

Galaxy Community Conference 2012

**Integration of S-MART, a toolbox to
aid RNA-seq data analysis in Galaxy**



URGI INRA Versailles
yufei.luo@versailles.inra.fr

ALIMENTATION
AGRICULTURE
ENVIRONNEMENT

INRA

High-throughput sequencing technologies



- ✧ Millions of reads
- ✧ Low cost
- ✧ cDNA samples (RNA-seq) :
 - transcriptome of eukaryotic genomes
 - gene expression measurement
 - unbiased and comprehensive manner for analyzing transcriptome
- ✧ Mapping high-throughput sequencing tools
- ✧ Mapped reads analysis tools?

S-MART[1]

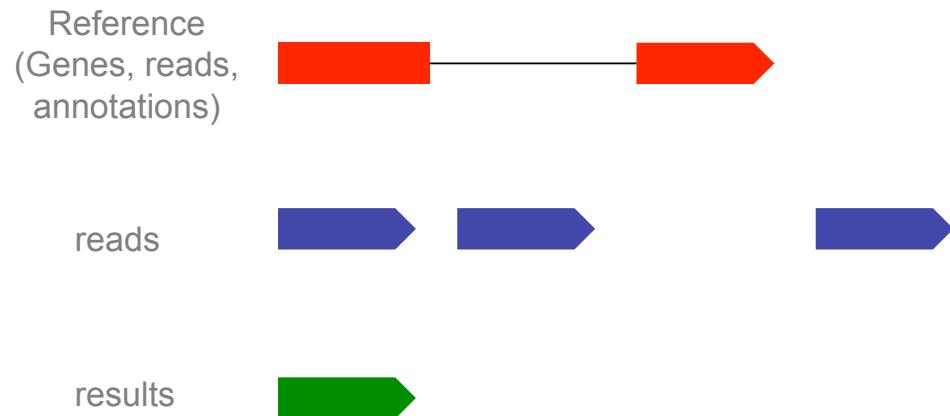
--analysis of mapped RNA-seq and CHIP-seq

- **Conversion : Gff*, csv, sam, fastq, fasta, ...**
 - **WIG : exploit Wig information**
 - **Merge**
 - **Comparison**
 - **Selection : Exon, Intron, Flanking, ...**
 - **Modification : genomic coordinates, sequence, adaptor**
 - **Visualization**
-
- [1] : Zytnicki M, Quesneville H (2011) S-MART, A Software Toolbox to Aid RNA-seq Data Analysis. PLoS ONE 6 (10):e25988.doi:10.1371/journal.pone.0025988

S-MART

--Tools for RNA-seq

CompareOverlapping :

