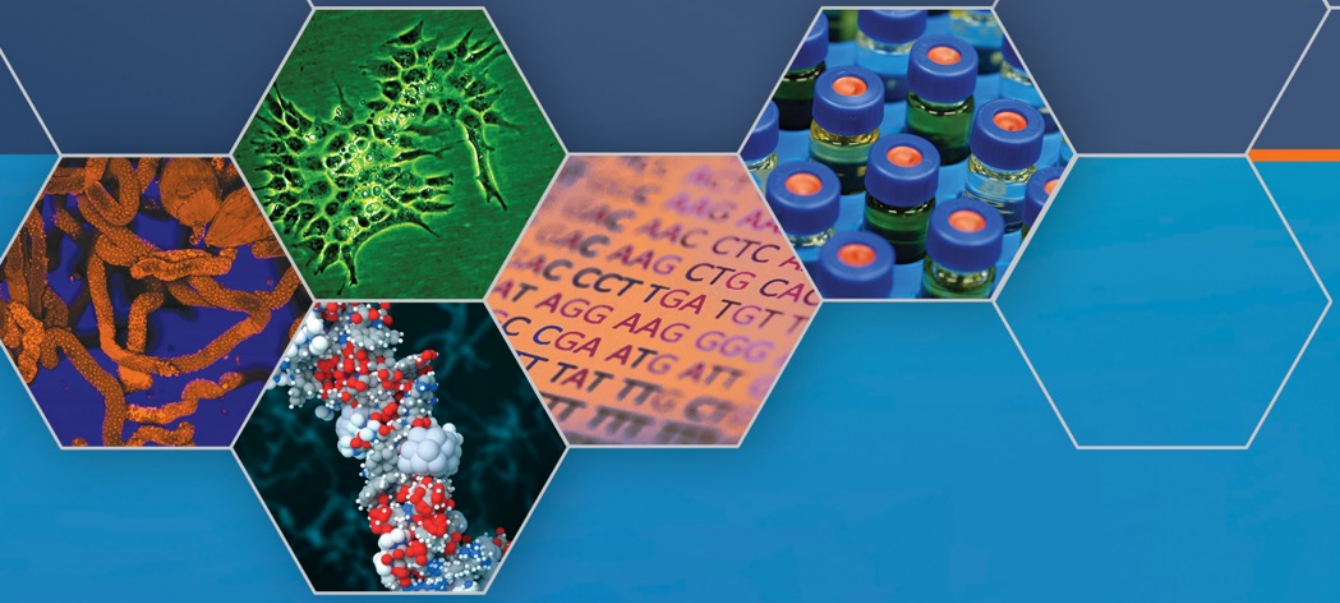
ICBR

Interdisciplinary Center
for Biotechnology Research



UF | UNIVERSITY of FLORIDA

ICBR

Interdisciplinary Center for Biotechnology Research

ISCB Workshop WK03: Bioinfo-core Workshop

Developing high-performance analysis pipelines in a core setting

Alberto Riva
ariva@ufl.edu

UF UNIVERSITY of FLORIDA

Outline

- The UF ICBR Bioinformatics Core
- What's "big" in big data?
- The Actor framework
- Conclusions / discussion points

ICBR

- ICBR: Interdisciplinary Center For Biotechnology Research. Founded in 1987 to create a common administrative structure for existing University of Florida (UF) core facilities.
- Enables molecular life sciences research by reducing barriers to implementation and practice of molecular technologies.
- Serves a very large and diverse scientific environment: colleges of Medicine, Sciences, Pharmacy, Dentistry, Veterinary Sciences, Genetics Institute, Cancer Center, Emerging Pathogens Institute, CTSI, Florida Museum of Natural History.

UF scientific environment

- Not just medicine:

