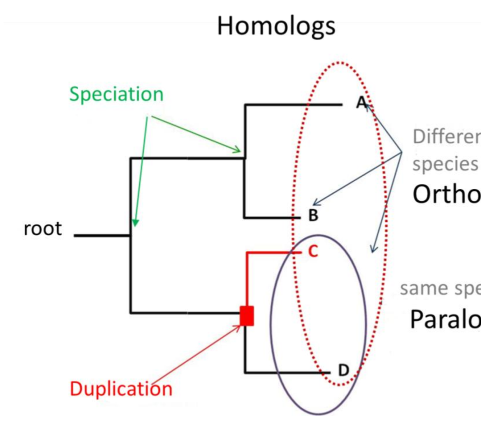
Expression of paralogous gene families in the human brain

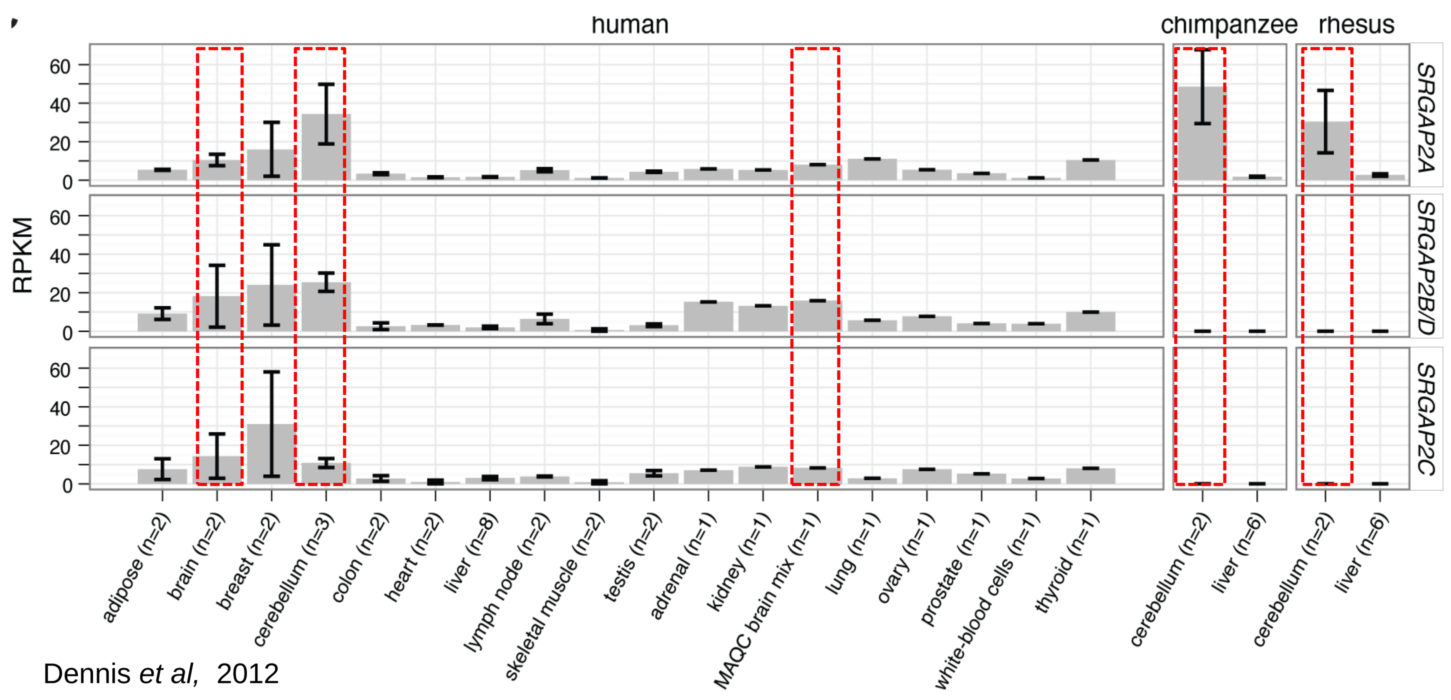
Solène JULIEN[1 2 3], Edith LE FLOCH[2], Christophe BATTAIL[2], François ARTIGUENAVE[2] and Vincent FROUIN[1]
[1]CEA/DSV/I[2]BM/Neurospin, [2]CEA/DSV/IG/CNG/LBI, [3]Phd Student julien@cng.fr

