Reproducible big data science: A case study in continuous FAIRness

Ravi K. Madduri
Data Science and Learning
Argonne National Laboratory and University of Chicago
Work made possible by...

Kyle Chard        Mike D’Arca        Segun Jung        Alex Rodriguez        Dina Sulakhe        Cory Funk

Matt Richards    Paul Shannon    Ben Heavner        Gustavo Glusman    Nathan Price        Carl Kesselman    Ian Foster
Last year researchers at one biotech firm, Amgen, found they could reproduce just six of 53 “landmark” studies in cancer research. Earlier, a group at Bayer, a drug company, managed to repeat just a quarter of 67 similarly important papers. A leading computer scientist frets that three-quarters of papers in his subfield are bunk. In 2000-10 roughly 80,000 patients took part in clinical trials based on research that was later retracted because of mistakes or improprieties.
“REPRODUCIBILITY IS LIKE BRUSHING YOUR TEETH. ONCE YOU LEARN IT, IT BECOMES A HABIT.”
Ten Simple Rules for Reproducible Computational Research

• Rule 1: For Every Result, Keep Track of How It Was Produced
• Rule 2: Avoid Manual Data Manipulation Steps
• Rule 3: Archive the Exact Versions of All External Programs Used
• Rule 4: Version Control All Custom Scripts
• Rule 5: Record All Intermediate Results, When Possible in Standardized Formats
• Rule 6: For Analyses That Include Randomness, Note Underlying Random Seeds
• Rule 7: Always Store Raw Data behind Plots
• Rule 8: Generate Hierarchical Analysis Output, Allowing Layers of Increasing Detail to Be Inspected
• Rule 9: Connect Textual Statements to Underlying Results
• Rule 10: Provide Public Access to Scripts, Runs, and Results

<table>
<thead>
<tr>
<th>Software: Services &amp; Tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>App store/User Interface</td>
</tr>
<tr>
<td>scientific analysis tools/workflows</td>
</tr>
<tr>
<td>Services: APIs, Containers, Indexing,</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Reference” Data Sets</td>
</tr>
<tr>
<td>User defined data</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Compute Platform:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cloud or HPC</td>
</tr>
</tbody>
</table>

NIH Commons
Our Approach: Big picture themes

- Continuous FAIRness: Make all data findable, accessible, interoperable, reusable at every stage, via pervasive use of simple identifier and exchange format conventions

- Build on proven security, data, and computation building blocks that have large user communities inside and outside biomedicine

- Solutions leverage industry best practices and professional services team to meet scalability, interoperability, sustainability, and reliability needs

There is no point in wasting good thoughts on bad data
- Sir John Sulston and the Human Genome Project
Globus Auth: A foundational service for an authentication and authorization ecosystem

• Globus Auth is a flexible security infrastructure that can be used across the commons
  • Enables federation across services using arbitrary linked identities (e.g., @gmail @xsede @uchicago)
• Facilitates secure/authorized communication between users, services, and clients
  • Supports arbitrary clients including REST, web, command line, software
• Flexible token management
  • Secure sharing between services
  • Fine-grain user consents and revocation

https://docs.globus.org/api/auth/
Standards-based, reliable, performant data management

- **Globus Connect Server**: S3-compatible HTTP/OAuth interface for secure client-server transfer
  - Endpoints have DNS names

- **Globus Transfer**: Managed, high-performance, secure, reliable bulk asynchronous transfer

- **In-place data sharing** with flexible and secure ACLs

- Standards compliant
  - S3, OAuth, OIDC, HTTP, GridFTP

https://docs.globus.org/api/transfer/
Interoperability: naming and exchange

**Minid**
- Lightweight identifiers for any product at any stage
- Easily created, dereferenced, validated
- Global integrity – validate content across the commons

**BDBag**
- Self-describing and flexible format for exchange
- Extended BagIt Specification
- Standard manifest representation that supports different protocols

[Diagram showing Minid and BDBag schemas with example identifiers and protocols]
Workspaces: Scalable compute for distributed data

