OPINION ARTICLE

Recommendations for the formatting of Variant Call Format (VCF) files to make plant genotyping data FAIR [version 1; peer review: 2 approved with reservations]

Sebastian Beier[1,2], Anne Fiebig[1], Cyril Pommier[3], Isuru Liyanage[4], Matthias Lange[1], Paul J. Kersey[5], Stephan Weise[1], Richard Finkers[6,7], Baron Koylass[4], Timothee Cezard[4], Mélanie Courtot[4,8], Bruno Contreras-Moreira[9], Guy Naamati[4], Sarah Dyer[4], Uwe Scholz[1]

[1]Breeding Research, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben, Seeland, 06466, Germany
[2]Institute of Bio- and Geosciences, Bioinformatics (IBG-4), Forschungszentrum Jülich GmbH, Jülich, 52425, Germany
[3]BioinfOmics, Plant bioinformatics facility, Université Paris-Saclay, INRAE, Versailles, France
[4]European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, UK
[5]Royal Botanic Gardens, Kew, Richmond, UK
[6]Plant Breeding, Wageningen University & Research, Wageningen, The Netherlands
[7]Gennovation B.V., Wageningen, The Netherlands
[8]Ontario Institute for Cancer Research, Toronto, Canada
[9]Laboratorio de Biología Computacional y Estructural, Estación Experimental Aula Dei-CSIC, Zaragoza, 50059, Spain

v1 **First published:** 24 Feb 2022, **11**(ELIXIR):231
https://doi.org/10.12688/f1000research.109080.1
Latest published: 19 May 2022, **11**(ELIXIR):231
https://doi.org/10.12688/f1000research.109080.2

Abstract

In this opinion article, we discuss the formatting of files from (plant) genotyping studies, in particular the formatting of (meta-) data in Variant Call Format (VCF) files. The flexibility of the VCF format specification facilitates its use as a generic interchange format across domains but can lead to inconsistency between files in the presentation of metadata. To enable fully autonomous machine actionable data flow, generic elements need to be further specified. We strongly support the merits of the FAIR principles and see the need to facilitate them also through technical implementation specifications. VCF files are an established standard for the exchange and publication of genotyping data. Other data formats are also used to capture variant call data (for example, the HapMap format and the gVCF format), but none currently have the reach of VCF. In VCF, only the sites of variation are described, whereas in gVCF, all positions are listed, and confidence values are also provided. For the sake of simplicity, we will only discuss VCF and our recommendations for its use. However, the part of the VCF standard relating to metadata (as opposed to the actual variant calls) defines a syntactic format but no vocabulary, unique identifier or recommended content. In practice, often only sparse (if any) descriptive metadata is included. When

Open Peer Review

Approval Status ✓ ✓

	1	2
version 2 (revision) 19 May 2022	✓ view	✓ view
version 1 24 Feb 2022	? view	? view

1. **Boas Pucker**, Institute of Plant Biology & BRICS, TU Braunschweig, Braunschweig, Germany
Alenka Hafner, Penn State University, University Park, USA

2. **Micha M. Bayer**, James Hutton Institute, Dundee, UK

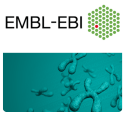
descriptive metadata is provided, proprietary metadata fields are frequently added that have not been agreed upon within the community which may limit long-term and comprehensive interoperability. To address this, we propose recommendations for supplying and encoding metadata, focusing on use cases from the plant sciences. We expect there to be overlap, but also divergence, with the needs of other domains.

Keywords
FAIR, plant, genotyping, snp, vcf, data management, phenotyping, ELIXIR

Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the Bioinformatics gateway.

This article is included in the ELIXIR gateway.



This article is included in the EMBL-EBI collection.

Corresponding author: Sebastian Beier (s.beier@fz-juelich.de)

Author roles: Beier S: Conceptualization, Investigation, Methodology, Project Administration, Supervision, Validation, Writing – Original Draft Preparation; **Fiebig A**: Data Curation, Investigation, Methodology, Writing – Review & Editing; **Pommier C**: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Supervision, Writing – Review & Editing; **Liyanage I**: Data Curation, Resources, Validation, Writing – Review & Editing; **Lange M**: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Resources, Validation; **Kersey PJ**: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Resources, Validation, Writing – Review & Editing; **Weise S**: Data Curation, Methodology, Resources, Validation, Writing – Review & Editing; **Finkers R**: Conceptualization, Data Curation, Investigation, Resources, Writing – Review & Editing; **Koylass B**: Data Curation, Investigation, Resources, Validation, Writing – Review & Editing; **Cezard T**: Data Curation, Formal Analysis, Investigation, Resources, Supervision, Writing – Review & Editing; **Courtot M**: Data Curation, Formal Analysis, Investigation, Methodology, Resources, Supervision, Writing – Review & Editing; **Contreras-Moreira B**: Data Curation, Formal Analysis, Investigation, Methodology, Resources, Supervision, Writing – Review & Editing; **Naamati G**: Data Curation, Formal Analysis, Methodology, Resources, Writing – Review & Editing; **Dyer S**: Data Curation, Resources, Supervision, Writing – Review & Editing; **Scholz U**: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This study received funding from ELIXIR, the research infrastructure for life-science data, through the ELIXIR Implementation Study: FONDUE - FAIR-ification of Plant Genotyping Data and its linking to Phenotyping using ELIXIR Platforms. ML and SW received funding for the AGENT project from the European Union's Horizon 2020 research and innovation programme under grant agreement No 862613. US received funding for the de.NBI project from the German BMBF under the FKZ 031A536A.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Beier S, Fiebig A, Pommier C *et al.* **Recommendations for the formatting of Variant Call Format (VCF) files to make plant genotyping data FAIR [version 1; peer review: 2 approved with reservations]** F1000Research 2022, **11**(ELIXIR):231 https://doi.org/10.12688/f1000research.109080.1

First published: 24 Feb 2022, **11**(ELIXIR):231 https://doi.org/10.12688/f1000research.109080.1

Introduction

As of today, there are several public repositories for genetic and genomic variation data. However, most of these repositories are exclusive to humans and do not include other organisms (NCBI Insights, 2017), such as dbSNP (Sherry *et al.*, 2001), dbGaP (Mailman *et al.*, 2007) and dbVar (Lappalainen *et al.*, 2013). There are two main resources for non-human variation data: The European Variation Archive (EVA) (Cezard *et al.*, 2021), hosted by EMBL-EBI, and the Genome Variation Map (GVM) (Song *et al.*, 2018), hosted by CNCB-NGDC. Submitting datasets to these two repositories works very similarly, but we will focus on the submission of genotyping datasets to EVA. Data and metadata are submitted to a File Transfer Protocol (FTP) file server and, after a quality check, are added to the database and displayed on their respective websites or kept hidden until a user-specified release date. Data are only checked for a few critical points: first, the VCF file must comply with the Variant Call Format (VCF) (Danecek *et al.*, 2011) specifications, second, the genome assembly used as reference must be registered with one of the databases of the International Nucleotide Sequence Database Collaboration (INSDC) (Cochrane *et al.*, 2011), i.e., GenBank (Benson *et al.*, 2013), the European Nucleotide Archive (ENA) (Leinonen *et al.*, 2011) or the DNA Data Bank of Japan (DDBJ) (Mashima *et al.*, 2017), respectively, and an accession number is available, and third, the VCF file must contain either allele frequencies and/or genotype information.

When a data submission is made to the EVA, samples are automatically registered in the associated BioSamples database (Courtot *et al.*, 2022), unless this has been explicitly done previously by the data submitter. Such automatically created samples possess only the minimum necessary attributes (name, domain, release date) and no other descriptive metadata. If pre-registering samples in BioSamples, metadata can be specified as key-value pairs. For some specific use cases, there are already predefined checklists that list which metadata should be supplied on sample registration, against which the metadata can be validated. Additional information, which is not yet available in a defined attribute, can also be submitted under a free text key. We recommend the manual registration of samples at BioSamples as this gives the greatest flexibility when editing and adding information.

