SEPTEMBRE 2014

CEA - DSV - NEUROSPIN

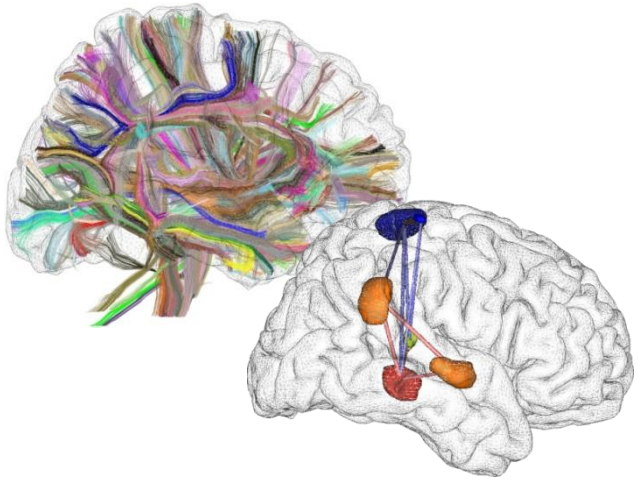
FROM RESEARCH TO INDUSTRY
cea

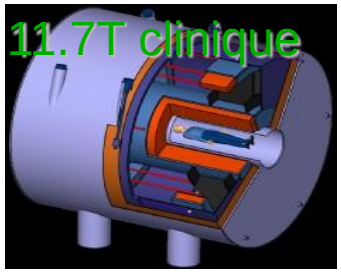
CEA: EPIC, 5 directions dont:
DSV: Direction des sciences de la vie

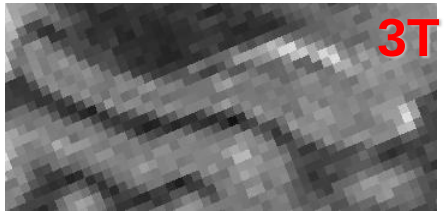
NeuroSpin

NeuroSpin: Centre de neuroimagerie
- IRM haut champ :

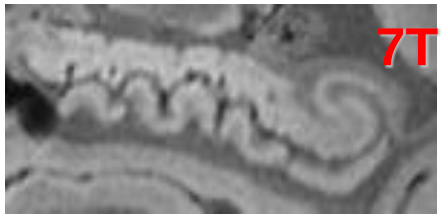
NeuroSpin
- IRM haut champs
- **UNATI**: Analyse d'image/statistique







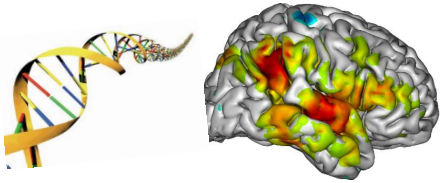
1000umx1000umx1000um



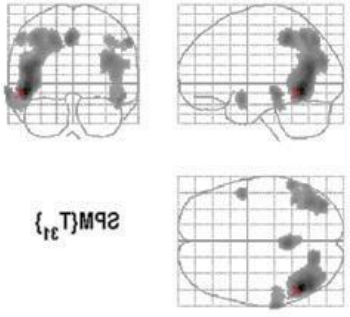
300umx300umx300um

SOMMAIRE

Introduction à l'imagerie génétique



Etudes d'association
génome-entier, cerveau-entier



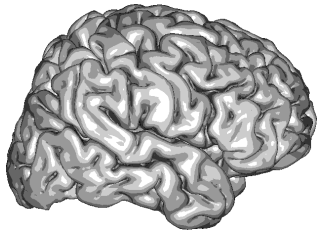
Eléments d'implantation des calculs

Résultats

PROBLÉMATIQUES ET ENJEUX

Compréhension des processus biologiques impliqués dans les maladies cérébrales

Génétique → Cellulaire → Cérébral → Phénotype



Pathologies
Syndrome
Comportement
Effet d'un médicament

Enjeux :

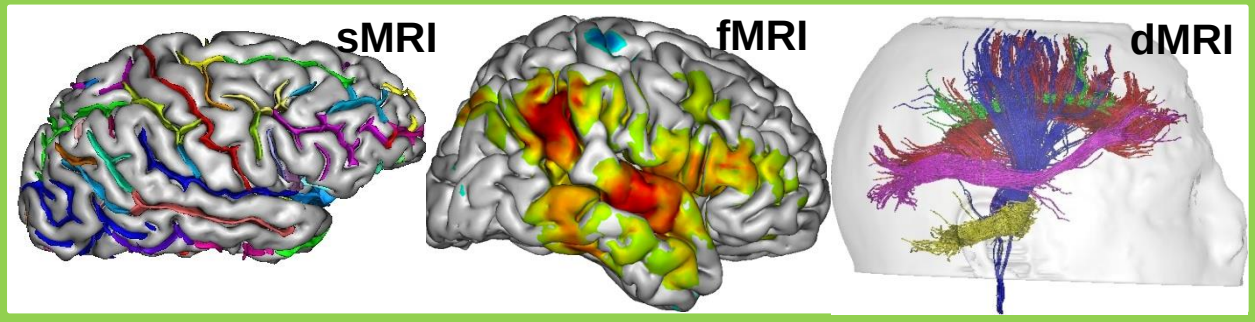
Diagnostic précoce
Développement de nouvelles thérapies
Médecine personnalisée

NEUROIMAGERIE GÉNÉTIQUE : LES FONDEMENTS

- Un outil pour l'identification de nouveau biomarqueurs (via les endophéntypes d'imagerie)
- Un outil pour aborder des questions de recherches fondamentales ou appliquées
  - Intégration des mesures obtenues à différents niveaux d'organisation
  - Révéler les réseaux d'interactions moléculaires sous-jacents

Patient file, Behavior, Environment

SNPs, CNVs, CGH Expression, Methylation



COHORTE DE NEUROIMAGERIE GENETIQUE À NEUROSPIN

FP6 IMAGEN: PI G Schumann KCL

- Partenaires: Hamburg, Dresden, Dublin, Paris
  Mannheim, Berlin, Nottingham, London,

http://www.imagen-europe.com/

Projet intégratif et translationnel

(modèles animaux)

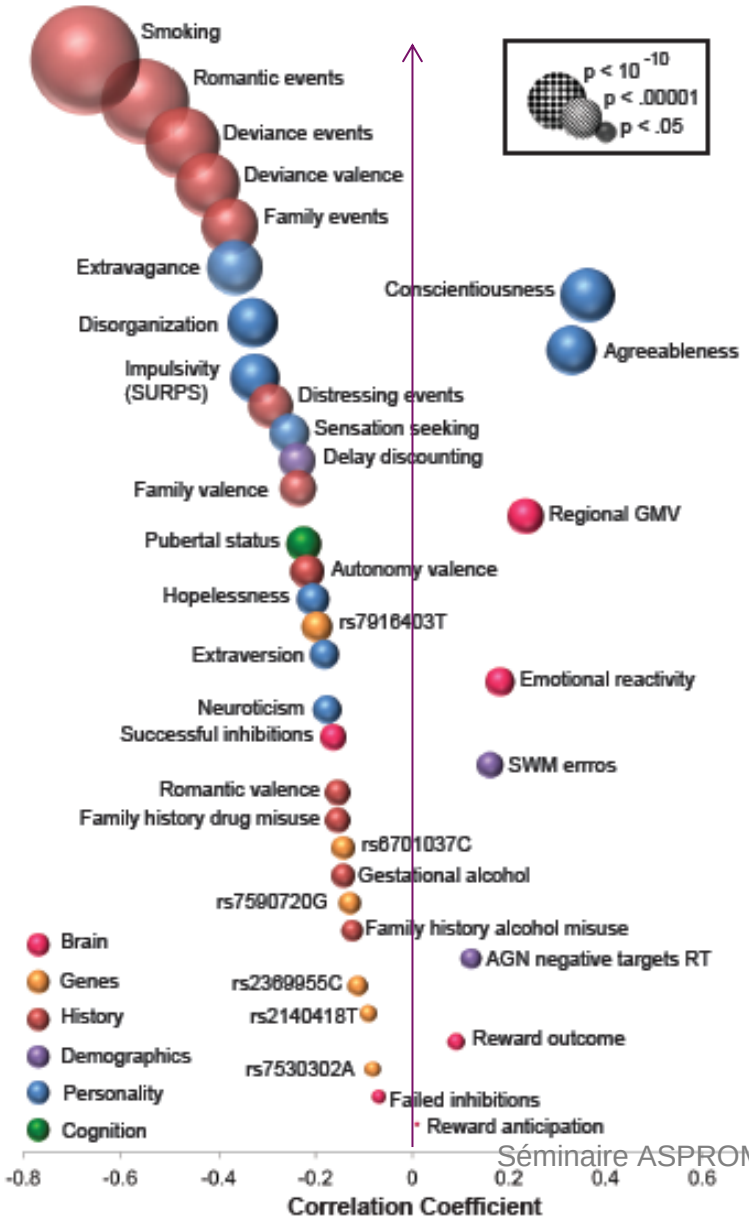
- Cohorte 2000 adolescents (14 ans à l'inclusion)
  - "**Brain**" : IRM anatomique,
    IRM fonctionnelle (MID, SST, Faces, RS)
    IRM Diffusion
  - Questionnaires "**Personality**", "**Life History**", '**Demographics**"
  - Genotypage : **SNPs** (+ partiel GE, Methylation)