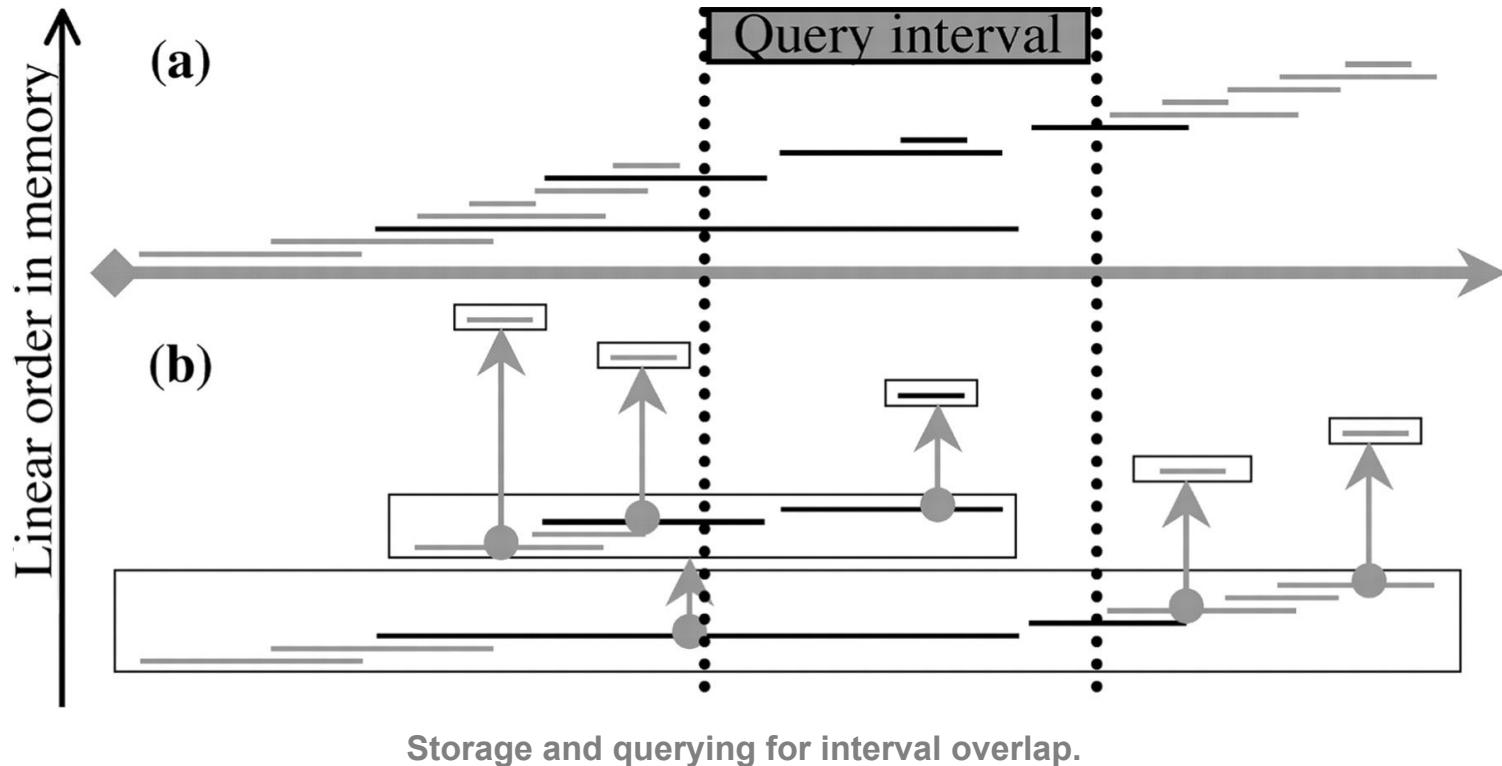


Faster, Lighter Algorithm

S-MART

--Tools for RNA-seq

Nested Containment List (NCLList)^[2]



Alekseyenko A V , Lee C J Bioinformatics
 2007;23:1386-1393

[2] A.Alekseyenko & J.Lee Nested; Containment List (NCLList) : a new algorithm for acceleration interval query of genome alignment and interval databases; Jan 18, 2007; Bioinformatics

S-MART

--Tools for RNA-seq

CompareOverlapping : O(N), where N is the number of short reads. 1h30->20 million reads VS. 250 thousand ref transcripts.

Options :

Restrict to N first nucleotides :



Extension on 5' or 3' direction :



Report introns :



Invert selection :



Colinear/Anti-sense :



Included :