Another useful resource for the analysis of plant variation data is Ensembl Plants (Howe *et al.*, 2020). This database, also hosted by EMBL-EBI, is a platform for displaying and visualising plant genomes. If the reference genome associated with the data submitted to EVA is supported in Ensembl, then it is possible to display genetic variants in their genomic context in the Ensembl browser, each linked to its sample. Data submitters should contact Ensembl helpdesk to request it. VCFs in EVA should be available as browsable files, as seen for example in soybean.

Lessons learned from studies on plant phenotyping and its application to metadata information in genotyping

The standardisation of plant variation data is still in its infancy. Therefore, it is beneficial to look to other data types for guidance and improvement. One particular data type where a lot of standardisation work has been done in recent years is plant phenotyping. Plant phenotyping has developed rapidly with the introduction of high-throughput technologies such as fully automated greenhouses, full-time sensor recording and aerial observation drones. The need to record data points and the method of observation has led the community to implement a standard for describing such experiments: MIAPPE (Papoutsoglou *et al.*, 2020) (Minimal Information About a Plant Phenotyping Experiment). Since its introduction in 2014, the standard has been extended to describe sample material (including the anatomical part sampled) through the use of specialised ontologies. MIAPPE-compliant data can be represented in the Investigation-Study-Assay (ISA) framework for structured data representation (Rocca-Serra *et al.*, 2010) and exposed programmatically via the Breeding API (BrAPI) (Selby *et al.*, 2019). The format is maintained and regularly updated by an active community. Fully MIAPPE-compliant data is rich in metadata that describes and identifies in detail both the sample material and the experiment performed. One aim is to allow machine access to the data via application programming interfaces (APIs). Therefore, the use of controlled vocabularies is encouraged by supporting different ontologies, with AgroPortal (Jonquet *et al.*, 2018) serving as a reference repository.

In contrast, genotyping data is often published and shared without sufficient metadata to ensure interoperability and reuse, as seen with other data formats (Bernstein *et al.*, 2017). Current automated tools do not fill in the metadata fields very well, leaving the user to take care of it. Some information that should be recorded cannot be easily retrieved from the analysis results, such as the identification of biological material studied, or the reference genome assembly and version used. Depending on who is handling the data and what skills are associated with the role, the difficulty of providing well-formatted metadata will vary. Bioinformaticians who have directly performed the genotyping analyses and thus the creation of the VCF files will consider it a comparatively simple task to enter metadata directly into the file. Similarly, a data steward who may not have previously been directly familiar with the data but with the structure itself should have no problems. Gathering experimental data from conversations with wet lab colleagues or in a laboratory information management system (LIMS) search will be the more laborious activity for individuals in either role. However,

experimentalists who have little or no experience with the required metadata formats are most likely to be overwhelmed without a simple GUI or input template. Principal investigators who want to submit the data at the end of an experiment may have similar difficulties. From these observations, we recommend performing both metadata and data validation. For the validation of VCF files, we recommend EBI's VCF validator.

Data and metadata formatting

The *de facto* standard data format for genotyping studies is the Variant Call Format (VCF). The following statements are based on the current version 4.3. A VCF file comprises a single text file that consists of three parts: (i) one or more meta-information lines, initiating with a ## describing the settings, samples and general experimental design of the genotyping study. File meta-information is included after the ## string and must be key=value pairs. There are currently no guidelines on how these are used or what they may contain (ii) a header line initiating with a single #, and (iii) one or more data lines, each recording the genotype calls at each varying position in the reference genome for a single sample. Both the header and data lines use tab stops to delineate separate fields. Meta-information lines are considered optional; however, they need to be well-formed if present. All structured lines that have their value enclosed within "<>" require an ID which must be unique within their type (Figure 1).

A critical aspect of VCF specifications is that sample naming within the VCF file does not follow any standard specifications, i.e. the user can name their samples without reference to any real biological material. Even worse, phenotyping and genotyping data from the same experimental setup often use different sample identifiers even when the same biological material has been used, which makes it difficult to reconstruct later which datasets were derived from a common sample. To be able to represent such relationships, descriptive metadata is required that relates these different sample identifiers to each other.

In response to the points discussed previously, we propose a minimal list of metadata fields, recommend an identifier schema and guidelines for vocabulary and data format within a VCF file. Our suggestions are divided into recommended and optional changes. Although, we are primarily addressing data submissions to the EMBL-EBI repositories BioSamples and EVA (and implicitly ENA through the submission of sequence information), subsequent formatting guidelines should be applied regardless of the specific deposition repository and should also be considered when designing databases and APIs.

In our view, these additional fields should be required for a valid VCF:

One meta-information line, ##fileformat, is obligatory in VCF. We also recommend using the additional lines ##filedate, ##bioinformatics_source, ##reference_ac, ##reference_url, ##contig and ##SAMPLE. To ensure permanent unique and stable IDs for samples and genotypes, we recommend the registration of used genotypes and samples in the BioSamples database. This enables the publishing of biological material used in variation studies, and we explicitly recommend the use of long-term stable BioSamples identifiers as primary IDs for material description in VCF files (Table 1).

File date field format

The creation date of the VCF should be specified in the metadata via the field ##fileDate, the notation corresponds to ISO 8601 (Kuhn, 1995) (in the basic form without separator: YYYYMMDD).

```
##fileformat=VCFv4.3
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

Figure 1. Example Variant Call Format (VCF) file structure, including meta-information lines and data lines (from https://samtools.github.io/hts-specs/VCFv4.3.pdf).

Table 1. Summary of recommendations for metadata formatting.

Metadata field	Definition	Format	Example	Cardinality
##fileDate	Creation date of the VCF file	Date (ISO 8601, YYYYMMDD)	##fileDate=20120921	1
##bioinformatics_source	Chains of bioinformatics tools for creating the VCF file	URL, DOI	##bioinformatics_source="doi.org/10.1038/s41588-018-0266-x"	1
##reference_ac	Accession number of reference genome assembly used in the VCF file	/[(GCA/GCF)_(d){9}\.(0-9)*]/	##reference_ac=GCA_902498975.1	1
##reference_url	URL of the reference genome assembly used in the VCF file	URL, DOI	##reference_url="ftp.ncbi.nlm.nih.gov/genomes/all/GCA/902/498/975/GCA_902498975.1_Morex_v2.0/GCA_902498975.1_Morex_v2.0_genomic.fna.gz"	1
##contig	Metadata about a single sequence in the reference genome assembly	Composite (see below)	##contig=<ID=chr1H,length=522466905, assembly=GCA_902498975.1, md5=8d21a35cc68340ecf40e2a8dec9428fa, species=NCBITaxon:4513>	1:N
	The primary identifier of the sequence	String	ID=chr1H	1
	The length in base pairs (bp) of the sequence	Integer	length=522466905	1
	The assembly accession number this sequence belongs to	/[(GCA/GCF)_(d){9}\.(0-9)*]/	assembly=GCA_902498975.1	1
	The md5 checksum of the sequence	MD5	md5=8d21a35cc68340ecf40e2a8dec9428fa	1
	The species of the sequence (NCBI Taxon ID)	/[(NCBITaxon):(\d+)]/	species=NCBITaxon:4513	1
##SAMPLE	Metadata about a single sample genotype that is part of the genotyping experiment in the VCF file	Composite (see below)	##SAMPLE=<ID=SAMEA104646767,DOI="doi.org/10.25642/IPK/GBIS/7811152">	1:N
	The primary identifier (BioSamples Database identifier) of the genotyping sample	/[(SAM)(E|N|D)(A|G)(\d+)]/	ID=SAMEA104646767	1
	The DOI of the genotyping sample (if available)	URL, DOI	DOI="doi.org/10.25642/IPK/GBIS/7811152"	0-1
	The external identifiers under which this genotyping sample is registered in other databases (either 'FAO-WIEWS_instcode:genus:accession_number' or 'DNS: database_identifier:identifier_scheme:identifier')	See Definition	ext_ID="DEU146:Hordeum:HOR 1361 BRG" or ext_ID="ipk-gatersleben.de:GBIS:akzessionId:7811152"	0:N

##fileDate=date

Example:

Description of a VCF file that was created on September 21st in 2012.