Suivi Longitudinal

Initial (images, snp, tests), Suivi1 à 16 ans (tests only), Suivi2 à 20 ans (images, tests)

- … "The IMAGEN study is the first multicentre genetic-neuroimaging study aimed at identifying the genetic and neurobiological basis of individual variability in impulsivity, reinforcer sensitivity and emotional reactivity, and determining their predictive value for the development of frequent psychiatric disorders." 2010 G Schumman *et al.* Mol Psy

COHORTE DE NEUROIMAGERIE GÉNÉTIQUE À NEUROSPIN



LETTER nature

Neuropsychosocial profiles of current and future adolescent alcohol misusers

Nature,512(7513) :185-189, 2014

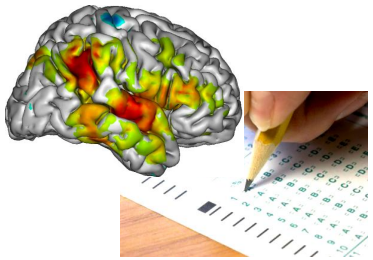
Bindge Drinking à 14 ans.

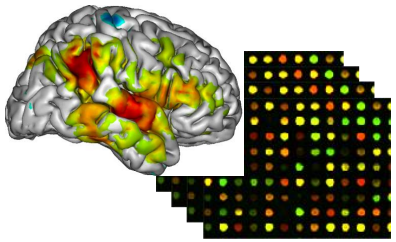
- Identification de variables caractéristiques (Brain, Personality, History, Genetics)
- Relation entre le fait d'appartenir à une catégorie (buveur/non-buveur) et certaines variables caractéristiques de l'étude IMAGEN.
- Approche : Classification par Machine Learning, Stabilité évaluée par bootstraps

COHORTE DE NEUROIMAGERIE GÉNÉTIQUE À NEUROSPIN

FP6 IMAGEN and FollowUps: PI G Schumann KCL

- Voir publications (Pubmed keyword « IMAGEN CONSORTIUM »)



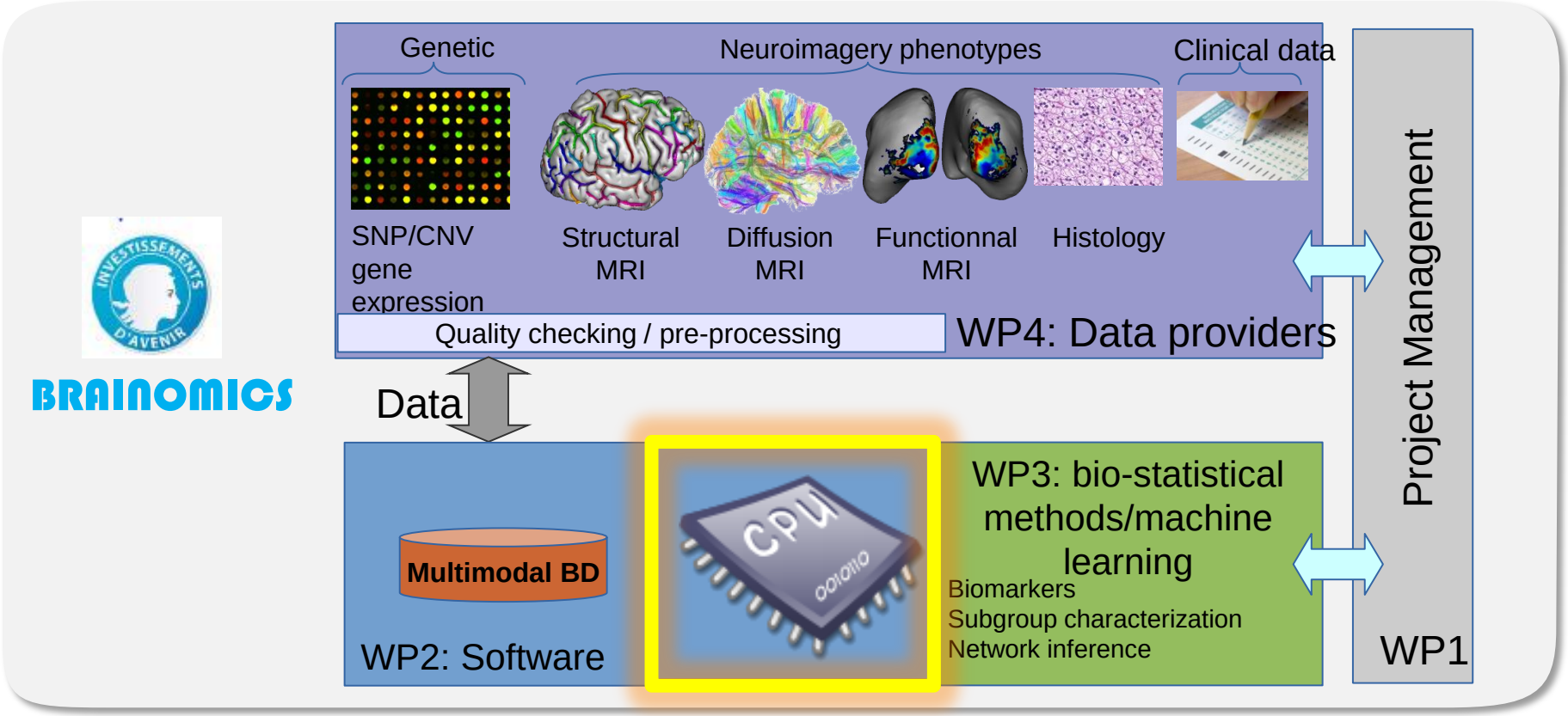


- **Associations** entre traits neuropsycho. et imagerie
  - patterns structuraux, réseaux fonctionnels,
- **Associations** traits neuropsycho./imagerie et génétique
  - Gene candidat, Etude Génome entier
  - Héritabilité : meta-analyse (ENIGMA1), et analyse multivariée
- **Méthodologie** (Collab. CEA-INRIA Parietal, Supélec, ICM):
  - Détection d'outliers
  - Approches univariées pour imagerie génétique,
  - Approches multivariées pour imagerie génétique,
  - Biostatistiques dédiées aux données p >> n

LE PROJET BRAINOMICS

Méthodes mathématiques et outils informatiques pour la découverte des liens imagerie / génétiques

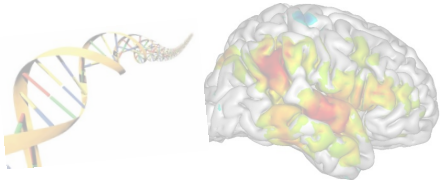
- Appel d'offre IA Bioinformatique 2010 →



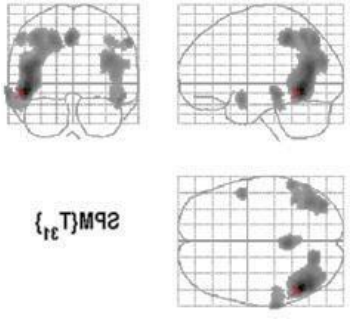
WP1: Neurospin; Keosys
WP2: Logilab; EOLEN; NeuroSpin;
WP3: Neurospin; Supelec
WP4: St Anne; IGR; Neurospin