UF scientific environment

- Not just medicine: from viruses

UF scientific environment

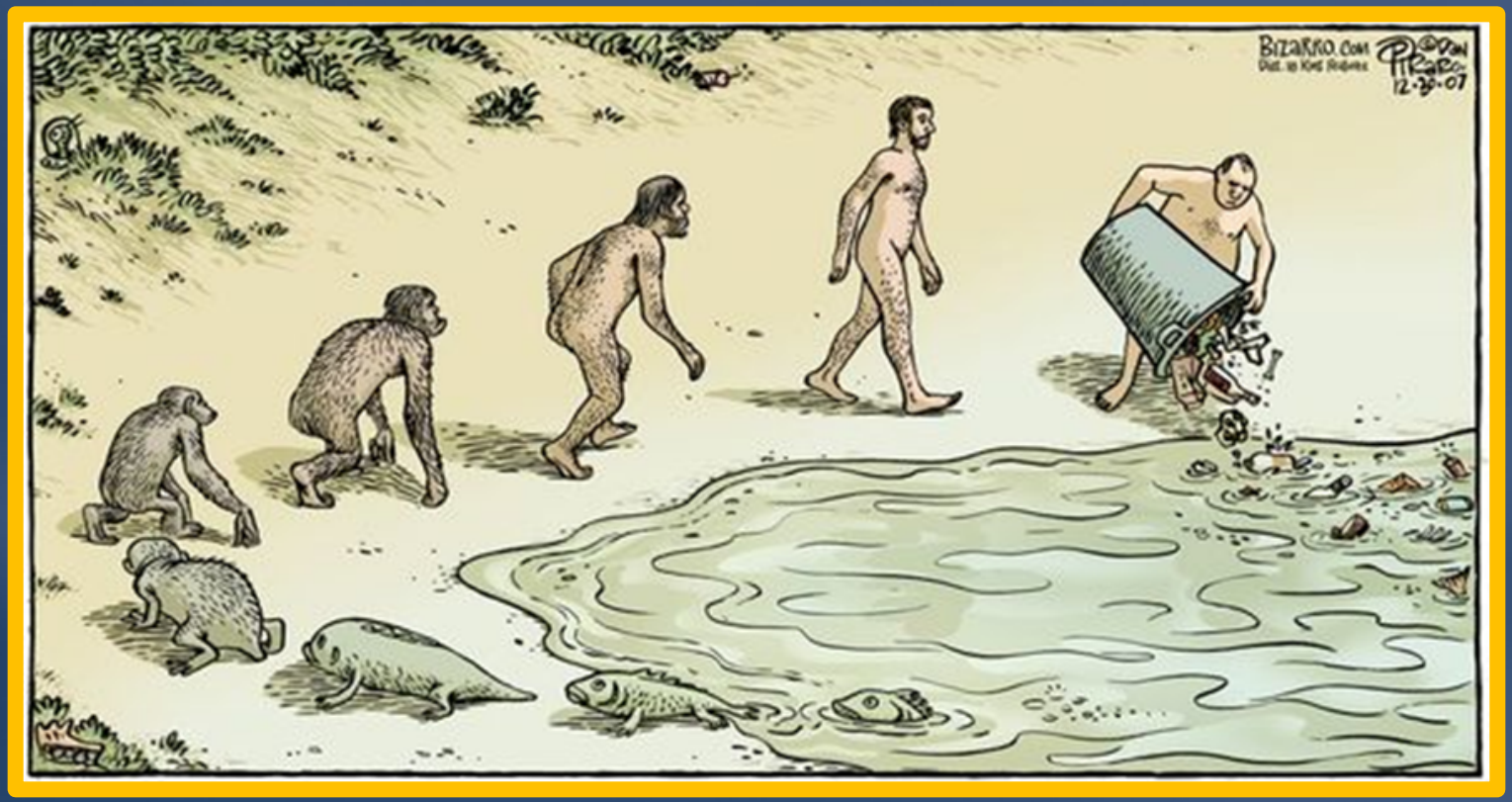
- Not just medicine: from viruses to forests,

UF scientific environment

- Not just medicine: from viruses to forests, from molecular evolution

UF scientific environment

- Not just medicine: from viruses to forests, from molecular evolution to human evolution,



UF scientific environment

- Not just medicine: from viruses to forests, from molecular evolution to human evolution, from bacteria

UF scientific environment

- Not just medicine: from viruses to forests, from molecular evolution to human evolution, from bacteria to horses,

UF scientific environment

- Not just medicine: from viruses to forests, from molecular evolution to human evolution, from bacteria to horses, cows,

UF scientific environment

- Not just medicine: from viruses to forests, from molecular evolution to human evolution, from bacteria to horses, cows, manatees,

UF scientific environment

- Not just medicine: from viruses to forests, from molecular evolution to human evolution, from bacteria to horses, cows, manatees, alligators…

UF scientific environment

- Not just medicine: from viruses to forests, from molecular evolution to human evolution, from bacteria to horses, cows, manatees, alligators… often under the same roof!

