

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



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*

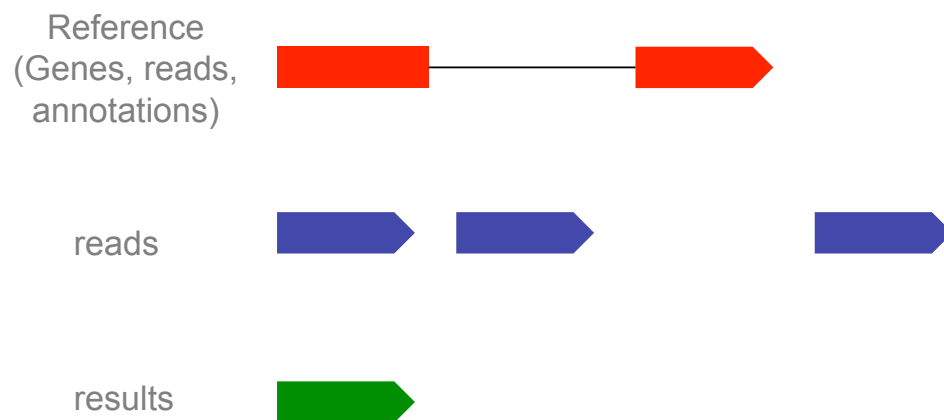




S-MART

--Tools for RNA-seq

CompareOverlapping :