SOMMAIRE

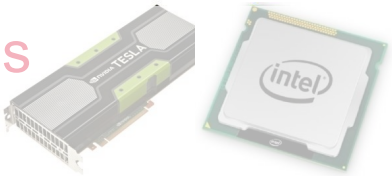
Introduction à l'imagerie génétique



Etudes d'association
génome-entier, cerveau-entier

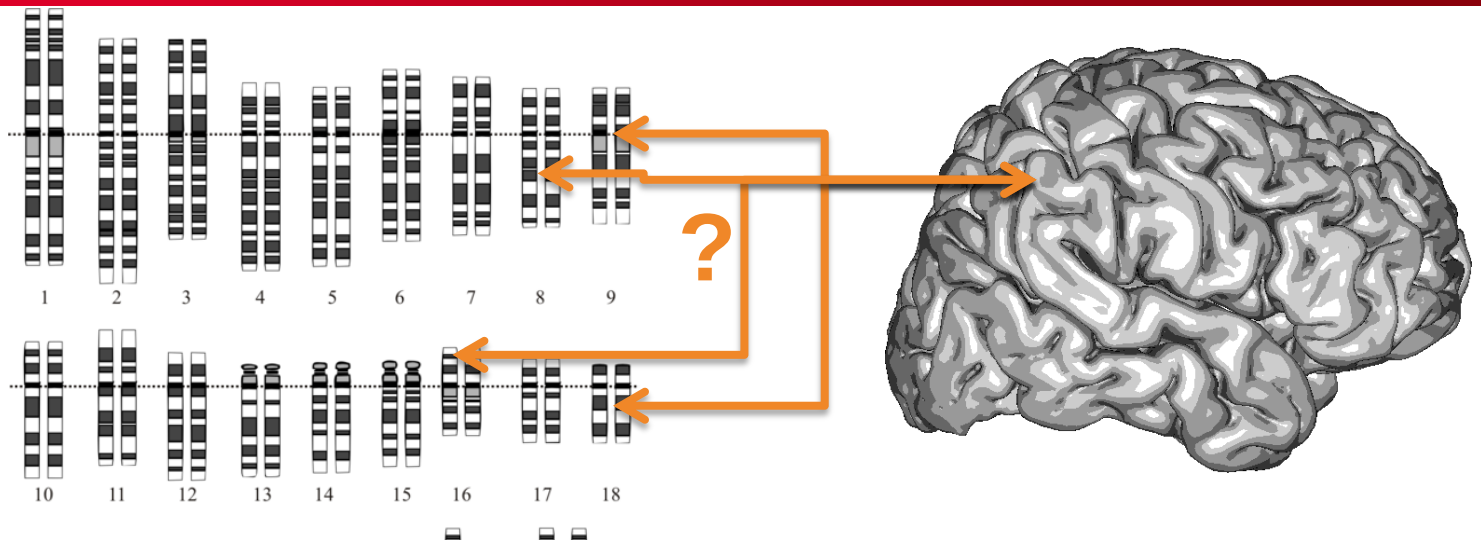


Eléments d'implantation des calculs



Résultats

ETUDE GENOME-CERVEAU ENTIER : POURQUOI ?



Applications à l'épaisseur corticale

- Utilisée pour l'étude des pathologies (MA) de phénotypes (IMC, exposition à l'alcool)
- Avec ~2000 sujets : Puissance statistique permet d'aller au-delà de Gene Candidat (~80)

➡ **Existe-t-il des variants de prédisposition ?**

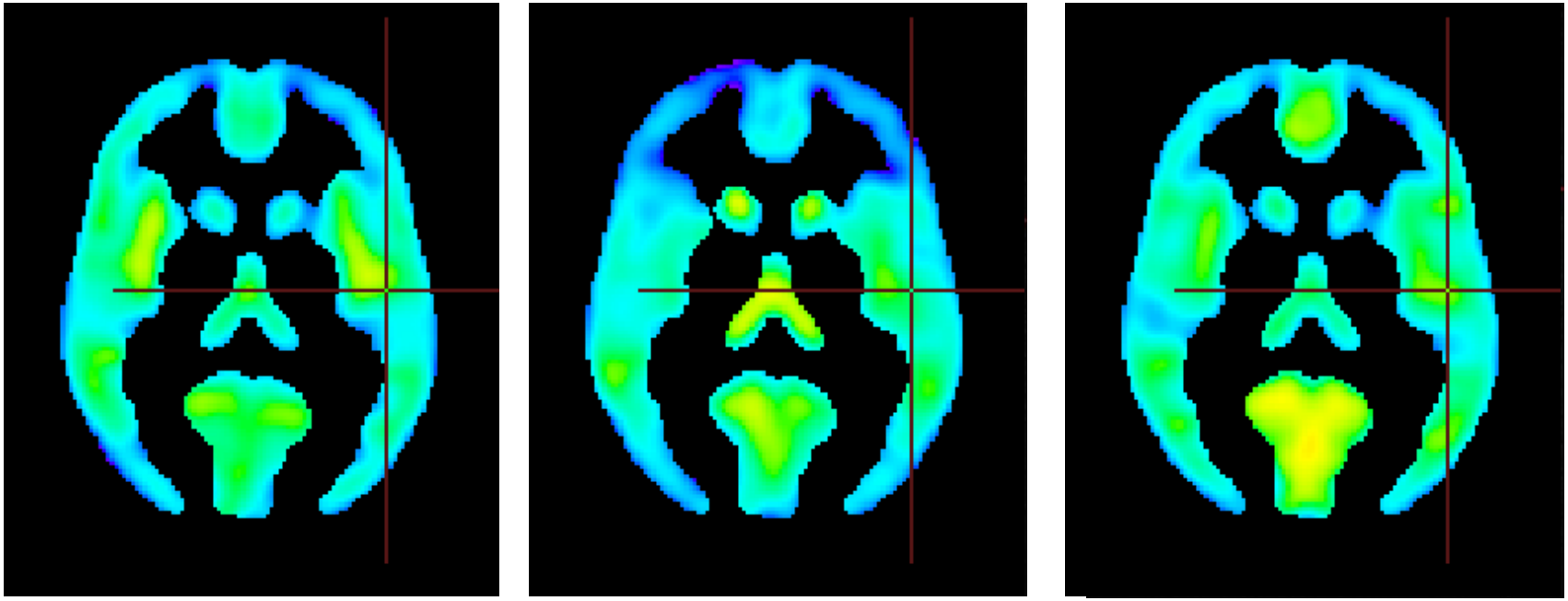
- Permettrait de stratifier les sujets dans les études cliniques ou fondamentales.

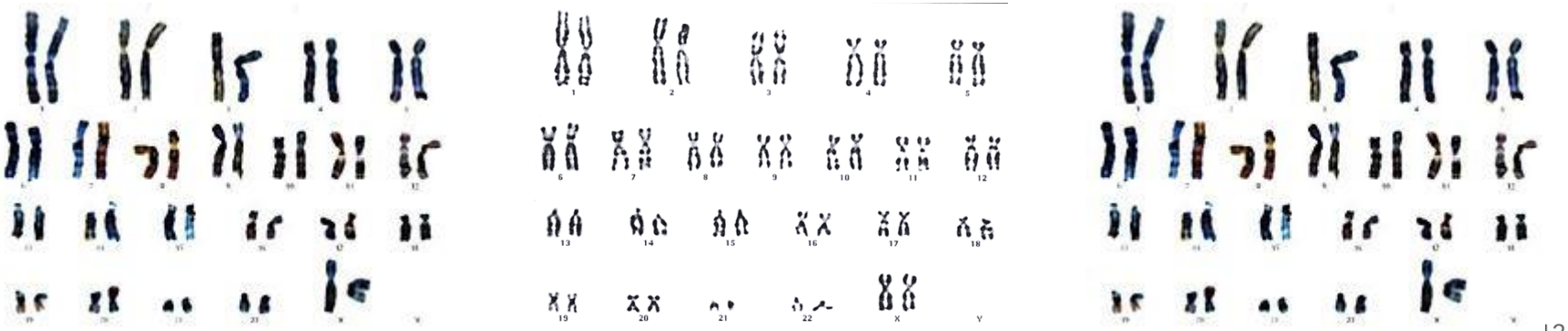
Méthodologie princeps : Stein 2010 NeuroImage

- Problème 1 : Met en œuvre un code non optimisé sur cluster
- Problème 2 : Pas de correction empirique pour les tests multiples.

ETUDE GENOME-CERVEAU ENTIER : POURQUOI ?

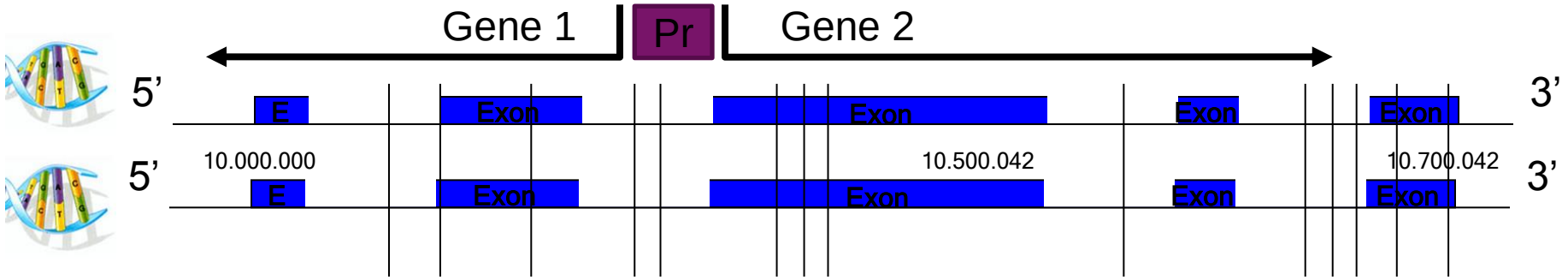
VBM matière grise : variation individuelle rapportée à un atlas (MNI)



