# Integrating multimodal databases into genomics analysis workflows

Xavier Lecours Boucher[1], Samir Das[1], Christine Rogers[1], David MacFarlane[1], Dahlia Snaiderman[1], Kathleen Klein Oros[2], Natacha Beck[3], Marie Forest[2], Lawrence Chen[2], Nicolas Brossard[4], Eva Unternaehrer[4], Kieran O'Donnell[4], Helene Gaudreau[4], Michael Meaney[4], Celia Greenwood[2], Alan Evans[1]

[1]Montreal Neurological Institute, McGill University, Montreal, Quebec, [2]Lady Davis Institute for Medical Research, Montreal, Quebec, [3]McGill Centre for Integrative Neuroscience, Montreal, Quebec, [4]Douglas Hospital Research Center, Montreal, Quebec

http://mcin.ca

## INTRODUCTION

The Ludmer Centre for Neuroinformatics & Mental Health (ludmercentre.ca) has created a (epi)genetics working group composed of three member institutions.
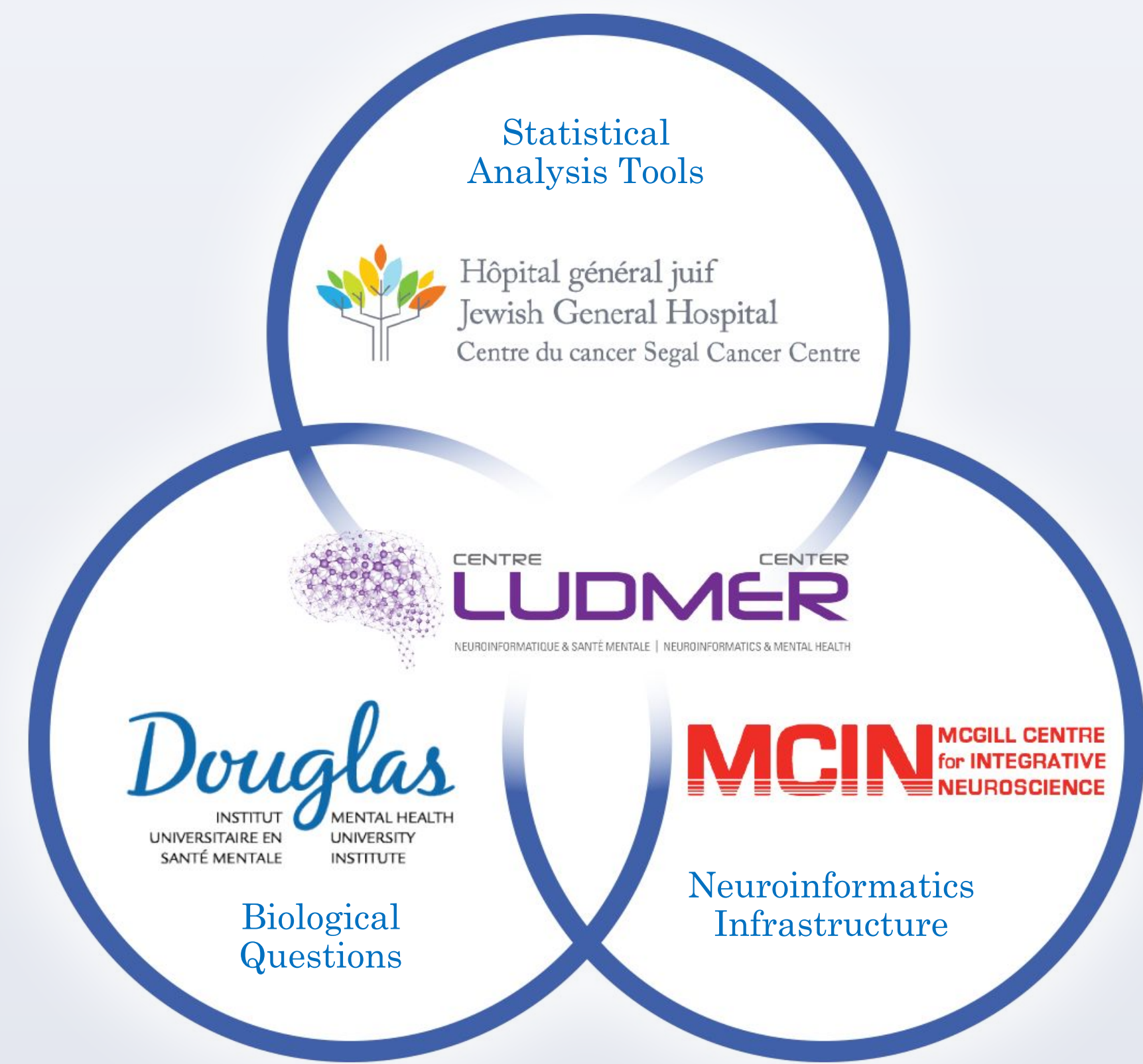