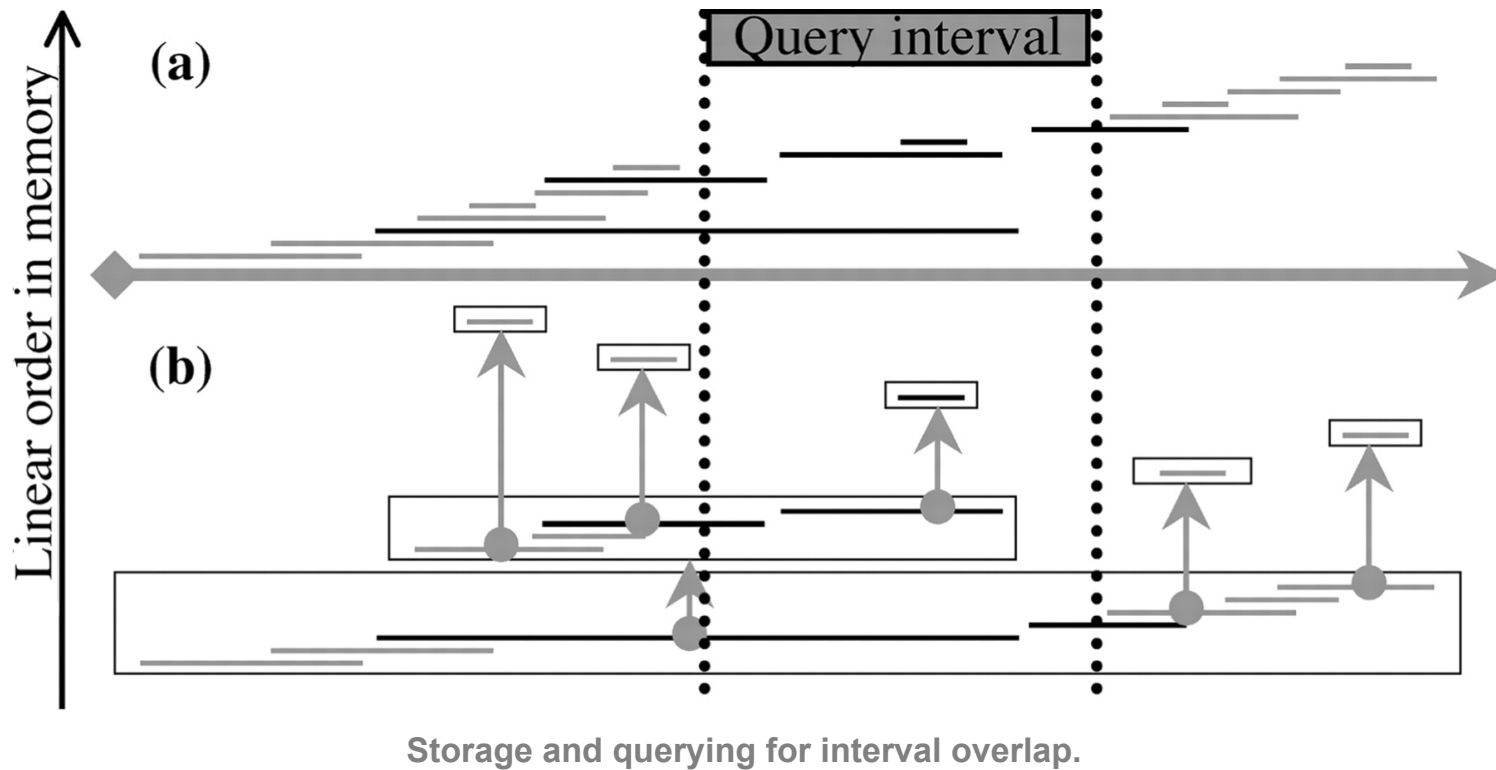
Faster, Lighter Algorithm



S-MART

--Tools for RNA-seq

Nested Containment List (NCList)^[2]



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

[2] A.Alekseyenko & J.Lee Nested; Containment List (NCList) : 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 :

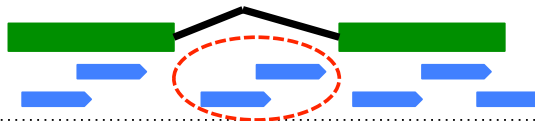
Restrict to N first nucleotides :



Extension on 5' or 3' direction :



Report introns :



Invert selection :



Colinear/Anti-sense :



Included :