- Workspaces bring together data and tools
- Infrastructure designed for scalability and portability
- Leverages
  - Federated identities and access control
  - Secure access to distributed data
  - Data interoperability and exchange
- Provenance
  - Tracking the activity around data
  - By whom? With what?
- Publication & sharing of Tools and Workflows
- Cost aware resource allocation for both compute and data movement
Search, navigation, and virtual cohorts

• DERIVA: Digital asset management for heterogeneous data
  • Organize, navigate and discover interrelated objects (e.g., assays from a sample over time)
  • REST interface

• Entity/Relation model for organizing data
  • Supports various DCPPC metadata models
  • Fine grain access control to support diverse collaboration models
  • Model evolution to enable continuous publication, diverse, heterogeneous use cases
  • Model driven user interface that self-configures to current data model

• Integration with Globus Auth, Minids, BDBags, and other components
Globus Search platform service

• Search service:
  • **Scalable**: to billions of entries
  • **Schema agnostic**: can use standard (e.g., DataCite) or custom metadata
  • **Fine grain access control**: only returns results that are visible to user
  • **Plain text search**: ranked results
  • **Faceted search**: for data discovery
  • **Rich query language**: ranges, expressions, regex, fuzzy, stemming, etc.

• Limited production, generally available target year end
A Commons Platform for Promoting Continuous FAIRNESS: UChicago and USC

Workspace Manager

- Bags
- Workspaces
- Pipelines

Virtual cohorts in standard manifest with lightweight ID

Uniform, secure, reliable access to storage

Search

Discover

Discover

Virtual cohorts in standard manifest with lightweight ID

Uniform search across multiple data sources

Publication assigns DOIs and indexes datasets

Workspaces support Jupyter and Galaxy on different clouds

All results tracked via standard manifest and lightweight IDs

Workspace Manager

- Bags
- Workspaces
- Pipelines

- minid_1
  - Galaxy
  - GTExRNA

- minid_2
  - Jupyter
  - GATKVar

- minid_3
  - RStudio
  - UCSC
Creating a atlas of putative transcription factor binding sites from terabytes of ENCODE DNase I hypersensitive sites sequencing data

A case study in reproducible big data science
Atlas of Transcription Factor Binding Sites from ENCODE DNase Hypersensitivity Data Across 27 Tissue Types

Cory C Funk, Segun Jung, Matthew A Richards, Alex Rodriguez, Paul Shannon, Rory Donovan, Ben Heavner, Kyle Chard, Yukai Xiao, Gustavo Glusman, Nilufer Erteskin-Taner, Todd Golde, Arthur Toga, Leroy Hood, John D Van Horn, Carl Kesselman, Ian Foster, Seth Ament, Ravi Madduri, Nathan D Price

doi: https://doi.org/10.1101/252023

Reproducible big data science: A case study in continuous FAIRness

Ravi K Madduri, Kyle Chard, Mike D’Arcy, Segun C Jung, Alexis Rodriguez, Dinanath Sulakhe, Eric W Deutsch, Cory Funk, Ben Heavner, Matthew Richards, Paul Shannon, Gustavo Glusman, Nathan Price, Carl Kesselman, Ian Foster

doi: https://doi.org/10.1101/268755
Analysis plan for generation of TFBS Atlas

- Uniform processing of next generation sequencing data
  - Align to reference genome
  - Identify DNase hypersensitive regions
- Apply multiple footprinting algorithms to locate putative transcription factor binding sites (TFBSs)
- Evaluate confidence in putative TFBSs
- Use TFBSs as features for machine learning approaches applied to disease-specific research
Not all Footprinting algorithms are the same

• Wellington—oldest and most widely used, somewhat conservative

• HINT—new approach has high sensitivity (lower specificity)

• PIQ—very different approach, starts with TF motif, good for low affinity receptors
Data and analysis products