Figure #1: *Neuroinformatics conglomerate*

## OBJECTIVES & METHODS

The group's goal is to 1) integrate heterogeneous datasets (behavioral/clinical, imaging, and genomic) within the LORIS[1][3] (loris.ca), 2) streamline analysis, 3) integrate results and 4) facilitate structured and versioned sharing for optimized reproducibility using high performance computing (HPC) on CBRAIN[9].

Biospecimens are collected during a participant's visit. DNA is extracted, undergoes quality control, and shipped for genotyping and methylation assessment using a bead chip array. Raw data are transferred to CBRAIN where a normalization pipeline can be launched, returning results back into the LORIS database along with metadata.
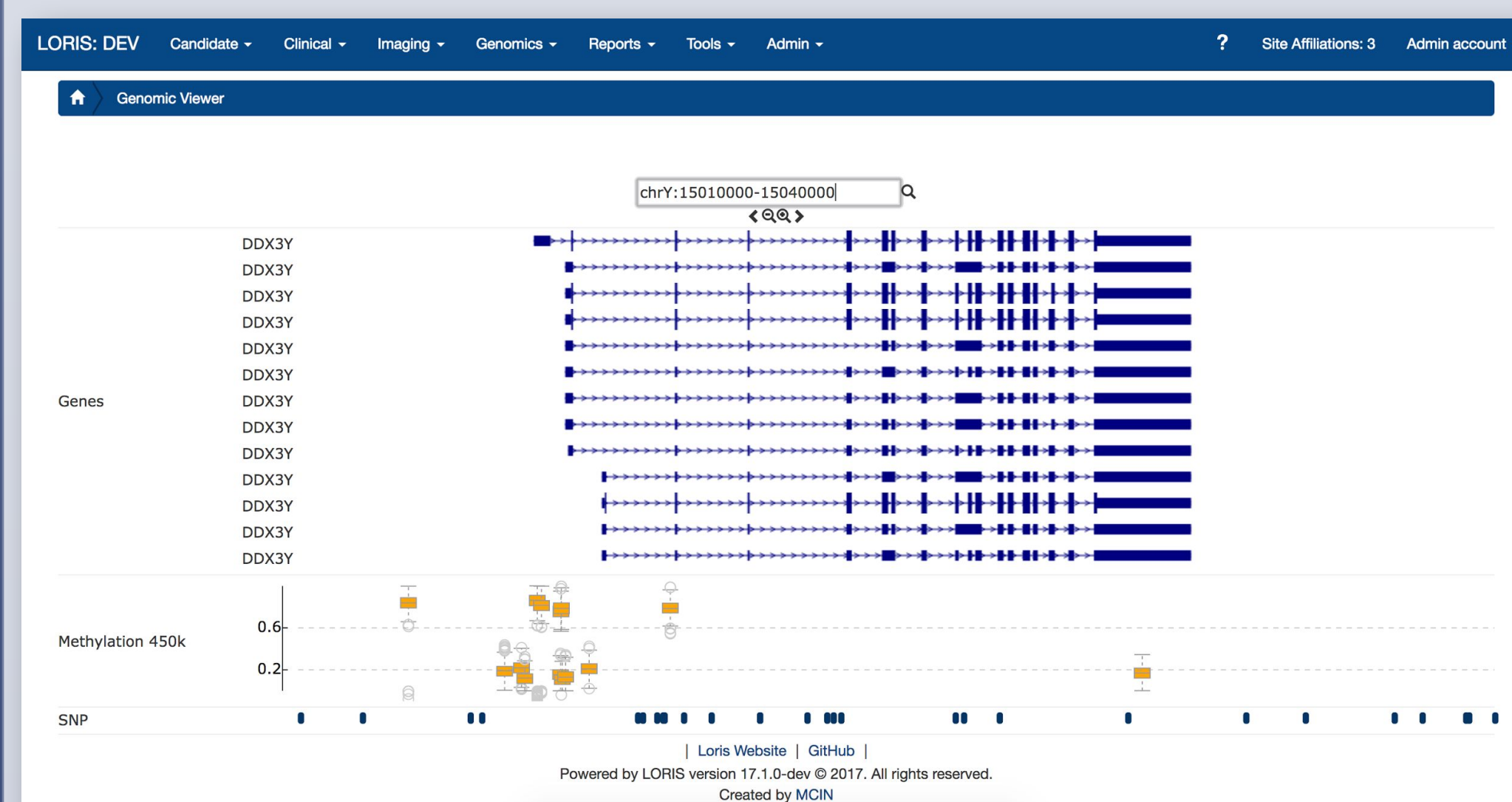