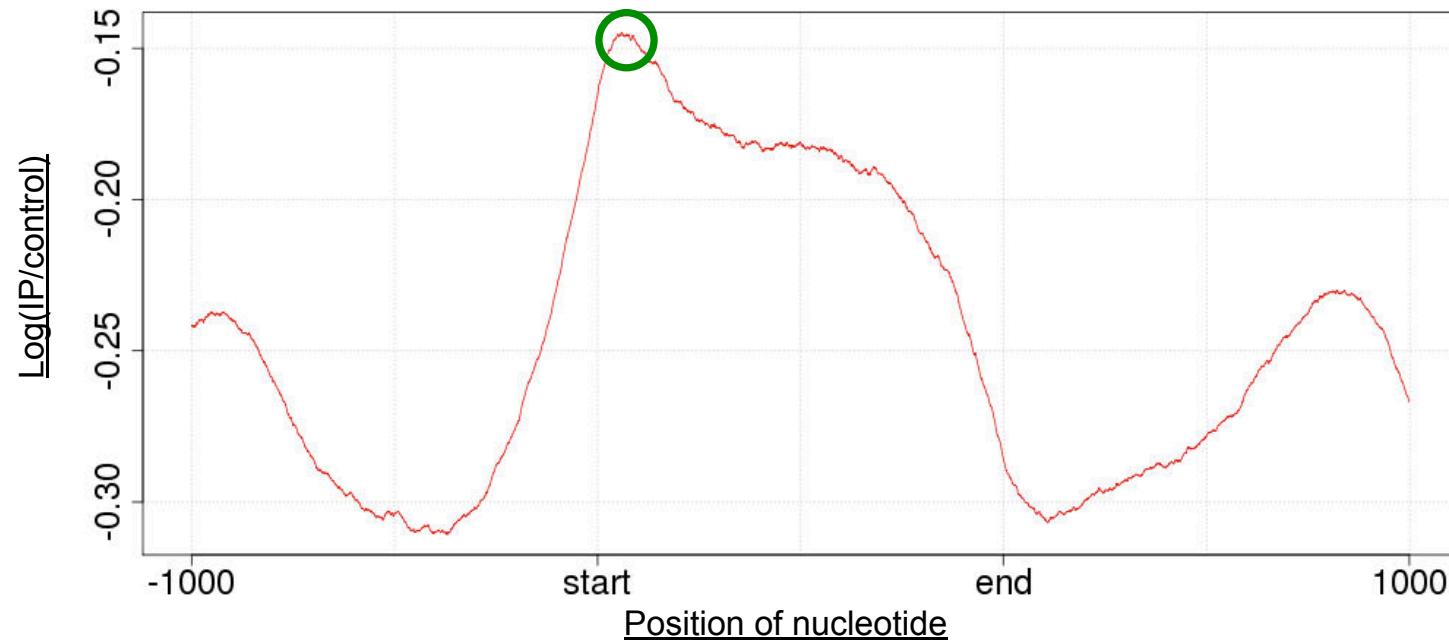


S-MART

--Tools for ChIP-seq

- **getWigProfile**

Wiggle (WIG): display of dense, continuous data (GC percent, probability scores, transcriptome data), the value of each genome nucleotide





S-MART in Galaxy --pipelines

Galaxy

Analyze Data Workflow Shared Data Visualization Help User Using 3%

Tools Options

Tools

URGI-S-MART CONVERSION TOOLS

Gff3 -> Wig Convert Gff3 File to Wig File.

Bed -> Csv Convert Bed File to Csv File.

Bed -> Gff2 Convert Bed File to Gff2 File.

Bed -> Gff3 Convert Bed File to Gff3 File.

Bed -> Sam Convert Bed File to Sam File.

Blast (-m 8) -> Csv Convert Blast (-m 8) File to Csv File.

Blast (-m 8) -> Gff2 Convert Blast (-m 8) File to Gff2 File.

Blast (-m 8) -> Gff3 Convert Blast (-m 8) File to Gff3 File.

Blast (-m 8) -> Sam Convert Blast (-m 8) File to Sam File.

Gff2 -> Csv Convert Gff2 File to Csv File.

Gff2 -> Gff3 Convert Gff2 File to Gff3 File.

Gff2 -> Sam Convert Gff2 File to Sam File.

Gff3 -> Csv Convert Gff3 File to Csv File.

Gff3 -> Gff2 Convert Gff3 File to Gff2 File.

Gff3 -> Sam Convert Gff3 File to Sam File.

Sam -> Csv Convert Sam File to Csv File.

The Galaxy team is a part of BX at Penn State.
This project is supported in part by NSF, NHGRI, and the Huck Institutes of the Life Sciences.

URGI INRA

History Options

upload files 1.2 Gb

210: Trim sequences on data 208

209: FastQC.html

208: Trim sequences on data 206

207: FastQC.html

206: Trim sequences on data 179

205: FastQC.html

179: FASTQ Groomer on data 173

173: 454reads test.fastq

172: s 4 cut 100000.mfq

171: s 3 cut 100000.mfq

170: s 3 cut.mfq

169: s 4 cut.mfq

155: f1cond2.tsv

154: f1cond1.tsv

62: reference.mfa

61: annotation.off

Galaxy is installed on URGI cluster with:

- CPU: 912 (Intel Xeon) / 79 nodes
- RAM max: 512 Gb per job
- Entry point 1: node « www »
- Entry point 2: node « ssh »

Using Sun Grid Engine (for job management) and a PostgreSQL Database (for Galaxy).

<http://urgi.versailles.inra.fr/galaxy>

Contact to urgi-support@versailles.inra.fr to have a count.