<table>
<thead>
<tr>
<th>#</th>
<th>Name</th>
<th>Identifier</th>
<th>Role</th>
<th>Description</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DNase-Seq</td>
<td>minid:b9dt2t</td>
<td>In</td>
<td>BDBag of 27 BDBags extracted from ENCODE by encode2bag, one per tissue: 1,591 FASTQ files in all.</td>
<td>2.40 TB</td>
</tr>
<tr>
<td>2</td>
<td>Alignment</td>
<td>minid:b9vx04</td>
<td>Out</td>
<td>BDBag of 54 BDBags produced by encode2bag, one per tissue, seed: 386 BAM files in all.</td>
<td>5.30 TB</td>
</tr>
<tr>
<td>3</td>
<td>Footprints</td>
<td>minid:b9496p</td>
<td>In</td>
<td>BDBag of 54 BDBags containing footprints computed by encode2bag, one per tissue, seed. Each BDBag contains two BED files per biosample, one per footprinting method.</td>
<td>0.04 TB</td>
</tr>
<tr>
<td>4</td>
<td>Motifs</td>
<td>minid:b97957</td>
<td>In</td>
<td>Database dump file containing the non-redundant motifs provided by Funk et al. [13].</td>
<td>31.5 GB</td>
</tr>
<tr>
<td>5</td>
<td>Hits</td>
<td>minid:b9p09p</td>
<td>Out</td>
<td>Database dump file containing the hits produced by encode2bag</td>
<td>0.04 TB</td>
</tr>
<tr>
<td>6</td>
<td>TFBSs</td>
<td>minid:b9v398</td>
<td>Out</td>
<td>BDBag of 54 BDBags containing candidate TFBSs produced by encode2bag, one per tissue, seed. Each BDBag contains two database dump files, one per footprinting method.</td>
<td>0.35 TB</td>
</tr>
</tbody>
</table>

**Identifiers for software**

1. extractDNaseSeq (encode2bag)
   - encode2bag service: https://github.com/inn-bdss/encode2bag-service
   - encode2bag client: https://github.com/inn-bdss/encode2bag
2. alignDNaseFootprints
   - Galaxy pipeline: minid:b93e4q
   - Dockerfile: minid:b9j46f
   - Docker image: minid:b97x0j
3. hits
   - R script: minid:b9azh5t
4. TFBSs
   - R script: minid:b9fxis
Digital Object Creation using minids
Table 1. Details of the per-tissue computations performed in the ensemble footprinting phase. Data sizes are in GB. Times are in hours on a 32-core AWS node; they sum to 2,149.1 node hours or 68,771 core hours. DNase: DNase Hypersensitivity (DNase-seq) data from ENCODE. Align: Aligned sequence data. Foot: Footprint data and footprint inference computation. Numbers may not sum perfectly due to rounding.