Figure #2: *LORIS web interface showing genomic data (Box plot methylation beta-values , SNP and genes aligned on the genome ) in the visualization tool (beta).*

## RESULTS

The LORIS Genomic Browser[8] enables viewing, filtering, and linking of summary genetic data (CNV, SNP, GWAS). Using the MAVAN study[7], a subset of subject SNPs from PLINK files have been loaded into LORIS. An uploader and genomic viewer were created so researchers can view aggregated CPG beta value distributions visually aligned with their SNP data, gene features and ENCODE H3K4me1 Histone mark data for a given genomic range. Each displayed element can be clicked to access external reference databases (dbSNP, 1000 Genomes), or provide more information, such as methylation level boxplot distributions of beta values grouped by sex and/or genotypes for a specific CpG and a given SNP. The design of complex workflows was also a significant part of the methodology, enabling requisite tool building and porting at various stages of development (Figure 3).
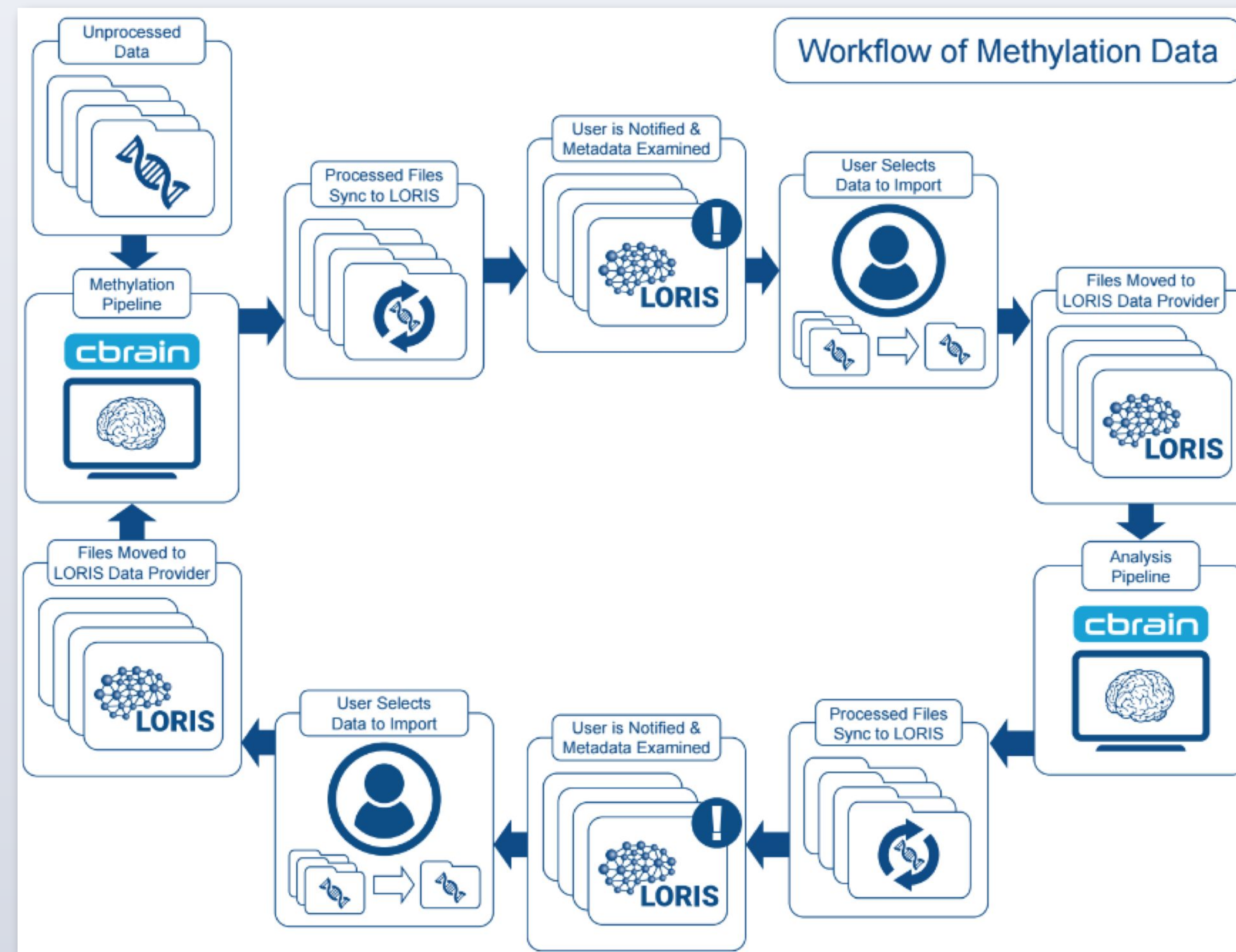