The **tissue-specificity** of **paralogous gene** functions can be brought out by measuring gene expression and members of a **gene family** can show **specific expression profiles** in **different tissues**. Our **main objective** is to **identify gene families** having specific **expression patterns** between **brain tissues** or shared between brain and accessible tissues. To reach this goal we have to deal with genomic regions of paralogs with a **high sequence homology** that can impact RNA-seq expression measurements. By looking to the **mappability** of duplicated genes we estimate the proportion of **problematic genes** and genomic regions. Moreover we **extract representative expression profiles** of gene families in **brain** and **blood** by using expression **correlation** and **differential expression** analyses.

Definition & Context

Paralogy
Phylogenetic link between two **duplicated genes.**
Gene family
Group of **paralogous genes.**
Subfunctionalization
Same function between two copies of one duplicated gene but each one is **specific** to a particular **tissue.**

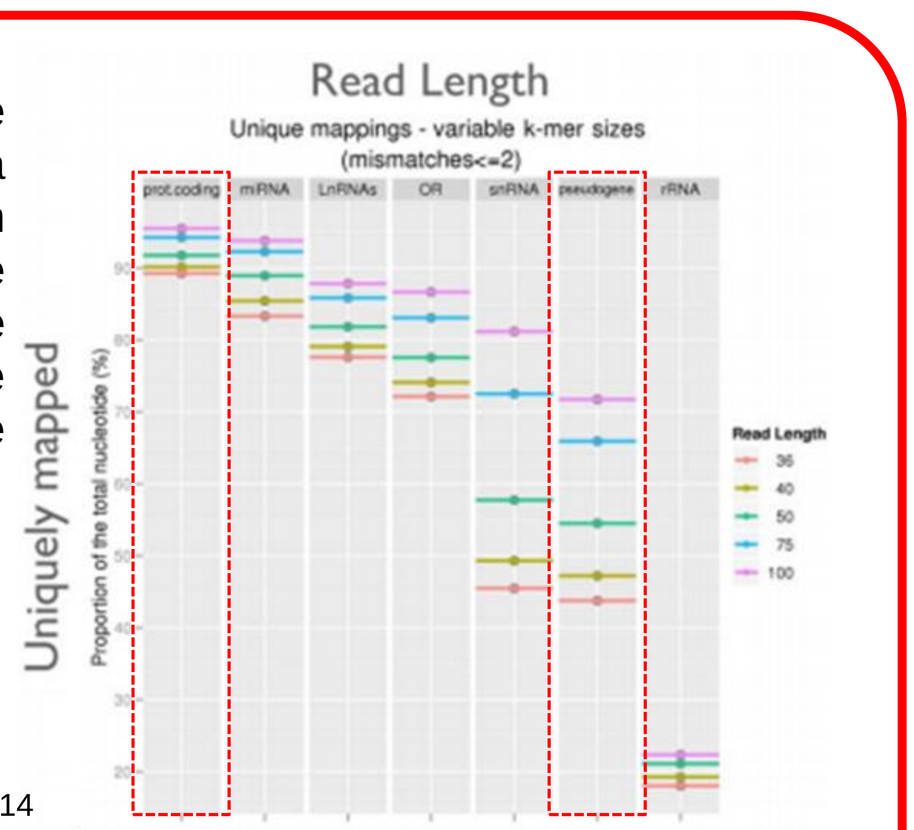




Dennis *et al*, 2012

- ❑ Copies of **SRGAP2 gene** are expressed in **brain tissues.**
- ❑ Highlight **tissue-specificity** in human brain of the **SRGAP2** duplicated gene with gene **expression** measures (cerebellum).

Mappability
The **inverse** of the **number of times** that a read originating from any position in the reference sequence **maps** to the sequence itself (identify unique mappings).