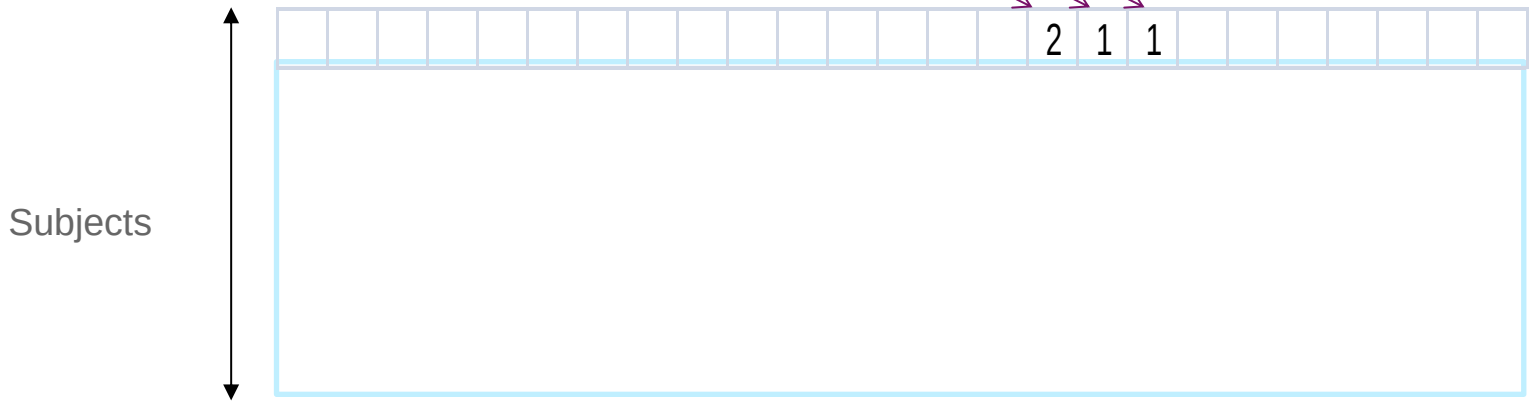


NEUROIMAGERIE-GÉNÉTIQUE : GÉNOTYPE

Single Nucleotide Polymorphism : 90% des variations de l'ADN (~10M loci)

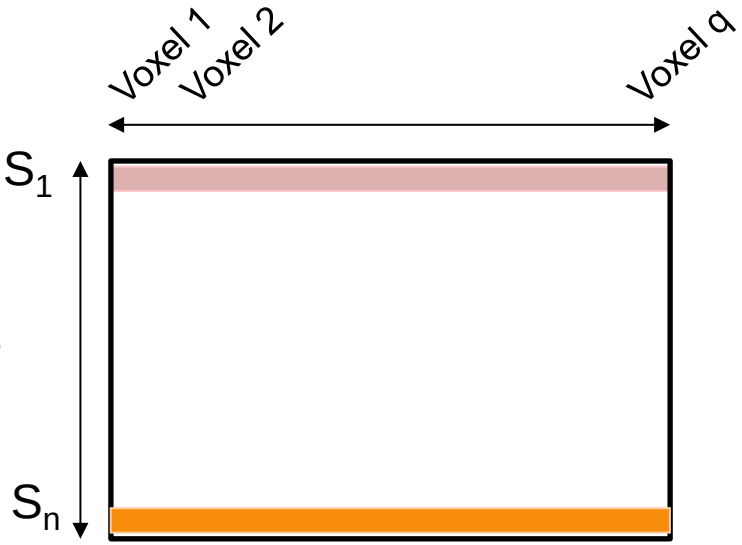


Configuration HOMOZYGOTE, homozygote ou hétérozygote : **0,2,1**



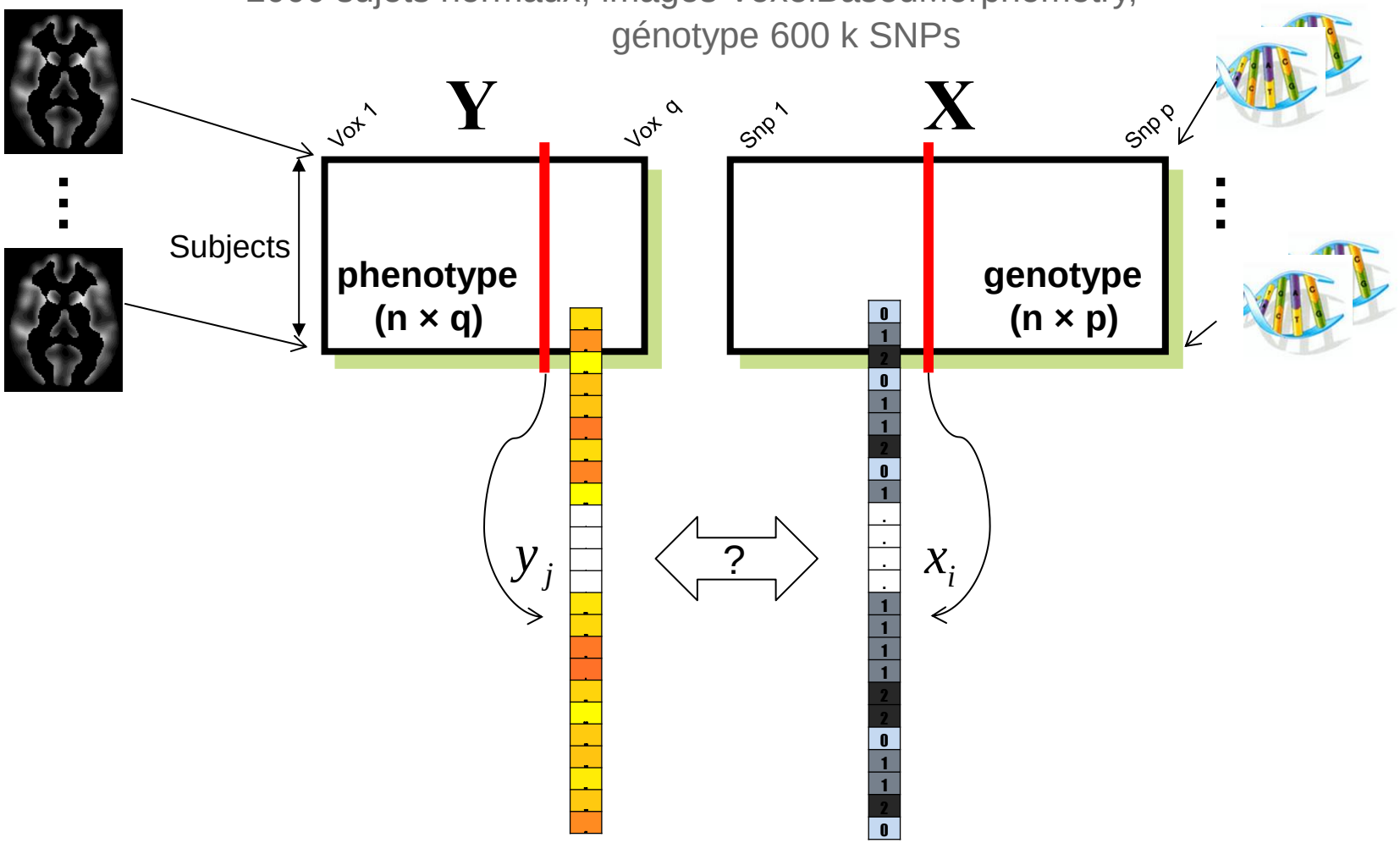
NEUROIMAGERIE-GÉNÉTIQUE : PHÉNOTYPE

Subjects



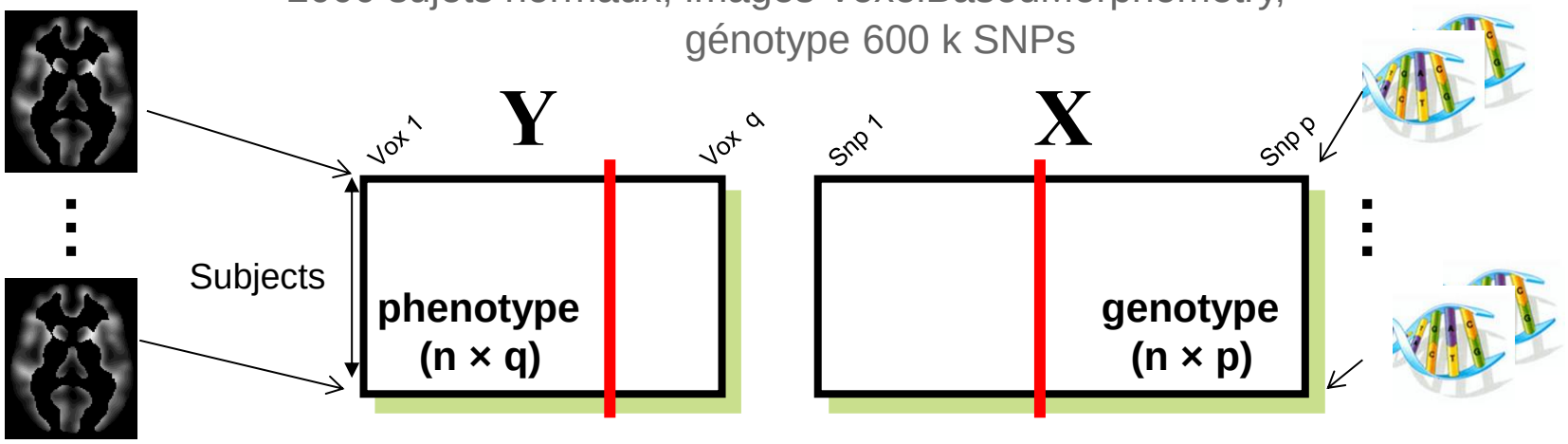
Construction des cartes statistiques des associations SNPs ↔ caractéristiques d'imagerie

