Wheat data management and sharing guidelines

Esther Dzale Yeuomo Kabore
Chair
French National Institute for Agriculture Research
Wheat Data Interoperability

Esther Dzalé Yeumo
Co-chair RDA Wheat Data Interoperability WG
Chair INRA competence center for data management and sharing services
An international research partnership for wheat improvement

• Created in 2011 following endorsement by G20 Agriculture Ministries to improve food security
• A framework to identify synergies and facilitate collaborations for wheat improvement at the international level
• The Wheat Initiative members
  – **Countries**: Argentina, Australia, Brazil, Canada, China, France, Germany, Hungary, India, Reland, Italy, Japan, Spain, Turkey, UK, USA
  – **International organizations**: CIMMYT, ICARDA
  – **Private companies**: Arvalis, Bayer CropScience, Florimond Desprez V&F, KWS UK, Limagrain, Monsanto Company, RAGT 2n Saateen Union Research, Syngenta Crop Protection

http://www.wheatinitiative.org/
The WheatIS Expert Working Group

- Build projects
- Build infrastructure
- Report to the Wheat Initiative
The WheatIS EWG goals

- Provide the wheat research community with a *single* entry point of access to wheat research data.
- Promote the development of services on top of current wheat / Triticeae databases.
- Authority to define guidelines for data curation, nomenclature, standards and integration.
- Registry for bioinformatics tools.
Most of the participants supported the data reuse policy promoted by the Bermuda/Fort Lauderdale/Toronto agreements (Nature 461, 168F170, doi:10.1038/461168a), that promotes the early dissemination of whole genome datasets but preserves the rights for the data generators to lead the analysis and publication of their data in peer reviewed journals.
WheatIS Expert Worging Group

State of the art

Studies

- Evolutionary: 56
- Genome structure: 89
- Map based cloning: 100
- Pop. genetic, diversity: 104
- Breeding: 116
- Comparative genomics: 119
- Functional: 138
- Genotype to phenotype: 148

Data types

- Segmental duplications
- Transposable elements
- Non coding RNA
- Simple repeats
- Metatranscriptomes
- Epigenome
- Protein profiles
- Alternative splicing
- Q-PCR
- Copy Number Variation
- Microarrays
- Presence Absence Variation
- Microarray
- Allele specific transcripts
- Orthology
- Protein families
- Wild relatives
- Association panels
- SSR
- Biparental populations
- RNAseq
- Molecular resources
- Mutants insertion lines
- Gene atlas expression
- Varieties
- Physical maps
- QTL maps
- Gene coding annotations
- Phenotypes
- Molecular markers
- Assemblies
- Genetic maps
- SNP

Repositories
The interoperability challenge illustrated

Top management or breeders
- What are the sources of resistance to stem rust (UG99) and tolerance to drought conditions in bread wheat?

Data scientists, bioinformaticians
- How do I extract information from all these databases?
- Do I have well-documented metadata to make queries?
- How do I link the data to get smarter information?

Data manager, data provider
- I want to make my data findable, reusable and linkable to other data:
  - What ontologies and metadata elements are commonly used to describe the types of data I am dealing with?
  - What data formats could I use to share my data
  - Where could I deposit my data?

Data are
- Dispersed
- Heretogeneous
- Abundant
The Wheat Data Interoperability WG

- Created in March 2014 within the frame of RDA
- Aims: contribute to the improvement of Wheat related data interoperability by
  - Building a common interoperability framework (metadata, data formats and vocabularies)
  - Providing guidelines for describing, representing and linking Wheat related data
The achievements

Surveys
- Landscape of Wheat related standards and their use by the community
- Comprehensive overview of Wheat related ontologies and vocabularies

Workshops
- Recommendations
- Mappings between different data formats
- Actions to conduct in order to improve the current level of Wheat related data interoperability
- Interoperability use cases

Implementation
- Interactive cookbook: recommendations + guidelines
- A repository of Wheat related linked vocabularies (Bioportal)
Data management practices survey

• Objective: identify
  – Data storage practices
  – Data management policy or guidelines in use
  – Data formats in use
  – Ontologies and vocabularies in use

• Complete results
Data management practices survey

- Total number of answers: 201
- Number of complete answers: 125
- Total number of incomplete answers: 77 (6 doubles removed: people who answered twice)
- Number of answers considered: 196
Ontologies & vocabularies survey

• Objective
  – Assess the level of visibility and interoperability of Wheat related vocabularies and ontologies
  – Identify the domain covered by the ontologies and vocabularies
  – Collect some technical details
Welcome

These recommendations have been drafted by the Wheat Genomics Working Group (WG), one of the Wheat Genomics Interoperability Interest Group (WG-III) initiatives, and aim to reinforce research programs to include societal demands for sustainable agriculture.

PROMOTE
the adoption of common standards, vocabularies, best practices for Wheat management

Wheat Data Interoperability Guidelines

Sequence variations

The sequence variations are the nucleotide differences between two or more sequences at the same locus (usually between a reference sequence and another sequence). These types of sequence variations include single nucleotide polymorphisms (SNPs), insertions and deletions (INDELs), and short tandem repeats (STRs). They are mainly reported in plant genomes. The most commonly available sequence variations for wheat are SNPs.

Recommendations

Summary

For Variant (e.g. SNP) calling performed by biologists:
1. Use a reference wheat genome sequence.
2. Data format: Use the VCF format.
3. Provide associated metadata.