URGI INRA Versailles yufei.luo@versailles.inra.fr



S-MART in Galaxy --pipelines

Galaxy

Analyze Data Workflow Shared Data Visualization Help User Using 3%

Tools Options

NGS: Peak Calling
NGS: Simulation
SNP/WGA: Data; Filters
SNP/WGA: QC; LD; Plots
SNP/WGA: Statistical Models
Human Genome Variation
Genome Diversity
VCF Tools

URGI TOOLS

URGI: Get Data for grapevine
URGI: BAC analysis
URGI: MAPHITS - PreProcess
Tools
URGI: MAPHITS - Tools
URGI: MAPHITS - PostProcess
Tools
URGI: MAPHITS - SNPs Chip Tools
URGI: S-MART

APLIBIO TOOLS

APLIBIO: Pipeline

Workflows

- ANTISENS detection pipeline
- CIS detection pipeline
- Plot the H3K36me3 histone modification around gene TSS (CHIP-Seq)
- compare RNA-Seq with tiling arrays using sliding windows
- 2 mappers
- piRNA clusters

- TRANS detection pipeline
- Differential expression DESeq (without replicates)

The Galaxy team is a part of BX at Penn State.
This project is supported in part by NSF, NHGRI, and the Huck Institutes of the Life Sciences.

upload files 1.2 Gb

History Options

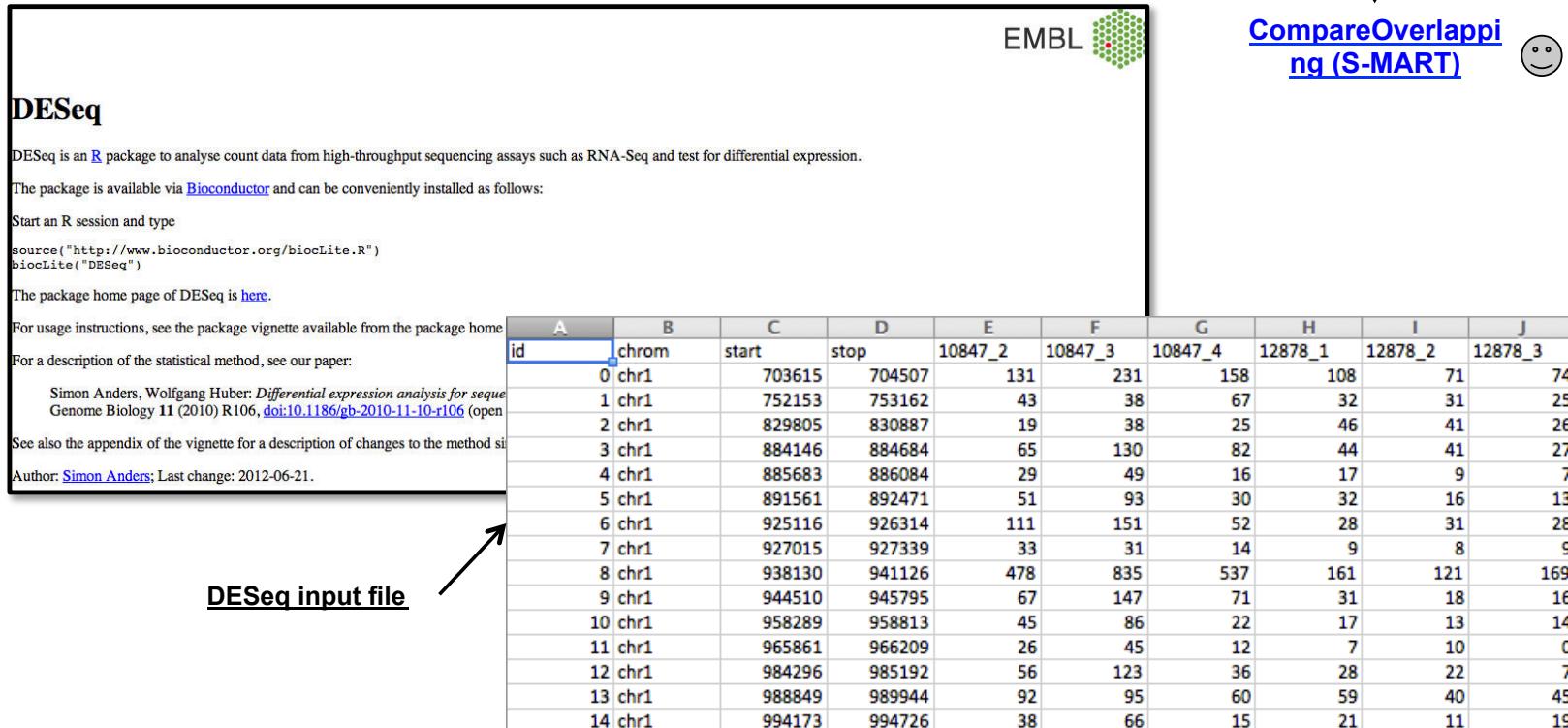
210: Trim sequences on data 208
209: FastQC.html
208: Trim sequences on data 206
207: FastQC.html
206: Trim sequences on data 179
205: FastQC.html
179: FASTQ Groomer on data 173
173: 454reads test.fastq
172: s 4 cut 100000.mfq
171: s 3 cut 100000.mfq
170: s 3 cut.mfq
169: s 4 cut.mfq
155: f1cond2.tsv
154: f1cond1.tsv
62: reference.mfa
61: annotation.gff