<table>
<thead>
<tr>
<th>Tissue</th>
<th>Biosamples</th>
<th>Replicates</th>
<th>DNase</th>
<th>Data size</th>
<th>Align</th>
<th>Foot</th>
<th>Compute time</th>
<th>Align</th>
<th>Foot</th>
</tr>
</thead>
<tbody>
<tr>
<td>adrenal gland</td>
<td>3</td>
<td>8</td>
<td>31</td>
<td>68</td>
<td>0.5</td>
<td>7.4</td>
<td>18.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>blood vessel</td>
<td>10</td>
<td>129</td>
<td>117</td>
<td>234</td>
<td>2.1</td>
<td>32.7</td>
<td>68.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>bone element</td>
<td>1</td>
<td>7</td>
<td>4</td>
<td>6</td>
<td>0.2</td>
<td>1.4</td>
<td>3.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>brain</td>
<td>29</td>
<td>185</td>
<td>402</td>
<td>840</td>
<td>6.2</td>
<td>120.0</td>
<td>160.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>bronchus</td>
<td>2</td>
<td>9</td>
<td>18</td>
<td>36</td>
<td>0.4</td>
<td>3.1</td>
<td>9.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>esophagus</td>
<td>2</td>
<td>41</td>
<td>35</td>
<td>64</td>
<td>0.3</td>
<td>12.6</td>
<td>7.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>extraembryonic</td>
<td>11</td>
<td>66</td>
<td>193</td>
<td>412</td>
<td>3.0</td>
<td>46.5</td>
<td>89.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>eye</td>
<td>8</td>
<td>53</td>
<td>129</td>
<td>252</td>
<td>2.3</td>
<td>26.9</td>
<td>109.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gonad</td>
<td>2</td>
<td>7</td>
<td>27</td>
<td>56</td>
<td>0.4</td>
<td>4.9</td>
<td>12.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>heart</td>
<td>8</td>
<td>69</td>
<td>169</td>
<td>342</td>
<td>2.0</td>
<td>38.8</td>
<td>76.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>kidney</td>
<td>8</td>
<td>29</td>
<td>84</td>
<td>174</td>
<td>2.0</td>
<td>19.5</td>
<td>96.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>large intestine</td>
<td>5</td>
<td>18</td>
<td>96</td>
<td>184</td>
<td>0.9</td>
<td>23.2</td>
<td>50.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>liver</td>
<td>3</td>
<td>8</td>
<td>26</td>
<td>62</td>
<td>0.5</td>
<td>4.9</td>
<td>19.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>lung</td>
<td>7</td>
<td>94</td>
<td>142</td>
<td>300</td>
<td>1.8</td>
<td>19.3</td>
<td>27.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>lymphatic vessel</td>
<td>2</td>
<td>30</td>
<td>21</td>
<td>286</td>
<td>0.4</td>
<td>3.6</td>
<td>11.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>lymphoblast</td>
<td>21</td>
<td>71</td>
<td>150</td>
<td>320</td>
<td>2.9</td>
<td>70.0</td>
<td>153.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mammary gland</td>
<td>2</td>
<td>5</td>
<td>18</td>
<td>36</td>
<td>0.4</td>
<td>2.8</td>
<td>10.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mouth</td>
<td>4</td>
<td>18</td>
<td>81</td>
<td>164</td>
<td>1.1</td>
<td>19.5</td>
<td>57.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>muscle organ</td>
<td>4</td>
<td>13</td>
<td>54</td>
<td>110</td>
<td>0.8</td>
<td>9.6</td>
<td>49.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pancreas</td>
<td>2</td>
<td>13</td>
<td>38</td>
<td>84</td>
<td>0.5</td>
<td>8.1</td>
<td>30.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>prostate gland</td>
<td>2</td>
<td>8</td>
<td>23</td>
<td>50</td>
<td>0.3</td>
<td>4.2</td>
<td>78.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>skin</td>
<td>48</td>
<td>401</td>
<td>377</td>
<td>780</td>
<td>8.8</td>
<td>220.0</td>
<td>181.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>spinal cord</td>
<td>2</td>
<td>34</td>
<td>66</td>
<td>128</td>
<td>0.7</td>
<td>9.2</td>
<td>28.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>stomach</td>
<td>1</td>
<td>5</td>
<td>24</td>
<td>52</td>
<td>0.4</td>
<td>3.5</td>
<td>3.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>thyroid gland</td>
<td>3</td>
<td>24</td>
<td>63</td>
<td>136</td>
<td>0.7</td>
<td>13.6</td>
<td>27.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tongue</td>
<td>2</td>
<td>8</td>
<td>51</td>
<td>106</td>
<td>0.7</td>
<td>13.7</td>
<td>26.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>urinary bladder</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>0.2</td>
<td>0.8</td>
<td>1.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>193</td>
<td>1,355</td>
<td>2,443</td>
<td>5,291</td>
<td>40.8</td>
<td>739.7</td>
<td>1,409.4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Detailed, validated instructions for reproducibility and reusability

https://github.com/globusgenomics/genomics-footprint
Lessons Learned

• Reproducibility requires discipline, patience
• Complexity and level of effort increases with size of the analysis team and data sources
• Data lifecycle and economics are still open issues
• Sustainability of infrastructure built
Acknowledgements

We thank the Galaxy and Globus teams, and other developers of tools on which we build here, for their work. We also thank the NIH Commons Cloud Credits program and the Amazon Web Services Research Cloud Credits program for their support. This work was supported in part by NIH contracts 1U54EB020406-01: Big Data for Discovery Science Center, 1OT3OD025458-01: A Commons Platform for Promoting Continuous Fairness, and 5R01HG009018: Hardening Globus Genomics; and DOE contract DE-AC02-06CH11357.