```
##fileDate=20120921
```

Bioinformatics source field format

The analytic approach (usually consisting of chains of bioinformatics tools) for creating the VCF file is specified in the ##bioinformatics_source field. Such approaches often involve several steps, like read mapping, variant calling and imputation, each carried out using a different program. Every component of this process should be clearly described, including all the parameter values.

##bioinformatics_source=url

This is ideally specified as the DOI of a publication, or more generally as URL/URI (like a public repository for the scripts and parameters used).

Examples:

1) Description of a GBS experiment in barley and subsequent read alignment and variant calling using a bioinformatics analysis pipeline consisting of cutadapt, BWA-MEM, SAMtools, NovoSort, Picard, BCFtools and seqArray.

```
##bioinformatics_source="doi.org/10.1038/s41588-018-0266-x"
```

2) Modified version of Tassel4 (v.4.3.7) for running the Tassel-GBS pipeline modified for polyploid species with high read depths used in (Pereira *et al.*, 2018).

```
##bioinformatics_source="github.com/gramarga/tassel4-poly"
```

Reference_ac field format

This field contains the accession number (including the version) of the reference sequence on which the variation data of the present VCF is based.

##reference_ac=assembly_accession

The NCBI page on the Genome Assembly Model states (NCBI, 2002): "The assembly accession starts with a three letter prefix, GCA for GenBank assemblies […]. This is followed by an underscore and 9 digits. A version is then added to the accession. For example, the assembly accession for the GenBank version of the public human reference assembly (GRCh38.p11) is GCA_000001405.26". Note these accessions are shared by all INSDC archives.

Example:

Reference genome assembly for barley (*Hordeum vulgare*) cultivar Morex version 2.

```
##reference_ac=GCA_902498975.1
```

Reference_url field format

While the ##reference_ac field contains the accession number of the reference genome, the ##reference_url field contains a URL (or URI/DOI) for downloading of this reference genome, preferably from one INSDC archive.

##reference_url=url

The reference genome should be in FASTA format; the user is free to provide a packed or unpacked publicly available version of the genome.

Example:

Reference genome assembly for barley (*Hordeum vulgare*) cultivar Morex version 2 download link on NCBI FTP.

```
##reference_url="ftp.ncbi.nlm.nih.gov/genomes/all/GCA/902/498/975/
GCA_902498975.1_Morex_v2.0/GCA_902498975.1_Morex_v2.0_genomic.fna.gz"
```

Contig field format

The individual sequence(s) of the reference genome are described in more detail in the #contig field(s).

##contig=<ID=ctg1, length=sequence_length, assembly=gca_accession, md5=md5_hash, species=NCBI Taxon ID>

Each contig contains at least the attribute ID, and typically also include length, assembly, md5 and species. The ID is the identifier of the sequence contig used in the reference genome assembly. Length contains the base pair length of the sequence contig in the reference genome. The assembly is the accession number of the reference genome. If the md5 parameter is given, please note that the individual sequence contigs MD5 checksum is expected, not the MD5 sum of the complete reference genome. The species is the taxonomic name of the species of the reference genome.

Examples:

1) Chromosome 1H of barley (*Hordeum vulgare*) cultivar Morex version 2.

```
##contig=<ID=chr1H,length=522466905,assembly=GCA_902498975.1,md5=8d21a35c-
c68340ecf40e2a8dec9428fa,species=NCBITaxon:4513>
```

2) Chromosome 1 of maize (*Zea mays*) cultivar B73 version 3.

```
##contig=<ID=GK000031.3,length=301433382,assembly=GCA_000005005.5,
md5=74dfe85ad898416814fa98e8d7048f76,species=NCBITaxon:4577>
```

Sample field format

The ##SAMPLE fields describe the material whose variants are given in the genotype call columns in greater detail and can be extended using the specifications of the VCF format.

##SAMPLE=<ID=BioSample_accession, DOI=doi, ext_ID=registry:identifier>

Genotyped samples are indicated in the VCF by the BioSample accession, which is formed as follows (based on information from the BioSamples documentation): “BioSample accessions always begin with SAM. The next letter is either E or N or D depending if the sample information was originally submitted to EMBL-EBI or NCBI or DDBJ, respectively. After that, there may be an A or a G to denote an Assay sample or a Group of samples. Finally, there is a numeric component that may or may not be zero-padded.” Additional information (like complete Multi-Crop Passport Descriptor (Alercia *et al.*, 2015) records) on the sample material is provided under the DOI (Alercia *et al.*, 2018). In case no DOI exists and the material is held by a FAO-WIEWS recognised institution, the external ID consists of the FAO-WIEWS instcode, the genus and the accession number (see example 2). If the database is not registered with FAO-WIEWS and is not available under a DOI, the DNS of the holding institution, the database identifier, the identifier scheme

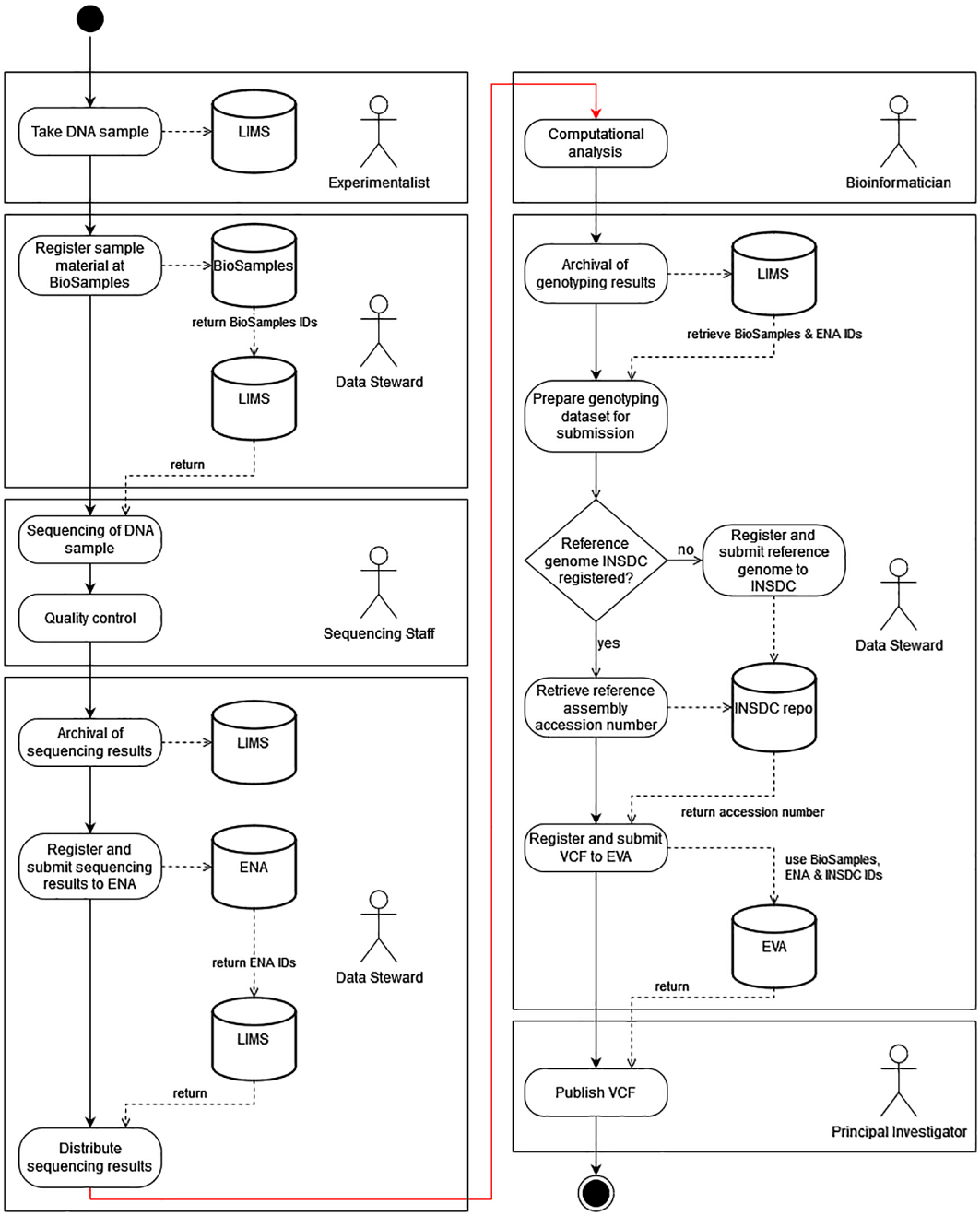


Figure 2. The recommended workflow for the submission of genotypic data to public databases. DNA samples are collected by an Experimentalist and their metadata are stored in a Laboratory Information Management System (LIMS). The Data Steward then registers these samples with BioSamples and in return receives unique BioSamples IDs back, which he adds to the created samples in the LIMS. The sequencing and quality control of these samples is then carried out by the Sequencing Staff and the primary sequence data is fed into the LIMS and linked to the sample data by the Data Steward. The sequencing results are then registered and submitted to the European Nucleotide Archive (ENA) using the BioSamples IDs to link the initially submitted samples to the generated sequencing reads. The study identifiers (ENA IDs) are assigned by ENA and added to the samples by the Data Steward in LIMS. The Bioinformatician then analyses the data and produces the genotyping results. Afterwards, the Data Steward prepares these data for transmission by linking them to the already created sample data from the LIMS and extracting the required metadata and adding it to the header of the Variant Call Format (VCF) file. If the reference genome used for genotyping is not yet available in public repositories, it will now be transferred by the Data Steward to one of the International Nucleotide Sequence Database Collaboration (INSDC) databases. Otherwise, the metadata-enriched VCF file can be registered and submitted to the European Variation Archive (EVA). The identifiers assigned by EVA are then transmitted back and the Principal Investigator can approve the publication of the data.

and the identifier value should be provided (see example 3). For multiple external IDs the field should be used multiple times (delimited by commas).

Examples (Please note that all examples here represent the same genotype. To avoid misunderstandings, if available, the preferred method of describing the data is by DOI.):

1) One genotype from the barley (*Hordeum vulgare*) GBS experiment with a DOI registered.