Figure #3: Methylation data processing as it cycle between storage in LORIS's database and dockerized processing pipelines hosted on CBRAIN. The outputs are re-inserted into LORIS where it can be visualization and cross modale querying occurs to create CBRAIN input datasets. In CBRAIN, a second pipeline can be launch to add derived variables back in LORIS.

Within LORIS, a prototype Dataset Builder creates new datasets by joining filtered genomic data with phenotypic data and/or imaging files queried from the Data Querying Tool (DQT)[6]. These new datasets are processed by CBRAIN statistical pipelines. Several prototype HPC workflows have been constructed and mounted in CBRAIN for the MAVAN project, such as methylation450KPipeline[10] and imputePrepSanger[5] running in a Docker container[4] along with weighted correlation network analysis (WGCNA) and principal component analysis (PCA) pipelines.

## ACKNOWLEDGEMENTS

Derek Lo (diagrams), Pierre Rioux (CBRAIN support), Alex Ilea (React.js), Jacob Penny (React.js)

## CONCLUSION

The goals of this working group are to facilitate brain research discovery by:

- Reducing human error with processing automation and seamless linking of multimodal data.
- Creating format definitions between components of this workflow will improve the integrative impact of these cross-modal tools.
- Deploying analysis pipelines on HPCs, focusing investigators on research instead of data handling.
- Optimizing performance, flexibility, and scalability using NoSQL and structured schema databases.
