Developing high-performance analysis pipelines in a core setting
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.

- Strong links with Sequencing, Gene Expression, and Proteomics cores. Part-time faculty director to increase visibility with research faculty.

- 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, ...)

ICBR | Interdisciplinary Center for Biotechnology Research
The three *Rs*…

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.

- Actor scripts are similar to shell scripts, but:
  - They automatically generate an HTML report containing a description of all analysis steps;
  - They allow for easy inclusion of tables, images, plots, downloadable files;
  - Input and output files, analysis scripts, and the HTML report can be automatically packaged in a ZIP file and published on 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:

<table>
<thead>
<tr>
<th>Description</th>
<th>Tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNA-Seq processing</td>
<td>Trimmomatic / sickle, STAR, Picard, cufflinks / cuffdiff / rsem, FastQC, counts, coverage, tracks.</td>
</tr>
<tr>
<td>ChIP-Seq processing</td>
<td>Trimmomatic / sickle, STAR / Bowtie, Picard, Homer, FastQC, counts, coverage, tracks.</td>
</tr>
<tr>
<td>Differential methylation analysis</td>
<td>Trimmomatic / sickle, bsmmap, Picard, cscall, mcomp, FastQC, counts, coverage, plots.</td>
</tr>
<tr>
<td>Regulatory network reconstruction</td>
<td>ARACNE, apple.py.</td>
</tr>
<tr>
<td>Multi-sample SNP calling</td>
<td>Trimmomatic / sickle, Bowtie, GATK / freebayes, SnpEff, FastQC</td>
</tr>
<tr>
<td>De-novo</td>
<td>Trimmomatic / sickle, spades, prokka, roary, mauve, FastQC.</td>
</tr>
<tr>
<td>Microarray analysis</td>
<td>R/Bioconductor (limma).</td>
</tr>
</tbody>
</table>
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.
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.
Automatically generated HTML report.

<table>
<thead>
<tr>
<th>pUTXDox_3d</th>
<th>pUTXDox_6d</th>
<th>KO-3</th>
<th>WT-5</th>
<th>KO-4</th>
<th>WT-5</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>14</td>
<td>0</td>
<td>0</td>
<td>59</td>
<td>42</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Control</th>
<th>Test</th>
<th>Isoform</th>
<th>Promoters</th>
<th>Splicing</th>
<th>TSS group</th>
</tr>
</thead>
<tbody>
<tr>
<td>KO-1</td>
<td>WT-1</td>
<td>908 (372 / 536)</td>
<td>0 (0 / 0)</td>
<td>0 (0 / 0)</td>
<td>1611 (690 / 921)</td>
</tr>
<tr>
<td>KO-1</td>
<td>WT-2</td>
<td>598 (209 / 389)</td>
<td>0 (0 / 0)</td>
<td>0 (0 / 0)</td>
<td>1153 (434 / 719)</td>
</tr>
<tr>
<td>KO-1</td>
<td>WT-3</td>
<td>1869 (813 / 1056)</td>
<td>0 (0 / 0)</td>
<td>0 (0 / 0)</td>
<td>2806 (1275 / 1530)</td>
</tr>
</tbody>
</table>
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.