Derrien *et al*, 2014

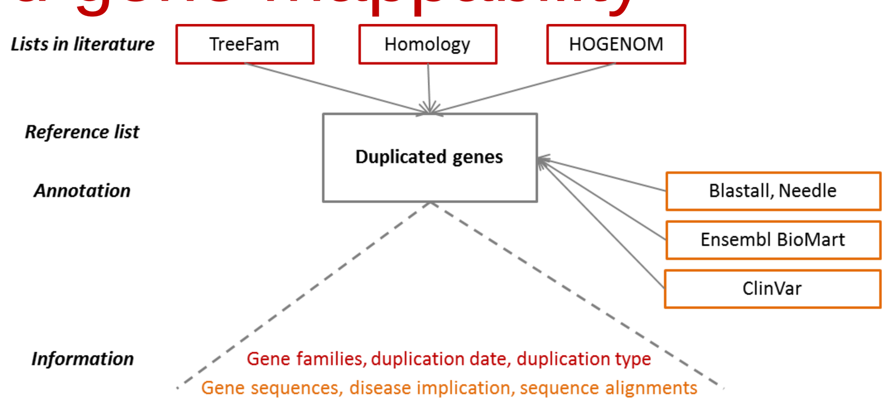
- ❑ ~5% of protein coding genome is not uniquely mapped.
- ❑ Duplicated genes can be difficult to map (only 65% of pseudogenes is “mappable”).

Objectives

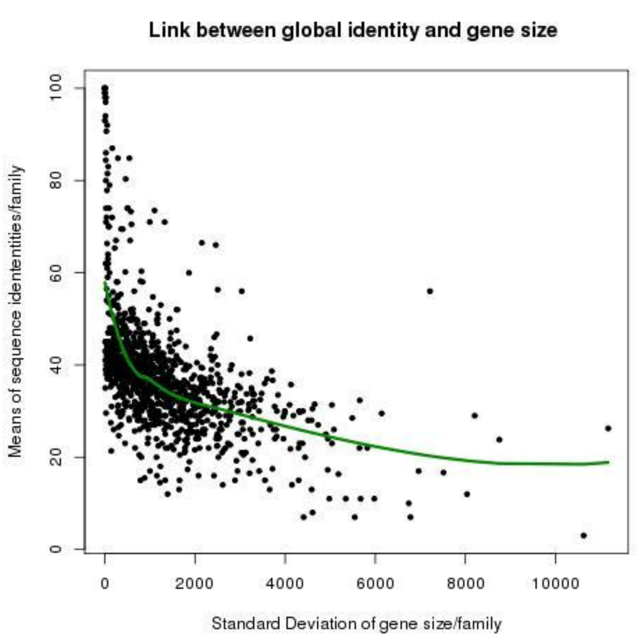
1) Study on the mappability of duplicated gene transcriptome
2) Identification of interesting expression profiles of gene families by differential expression studies

1) The duplicated gene mappability

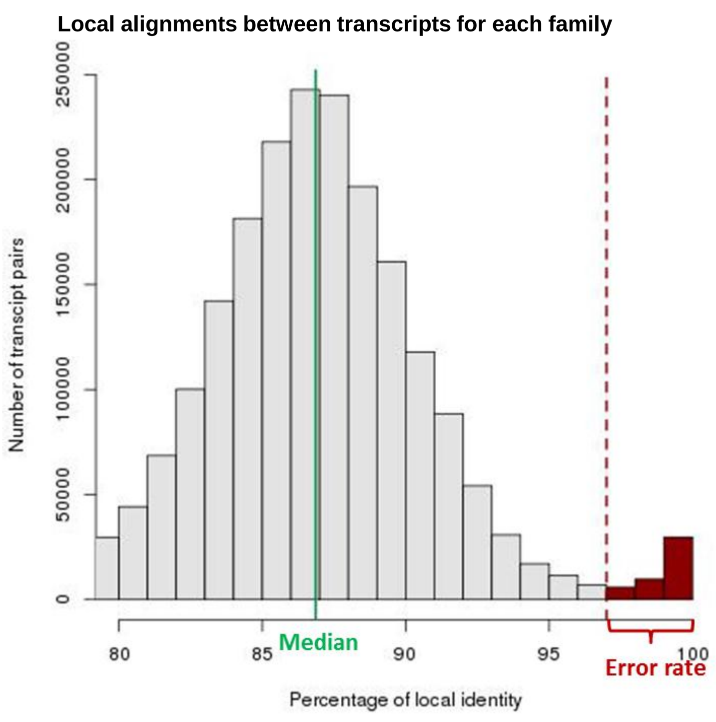
✓ Duplicated gene reference list:



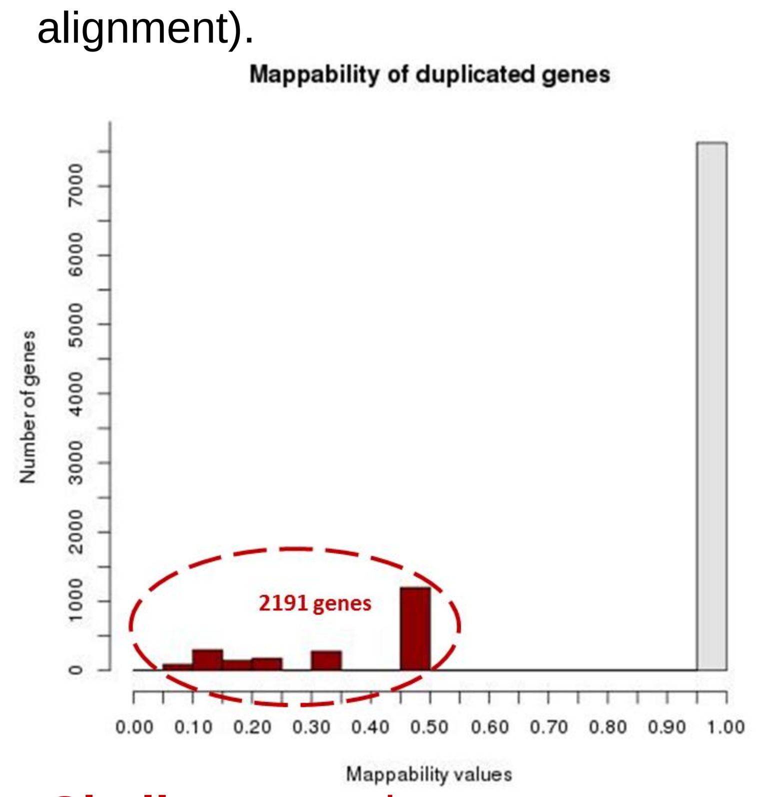
Identification of problematic genes for the mapping of sequencing reads:



- ❑ Families contain genes with **homogenous** sizes.
- ❑ The **higher** is the **gene size variance** per family, the **weaker** is the percentage of **identity** per family.
(Correlation=-0,33 and p-value<2.10^{-16})



- ❑ **915** duplicated genes with regions **>98%** of identity and **>80pb** (39496 transcript pairs).
- ❑ Regions with high sequence identity.
=> difficulties to map reads on the right gene.
- ❑ **All genes** with **regions > 80%** of identity can be **problematic** for the **mapping** step (Depending on the region size for the local alignment).



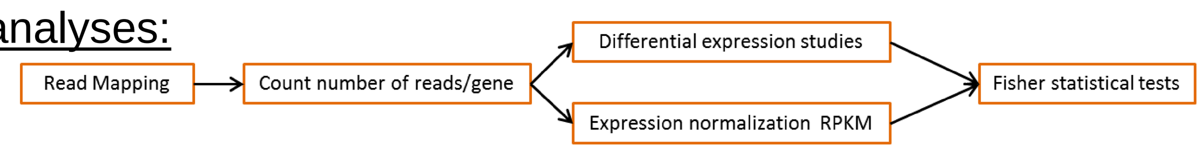
- ❑ **Mappability** of regions of **100pb** in the duplicated gene transcriptome **without mismatches** allowed.
- ❑ In reality **2191 genes** have at least one **region** that is **not uniquely mapped** on the duplicated gene transcriptome.
- ❑ **88%** of the total **nucleotides** on the duplicated gene transcriptome are **uniquely mapped.**

➢ **Challenge** on the measurement of gene **expression** for genes of same family with strong sequence **homology.**

2) Expression analyses of gene families

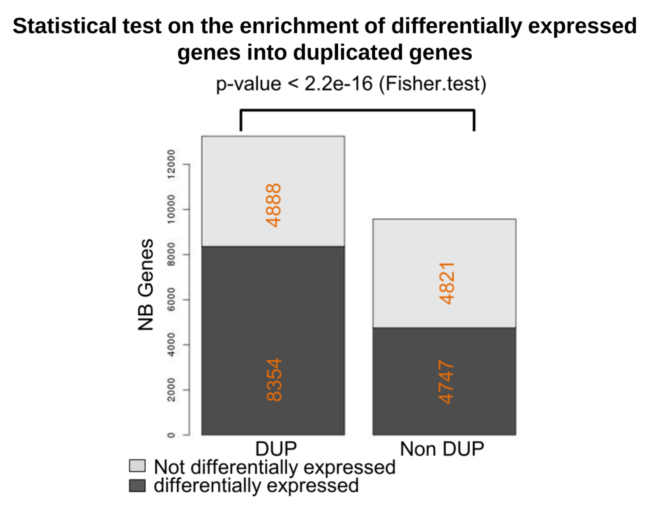
✓ RNA-Seq Sequencing Data:
- ➢ **Human Brain Reference** sample of 23 pooled patients (SEQC consortium, 2014).
- ➢ **Whole blood** sample of 6 pooled patients (Shin *et al*, 2014).