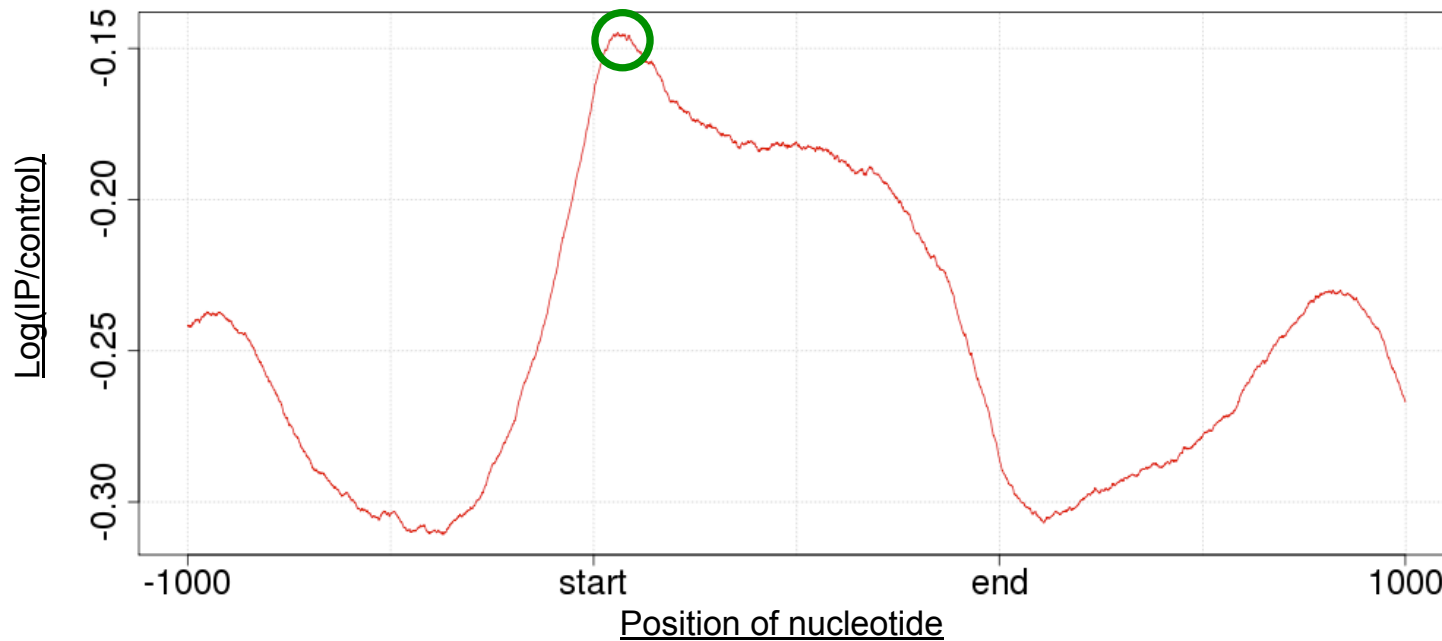


S-MART

--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

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.

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://urqi.versailles.inra.fr/galaxy>
[Contact to urqi-support@versailles.inra.fr to have a count.](mailto:urqi-support@versailles.inra.fr)





S-MART in Galaxy

--pipelines

The screenshot shows the Galaxy web interface. The top navigation bar includes 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. The 'Tools' sidebar on the left lists various categories: 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, and APLIBIO TOOLS. Under URGI TOOLS, the 'URGI: S-MART' pipeline is highlighted with a red box. The main workspace displays the URGI logo and the INRA logo. The 'History' sidebar on the right shows a list of recent jobs, including '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', and '61: annotation.gff'. Below the logos, a text box states: '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.'





S-MART in Galaxy --pipelines

Pipeline of Differential expression analysis using DESeq

[MATLAB](#)



[CompareOverlapping \(S-MART\)](#)

EMBL

DESeq

DESeq is an [R](#) package to analyse count data from high-throughput sequencing assays such as RNA-Seq and test for differential expression.

The package is available via [Bioconductor](#) and can be conveniently installed as follows:

Start an R session and type

```
source("http://www.bioconductor.org/biocLite.R")
biocLite("DESeq")
```

The package home page of DESeq is [here](#).

For usage instructions, see the package vignette available from the package home

For a description of the statistical method, see our paper:

Simon Anders, Wolfgang Huber: *Differential expression analysis for sequence data using DESeq*. *Genome Biology* 11 (2010) R106, [doi:10.1186/gb-2010-11-10-r106](https://doi.org/10.1186/gb-2010-11-10-r106) (open access)

See also the appendix of the vignette for a description of changes to the method since version 1.10.0.

Author: [Simon Anders](#); Last change: 2012-06-21.

DESeq input file

	A	B	C	D	E	F	G	H	I	J
id	chr1	start	stop	10847_2	10847_3	10847_4	12878_1	12878_2	12878_3	
0	chr1	703615	704507	131	231	158	108	71	74	
1	chr1	752153	753162	43	38	67	32	31	25	
2	chr1	829805	830887	19	38	25	46	41	26	
3	chr1	884146	884684	65	130	82	44	41	27	
4	chr1	885683	886084	29	49	16	17	9	7	
5	chr1	891561	892471	51	93	30	32	16	13	
6	chr1	925116	926314	111	151	52	28	31	28	
7	chr1	927015	927339	33	31	14	9	8	9	
8	chr1	938130	941126	478	835	537	161	121	169	
9	chr1	944510	945795	67	147	71	31	18	16	
10	chr1	958289	958813	45	86	22	17	13	14	
11	chr1	965861	966209	26	45	12	7	10	0	
12	chr1	984296	985192	56	123	36	28	22	7	
13	chr1	988849	989944	92	95	60	59	40	45	
14	chr1	994173	994726	38	66	15	21	11	19	





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
- URGI: MAPHITS - Tools

Running workflow "Differential_expression_DESeq (without replicates)"

Expand All Collapse

This pipeline allows an analysis of differential expression with two different condition samples (For now, only with no biological replicates).

