RESEARCH DATA ALLIANCE

Wheat Data Interoperability WG outputs

research data sharing without barriers

rd-alliance.org

The problem

Top management or breeders

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

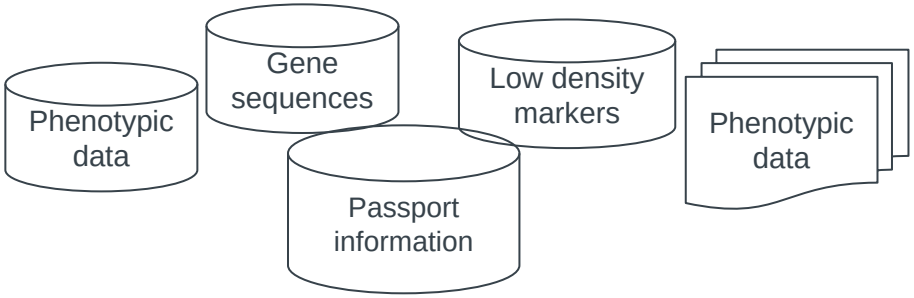
Data scientists, bioinformaticians

- 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

- 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

Contributors

Sponsors

Active members: Alaux Michael (INRA, France), Aubin Sophie (INRA, France), Arnaud Elizabeth (Bioversity, France), Baumann Ute (Adelaide Uni, Australia), Buche Patrice (INRA, France), Cooper Laurel (Planteome, USA), Fulss Richard (CIMMYT, Mexico), Hologne Odile (INRA, France), Laporte Marie-Angélique (Bioversity, France), Larmand Pierre (IRD, France), Letellier Thomas (INRA, France), Lucas Hélène (INRA, France), Pommier Cyril (INRA, France), Protonotarios Vassilis (Agro-Know, Greece), Quesneville Hadi (INRA, France), Shrestha Rosemary (INRA, France), Subirats Imma (FAO of the United Nations, Italy), Aravind Venkatesan (IBC, France), Whan Alex (CSIRO, Australia)

Co-chairs: Esther Dzalé Yeumo Kaboré (INRA, France), Richard Allan Fulss (CIMMYT, Mexico)

The deliverables

- Guidelines (http://wheatis.org/DataStandards.php)
  - Data exchange formats
    - Example: VCF (Variant Call Format) for sequence variation data, GFF3 for genome annotation data, etc.
  - Data description best practices
    - Consistent use of ontologies, consistent use of external database cross references
  - Data sharing best practices
    - Share data matrices along with relevant metadata (example: trait along with method, units and scales or environmental ones)
  - Useful tools and use cases that highlight data formats and vocabularies issues
- A portal of wheat related ontologies and vocabularies (http://agroportal.lirmm.fr/ontologies?filter=WHEAT)
  - Allows the access to the ontologies and vocabularies through APIs.
- A prototype
  - Implementation of use cases of wheat data integration within the AgroLD (Agronomic Linked Data) tool: http://volvestre.cirad.fr:8080/agrold/

Wheat Data Interoperability Guidelines

Home Guidelines Ontolog

Photo banner: Daniel Jolivet - https://flic.kr/p/usuQ7z

Welcome

These recommendations hav
Group (WG), one of the WGs
Interoperability Interest Grou
initiative that aims to reinfor
research programmes to inc
societal demands for sustai

PROMOTE
the adoption of commo
standards, vocabularies a
best practices for Wheat d
management

Home › Sequence variations

Sequence variations

The sequence variations are the nucleotides differences between two (or several) sequences at the same locus (usually between a reference sequence and another sequence). Three types of sequence variations— single-nucleotide polymorphisms (SNPs), insertions and deletions (indels), and short tandem repeats (STRs) — have been mainly reported in plant genomes.
The most currently available sequence variations for wheat are SNPs.

Recommendations

Summary

For Variant (e.g. SNP) calling performed by bioint

1. Use a reference wheat genome sequence
2. Data format: Use the VCF
3. Provide associated metadata

1. Reference sequence

The currently most commonly used reference bread wheat seque
Chinese Spring), available at the IWGSC Sequence Repository and

When available, we encourage the use of the chromosomes refer

2. Data format

We recommend to use the latest VCF file format.

Description
The Variant Call Format (VCF) is a text file used in bioinformatics
format has been developed with the advent of large-scale genoty
the 1000 Genomes Project. VCF format specifications can be fou

Warning: The VCF files generated for exome capture need to be
with those from IWGSC context.

3. Metadata

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

Data sharing
For data sharing, the following information should be provided in
lines have to be preceded by "##" characters) or as a separate ta

Name	Description
RUN NAME	Name of the sequencing run that produ
RUN DESCRIPTION	Description of this run.
SUB RUN NAME	Part of a sequencing run that produced to the sequencing technology involved, sequencers), a flowcell for (Ilumina seq
ANALYSIS NAME	Name of the SNP calling analysis
ANALYSIS SOFTWARE NAME	Software used for the SNP calling analy
ANALYSISCONTACT NAME	Person who performed the analysis
PROTOCOL NAME	Name of the sequencing protocol
MAPPING GENOME NAME	Name and version of the reference geno
MAPPING GENOME TAXON NAME	Taxon of the reference genome used to
MAPPING_GENOME DESCRIPTION	Description of the reference genome used to call the variations
GENOTYPE NAME	Name of the sample/individual that has been sequenced.
GENOTYPE TAXON	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 (ex: DP > 10)

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

Data submission
For data submission in international repositories (EBI, NCBI), we advise to fill the dedicated XML format (http://www.ebi.ac.uk/ena/submit/preparing-xmls#vcf).