```
##SAMPLE=<ID=SAMEA104646767,DOI="doi.org/10.25642/IPK/GBIS/7811152">
```

2) One genotype from the barley (*Hordeum vulgare*) GBS experiment with the FAO-WIEWS code available but no DOI.

```
##SAMPLE=<ID=SAMEA104646767,ext_ID="DEU146:Hordeum:HOR 1361 BRG">
```

3) One genotype from the barley (*Hordeum vulgare*) GBS experiment with no DOI and no FAO-WIEWS code available.

```
##SAMPLE=<ID=SAMEA104646767,ext_ID="ipk-gatersleben.de:GBIS:akzessio-
nId:7811152">
```

Recommendations for data fields

In order to allow the highest degree of interoperability, we suggest using BioSamples IDs as the column headers for each sample. In the header line, they should be provided after the 9 mandatory column headings (#CHROM, POS, ID, REF, ALT, QUAL, FILTER, INFO, FORMAT).

In addition, ensure that the genomic positions in the data lines (consisting of the #CHROM and POS tuple) use the same nomenclature as in the reference genome FASTA file and that the positions of the variations are within the start and end positions of the respective chromosome or contig. Watch out for programmes that change these values automatically (especially during imputation).

Additional meta-information fields

On top of the preceding recommendations to improve findability and interoperability, we encourage everyone to describe their data in as much detail as possible in the metainformation lines. Before introducing new fields, please check the official format specifications (in VCFv4.3 this would be under 1.4 Meta-information lines) to avoid redundancy and possible incompatibilities.

Conclusion

With the data and metadata recommendations for VCF files presented here, we hope to make a contribution to linking genotypic and other data for plants. In our view, the minimum to achieve this is to have traceable material and sample management. Analytical results should be linked out to the respective sample(s) and defined in the context of the study being reported. One way to ensure this is to generate long-term stable identifiers at an early stage, ideally when the sample is taken, and to document all work steps accurately. Reproducibility is also an important aspect, which has recently been criticised more frequently in various studies (Baker, 2016; Miyakawa, 2020). Technologies such as containers or the provision of the entire data set and the analytical computing pipeline in a cloud environment could be a further step towards overcoming such problems (Grüning *et al.*, 2018).

The BioSamples database at EMBL-EBI stores samples metadata and allows their pre-registration; it provides unique, stable identifiers for each sample. BioSamples connects to other archives, enabling consistent tracking through time and assays of the samples and derived data. It supports validation of plant phenotypic metadata according to the MIAPPE standard, ensuring data FAIRness (Wilkinson *et al.*, 2016) at submission time as well as keeping metadata on hold pending publication of results. It is recognised by ELIXIR as a recommended Deposition Database for Biomolecular Data. This ensures that comprehensive, validated metadata can be captured at all stages of sample and data generation and that relationships between samples and derived data can be tracked across molecular archives.

The responsibilities of the people involved may vary from research institution to research institution, but the general tasks for the generation of plant genotyping data and the subsequent publication of these data follow a common pattern. To highlight how the complete data management of a genotyping project could be structured, we have designed an exemplary Unified Modeling Language (UML) diagram (Figure 2) as a best practice proposal. We assume that the research institution has a LIMS and that sample collection, sample preparation, sequencing and all bioinformatic analyses are carried out in house. Even if one or more of these activities are outsourced, most data management activities (indicated in the figure by the actor "Data Steward") and thus also the primary communication with public repositories remain the scientific responsibility of the research institution. It is relatively obvious that the timing of interaction with public repositories varies greatly depending on the purpose (registration of datasets, retrieval of identifiers, or updating of datasets) and is recommended to occur at the earliest possible date in order to use the persistent identifiers of the datasets in the further course of the analyses and thus avoid errors due to the use of short-lived internal identifiers.

This approach to data management facilitates the submission of data for publication or at the end of the research project. Here, the situation often arises that the data steward, under time pressure, fails to submit the necessary (meta-)information to the public repositories. The submitted dataset therefore only consists of very generic and not meaningful metadata (Toczydlowski *et al.*, 2021). Such behaviour is the lesser evil compared to not publishing the dataset but can hinder its interoperability and reusability. During the peer review process, large and complex datasets often cannot be checked in depth by the reviewers. A wider use of automatic validations or checklists (such as those supported by BioSamples) that the metadata adhere to would enable reviewers and users to identify well-annotated datasets suitable for re-use.

Once well-defined metadata is submitted, it can be used by search engines. For example, plant material and sample identifications, as recommended here, are used as germplasm filters in the FAIDARE search portal, allowing discovery of genotyping and phenotyping data containing the same plant material. Adoption of these guidelines and best practices will help make plant genotyping data FAIR and provide new opportunities to advance our understanding of relationships between genotypic and phenotypic data.

Data availability

No data are associated with this article.

References

Alercia A, Diulgheroff S, Mackay M: *FAO/bioversity multi-crop passport descriptors V. 2.1 [MCPD V. 2.1].* Rome (Italy): Food and Agriculture Organization of the United Nations (FAO); Bioversity International; 2015.

Alercia A, López FM, Sackville Hamilton NR, *et al.*: *Digital Object Identifiers for food crops - Descriptors and guidelines of the Global Information System.* Rome: Food and Agriculture Organization of the United Nations; 2018.

Baker M: **1,500 scientists lift the lid on reproducibility.** *Nature.* 2016; **533**: 452–454.
PubMed Abstract | Publisher Full Text

Bernstein MN, Doan A, Dewey CN: **MetaSRA: normalized human sample-specific metadata for the Sequence Read Archive.** *Bioinformatics.* 2017; **33**: 2914–2923.
Publisher Full Text

Benson DA, Cavanaugh M, Clark K, *et al.*: **GenBank.** *Nucleic Acids Res.* 2013; **41**: D36–D42.
PubMed Abstract | Publisher Full Text

Cezard T, Cunningham F, Hunt SE, *et al.*: **The European Variation Archive: a FAIR resource of genomic variation for all species.** *Nucleic Acids Res.* 2021; **50**: D1216–D1220.
PubMed Abstract | Publisher Full Text

Cochrane G, Karsch-Mizrachi I, Nakamura Y, *et al.*: **The International Nucleotide Sequence Database Collaboration.** *Nucleic Acids Res.* 2011; **39**: D15–D18.
PubMed Abstract | Publisher Full Text

Courtot M, Gupta D, Liyanage I, *et al.*: **BioSamples database: FAIRer samples metadata to accelerate research data management.** *Nucleic Acids Res.* 2022; **50**: D1500–D1507.
PubMed Abstract | Publisher Full Text

Danecek P, Auton A, Abecasis G, *et al.*: **The variant call format and VCFtools.** *Bioinformatics.* 2011; **27**: 2156–2158.
PubMed Abstract | Publisher Full Text

Grüning B, Chilton J, Köster J, *et al.*: **Practical Computational Reproducibility in the Life Sciences.** *Cell Syst.* 2018; **6**: 631–635.
PubMed Abstract | Publisher Full Text

Howe KL, Contreras-Moreira B, De Silva N, *et al.*: **Ensembl Genomes 2020—enabling non-vertebrate genomic research.** *Nucleic Acids Res.* 2020; **48**: D689–D695.
PubMed Abstract | Publisher Full Text

Jonquet C, Toulet A, Arnaud E, *et al.*: **AgroPortal: A vocabulary and ontology repository for agronomy.** *Comput. Electron. Agric.* 2018; **144**: 126–143.
Publisher Full Text

Kuhn M: **A summary of the international standard date and time notation.** 1995. (accessed 9.1.21).
Reference Source

Lappalainen I, Lopez J, Skipper L, *et al.*: **dbVar and DGVa: public archives for genomic structural variation.** *Nucleic Acids Res.* 2013; **41**: D936–D941.
PubMed Abstract | Publisher Full Text

Leinonen R, Akhtar R, Birney E, *et al.*: **The European Nucleotide Archive.** *Nucleic Acids Res.* 2011; **39**: D28–D31.
PubMed Abstract | Publisher Full Text

Mailman MD, Feolo M, Jin Y, *et al.*: **The NCBI dbGaP database of genotypes and phenotypes.** *Nat. Genet.* 2007; **39**: 1181–1186.
PubMed Abstract | Publisher Full Text

Mashima J, Kodama Y, Fujisawa T, *et al.*: **DNA Data Bank of Japan.** *Nucleic Acids Res.* 2017; **45**: D25–D31.
PubMed Abstract | Publisher Full Text

Miyakawa T: **No raw data, no science: another possible source of the reproducibility crisis.** *Mol. Brain.* 2020; **13**: 24.