1988 ICBR Core Portfolio

- Cytometry
- Electron Microscopy
- Monoclonal Antibody
- Protein Analysis
- Sanger Sequencing

2016 ICBR Core Portfolio

- Cytometry
- Electron Microscopy
- Monoclonal Antibody
- Protein Analysis
- Sanger Sequencing
- Gene Expression and Genotyping
- Next-Gen DNA Sequencing
- Bioinformatics

Bioinformatics Core

- Four full-time bioinformatics specialists, with complementary backgrounds and several decades of combined expertise in the field.
- Computational resources: in-house equipment; large allocation on UF's cluster computer (HiPerGator, 50,000+ cores).
- Mission: support genomic research at UF by providing data analysis services for large scale DNA sequencing, genotyping, methylation analysis, gene expression (microarray, RNAseq), genome assembly and annotation, etc…

What is “big” in big data?

- The absolute size of datasets is not the main issue. Disk space is cheap and getting cheaper.
- Computational complexity: get a more powerful computer (UF just did that)! Or rent it.
- Other dimensions of complexity:
  - Wide range of research areas, scientific questions;
  - No two projects are ever identical;
  - Field in constant evolution (technology, tools, methods, questions, standards, requirements, …)

The three *R*s…

1. *Reliability*
   Clients want *correct* results!

2. *Reproducibility*
   We should be able to re-run an analysis six months later and get the same results. Or run the same analysis on similar input datasets and get consistent results.

3. *Reusability*
   Pipelines share basic components (e.g. alignment). We don't have time and resources to re-write pipelines from scratch every time, and it does not make sense anyway.

Actor

- **Actor** is a meta-scripting tool for reproducible computing in a high-performance computing setting.

- Actor scripts:
  - Automatically generate an HTML report containing a description of all analysis steps;
  - Allow for easy inclusion of tables, images, plots, downloadable files;
  - Automatically package input and output files, analysis scripts, and the HTML report in a ZIP file (and publish it the web).

Actor

- Actor is implemented as a Python library. Actor calls can be freely mixed with standard Python code.
- The library is divided into sections dealing with: initialization and setup, execution of programs and scripts, report generation.
- The library provides data structures to represent conditions, samples, technical and biological replicates.

Actor

- Actor has been used to implement the following pipelines so far:

Description	Tools
RNA-Seq processing	Trimmomatic / sickle, STAR, Picard, cufflinks / cuffdiff / rsem, FastQC, counts, coverage, tracks.
ChIP-Seq processing	Trimmomatic / sickle, STAR / Bowtie, Picard, Homer, FastQC, counts, coverage, tracks.
Differential methylation analysis	Trimmomatic / sickle, bsmap, Picard, cscall, mcomp, FastQC, counts, coverage, plots.
Regulatory network reconstruction	ARACNE, apple.py.
Multi-sample SNP calling	Trimmomatic / sickle, Bowtie, GATK / freebayes, SnpEff, FastQC
De-novo	Trimmomatic / sickle, spades, prokka, roary, mauve, FastQC.
Microarray analysis	R/Bioconductor (limma).

Actor

- Script execution is controlled by a simple configuration file.