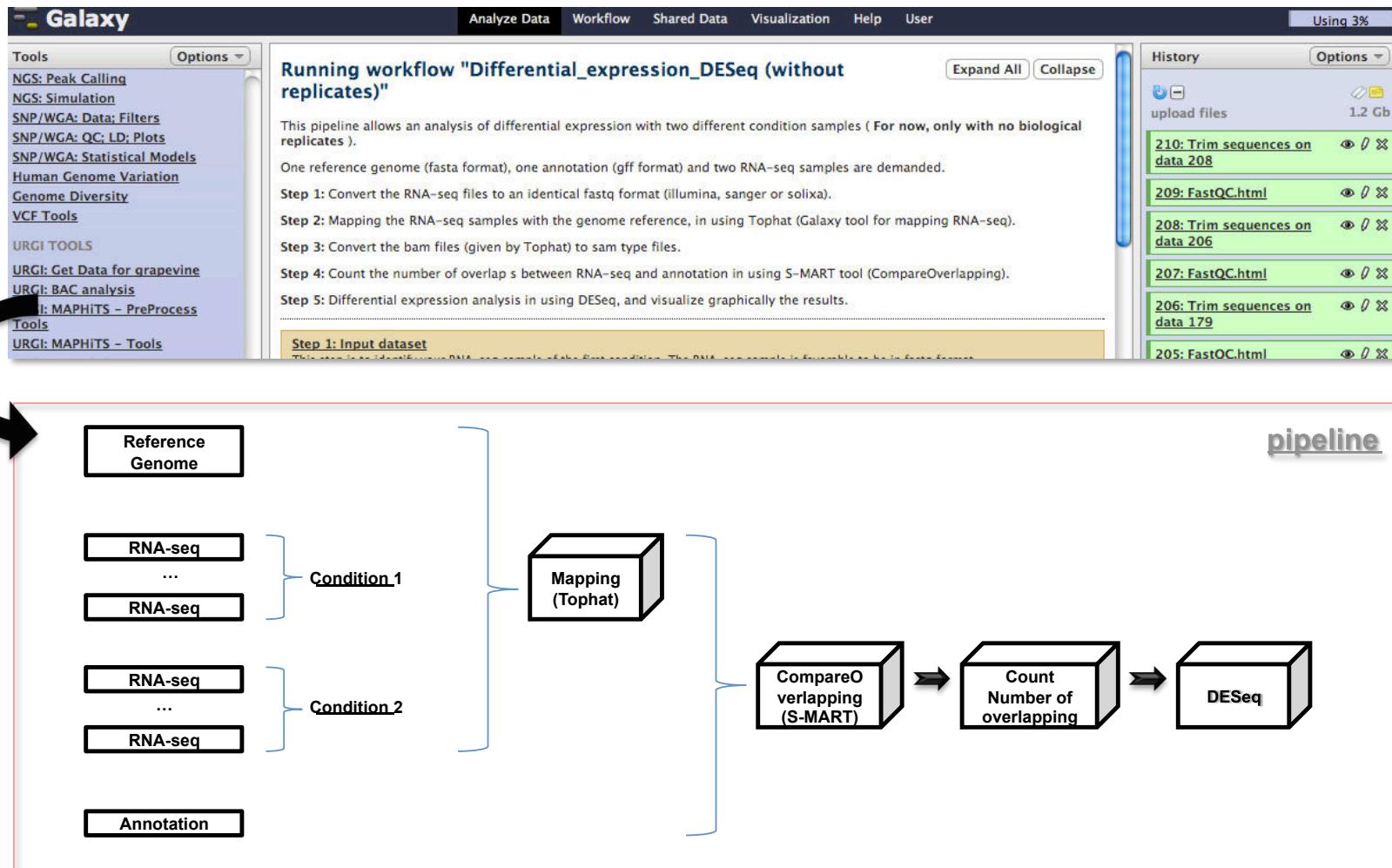
URGI INRA Versailles yufei.luo@versailles.inra.fr