PubMed Abstract | Publisher Full Text

NCBI: **NCBI Genome Assembly Model.** 2002. (accessed 9.1.21).
Reference Source

NCBI Insights: **NCBI Insights: Phasing out support for non-human genome organism data in dbSNP and dbVar.** *NCBI Insights.* 2017. (accessed 8.31.21).
Reference Source

Papoutsoglou EA, Faria D, Arend D, *et al.*: **Enabling reusability of plant phenomic datasets with MIAPPE 1.1.** *New Phytol.* 2020; **227**: 260–273.
PubMed Abstract | Publisher Full Text

Pereira GS, Garcia AAF, Margarido GRA: **A fully automated pipeline for quantitative genotype calling from next generation sequencing data in autopolyploids.** *BMC Bioinformatics.* 2018; **19**: 398.
PubMed Abstract | Publisher Full Text

Rocca-Serra P, Brandizi M, Maguire E, *et al.*: **ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level.** *Bioinformatics.* 2010; **26**: 2354–2356.
PubMed Abstract | Publisher Full Text

Selby P, Abbeloos R, Backlund JE, *et al.*: **BrAPI—an application programming interface for plant breeding applications.** *Bioinformatics.* 2019; **35**: 4147–4155.
PubMed Abstract | Publisher Full Text

Sherry ST, Ward M-H, Kholodov M, *et al.*: **dbSNP: the NCBI database of genetic variation.** *Nucleic Acids Res.* 2001; **29**: 308–311.
PubMed Abstract | Publisher Full Text | Free Full Text

Song S, Tian D, Li C, *et al.*: **Genome Variation Map: a data repository of genome variations in BIG Data Center.** *Nucleic Acids Res.* 2018; **46**: D944–D949.
PubMed Abstract | Publisher Full Text

Toczydlowski RH, Liggins L, Gaither MR, *et al.*: **Poor data stewardship will hinder global genetic diversity surveillance.** *Proc. Natl. Acad. Sci.* 2021; **118**: e2107934118.
PubMed Abstract | Publisher Full Text

Wilkinson MD, Dumontier M, Aalbersberg IjJ, *et al.*: **The FAIR Guiding Principles for scientific data management and stewardship.** *Sci. Data.* 2016; **3**: 160018.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ? ?

Version 1

Reviewer Report 19 April 2022

https://doi.org/10.5256/f1000research.120539.r128674

?

Micha M. Bayer
Information & Computational Sciences (ICS) group, James Hutton Institute, Dundee, UK

In this opinion article, Beier and co-authors review the current procedures for submitting genotypic variation data to public archives and make a number of recommendations to improve the standardisation of submitted data in order to advance the FAIR principles in this area. They show with specific examples how metadata in VCF files could be improved and suggest additional fields that should be included in VCF-based data submissions.

This is a very useful paper that should help stimulate further discussion and progress in this area. The presentation is systematic, clear, and thorough. I have a number of suggestions for improvement that should be implemented before indexing, but otherwise, I am very happy to recommend indexing. This will be a very useful contribution to the community.

- What are the specific differences between plants and other organisms in this context? What requirements does plant data have that other data doesn't? It would be good to expand on this a little.

- Should a standard such as MIAPPE be established for genotyping experiments/data? This would be a good place to discuss this and perhaps start describing what it could look like.

- The age of pan-genomes is now firmly upon us, and it might be a good idea to add some thoughts on how graph-based genotyping might affect any of the suggestions proposed here - for example, graph-based reference genomes.

- "*Bioinformaticians who have directly performed the genotyping analyses and thus the creation of the VCF files will consider it a comparatively simple task to enter metadata directly into the file*." It would be helpful to specify what tools can be used (and how) to add metadata to VCF headers and how this can be done with minimal risk of getting things wrong (e.g. mislabelling samples). Is manual editing of a header and replacing it with bcftools the best we have, or are there tools out there that can achieve this in a more foolproof/refined fashion?

- It would be beneficial to mention the ##META header lines for defining phenotype metadata that were introduced in the VCF v4.3 revision and to provide some examples of how they could be used. How do they fit into the existing framework of data in EVA and ENSEMBL?

- P.7, section "Sample field format": Please include some detail on what the "registry" entails – is there a standardised way of creating/naming/referring to registries?

- P.9, bottom: Please provide a reference and/or URL for ELIXIR in the text here.

- Figure 2 is an idealised scenario that assumes an institution has both a LIMS and a data steward. Neither is a given – this very much depends on the organisational structure and the levels of funding available to an institution. It would be helpful to have – in addition – an alternative workflow that is more realistic/flexible and describes what can be done when these resources aren't available.

- Figure 2: Replace pronouns like "he" with gender-neutral equivalents like "they" or use the passive voice.

- Figure 2: The header says "*...submission of genotypic data to public databases*" but then the legend only mentions the EVA specifically.

Is the topic of the opinion article discussed accurately in the context of the current literature?
Yes

Are all factual statements correct and adequately supported by citations?
Yes

Are arguments sufficiently supported by evidence from the published literature?
Yes

Are the conclusions drawn balanced and justified on the basis of the presented arguments?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Crop plant bioinformatics, variomics, genomics, transcriptomics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 17 May 2022

Sebastian Beier, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK)

Gatersleben, Seeland, Germany

Dear Micha M. Bayer,
thank you for taking the time to review our article and providing valuable feedback on our project. We would like to address your comments below:

- What are the specific differences between plants and other organisms in this context? What requirements does plant data have that other data doesn't? It would be good to expand on this a little.

There are differences with the precise identification, especially for interoperability and genealogy, as well as the chloroplast genome. We have added some clarification in the introduction part of the manuscript. In contrast to plant species other data domains need a different set of metadata, for example animal sciences in particular need information about the breed and gender of the animal that was genotyped.

- Should a standard such as MIAPPE be established for genotyping experiments/data? This would be a good place to discuss this and perhaps start describing what it could look like.

Such an attempt has indeed been made in the past (Huang et al., 2011, 10.4056/sigs.1994602). However, this was apparently not given much attention by the community, so that no critical mass of users could be generated. MIAPPE itself emphasizes phenotyping and is not currently planning to expand into the genotyping domain. If another attempt is made to introduce a standard for genotyping experiments, MIAPPE can serve as an incubator and provide advice and suggestions, as well as interface with other standards or APIs such as BrAPI.

- The age of pan-genomes is now firmly upon us, and it might be a good idea to add some thoughts on how graph-based genotyping might affect any of the suggestions proposed here - for example, graph-based reference genomes.

Thanks a lot for this question, yes there is indeed a need for a better description about both which reference genome was used for genotyping experiments as well as graph-based genotyping when pan-genomes could be utilized. However, the use of VCF has been discussed in various pan-genome contexts, with the conclusion that VCF is not suitable for structural variants, as it is not appropriate for representing nested or complex variants (Hickey et al., 2020, 10.1186/s13059-020-1941-7). Similar conclusions were reached by Li et al. (2020, 10.1186/s13059-020-02168-z), who found that it is not possible to define coordinates for insertions in VCF, which limits its use for simple variations. In the context of this paper, we believe that this is not the best position to talk about other formats, but to acknowledge the shortcomings of VCF and try to make the format in its current form the most FAIR version it can be. Finally, it should be noted that in future versions of VCF (v4.4 and later) there will be efforts to overcome these shortcomings of the specifications to better represent structural variants in all their complexity (see the following pull requests on github: https://github.com/samtools/hts-specs/pull/465, https://github.com/samtools/hts-specs/pull/553).

- "*Bioinformaticians who have directly performed the genotyping analyses and thus the creation of the VCF files will consider it a comparatively simple task to enter metadata directly into the file*." It would be helpful to specify what tools can be used (and how) to add metadata to VCF headers and how this can be done with minimal risk of getting things wrong (e.g. mislabelling samples). Is manual editing of a header and replacing it with bcftools the best we have, or are there tools out there that can achieve this in a

