FAIR data stewardship and tool development in X-omics

X-omics strives to adhere to the FAIR principles for data stewardship [doi:10.1038/sdata.2016.18], uses the FAIRmetrics [www.github.com/FAIRmetrics and doi:10.1038/sdata.2018.118] to test the level of FAIRness of their tools and resources, and translates the FAIR principles in the following guidelines and activities:

F(indable):

1. X-omics data are associated with rich, ontology-supported and human and machine understandable metadata. Together with FAIRgenomes, we have worked on a semantic model for the metadata of genomics experiments (study, consent, personal data, clinical data, material, sample preparation, sequencing and analysis methods) [https://github.com/fairgenomes/fairgenomes-semantic-model]. This model will be extended to other types of -omics and X-omics experiments and aligned to the ontologies defined by standard initiatives from the different -omics domains (see I.1) and used by the different repositories (see F.3).
2. Metadata schemes are shared via bioschemas.org[https://bioschemas.org and fairsharing.org [https://fairsharing.org/].
3. X-omics data are submitted to the relevant repositories at EBI (ENA/EGA [https://www.ebi.ac.uk/ena/submit/sra/] for genomics; PRIDE [https://www.ebi.ac.uk/pride/] for proteomics and Metabolights [https://www.ebi.ac.uk/metabolights/] for metabolomics. The submission process is facilitated by the data generating facilities in X-omics, by providing data and associated metadata in a format such as ISA-TAB that can readily be submitted to the repositories.

A(ccessible):

1. X-omics data are made accessible to the scientific community at the time of first publication (at the latest, can be earlier). Depending on the sensitivity of the data, this can be fully open or under controlled access. The preferred repositories to guarantee long term availability of the data are the EBI repositories indicated above. EGA (personal genome data) is currently the only EBI repository that allows for controlled access, but EBI is working with us on controlled access of proteomics and metabolomics data
2. X-omics data and analysis tools can be made available for analysis in the collaborative, cloud analysis environments available to the X-omics community (anDREa [https://www.andrea-cloud.eu/] and Surf Research cloud [https://www.surf.nl/en/surf-research-cloud-collaboration-portal-for-research]

I(nteroperability)

1. X-omics uses community standards for data formats (proteomics: [https://www.psidev.info/specifications); metabolomics: [https://github.com/MSI-Metabolomics-Standards-Initiative/CIMR]; genomics: [https://www.ga4gh.org/genomic-data-toolkit/]. A table with standard formats is given below:

Omics level	Raw data format	Processed data format
Genomics (including methylation and transcriptomics data)	Fastq	BAM, CRAM, VCF, gVCF, GFF
Proteomics	mzML	mzIdentML, mzQuantML, mzTab
Metabolomics	mzML	mzTab-M, m-file

2. X-omics metadata are mapped to existing ontologies, where possible. Examples of X-omics preferred ontologies are: Semantic science Integrated Ontology (SIO), Data Use Ontology (DUO), Experimental Factor Ontology (EFO), Human Phenotype Ontology, MIABIS, NCI Thesaurus, and EDAM.

3. For data interoperability, X-omics develops the FAIR data cube, a FAIR data object that contains a FAIR data point (architecture described here: [https://fairdatapoint.readthedocs.io/en/latest/]) with linked meta-data (in Resource Description Framework [https://www.w3.org/RDF/]) about a X-omics experiment, defines the access conditions and provides an API to query the data, and the data itself or pointers to the data (in the standard formats mentioned in I.1). The formats of the actual -omics data in the FAIR data cube is maintained in the established data formats described above.

R(eusability)

1. X-omics data, software and other resources come with a license. Software/tools/workflows developed within the context of X-omics comes with one of the open source licenses listed here: [https://opensource.org/licenses]

2. X-omics software/tools/workflows tools and workflows are version-controlled, citable (URL or DOI), preferably containerized for easy deployment on diverse infrastructure, and packaged with example data.

3. X-omics provides transparency on and gives credits to the data producers and processors. They are preferably listed with their ORCID in the metadata accompanying a X-omics experiment.