2000 sujets normaux, images VoxelBasedMorphometry, génotype 600 k SNPs



Construction des cartes statistiques des associations SNPs ↔ caractéristiques d'imagerie

2000 sujets normaux, images VoxelBasedMorphometry, génotype 600 k SNPs

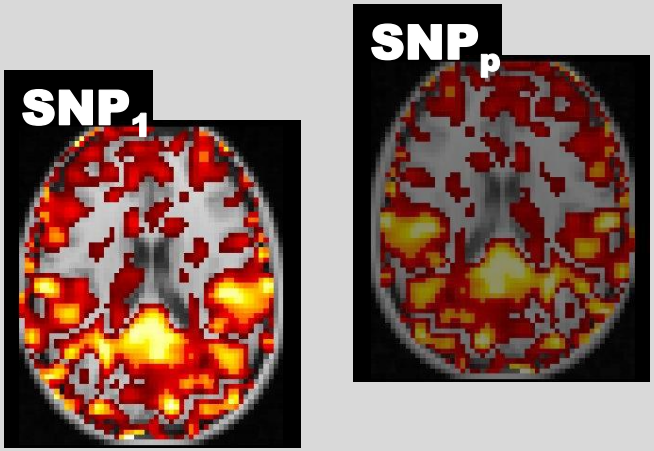


Carte → Cerveau-Entier,
SNPs → Génome-Entier

Dosage allélique

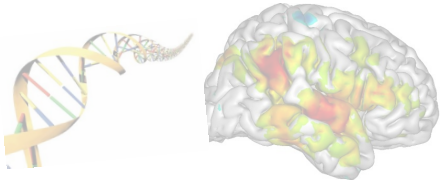
$$y_j = \beta_{jk} x_k + \epsilon$$

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

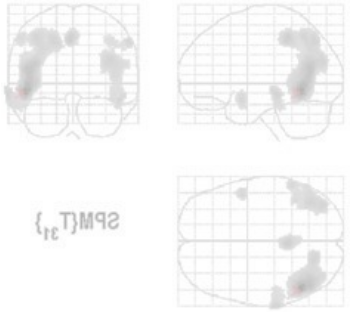


SOMMAIRE

Introduction à l'imagerie génétique



Etudes d'association
génome-entier, cerveau-entier



Eléments d'implantation des calculs

Résultats

MODÈLE UNIVARIÉ DE MASSE

$$Y = x\beta + Z\gamma + \epsilon$$

IRM génomique covariables "bruit"

Jeu de données ciblé 1292 sujets :

- 336,188 voxels (matière grise cérébrale), 466,125 SNPs , 10 covariables
- Correction empirique pour les tests multiples : distrib. des scores sous H_0

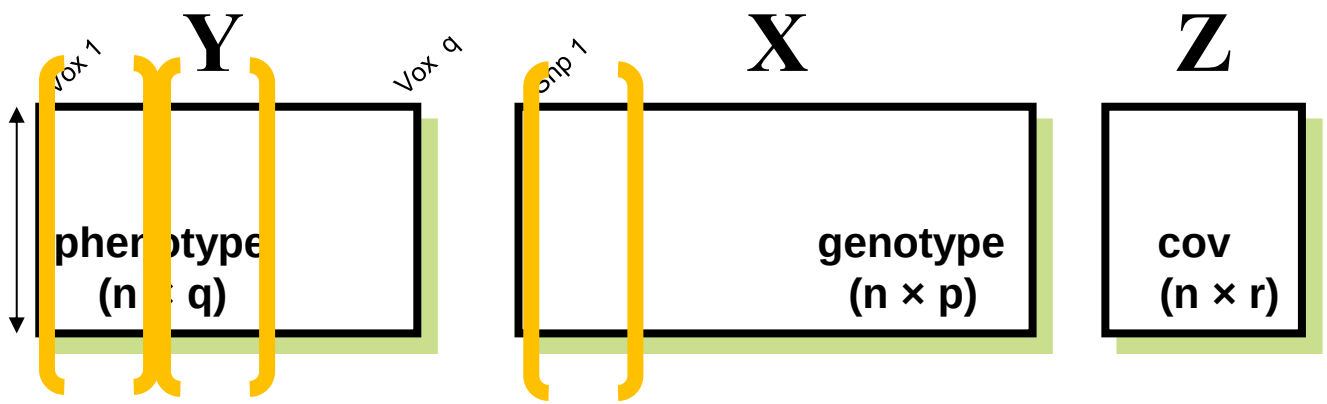
→**Corrélations** 1 à 1 sur données observées ET
Refaire ce calcul sur 10 000 permutations

- 1,5 x 10^{15} corrélations permettant de calculer un F-score d'association.
- 8 x 1,5 x 10^{15} = 12 Po (12000 disques durs de 1To ou 2.5M de DVDs)
- Seuls les meilleurs scores devront être gardés (0.1 à 0.01%)
- Filtrage sur le GPU

ALGORITHMIE : ASPECTS SÉQUENTIELS

$$Y = x\beta + Z\gamma + \epsilon$$

IRM génomique covariables “bruit”



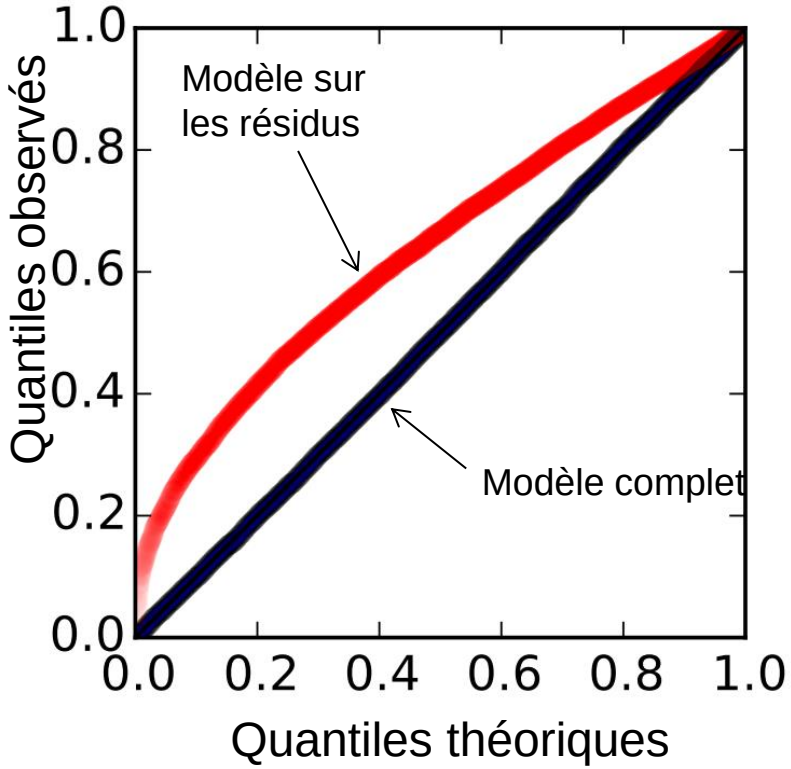
1. Sur CPU calcul du résidu de Y par rapport à Z

2. Pour chaque permutation, calcul d’un terme correctif.
M. J. Anderson and J. Robinson. (2001),

3. Corrélations calculées sur des sous-blocs de X et Y, 10,000 permutations et calcul de corrélation correspondant

PERMUTATIONS EN PRÉSENCE DE COVARIABLES

- Modèle complet:
  trop couteux en temps de calcul
- Modèle sur les résidus:
  mauvais contrôle des faux positifs
- Solution:
  approximation de Freedman and Lane



Permutation tests for linear models. M. J. Anderson and J. Robinson. (2001), Australian and New Zealand Journal of Statistics, (43):75–88, 2001.