One reference genome (fasta format), one annotation (gff format) and two RNA-seq samples are demanded.

Step 1: Convert the RNA-seq files to an identical fastq format (illumina, sanger or solixa).

Step 2: Mapping the RNA-seq samples with the genome reference, in using Tophat (Galaxy tool for mapping RNA-seq).

Step 3: Convert the bam files (given by Tophat) to sam type files.

Step 4: Count the number of overlap s between RNA-seq and annotation in using S-MART tool (CompareOverlapping).

Step 5: Differential expression analysis in using DESeq, and visualize graphically the results.

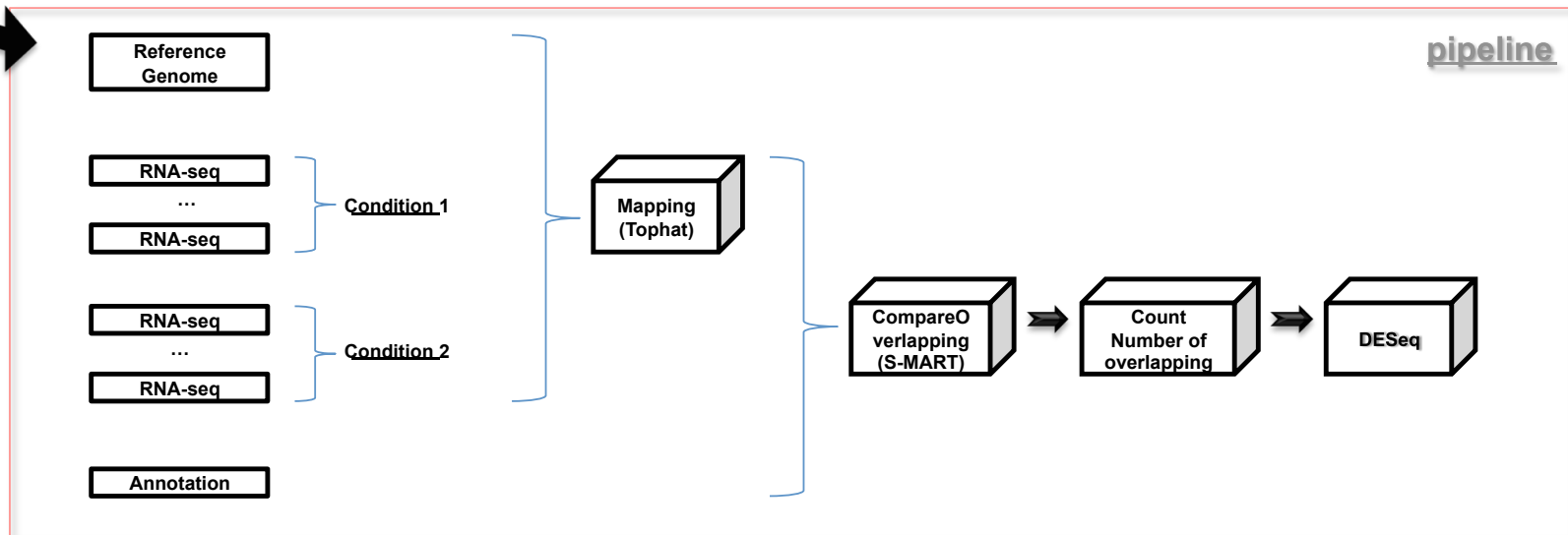
Step 1: Input dataset

This step is to identify unique RNA-seq samples of the first condition. The RNA-seq sample is formatted to be in fastq format.