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 unrelevant results regarding mainly depth and sequence quality and mapping quality.

Mapping tools

- BWA
- Bowtie
- Bowtie 2

SNP calling tools

- GATK
- SAM tools

Filter tools

- VCF tools
- VCF utils
- SAM tools

Example

Example of a VCF file dedicated to wheat data:

```
##fileformat=VCFv4.1
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT 102 4
labasskaja CS Estacao M6 Marquis Neepawa PI153785 P
PI185715 PI192001 PI192147 PI192569 PI210945 PI2226
297 PI349512 PI366716 PI366905 PI382150 PI4
I481718 PI481923 PI565213 PI82469 PI8813 PR267 Roem
cc3 acc4 acc5 berkut chakwal86 cham6 clear_white dh
maco opata pavon pbw343 rac875 vorobey
3929455_1al 1623 . T C 245.53 . AC=18;AF=0.196;AN=9
;Dels=0.00;FS=0.000;HaplotypeScore=0.1087;Inbreedin
AF=0.196;MQ=100.00;MQ0=0;MQRankSum=-1.426;QD=27.28;
D:DP:GQ:PL 0/0:1,0:1:3:0,3,41 0/0:1,0:1:3:0,3,41 1/
:3:41,3,0 ./. 0/0:1,0:1:3:0,3,41 0/0:1,0:1:3:0,3,39
. ./. 1/1:0,1:1:3:39,3,0 0/0:1,0:1:3:0,3,39 ./. 1/1
```

```
0:1:3:0,3,36 0/0:1,0:1:3:0,3,38 0/0:1,0:1:3:0,3,39 0/0:1,0:1:3:0,3,39 0/0:1,0
:1:3:0,3,38 0/0:1,0:1:3:0,3,38 0/0:1,0:1:3:0,3,38 1/1:0,1:1:3:39,3,0 1/1:0,1:
1:3:38,3,0 1/1:0,1:1:3:38,3,0 0/0:1,0:1:3:0,3,39 0/0:1,0:1:3:0,3,38 1/1:0,1:1
:3:38,3,0
```

```
Writing: WDI working group
Creation date: 02 October 2014
Update: 30 June 2015
```

No Comments Yet

Leave a Reply

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

Name *

Email *

Website

Comment

Post Comment

Benefits for many target users

- For data managers, data providers
  - One stop shop for relevant information related to 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 datasets
- 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

wheatis.org

About | Collaborators | Search | Data Standards | Links | WheatIS Nodes

WheatIS

@ PRATT J.C. / INRA

Wheat Data Interoperability Guidelines

Home | Guidelines | Ontologies & Vocabularies | Use cases | Getting involved | About

Welcome

These recommendations have been prepared by members of the Wheat Data Interoperability Working Group (WG), one of the WGs of the Research Data Alliance and the only WG of the Agriculture Data Interoperability Interest Group. The group is coordinated by members of the Wheat Initiative, a global initiative that aims to reinforce synergies between bread and durum wheat national and international research programmes to increase food security, nutritional value and safety while taking into account societal demands for sustainable and resilient agricultural production systems.

More specifically, the WG aims to:

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

FACILITATE access, discovery and reuse of wheat data

FACILITATE wheat data integration

GETTING INVOLVED

RDA RESEARCH DATA ALLIANCE | WheatIS

Guidelines | Ontologies & Vocabularies | Use Cases

evolve theme by Theme4Press • Powered by WordPress

If we have missed a link
The site is supported by

About

This project aims at building an International Wheat Information System, called hereafter WheatIS, to support the wheat research community. The main objective is to provide a single-access web base system to access to the available data resources and bioinformatics tools.

This project is based on the principles listed below:

- Collective building of the WheatIS to better respond to the needs of the international wheat community;
- Incremental implementation to offer rapidly an operational information system;
- Emphasis on Quality Assurance to serve as a framework for an approach with incremental implementation;
- Promotion of an open-access model for data exchange;
- Reliance on a distributed system;
- Use of Virtual Machine and Cloud Computing technologies to facilitate sharing data and tools;
- Promotion of the visibility of each participating platform to contribute to their sustainability.

How to access and use the deliverables

- The guidelines: http://wheatis.org/DataStandards.php
- The ontologies portal: http://agroportal.lirmm.fr/ontologies?filter=WHEAT
  - The API documentation: http://data.agroportal.lirmm.fr/documentation
- The semantic web prototype: http://volvestre.cirad.fr:8080/agrold/

Contact Information

- For feedback on the guidelines, please post a comment on the website: http://wheatis.org/DataStandards.php
- For any other question email to rda-wdinterop-wg@rda-groups.org

Thank you!