Integration of S-MART, a toolbox to aid RNA-seq data analysis in Galaxy
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\textsuperscript{[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

S-MART
--Tools for RNA-seq

CompareOverlapping:

Reference
(Genes, reads, annotations)

reads

results

Faster, Lighter Algorithm
Nested Containment List (NCLlist)$^2$

(a)

(b)

Storage and querying for interval overlap.

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
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

![Graph](image-url)
S-MART in Galaxy
--pipelines

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.
S-MART in Galaxy
--pipelines

The Galaxy team is a part of BX at Penn State.
This project is supported in part by NSF, INRA, and the Huck Institutes of the Life Sciences.
Pipeline of Differential expression analysis using DESeq

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 page.

DESeq input file

MATLAB

CompareOverlapping (S-MART)
S-MART in Galaxy
--pipelines

Reference Genome

RNA-seq...
RNA-seq

Condition 1
Mapping (Tophat)

RNA-seq...
RNA-seq

Condition 2

CompareO overlapping (S-MART)
Count Number of overlapping

DESeq
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)

Proportion of reads from most expressed gene

MAplot

Normalized data
Results
Acknowledgment
Thank you for your attention !!!