more foolproof/refined fashion?

To our knowledge, there is no tool that could do this. One possibility would be that read mapping tools are able to provide metadata on samples and write this data directly into the mapping files.To ensure that the metadata is not lost in the downstream analysis and calculation of all subsequent tools, these metadata fields would need to be known and would not be allowed to be overwritten/edited. In hindsight, this may be too ambitious to implement at this stage and it may make more sense to develop a tool capable of adding this metadata to the final VCF file before submission to public archives. We would like to highlight the role of VCF as a data exchange format in particular, this means that VCFs are in general the result of an analysis pipeline or exported resultset from a database (such as EMBL EVA). Manual maintenance of exchange formats, like JSON or XML-based formats can certainly not be excluded, but should rather be the exception. From our point of view this would mean VCF-generating pipelines should implement necessary components for the user-friendly, correct registration of metadata. We have added a paragraph in the text highlighting the need for appropriate supporting tools or APIs for the validation of VCFs and in particular the VCF metadata.

- It would be beneficial to mention the ##META header lines for defining phenotype metadata that were introduced in the VCF v4.3 revision and to provide some examples of how they could be used. How do they fit into the existing framework of data in EVA and ENSEMBL?

The ##META meta-information of the VCF 4.3 is used to define sample descriptors and not phenotypes in the sense of MIAPPE. They might be used for some of the elements listed in BioSamples checklists, such as the MIAPPE list used by the BioSamples Validator. But our approach was to include only the minimal identification metadata in the VCF (DOIs, ID lists) and rely on BioSamples for a detailed description of the sample. Regarding the ##META field: There is currently a discussion in the VCF community to remove this field completely from the specification (https://github.com/samtools/hts-specs/issues/558#issuecomment-829421610), and basically the trend is more moving towards outsourcing metadata to FAIR archives (like BioSamples).

- P.7, section “Sample field format”: Please include some detail on what the “registry” entails – is there a standardised way of creating/naming/referring to registries?

Thanks, we clarified this point in the paper.

- P.9, bottom: Please provide a reference and/or URL for ELIXIR in the text here.

We added a recent publication about ELIXIR to the manuscript.

- Figure 2 is an idealised scenario that assumes an institution has both a LIMS and a data steward. Neither is a given – this very much depends on the organisational structure and the levels of funding available to an institution. It would be helpful to have – in addition – an alternative workflow that is more realistic/flexible and describes what can be done when these resources aren’t available.

Genotyping of large panels of genotypes or pan-genome projects depend on a well-structured implementation of the necessary data handling processes according to the research data life cycle to allow quality and efficiency of resources as well as to ensure the long-term assured re-usability of the data. The use of process-oriented database-based solutions starting with material and sample description, the assignment of PUIDs for sequences and material, the machine-readable data processing in sequence laboratories and the ingestion into file storage systems, their registration in in-house databases and

their final publication in central repositories, such as EMBL-EBI EVA, is an important task for medium to large institutions. The definition of clear specifications for data exchange formats and APIs provides the necessary framework to allow sufficient freedom for technologies. These can range from open source to commercial solutions that are sufficiently available for the process described. A best practice paper has been referenced accordingly in the manuscript: https://doi.org/10.1093/bib/bbab010. We see our contribution to this discussion as a standardised workflow that can occur in many different variations. For example, it is indeed a problem that institutions do not have dedicated data stewards or an institution-wide LIMS, so that these sub-aspects either have to be outsourced or interim solutions have been created. In the context of this manuscript, we cannot and do not want to refer to all the possibilities for implementing good scientific practice, but rather provide a stylised best practice guide to be able to map such a workflow to one's own organisation.

- Figure 2: Replace pronouns like “he” with gender-neutral equivalents like “they” or use the passive voice.

We changed this to gender-neutral equivalents within the manuscript.

- Figure 2: The header says “*...submission of genotypic data to public databases*” but then the legend only mentions the EVA specifically.

We have changed the text in the figure header accordingly to make it clear that this workflow is intended for use with EMBL-EBI repositories.

Competing Interests: No competing interests were disclosed.

Reviewer Report 02 March 2022

https://doi.org/10.5256/f1000research.120539.r125389

?

Boas Pucker
Institute of Plant Biology & BRICS, TU Braunschweig, Braunschweig, Germany

Alenka Hafner
Intercollege Graduate Degree Program in Plant Biology, Penn State University, University Park, PA, USA

Beier *et al.* present their recommendations for the inclusion of metadata in VCF files. A defined structure is proposed to make VCF files more suitable for re-use. Suggested fields include file date, bioinformatics source, reference URL, contig, and sample. This opinion paper adds a novel perspective to the area. It is important for advancing FAIR and reproducibility in general if individual formats are analysed in this manner. The suggested metadata standards for VCF seem appropriate for users, with some clarification/additions (see comments). The paper is well and clearly written. The figures help the narrative and provide a quick overview.

Specific comments:

1. FAIR principles are not directly discussed in the text despite the abstract/key word emphasis. The term is only mentioned twice in the manuscript. The recommendations presented in this paper would improve many aspects of FAIR in VCF but this is not explicitly discussed. There is mention of individual components but a more direct discussion would be beneficial.

2. Why is gVCF excluded? Are there reasons that prevent the direct transfer of this VCF standard to gVCF?

3. The introduction mentions a few variant calling databases. What about 1001 genomes GMI-MPI vcf (https://1001genomes.org/data-center.html)? Several variant datasets for *Arabidopsis thaliana* are also hosted on https://jbrowse.arabidopsis.org/. It is only for *Arabidopsis* but considering the importance of this model it should be mentioned that genomic variation data is often hosted on organism-specific databases (which is in itself a problem of format unification). GVM, for example, does not host Arabidopsis VCFs presumably because they are found in their own database.

4. The introduction assumes that the reference genome sequence is supported by Ensembl, but what to do if it is not available?

5. Considering that phenotype standards are the only provided example, it does not seem like the most useful comparison. The data type/collection/prevalence etc. is very different to VCF. More examples would be needed to emphasize shared elements of various standards (for example, SAM/BAM flags).

6. One statement about VCFs should be checked: "one or more data lines". It does not make much sense to share such a file, but the VCF file could be empty i.e. no lines with data.

7. "well-formed" could be explained in more detail.

8. It might be better to use bioinformatics_source to include a complete description of the data processing in the VCF file. Otherwise, there is the risk that data sets cannot be re-used due to broken links. Some data sets might be the result of unpublished protocols hence it will be impossible to link to a publication. For example, the data set and the workflow might be part of the same paper.

9. RefSeq instead of GenBank might be preferentially entered by some groups so this needs to be screened/specified. Again, what happens if the reference genome sequence is not publicly available (yet)? Fig. 2 nicely points out that we must also consider a not-yet-published reference used by a lab to adhere to this standard by registering their reference. It would be good to mention this in the description of the respective field as well.