- Creating hooks between CBRAIN and LORIS for automated task launching to streamline setup, task creation, provenance capture, and re-insertion.
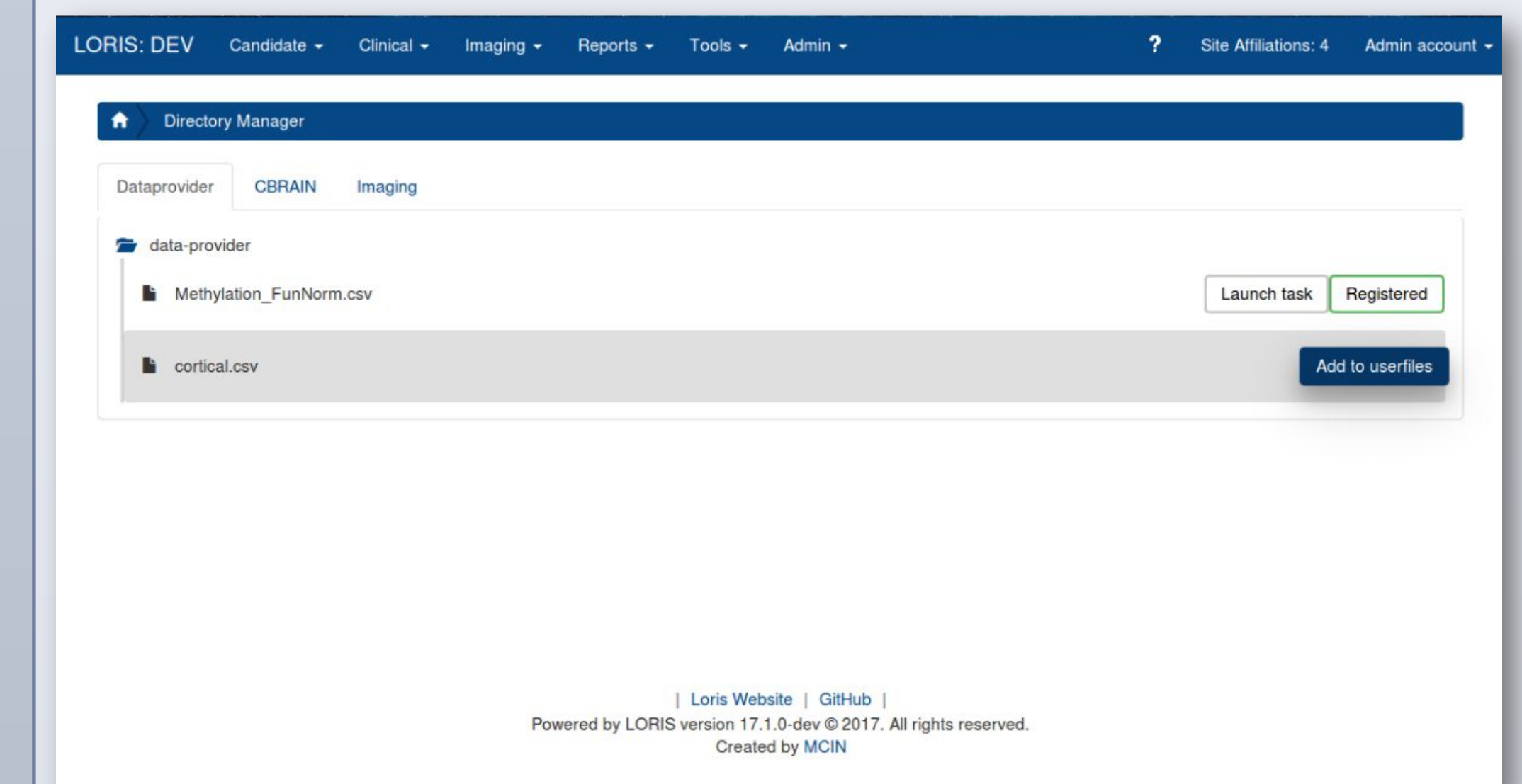


Figure 4: *LORIS web interface showing files inked between CBRAIN and LORIS.*

## REFERENCES

[1] Das, S. (2011), 'LORIS: A web-based data management system for multi-center studies', Frontiers in Neuroinformatics, 5:37.

[2] Das, S. (2015), 'The MNI data-sharing and processing ecosystem', NeuroImage 124:1188–1195.

[3] Das, S. (2016), 'Cyberinfrastructure for Open Science at the Montreal Neurological Institute', Frontiers in Neuroinformatics, DOI: 10.3389/fninf.2016.00053

[4] Glatard, T. (2015). 'Boutiques: an application-sharing system based on Linux containers', Frontiers in Neuroscience, DOI: 10.3389/conf.fnins.2015.91.00012

[5] Forest, M. (2016), Docker repository, https://hub.docker.com/r/eauforest/imputeprepsanger

[6] MacFarlane, D., OHBM 2014

[7] O'Donnell, K. A., et al., 2014. 'The maternal adversity, vulnerability and neurodevelopment project: theory and methodology', Canadian Journal of Psychiatry, 2014 Sep;59(9):497-508.

[8] Rogers, C., OHBM 2015

[9] Sherif, T., et al., 2014. CBRAIN: A web-based, distributed computing platform for collaborative neuroimaging research. Frontiers in Neuroinformatics, DOI: 10.3389/fninf.2014.00054

[10] Statistical Genetics at the Lady Davis Institute, (2016), GitHub repository, https://github.com /GreenwoodLab/methylation450KPipeline