History Options ▾

upload files 1.2 Gb

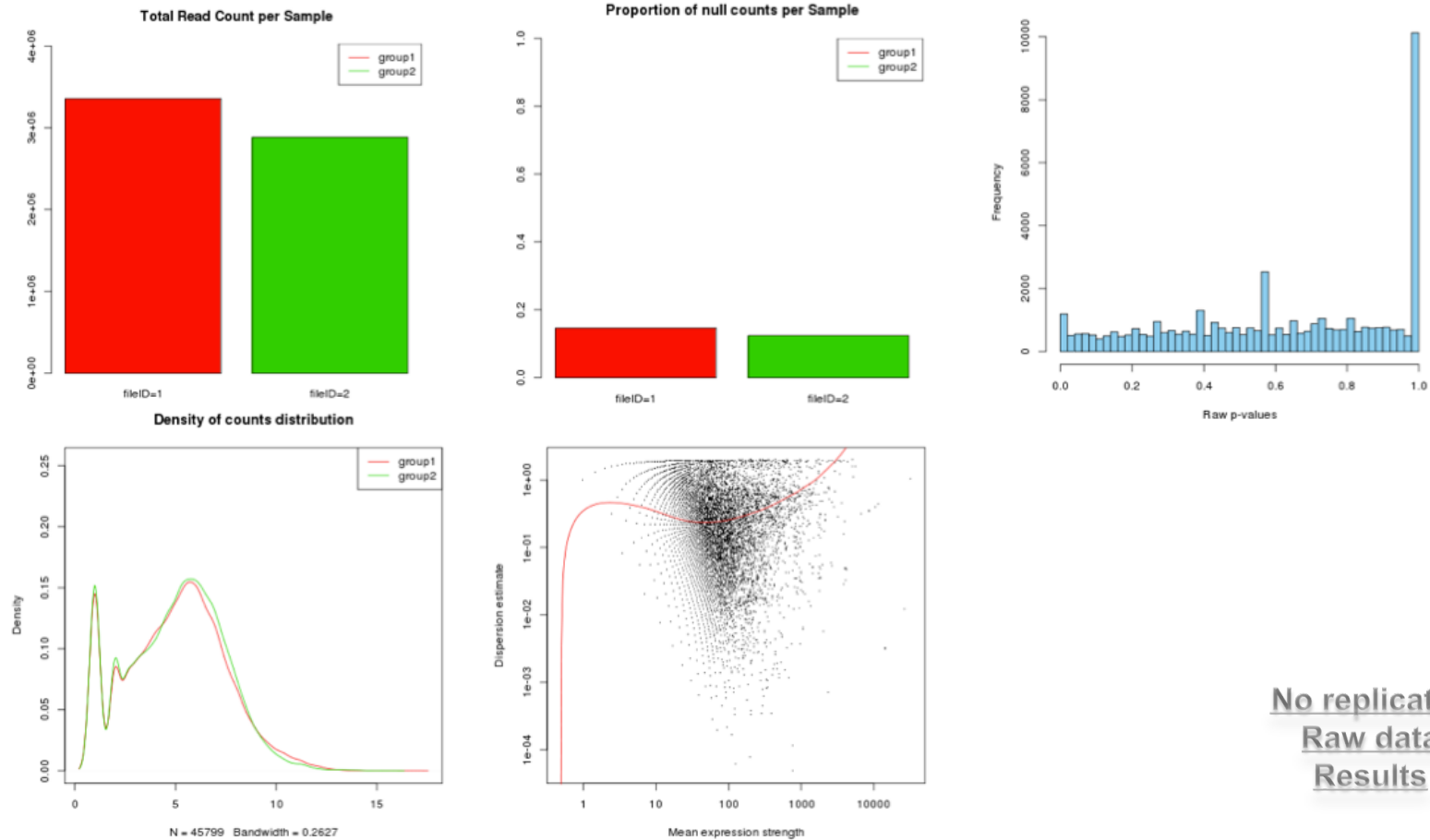
- 210: Trim sequences on data 208
- 209: FastQC.html
- 208: Trim sequences on data 206
- 207: FastQC.html
- 206: Trim sequences on data 179
- 205: FastQC.html