10. Some assumptions underlying Fig.2 might be too optimistic. The actual sequencing of many projects is conducted by external sequencing providers. Many institutions also lack a LIMS and a "Data Steward". Maybe it would be better to adjust the process in Fig.2 or display an alternative scenario. What would happen if, say, an inexperienced PhD student needs to

handle the entire process?

11. "genome" should be replaced by "genome sequence" at several places in manuscript. This is about different assembly version of the same accession so the genome remains the same, but the quality of the representation improves.

12. It would be good to mentioned other examples of VCF containing databases besides EVA and GVM. Even though they are the main ones, one of the biggest issues with metainformation is that you have “straggler” databases in niche fields/organisms that do not adhere to the format.

13. With respect to the tools automatically changing information about variant positions: Can this be monitored in any way? It is not reasonable to expect, e.g. EVA to check this and presumably it is databases that would need to enforce these standards. Would a list of common programs that change this help to avoid issues?

14. Additional meta-information fields:

    a) "findability and interoperability": To reiterate a comment above - reusability (at least, if not also accessibility) would be enhanced by these metadata standards. The impact/implications to VCF FAIR would fit the narrative in here and in the conclusion.

    b) "metainformation lines”: Indeed, and some some suggestions, e.g. from the barley example are needed.

15. "genotypic and other data from plants": What are these other data? Please specify if this would be phenotypic.

16. "Analytical results should be linked out to the respective sample(s) and defined in the context of the study being reported." ... this could be a bit more specific.

17. It does not become clear how the paragraph about BioSample database at EMBL-EBI contributes to the conclusion. Currently, it is only about phenotypic data and a connection to VCF would be helpful.

18. Re-use is a very important aspect of missing/insufficient metadata consequences. It could be emphasised more as a major benefit of adopting this type of VCF standardisation.

19. More specific portals for adoption would be beneficial to mention in the conclusion - should all databases adopt these guidelines, should journals make sure the data referenced in the paper adheres to them, etc.? Enforcement is a major hurdle of adopting metadata (and data) standards that adhere to FAIR (see our paper, Sielemann, Hafner & Pucker, 2020[1]).

20. The authors focus on plant VCF and EVA. However, considering these standards are sorely needed in other areas/databases and that VCF formatting should be fairly universal, the authors might want to consider a broader application of these standards. This could be suggested and explored more in the conclusion.

21. It could be helpful to include a sample of the improved VCF as supplementary file.

Minor comments:

1. The link to one reference is not correctly formatted: "NCBI Insights, 2017".

2. 'Each contig contains at least the attribute ID, and typically also include length' > "Each contig entry contains at least the attribute ID, and typically also includes length..."

References

1. Sielemann K, Hafner A, Pucker B: The reuse of public datasets in the life sciences: potential risks and rewards.*PeerJ*. 2020; **8**: e9954 PubMed Abstract | Publisher Full Text

Is the topic of the opinion article discussed accurately in the context of the current literature?

Yes

Are all factual statements correct and adequately supported by citations?

Yes

Are arguments sufficiently supported by evidence from the published literature?

Yes

Are the conclusions drawn balanced and justified on the basis of the presented arguments?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Plant genomics, plant specialized metabolism

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Author Response 17 May 2022

Sebastian Beier, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben, Seeland, Germany

Dear Boas Pucker and Alenka Hafner,
thank you for taking the time to review our article and providing valuable feedback on our project. We would like to address your comments below:

1. FAIR principles are not directly discussed in the text despite the abstract/key word emphasis. The term is only mentioned twice in the manuscript. The recommendations presented in this paper would improve many aspects of FAIR in VCF but this is not explicitly discussed. There is mention of individual components but a more direct discussion would be beneficial.

We added explicit FAIR principles discussion in the discussion to clearly state the improved Findability, Reusability and Interoperability.

2. Why is gVCF excluded? Are there reasons that prevent the direct transfer of this VCF standard to gVCF?

We added some remarks, showing these recommendations can be directly transferred to gVCF. There is nothing that would exclude the use of our recommendations with gVCF.

3. The introduction mentions a few variant calling databases. What about 1001 genomes GMI-MPI vcf (https://1001genomes.org/data-center.html)? Several variant datasets for *Arabidopsis thaliana* are also hosted on https://jbrowse.arabidopsis.org/. It is only for *Arabidopsis* but considering the importance of this model it should be mentioned that genomic variation data is often hosted on organism-specific databases (which is in itself a problem of format unification). GVM, for example, does not host Arabidopsis VCFs presumably because they are found in their own database.

We added some remarks regarding this in the manuscript.

4. The introduction assumes that the reference genome sequence is supported by Ensembl, but what to do if it is not available?

In the case where the reference genome sequence is not yet in Ensembl the VCF cannot be displayed (within Ensembl). However in such cases Ensembl would try to prioritize adding the missing genome if it is in the public archives. This should not be confused with EVAs requirement of needing a genome assembly supported by INSDC.

5. Considering that phenotype standards are the only provided example, it does not seem like the most useful comparison. The data type/collection/prevalence etc. is very different to VCF. More examples would be needed to emphasize shared elements of various standards (for example, SAM/BAM flags).

It is true that MIAPPE, and thus the phenotyping standard, was not intended to characterize genotyping analyses or VCF files. It has been used to enable interoperability with genotyping dataset through shared objects and IDs. It is also true that in order to create a VCF file, there must be one or more mapping files (SAM/BAM) that record the differences and similarities between the genotypes under study and the reference genome assembly. Since there are a large number of different workflows to create a VCF file, our intention was to make the final result FAIR and not necessarily cover all intermediate steps with our suggestions. What all genotyping experiments have in common is that samples are obtained from physical material that is currently poorly described. It is precisely this point that is better addressed and makes it possible to link different experiments and analyses based on the material used with the recommendations presented here. MIAPPE offers a well-designed framework for this and is also ideally suited for the plant domain.

6. One statement about VCFs should be checked: "one or more data lines". It does not make

much sense to share such a file, but the VCF file could be empty i.e. no lines with data.

We agree that sharing such a file would not make much sense, still it would be a perfectly valid VCF file according to the specifications. This wording is taken directly from the specifications.

7. "well-formed" could be explained in more detail.

We have adjusted the text to make this clearer.

8. It might be better to use bioinformatics_source to include a complete description of the data processing in the VCF file. Otherwise, there is the risk that data sets cannot be re-used due to broken links. Some data sets might be the result of unpublished protocols hence it will be impossible to link to a publication. For example, the data set and the workflow might be part of the same paper.

This is a valid concern, but we cannot see that embedded documentation will ensure a sustainable, comprehensive and FAIR documentation of a workflow that enables a re-processing in 10 years. For example, used software needs to be referred to using PUIDs or links. Here, we could also face broken links. If mentioned software or scripts cannot be resolved anymore, VCF embedded documentation could become ambiguous too. To get around this, a container (RO-Crate, Docker, Singularity, etc.) is one of the only solutions that can remedy this and make both the workflow and the data accessible to users. It should be noted here, however, that there are some scenarios where the disclosure of all programmes, scripts or data is not wanted or legally possible. However, until containers are common practice, explaining the workflow in as much detail as possible is a good start. There are some approaches that map this in a more structured way, such as Common Workflow Language (CWL, https://doi.org/10.1038/sdata.2018.118). In general, however, it should be said that here the workflows can also be referenced as DOI, which is fully compliant with our recommendations.