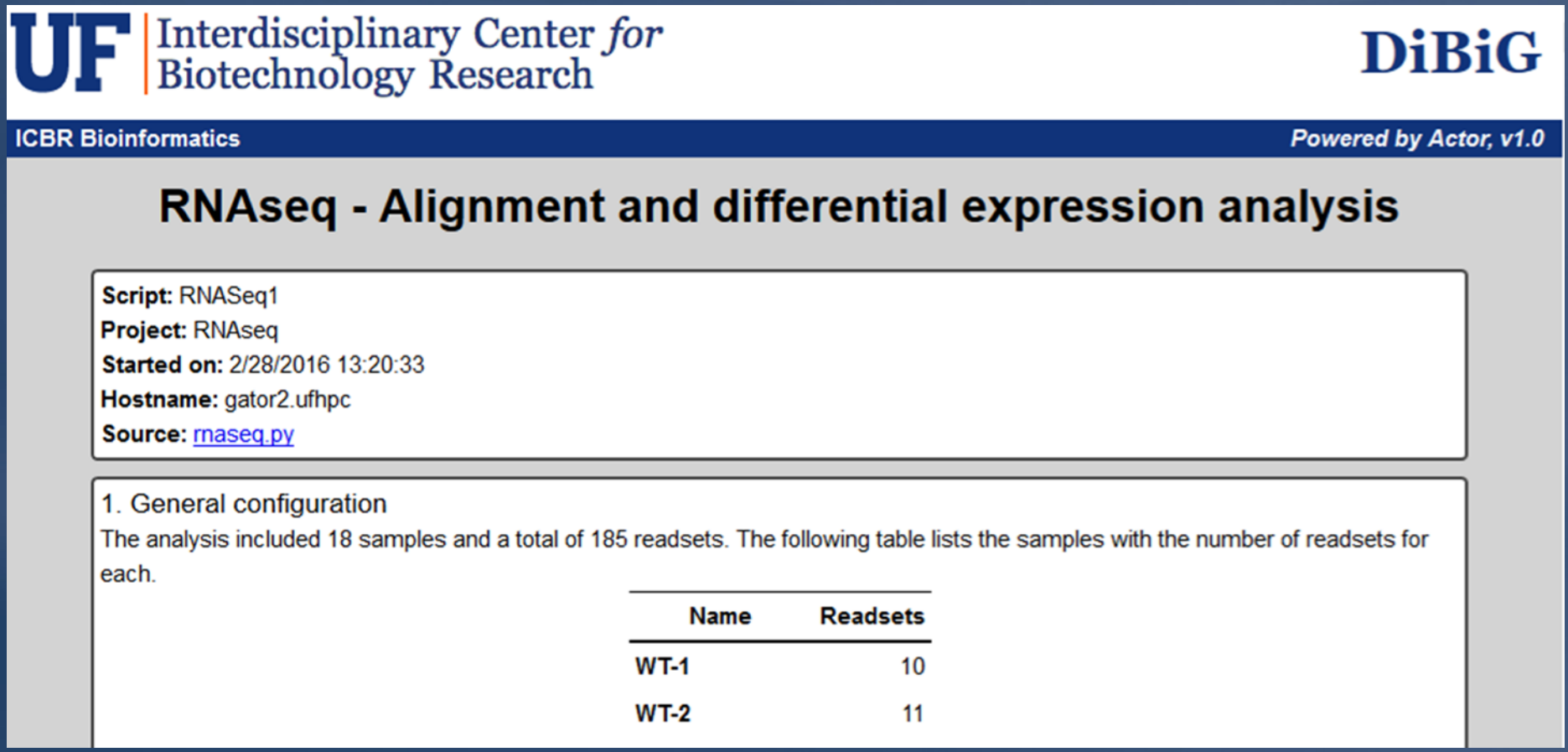
```
[General]
title = run1
staridx  = <<path to STAR index>>
cufflinksGTF = <<path to GTF annotations file>>

# We compare two experimental conditions, wildtype (WT) and knock-out (KO)
conditions = WT, KO
contrasts = KO^WT

# Each condition may have any number of samples (biological replicates)
[WT]
samples = WT-1, WT-2, WT-3, WT-4, WT-5
[KO]
Samples = KO-1, KO-2, KO-3, KO-4
```

Actor

- Automatically generated HTML report.



UF | Interdisciplinary Center *for* Biotechnology Research

DiBiG

ICBR Bioinformatics — *Powered by Actor, v1.0*

RNAseq - Alignment and differential expression analysis

Script: RNASeq1
Project: RNAseq
Started on: 2/28/2016 13:20:33
Hostname: gator2.ufhpc
Source: rnaseq.py

1. General configuration

The analysis included 18 samples and a total of 185 readsets. The following table lists the samples with the number of readsets for each.

