BioLite
A lightweight bioinformatics framework with automated tracking of diagnostics and provenance

Mark Howison¹², Nicholas A. Sinnott-Armstrong², Casey W. Dunn²

¹ Center for Computation and Visualization
² Department of Ecology and Evolutionary Biology

Brown University

Presented at USENIX TaPP ’12 | http://www.dunnlab.org/biolite
The Problem

Next-Gen Sequencing technologies produce big data:
• ~250GB per run for an Illumina HiSeq 2000

The data require complex analyses:
• Quality control and filtering of the raw ‘reads’
• Assembly of short reads into contiguous sequences
• Alignment and comparison to known sequences

Need a better solution than ad hoc analyses and one-time scripts!
Other ‘Workflow’ Solutions

- **Need lightweight + provenance!**

<table>
<thead>
<tr>
<th>Lightweight</th>
<th>Heavyweight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ruffus</td>
<td>Galaxy</td>
</tr>
<tr>
<td>PaPY</td>
<td>Taverna</td>
</tr>
</tbody>
</table>

With Provenance

Without
Lightweight Design Goals

• Command-line usage
• Easily extendable through scripting and programming
• Minimal administrative overhead and dependencies
• Portability and performance
BioLite

A Python framework and set of C++ tools for:
• building out customized analysis pipelines
• fault-tolerance, through built-in checkpointing
• automating the collection/reporting of diagnostics
• tracking the provenance of analyses:
  – resource usage
  – paths and parameters
  – program versioning
  – statistics
The `diagnostics` table has a complete non-executable history of the analysis:

\[
\text{diagnostics} + \text{pipeline script} = \text{reproducibility}
\]
Reports

• API for accessing raw diagnostics, generating custom reports
• Reporting code is integrated with analysis scripts
• Tabular reports show comparisons across data sets
Resource Profiling

For understanding how resources are used by different stages of a pipeline run.
Resource Profiling

For comparing computational requirements across analyses, forecast future requirements

![Graph showing input size vs. memory used in pre-assembly](chart.png)
Diagnostics

For answering questions about up-stream data collection, comparisons across analyses, etc.

Example application: how does purification method affect usable RNA content?
Applications of BioLite

- Development is driven primarily by Agalma, a *de novo* transcriptome pipeline
- Chlorox, a chloroplast genome assembly tool
- Other tools in progress at Brown…
Questions?