A nonstochastic interpretation of reported significance levels. D. Freedman and D. Lane. (1983)
Journal of Business & Economic Statistics, 1(4):292{98, 1983}.

ALGORITHMIE : ASPECTS SÉQUENTIELS

Algorithm Fit the model $Y = x\beta + Z\gamma + \epsilon$ for all x in $X = [x_1, .., x_q]$ and get a score for each pair (SNP, voxel)

Require: The data Y, the X and Z regressor matrices

if first regression **then**

$Z_w \leftarrow (Z^T Z)^{-1/2} Z$ {whitening}

$Y_{norm} \leftarrow Y {\Delta_1}^{-1}$ with $\Delta_1 = diag(\|Y_i\|, i = 1 \dots p)$ {p number of voxels}

$X_{norm} \leftarrow X {\Delta_2}^{-1}$ with $\Delta_2 = diag(\|x_i\|, i = 1 \dots q)$ {q number of SNPs}

$R_{Y|Z} \leftarrow Y_{norm} - Z_w \hat{\beta}_1$ with $\hat{\beta}_1 = {Z_w}^T Y_{norm}$ {residuals}

$R_{X|Z} \leftarrow X_{norm} - Z_w \hat{\beta}_2$ with $\hat{\beta}_2 = {Z_w}^T X_{norm}$ {residuals}

$cache \leftarrow Z_w, R_{Y|Z}, R_{X|Z}$

else

$Z_w, R_{Y|Z}, R_{X|Z} \leftarrow cache$

end if

CPU

$\hat{\beta} \leftarrow {R_{X|Z}}^T R_{Y|Z}$

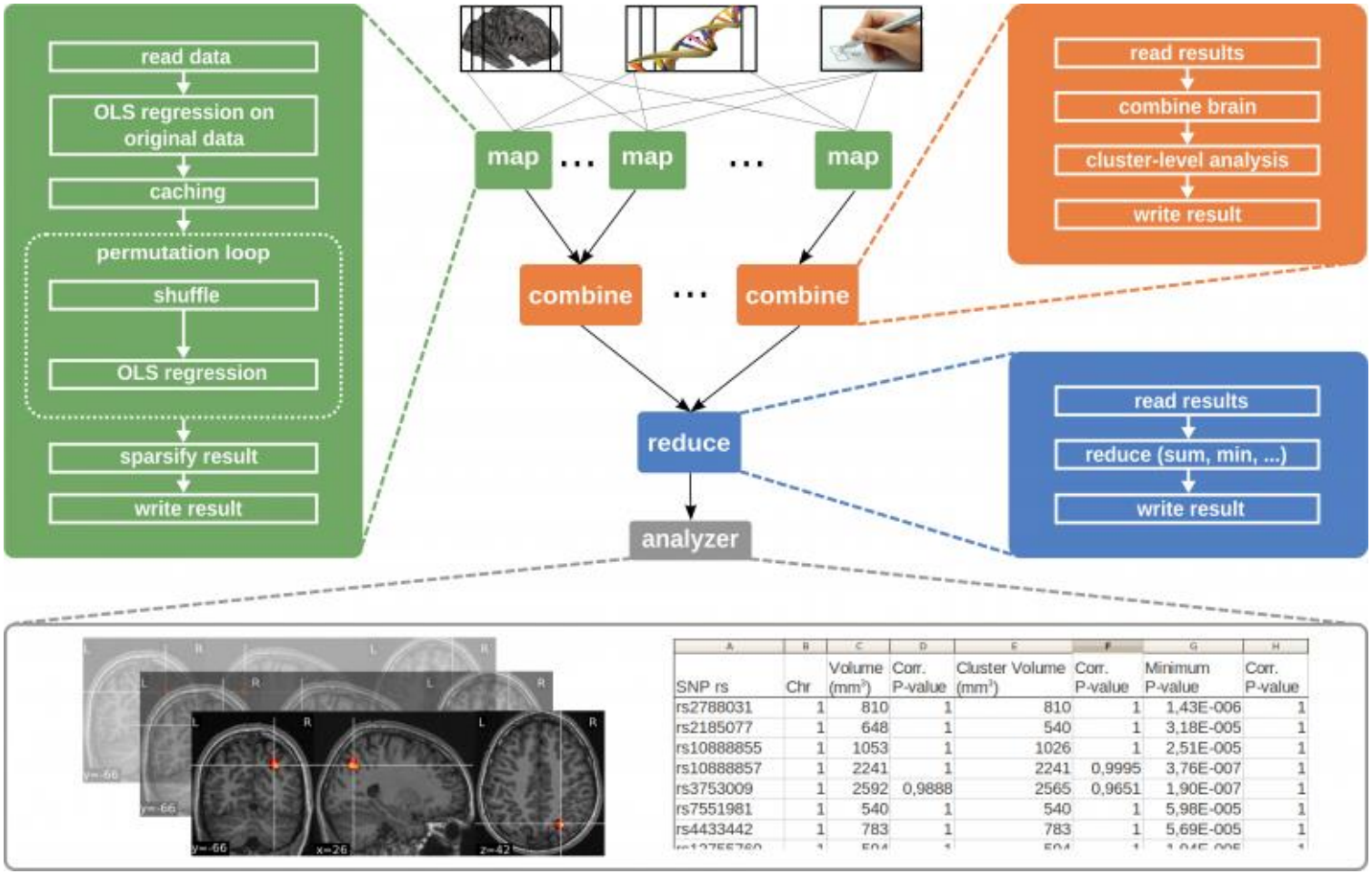
$\hat{\gamma} \leftarrow {Z_w}^T R_{Y|Z}$ {for Freedman and Lane approximation}

F-scores $\propto \frac{\hat{\beta}^2}{1 - \hat{\beta}^2 - \hat{\Gamma}^2}$ with $\hat{\Gamma}^2 = \sum_{i=1}^{r} \hat{\gamma}_i^2$ {r = number of confounding variables}

GPU

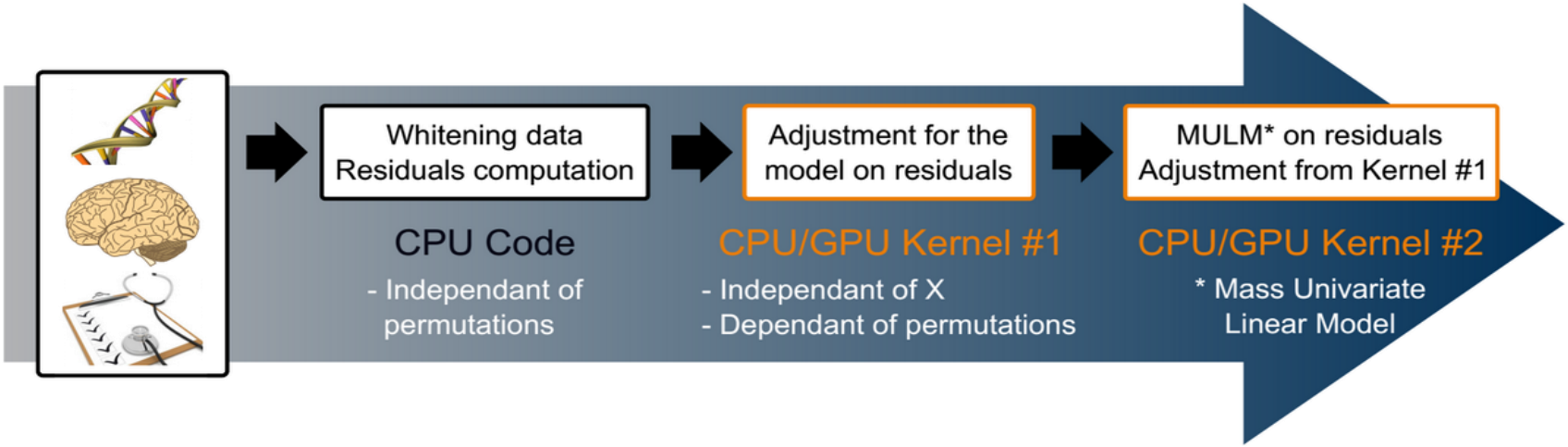
return $\hat{\beta}$, F-scores

Stratégie Map-Reduce



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

DÉTAILS D'IMPLÉMENTATION



Généralités

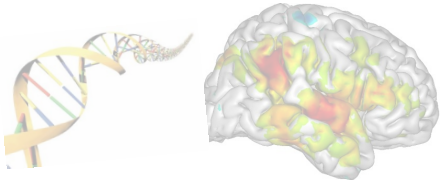
- Code CPU en Python
- Code GPU en CUDA C
- Kernel #1 et #2 en Python pour la version hétérogène
- MapReduce avec Soma-Workflow