Name	Readsets
WT-1	10
WT-2	11

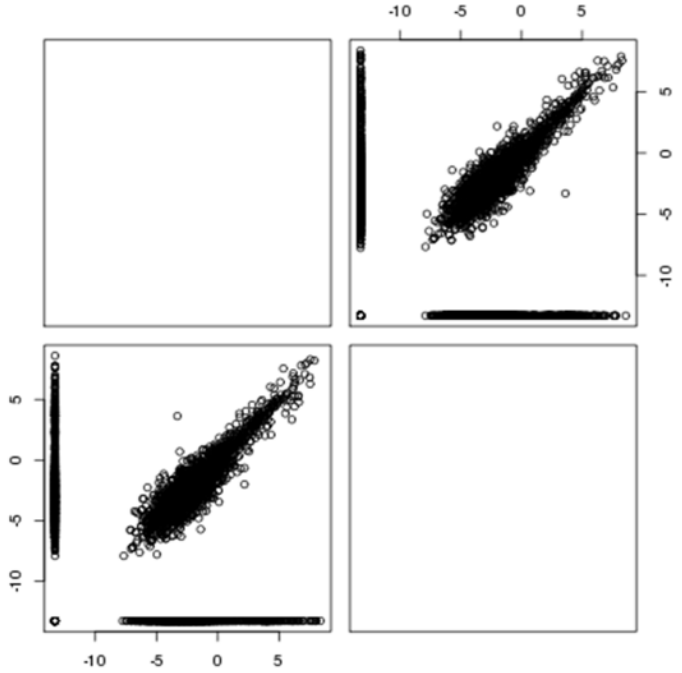
Actor

- Automatically generated HTML report.

4. Expression quantification

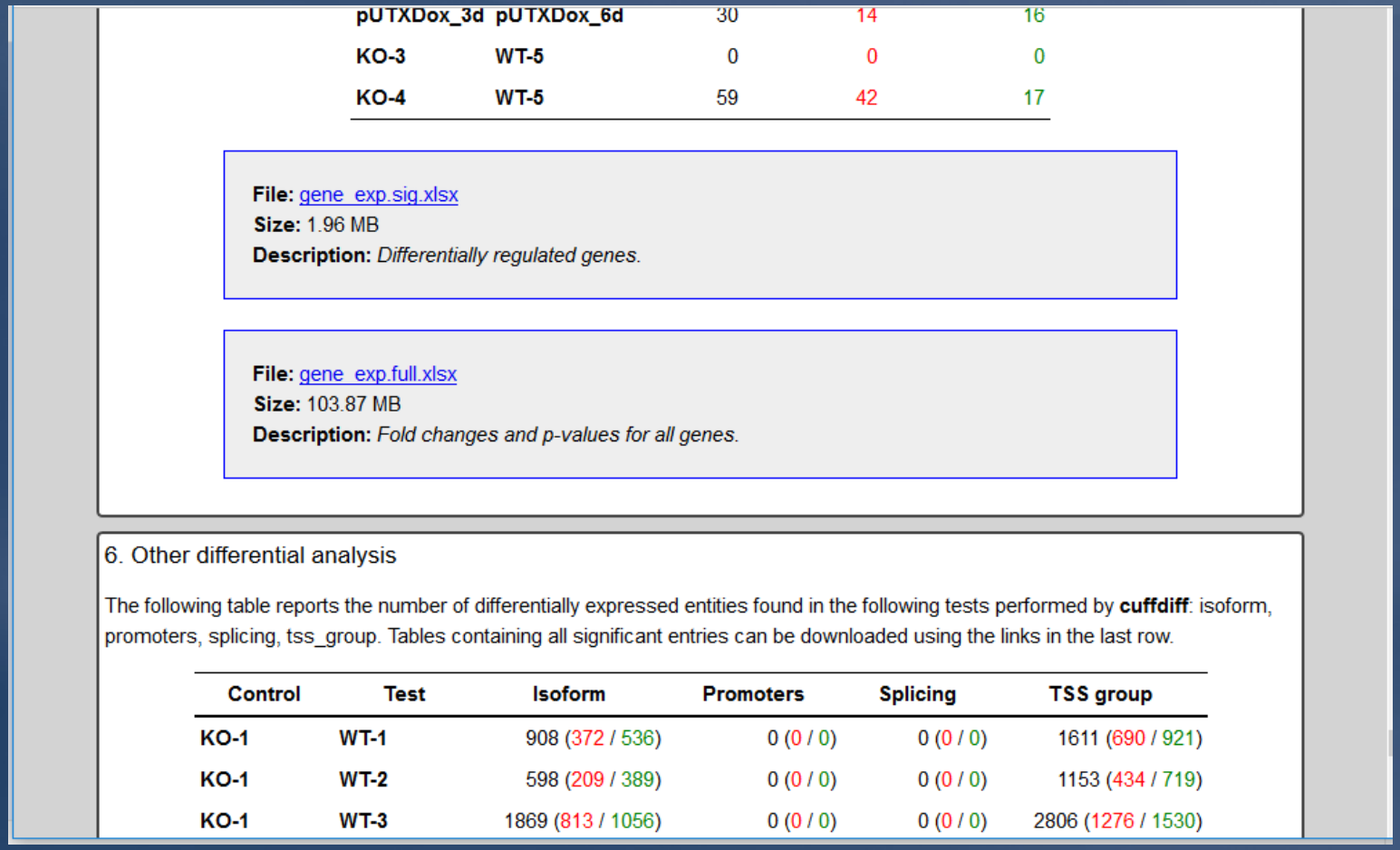
Aligned reads were processed with **cufflinks** to estimate gene expression levels. The file all-fpkms.xlsx contains the FPKM values for all genes in all conditions. The following plots display the pairwise correlation between replicates in each sample.