✓ Workflow of expression analyses:

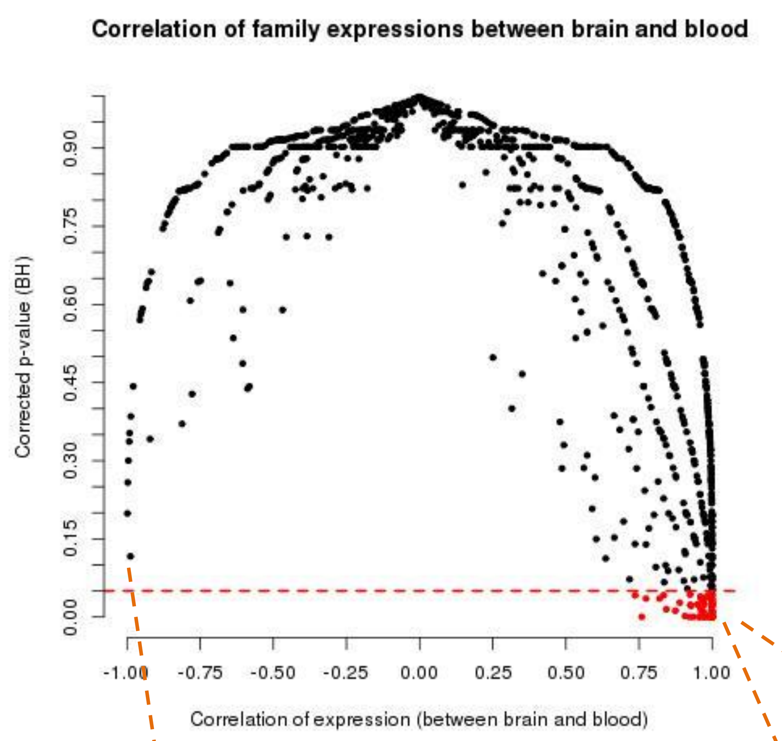


Gene expression analyses between brain and blood :

Duplicated genes are **enriched** in **differentially expressed genes** (63%) compared to all **non duplicated coding genes** (57%).



Identification of different expression profiles of families:

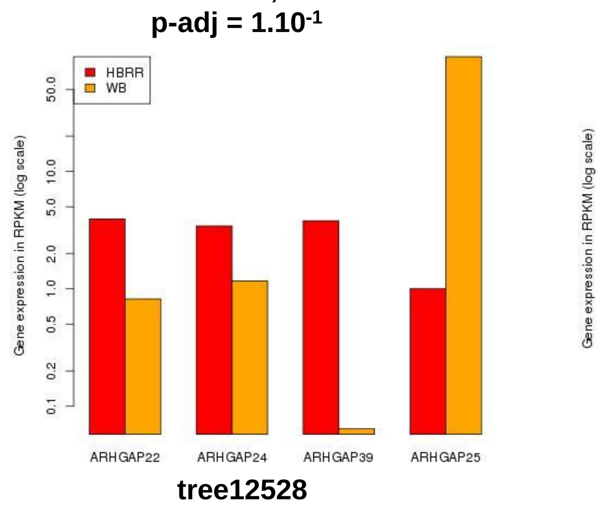


- ❑ **Correlation tests** of gene expressions between brain and blood tissues for **each family.**
- ❑ Find **families** with **significantly correlated** or **anti-correlated** expression profiles.

