DATA REUSE IN AGRICULTURAL GENOMICS RESEARCH: Challenges, solutions, and a case study

Alenka Hafner
PhD candidate in Plant Biology, Penn State University
AgBioData Data Reuse WG
MOTIVATION FOR DATA REUSE WG

- Sequence-based datasets + WWW + Open Science movement
  - a lot of data out there
- No dataset is perfect
- Data are often underutilized due to insufficient metadata and other challenges related to the FAIR data standards
- Objective for WG: to identify bottlenecks in data reuse and critical needs to propose solutions for agricultural genomics community
- Personal objective: understand the lack of reusability of methylome data
Barriers to data reuse and recommendations to overcome them

A case study of methylome data reuse

The future of data reuse

DATA REUSE =

the practice of utilizing existing data for

a novel scientific purpose beyond their original scope
1. **BARRIERS TO DATA REUSE & RECOMMENDATIONS TO OVERCOME THEM**

- Data quality
- Metadata
- Data availability
- Interoperability
- Data ownership
- User skill and resources
DATA QUALITY: STANDARDS AS A SOLUTION

• Data made publicly available regardless of quality → a (subjective) decision on suitability for reuse

• Factors to assess:
  ▪ coverage,
  ▪ depth,
  ▪ technical and biological replication,
  ▪ tissue type,
  ▪ sample collection method,

• Limited scope of existing standards (even for common data types and organisms)

  ▪ extraction method and library preparation,
  ▪ sequencing technology,
  ▪ other dataset properties

• More protocol, pipeline, and statistical standards needed in agricultural genomics field
ON THE ROAD TO COMPLETE METADATA: INCENTIVES

- Limited/incomplete/missing metadata submission templates
- Submission requires work □ Trade-off between collecting all some metadata via a lenient submission system and mandating comprehensive metadata
- Incentives are needed! E.g., data citations…
Survey of Science and Nature in 2021 found that an alarming less than 50% of data stated to be “available upon request” could be effectively obtained from the original authors (Tedersoo et al., 2021)
TOWARDS INTEROPERABILITY: DATA FORMATTING

• Our community has converged on (meta)data standards for data file types:
  FASTQ, SAM/BAM, VCF, GTF, GFF3, BED, …

• Issue: lack of standards consistency of use

• Reference genome mapping can be an issue down the line

• “Backwards compatibility”: outdated lab and sequencing methodology can be combated through extensive metadata (https://www.protocols.io)
DATA OWNERSHIP & SHARING REQUIREMENTS

• Challenge: Having access to relevant, affordable study populations from breeding companies that can also be shared publicly as sequence or genotype data

• Already many sharing requirements + 2026 mandate to make research funded by the USA government publicly available
A recent study (LaFlamme et al., 2022) shows that skill or perceived ability was identified by many participants as a major factor influencing reuse behavior.

2017-2018 global survey: most scientists exhibited “high and mediocre risk data practices” (Tenopir et al., 2020).

US-based institutions: computational resources likely not the limiting factor – it’s skill level

Education programs for awareness-raising and good practice training needed

Incentives (!): DataWorks! Prize (https://www.herox.com/dataworks)
2. A CASE STUDY OF METHYLOME DATA REUSE

- Sample provenance
- Type of replicate
- Consistency of formats
- .pdf
- Computational resources
- Barrier of entry

<table>
<thead>
<tr>
<th>Data quality</th>
<th>Metadata</th>
<th>Data availability</th>
<th>Interoperability</th>
<th>Data ownership</th>
<th>User skill and resources</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Depth of sequencing</td>
<td>• Pipeline intermediates</td>
<td>• Crops?</td>
<td>• Available on request”</td>
<td>• Code</td>
<td>• “Available on request”</td>
</tr>
<tr>
<td>• Experimental design</td>
<td>• .pdf</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Tissue type</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• REPLICATES!</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3. THE FUTURE OF DATA REUSE

The importance and benefits of equity and inclusion

Take-aways and looking ahead
THE IMPORTANCE AND BENEFITS OF EQUITY AND INCLUSION

- **Diversity breeds innovation**
- Reuse requires computational capacity, internet access, digital literacy, and proficiency in dominant languages
- Data sovereignty: [https://localcontexts.org](https://localcontexts.org)
TAKE-AWAYS AND LOOKING AHEAD

The future of data reuse is bright and exciting!

- Integration of datasets
- Emerging data types: phenomes, metabolomes, proteomes, interactomes, enviromes, microbiomes, lipidomes, and glycomes
- AI and ML