Spécifiques aux GPU

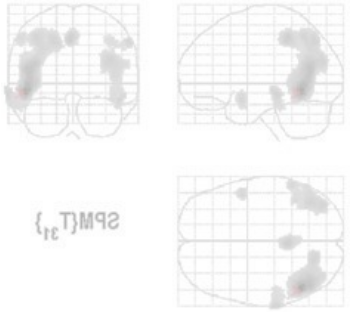
- Permutations réalisées en mémoire partagée (shared) pour maximiser la réutilisation (kernel #2)
- Optimisations : alignements en cache et accès coalescents
- Code réglé finement pour l'architecture cible

SOMMAIRE

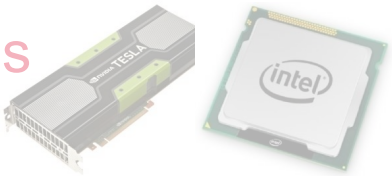
Introduction à l'imagerie génétique



Etudes d'association
génome-entier, cerveau-entier

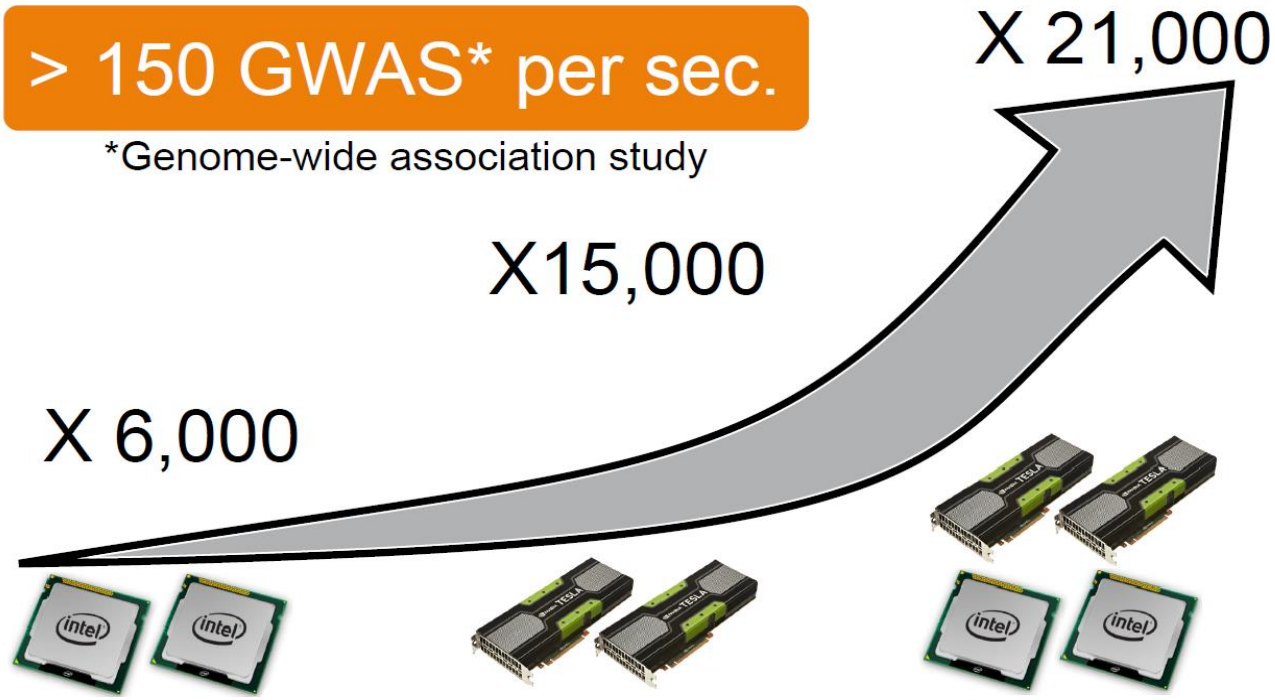


Eléments d'implantation des calculs



Résultats

PERFORMANCES SUR NŒUD GPU



Comparaison avec ***Voxelwise genome-wide association study (vGWAS)***, Stein J.L et al. (2010), NeuroImage 2010, 53(3), pp. 1160-74, PMID:20171287.

PERFORMANCES SUR CLUSTER GPU

Pour le jeu de données ciblé, nous avons utilisé la tranche hybride (GPU) du supercalculateur Curie (PRACE Preparatory Access)

- Plus grand calcul de neuroimagerie-génétique jamais réalisé

  Gestion multi-machines à l'aide de soma-workflow
  18 tâches de 3 h 20, jusqu'à 200 GPU simultanément
  Une durée totale de 60 h, l'équivalent de 12000 h sur 1 GPU
  Post-traitements également réalisés sur Curie

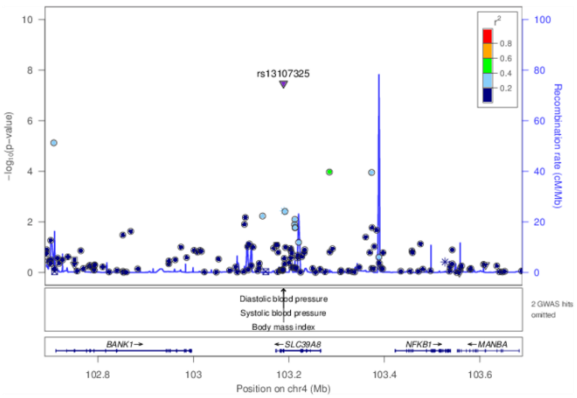
- 2 associations significatives qui demandent de plus amples investigations

Nœuds Curie Hybrid

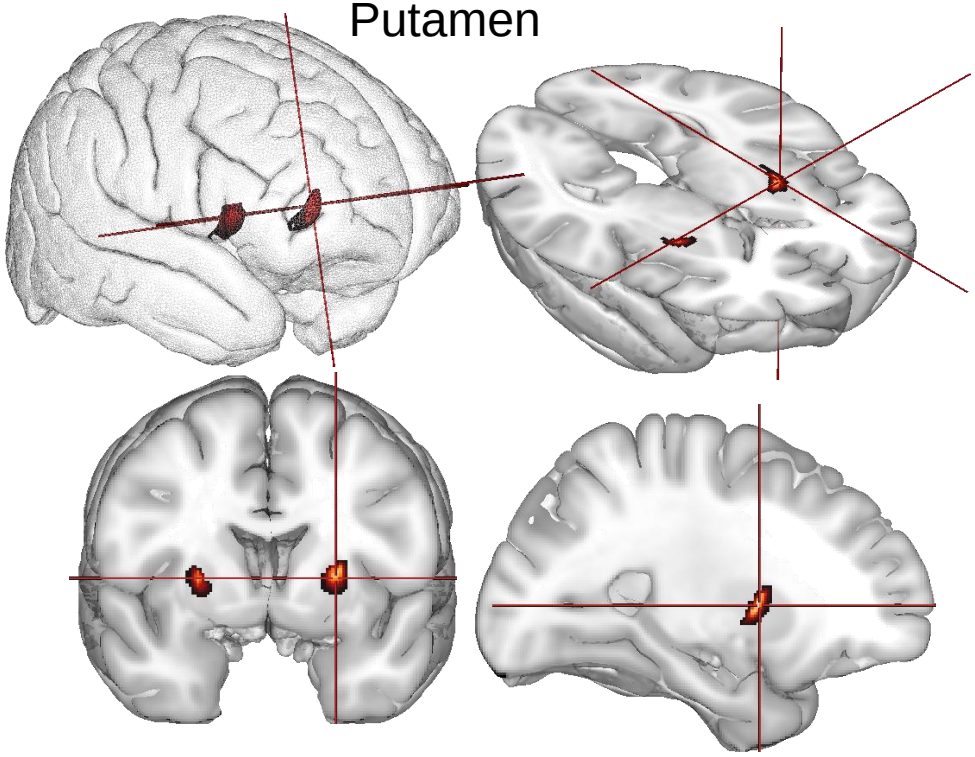
2 × Intel Xeon E5640
2 × Nvidia M2090

RÉSULTATS NEUROSCIENCES

SNP : rs13107325



Putamen



Chromosome 4q22 dans exon 8 du gene SLC39A8 (ZIP8) (metal ions transporter)

Obésité [Speliote 2011, Juonala 2011]
Schizophrénie [Carrera 2012]

TOC [De Wit 2014]
Schizophrénie [Van Erp 2014]

LE CALCUL GPU: UNE ÉVIDENCE ?