1. Reference sequence

The currently most commonly used reference wheat genome sequence (Chinese Spring) is available at the WSCC Sequence Repository (http://datastandards.wheatis.org).

When available, we encourage the use of the chromosome refer modeling tool.

2. Data format

We recommend to use the latest VCF file format.

Description

The VCF format (VCF for Variant Call File) is a file format used in bioinformatics that has been developed with the advent of large-scale genomics. The 1000 Genomes Project VCF format specifications can be found at http://www.1000genomes.org.

Warning: The VCF format has gained some capture capability, but still needs to be handled with care in terms of content.

3. Metadata

We recommend to provide a minimal set of metadata to context, provide information about the SNP quality analysis.

Data sharing

For data sharing, the following information should be provided in each run to be accepted by both SNPs and SNP tools:

Name
Description
RUN NAME
Name of the sequencing run that produced the data.
RUN DESCRIPTION
Description of this run.
SUB RUN NAME
Part of a sequencing run that produced the data. The sequencing technology involved (e.g., Ion Torrent's) is described for Illumina runs.
ANALYSIS NAME
Name of the SNP calling tool.
ANALYSIS SOFTWARE NAME
Software used for the SNP calling analysis.
ANALYSISCONTACT NAME
Name of the person who performed the analysis.
ANALYSISCONTACT EMAIL
Email of the person who performed the analysis.
PROTOCOL NAME
Name of the sequencing protocol.
MAPPING GENOTYPE NAME
Name of the genotype.
MAPPING GENOTYPE TAXON NAME
Taxon of the reference genome used to map the genotype.
MAPPING GENOTYPE DESCRIPTION
Description of the sample/individual that has been sequenced.
GENOTYPETAXON NAME
Taxon of the sample/individual that has been sequenced.
PROJECT NAME
Name of the project that funded the sequencing.
FILTERS
Filters applied to call SNPs (e.g., DP = 10).

Example

Example of a VCF file declared to wheat data:

#VCF file format is specified.

SNP calling tools

GATK
SAMBAM
SAMtools

Filter tools

VCFtools
VCFtab
SAMtools

Warning: BAM/SAM files should be kept for traceability of further analysis since they are not suitable for sharing.

Data submission

For data submission in international repositories (e.g. NCBI), we advise filling the dedicated XML format:

http://www.ebi.ac.uk/ena/index.html#submitting-data

Most popular Tools

Identification of sequence variations includes 3 steps:

1. Mapping of the reads on the reference genome.
2. Calling the sequence variations.
3. Filtering out incorrect results based on mapping depth and sequence quality and mapping quality.

Mapping tools

BWA
Bowie
Bowie 2

SNP calling tools

GATK
SAM tools

Filter tools

VCF tools
VCF tab
SAM tools

No Comments Yet

Leave a Reply

Your email address will not be published. Required fields are marked:

Name *
Email *
Website
Comment

Post Comment

Copyright © 2015

Wheat Data Interoperability guidelines

http://datastandards.wheatis.org
Wheat related vocabularies in Agroportal

- [http://wheat.agroportal.lirmm.fr/ontologies](http://wheat.agroportal.lirmm.fr/ontologies)
  - Access to, and retrieve the ontologies through the Web interface, an API and a Sparql Endpoint
  - Subscribe a RSS feed to receive alerts for submissions of new ontologies, new versions of ontologies, new notes, and new projects. You can subscribe to feeds for a specific ontology at the individual ontology page
  - Search for terms across multiple ontologies, browse mappings between terms in different ontologies, receive recommendations on which ontologies are most relevant for a corpus, annotate text with terms from ontologies
The benefits

For data producers, managers, providers

- One stop shop for relevant information related to wheat data management → arise awareness, avoid duplicated efforts, foster adoption of common practices
- Facilitate the use of common data exchange formats → easy data sharing/submission to international repositories
- Foster a standardized description of datasets with consistent use of ontologies and metadata → increase the identification, the findability and the usability of the dataset

For data scientists, bioinfomaticians

- Facilitate the access, integration and analysis of data from various sources
- Access to data of higher quality

For top management, researchers

- Increase the chance to answer complex questions
Acknowledgement

**WDI WG members**: Fulss Richard, co-chair (CIMMYT), Alaux Michael, Alaux Michael (INRA), Aubin Sophie (INRA), Arnaud Elizabeth (Bioversity), Baumann Ute (Adelaide University), Buche Patrice (INRA), Cooper Laurel (Planteome), Hologne Odile (INRA), Laporte Marie-Angélique (Bioversity), Larmand Pierre (IRD), Letellier Thomas (INRA), Pommier Cyril (INRA), Protonotarios Vassilis (Agro-Know), Shrestha Rosemary (CIMMYT), Subirats Imma (FAO of the United Nations), Aravind Venkatesan (IBC), Whan Alex (CSIRO)

And

Clément Jonquet (Lirmm, Agroportal), Hélène Lucas (Wheat Initiative) Hadi Quesneville (WheatIS EWG)
Thank you to our sponsors:

We will add this to the end of each presentation
Thank you!

University of Idaho

Washington State University

Oregon State University

USDA
NIFA

United States Department of Agriculture
National Institute of Food and Agriculture

Pacific Northwest Farmers Cooperative

CHS

Monsanto