WT-1
WT-2
WT-3
WT-4
WT-5
KO-1
KO-2
KO-3
KO-4



Actor

- Automatically generated HTML report.



pUTXDox_3d	pUTXDox_6d	30	14	16
KO-3	WT-5	0	0	0
KO-4	WT-5	59	42	17

File: gene_exp.sig.xlsx
Size: 1.96 MB
Description: *Differentially regulated genes.*

File: gene_exp.full.xlsx
Size: 103.87 MB
Description: *Fold changes and p-values for all genes.*

6. Other differential analysis

The following table reports the number of differentially expressed entities found in the following tests performed by **cuffdiff**: isoform, promoters, splicing, tss_group. Tables containing all significant entries can be downloaded using the links in the last row.

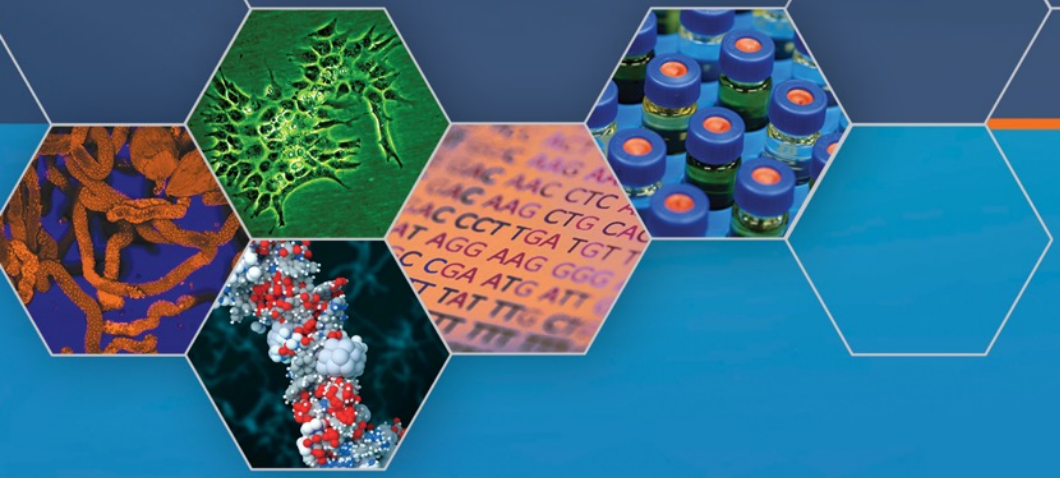
Control	Test	Isoform	Promoters	Splicing	TSS group
KO-1	WT-1	908 (372 / 536)	0 (0 / 0)	0 (0 / 0)	1611 (690 / 921)
KO-1	WT-2	598 (209 / 389)	0 (0 / 0)	0 (0 / 0)	1153 (434 / 719)
KO-1	WT-3	1869 (813 / 1056)	0 (0 / 0)	0 (0 / 0)	2806 (1276 / 1530)

Conclusions / discussion points

- Technology and analytical skills are not enough.
- *Process* and *infrastructure* are key to providing "big data" services in a reliable and efficient way.
- Process: a bioinformatics core facility should not be a "black box". Work is inherently exploratory and collaborative; we need to be nimble and adaptable.
- Infrastructure: investment in making work more efficient, reliable, reproducible. Short-term cost, long-term benefit.

ICBR

Interdisciplinary Center
for Biotechnology Research



Alberto Riva
ariva@ufl.edu
DiBiG, Bioinformatics Core
ICBR, University of Florida

www.biotech.ufl.edu

UF UNIVERSITY of FLORIDA