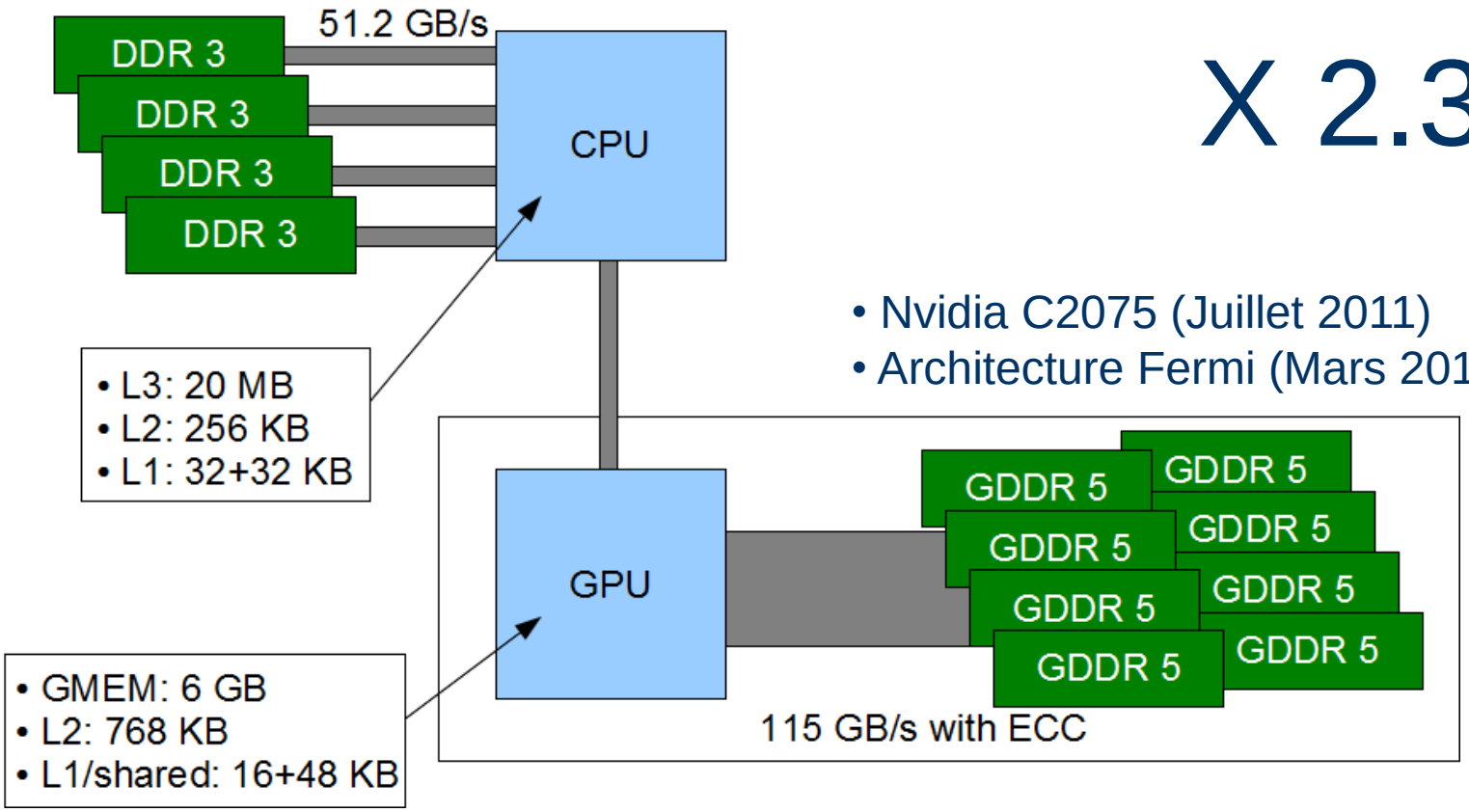
Green500 Rank	MFLOPS/W	Site*	Computer*	Total Power (kW)
1	4,503.17	GSIC Center, Tokyo Institute of Technology	TSUBAME-KFC - LX 1U-4GPU/104Re-1G Cluster, Intel Xeon E5-2620v2 6C 2.100GHz, Infiniband FDR, NVIDIA K20x	27.78
2	3,631.86	Cambridge University	Wilkes - Dell T620 Cluster, Intel Xeon E5-2630v2 6C 2.600GHz, Infiniband FDR, NVIDIA K20	52.62
3	3,517.84	Center for Computational Sciences, University of Tsukuba	HA-PACS TCA - Cray 3623G4-SM Cluster, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband QDR, NVIDIA K20x	78.77
4	3,208.83	CINECA	Eurora - Eurotech Aurora HPC 10-20, Xeon E5-2687W 8C 3.100GHz, Infiniband QDR, NVIDIA K20	30.70
5	3,185.91	Swiss National Supercomputing Centre (CSCS)	Piz Daint - Cray XC30, Xeon E5-2670 8C 2.600GHz, Aries interconnect , NVIDIA K20x	1,753.66
6	3,130.95	ROMEO HPC Center - Champagne-Ardenne	romeo - Bull R421-E3 Cluster, Intel Xeon E5-2650v2 8C 2.600GHz, Infiniband FDR, NVIDIA K20x	81.41
7	3,068.71	GSIC Center, Tokyo Institute of Technology	TSUBAME 2.5 - Cluster Platform SL390s G7, Xeon X5670 6C 2.930GHz, Infiniband QDR, NVIDIA K20x	922.54
8	2,702.16	University of Arizona	iDataPlex DX360M4, Intel Xeon E5-2650v2 8C 2.600GHz, Infiniband FDR14, NVIDIA K20x	53.62
9	2,629.10	Max-Planck-Gesellschaft MPI/IPP	iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband, NVIDIA K20x	269.94
10	2,629.10	Financial Institution	iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband, NVIDIA K20x	55.62
▪ ▪ ▪				
46	1,247.57	Météo France	Bullx DLC B710 Blades, Intel Xeon E5 v2 12C 2.700GHz, Infiniband FDR	401.00

Novembre 2013 : les 10 machines les plus efficaces utilisent des GPU Nvidia k20

QUEL GAIN POUR NOTRE APPLICATION ?

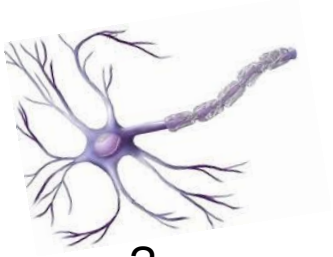
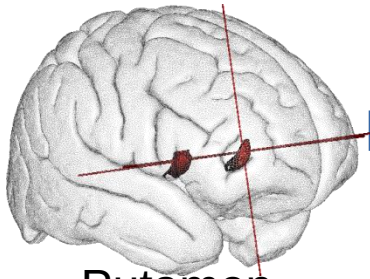
- Intel Xeon E5-2620 (Avril 2012)
- Architecture Sandy Bridge (Janvier 2011)

Accélération X 2.3



- Nvidia C2075 (Juillet 2011)
- Architecture Fermi (Mars 2010)

CONCLUSION ET PERSPECTIVES

Génétique	Cellulaire	Cérébral	Maladie
			
SLC39A8 (ZIP8)	?	Putamen	Obésité TOC Schizophrénie

Résultats :

- Marqueurs génétique → Imagerie → pathologies

Perspectives neursciences :

- Mécanismes cellulaires / effet sur le putamen
- Expression de ce gène dans le putamen (Allen atlas)
- Liens avec les pathologies

Perspectives méthodologiques : Code GW-BW sur GPU disponible opensource

BRAINOMICS
IA Bioinfo 2010

NeuroSpin

E Duchesnay, V Guillemot
V Frouin T Löfstedt

Supélec

A Tenenhaus

Centre Hospitalier Sainte Anne

LPMP,UMR 894, INSERM

A Cachia, MO Krebs

Institut de cancérologie GUSTAVE ROUSSY
VILLEJUIF - www.igr.fr

UMR CNRS 8203-IGR

C Philippe, J Grill

EOLEN
CONSEIL ET INGÉNIERIE

B DaMota, I Cadenne,
S Monot, V Ducrot

Logilab

V Michel, O Cayrol

KEOSYS
Medical Imaging

Ph Tervé, J Beranger

ANR GENIM Li J

PRACE Li J, B Da Mota

BrainOmics team:

E Duchesnay
D Papdopoulos
V Frouin

F Hadj-Selem, V Guillemot,
T Lofstedt, C Philippe, J Le

JB Poline
B. Thyreau, A Barbot, Y Schwarz
C Lanquetuit, D Goyard

Collaborations:

IMAGEN cons. EU-AIMS cons. (D Goyard)
ANR/Brainomics ANR/Genim

NeuroSpin PF. (A. Grigis)
LNAO(JF Mangin), PARIETAL (B Thirion)

Commissariat à l'énergie atomique et aux énergies alternatives
Centre de Saclay | 91191 Gif-sur-Yvette Cedex

Etablissement public à caractère industriel et commercial | RCS Paris B 775 685 019

DSV
I2BM
NeuroSpin