9. RefSeq instead of GenBank might be preferentially entered by some groups so this needs to be screened/specified. Again, what happens if the reference genome sequence is not publicly available (yet)? Fig. 2 nicely points out that we must also consider a not-yet-published reference used by a lab to adhere to this standard by registering their reference. It would be good to mention this in the description of the respective field as well.

The complete described pipeline builds upon deposition of genotyping information at EVA, which needs a published genome assembly accession number. That is one of the reasons Fig. 2 also tries to highlight that the genome assembly needs to be deposited and an accession number received before being able to proceed to the next step. We updated the text to make this more clear. In addition, the trend in genome sequencing and assembly is for both the read data and the assembly sequence to be published early, making it less likely in the future that the genome sequence will not be publicly available.

10. Some assumptions underlying Fig.2 might be too optimistic. The actual sequencing of many projects is conducted by external sequencing providers. Many institutions also lack a

LIMS and a "Data Steward". Maybe it would be better to adjust the process in Fig.2 or display an alternative scenario. What would happen if, say, an inexperienced PhD student needs to handle the entire process?

Please have a look at our response to Micha Bayer's comment referring to Figure 2. As additional clarification and help for more inexperienced scientists we have published a stepwise guide on how to submit data using the recommendations in this manuscript in the FAIR Cookbook under recipe https://w3id.org/faircookbook/FCB061.

11. "genome" should be replaced by "genome sequence" at several places in manuscript. This is about different assembly version of the same accession so the genome remains the same, but the quality of the representation improves.

Indeed, a good catch. We updated the manuscript to be more precise.

12. It would be good to mentioned other examples of VCF containing databases besides EVA and GVM. Even though they are the main ones, one of the biggest issues with metainformation is that you have "straggler" databases in niche fields/organisms that do not adhere to the format.

In this paper, we focused on the use case of EMBL-EBI submissions and their use of VCF in collaboration with BioSamples and EVA. We are not aiming at an extensive review of all possible VCF submissions and publishers. We added some remarks in the discussion regarding adoption of the metadata recommendations in the broader community.

13. With respect to the tools automatically changing information about variant positions: Can this be monitored in any way? It is not reasonable to expect, e.g. EVA to check this and presumably it is databases that would need to enforce these standards. Would a list of common programs that change this help to avoid issues?

We are not aware of any programs that change the position or other characteristics of variants, but programs such as PLINK are known to define the variant within a genotyping study with the major allele as the reference and the minor allele as the alternative, regardless of the base present in the reference genome assembly. However, PLINK has an option to use the reference allele, and EVA often asks to correct this at the time of submission. EVA also consistently validates each VCF file that is submitted to the archive. Validation compares the reference allele to the reference genome sequence, most likely detecting misplaced variants. It is technically possible for misplaced variants to still match the reference, but in practice this method will detect most of these errors. A list of common programs that work in a similar way to PLINK would be very welcome. However, we do not believe that this would completely avoid this problem, since such a list is not easy to maintain (based on new versions, programs, workflows) and could lull the user into a false sense of security if other programs were used. We would therefore rather urge caution and diligence when using programs that interact with VCF files.

14. Additional meta-information fields:

a) "findability and interoperability": To reiterate a comment above - reusability (at least, if not also accessibility) would be enhanced by these metadata standards. The impact/implications to VCF FAIR would fit the narrative in here and in the conclusion.
b) "metainformation lines": Indeed, and some some suggestions, e.g. from the barley example are needed.

a) We added reusability in the listing and expanded on the FAIRness of data in the discussion part. b) Information, such as further metadata about the person who analysed the data or collected the plant material and the prevailing environmental conditions (e.g. contact information, tools used, temperature or humidity), would be examples of additional metainformation fields that were not explicitly recommended by us, but in some scenarios make sense to include additionally.

15. "genotypic and other data from plants": What are these other data? Please specify if this would be phenotypic.

Since the sample description would be uniform, this could be applied to all kinds of different data. This includes phenotypic data but also other -omics layers, like transcriptomic data, metabolomic data and so forth.

16. "Analytical results should be linked out to the respective sample(s) and defined in the context of the study being reported." ... this could be a bit more specific.

We have updated the text to be more precise.

17. It does not become clear how the paragraph about BioSample database at EMBL-EBI contributes to the conclusion. Currently, it is only about phenotypic data and a connection to VCF would be helpful.

We have updated the wording to clarify that the plant metadata stored at BioSamples does not necessarily have to be of phenotypic origin.

18. Re-use is a very important aspect of missing/insufficient metadata consequences. It could be emphasised more as a major benefit of adopting this type of VCF standardisation.

We agree with this statement and have added this throughout the manuscript.

19. More specific portals for adoption would be beneficial to mention in the conclusion - should all databases adopt these guidelines, should journals make sure the data referenced in the paper adheres to them, etc.? Enforcement is a major hurdle of adopting metadata (and data) standards that adhere to FAIR (see our paper, Sielemann, Hafner & Pucker, 2020).

We agree with the concerns raised in Sielemann, Hafner & Pucker, 2020, and one of the solutions they offer is improvement of metadata standard which is what this paper tries to accomplish.
Enforcing these new metadata standards on many underfunded databases is unlikely. Adoption of new standards is always tricky and takes time but we believe the best way for

that is to enforce it in a central point like upon submission to the archives. Alternatively, journals could be approached to enforce compliance with the recommendations. However, one obstacle would be convincing not just one institution but many different publishers to do so, which is very time-consuming. In our view, critical mass is more likely to be achieved through community engagement via larger organisations (such as DivSeek, AgBioData, Australian BioCommons, etc.). We have started to do this and there is overall positive feedback and early signs of acceptance within these organisations.

20. The authors focus on plant VCF and EVA. However, considering these standards are sorely needed in other areas/databases and that VCF formatting should be fairly universal, the authors might want to consider a broader application of these standards. This could be suggested and explored more in the conclusion.

Indeed, we are very interested in bringing these ideas and recommendations to other groups. In particular we have been in discussions with scientists in the animal field to work on an adaptation to their field.

21. It could be helpful to include a sample of the improved VCF as supplementary file.

We have uploaded an example of this workflow to EVA and highlighted this in the manuscript. The VCF has been validated and accessioned successfully by EVA. The Study is now available on the EVA website at https://www.ebi.ac.uk/eva/?eva-study=PRJEB51851.

Minor comments:
1. The link to one reference is not correctly formatted: "NCBI Insights, 2017".

We have changed the reference to the EVA paper published last year where this was stated.

2. 'Each contig contains at least the attribute ID, and typically also include length' > "Each contig entry contains at least the attribute ID, and typically also includes length..."

Changed this accordingly in the manuscript.

Competing Interests: No competing interests were disclosed.