Differentially expressed
❑ **Tissue-specificity**

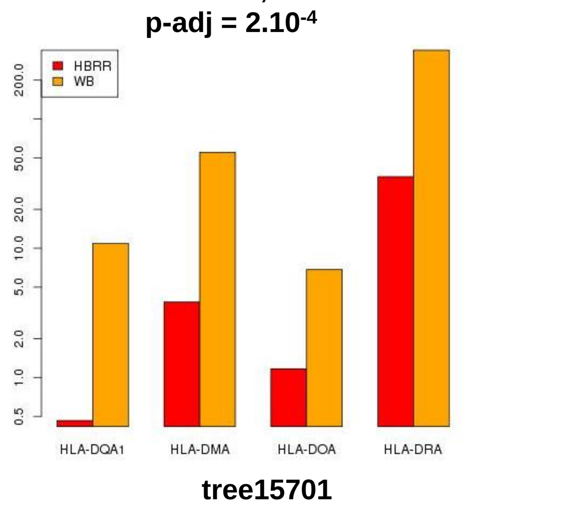
Not differentially expressed
❑ **Potential biomarkers**

Cor = -0,98
p-adj = 1.10^{-1}



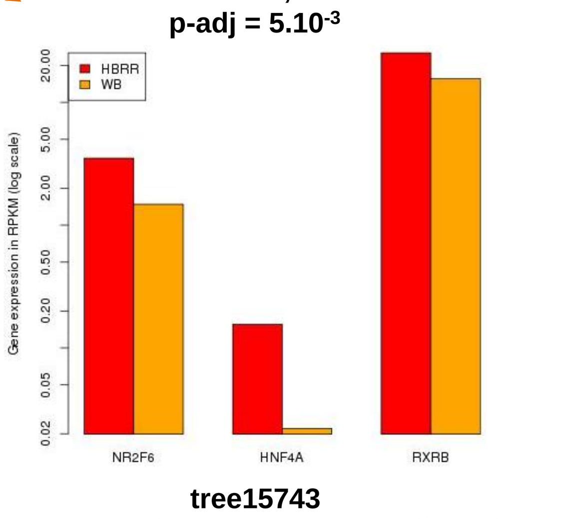
tree12528

Cor = 0,99
p-adj = 2.10^{-4}



tree15701

Cor = 0,99
p-adj = 5.10^{-3}



tree15743

➢ Find **specific expression** profiles for **biomarkers** and **tissue-specificity** between different tissues from **shared individuals.**

Perspectives

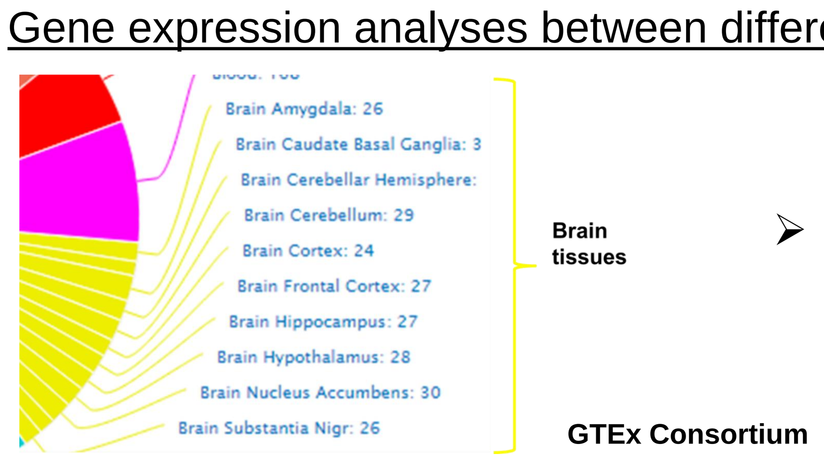
Manage to differentiate brain tissues from expression profiles of gene families

Improving the expression measurement of duplicated genes

- Bioinformatics method to improve **brain reference transcriptome** based on **hybrid** RNA sequencing.
- Cope with complex splicing and high sequence local homology with PacBio long reads.
- Accurate estimation of transcript abundance with RNA-Seq deep coverage.

PACBIO ISO-SEQ	ILLUMINA RNA-SEQ
Long reads	Short reads
Low coverage	Deep coverage
High error rate	Low error rate

Gene expression analyses between different brain tissues



GTEx Consortium

➢ Identification of **gene family expression patterns specific or shared between brain tissues.**

CEA
CNG Centre National de Génotypage
NeuroSpin
Université Paris Sud
Université Paris-Saclay
Laboratoire de Mathématiques et Modélisation d'Évry LaMME