S-MART in Galaxy

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



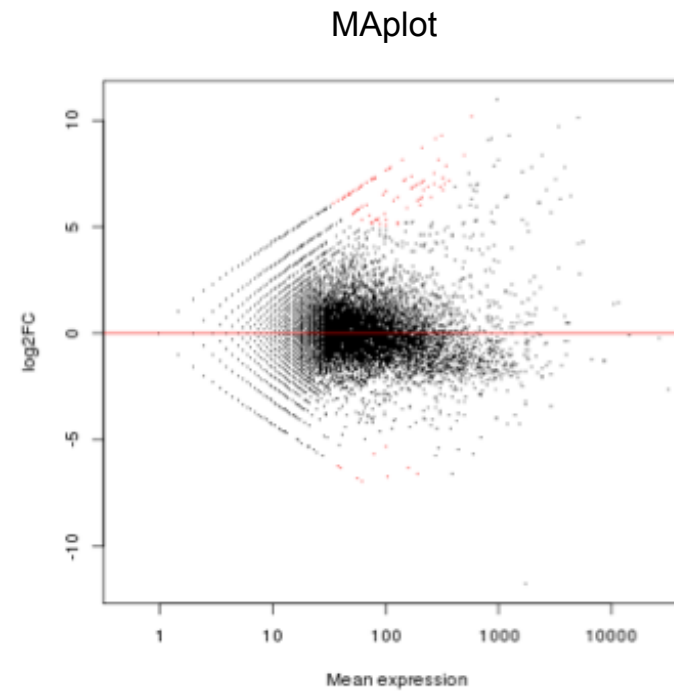
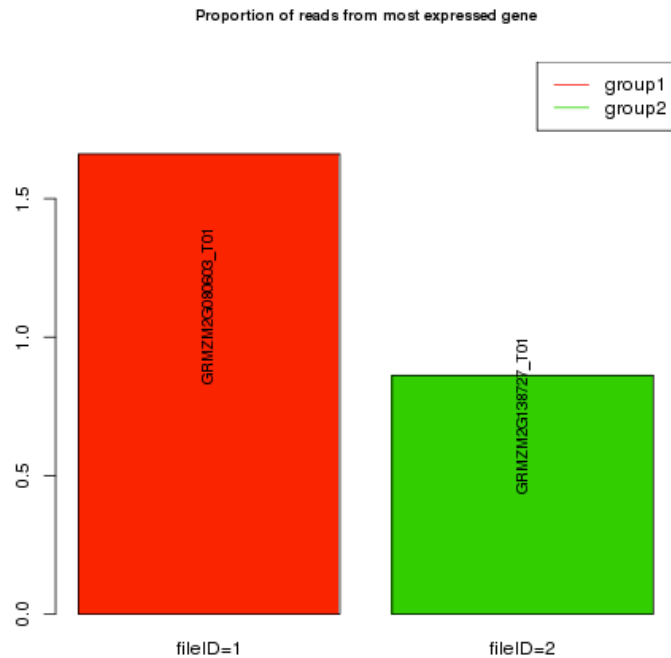
No replicates
Raw data
Results





S-MART in Galaxy

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

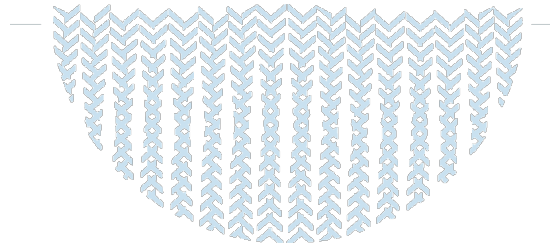




Acknowledgment



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



Thank you for your attention !!!



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

ALIMENTATION
AGRICULTURE
ENVIRONNEMENT