S-MART in Galaxy --pipelines

Pipeline of Differential expression analysis using DESeq

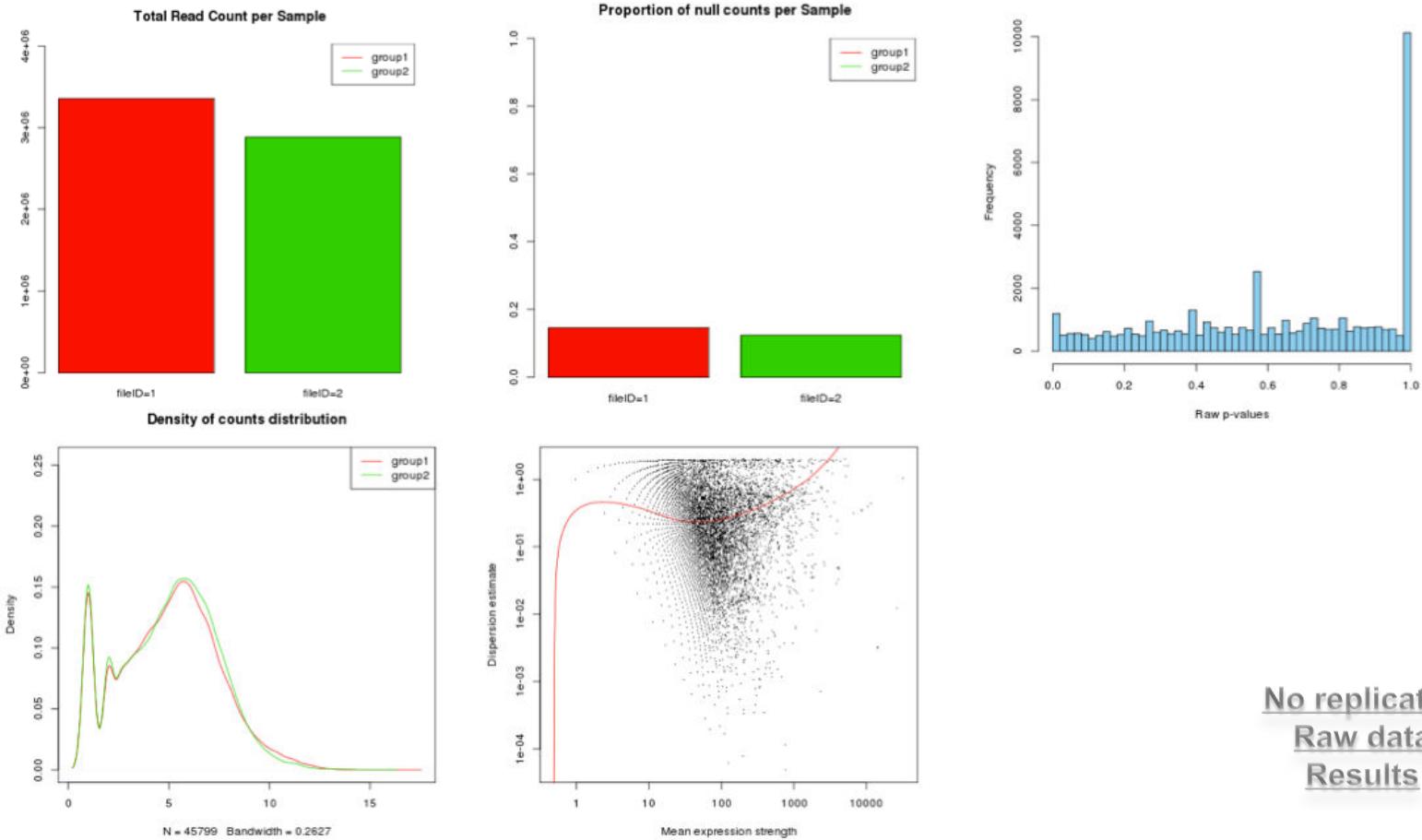


S-MART in Galaxy --pipelines



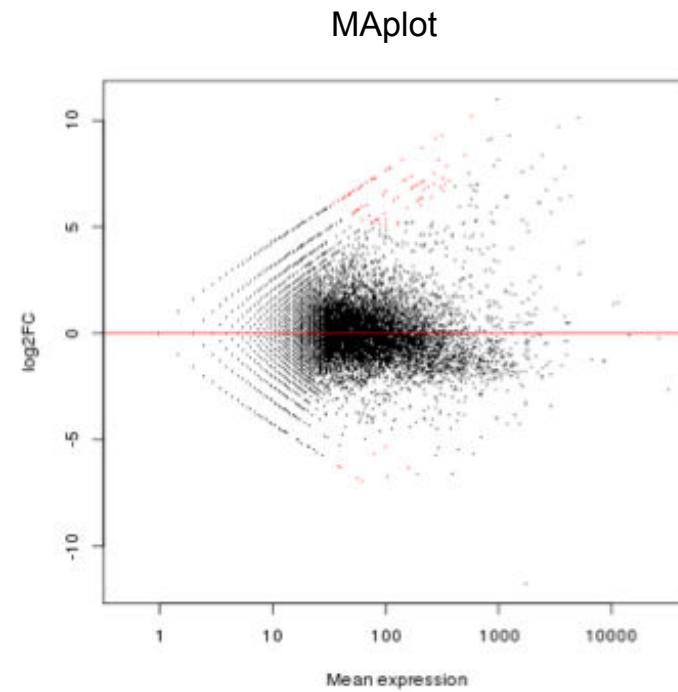
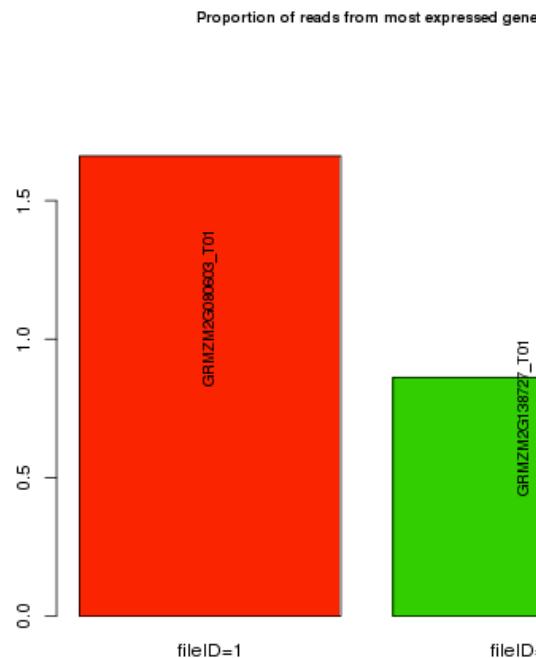
S-MART in Galaxy

--pipelines (Maize sample, 2 different DAP conditions, neither biological nor technical replicates)



S-MART in Galaxy

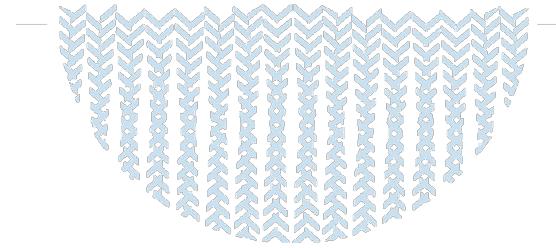
--pipelines (Maize sample, 2 different DAP conditions)



Normalized data
Results

Acknowledgment





Thank you for your attention !!!



URGI INRA Versailles
yufei.luo@versailles.inra.fr

ALIMENTATION
AGRICULTURE
ENVIRONNEMENT

INRA