INFRAVEC2 OPEN RESEARCH DATA MANAGEMENT PLAN

Authors: Andrea Crisanti, Gareth Maslen, Andy Yates, Paul Kersey, Alain Kohl, Clelia Supparo, Ken Vernick
Date: 10th July 2020
Version: 3.0

Overview

Infravec2 will align to Open Research Data, as follows:

Data Types and Standards

Infravec2 will generate a variety of data types, including molecular data types: genome sequence and assembly, structural annotation (gene models, repeats, other functional regions) and functional annotation (protein function assignment), variation data, and transcriptome data; arbovirus and malaria experimental infection data, linked to archived samples; and microbiome data (Operational Taxonomic Units), including natural virome composition.

All data will be released according to the appropriate standards and formats for each data type. For example, DNA sequence will be released in FASTA format; variant calls in Variant Call Format; sequence alignments in BAM (Binary Alignment Map) and CRAM (Compressed Read Alignment Map) formats, etc. We will strongly encourage the organisation of linked data sets as Track Hubs, a mechanism for publishing a set of linked genomic data that aids data discovery, sharing, and selection for subsequent analysis. We will develop internal standards within the consortium to define minimal metadata that will accompany all data sets, following the template of the Minimal Information Standards for Biological and Biomedical Investigations (http://www.dcc.ac.uk/resources/metadata-standards/mibbi-minimum-information-biological-and-biomedical-investigations).

Data Exploitation, Accessibility, Curation and Preservation

All molecular data for which existing public data repositories exist will be submitted to such repositories on or before the publication of written manuscripts, with early release of data (i.e. as soon as quality control is completed) the default option in the absence of compelling arguments otherwise. Suitable repositories include the databases of the International Nucleotide Sequence Database Collaboration (INSDC - composed of the European Nucleotide Archive, GenBank, and the DNA Database of Japan) for nucleotide sequences, the European Variation Archive for data on polymorphism, ArrayExpress for gene expression data, etc. Data for which no appropriate repository exists will be deposited in institutional or other public, persistent repositories for unstructured data (e.g. figshare), identified using Document Object Identifiers (DOIs), and advertised via the Infravec2 web portal. All repositories used will conform to the FAIR principles (https://www.force11.org/group/fairgroup/fairprinciples): data will be Findable, Accessible, Identifiable and Re-usable.

This project has received funding from the European Union's Horizon 2020 research innovation programme under grant agreement No 731060.

We will also release an integrated and interpreted view on the assembly and sequence annotation generated through Ensembl Metazoa (http://metazoa.ensembl.org). Ensembl Metazoa is a leading global resource for access to genome-scale data from insect species. In Ensembl Metazoa, data generated in the as part of Infravec2 will be visible in the context of reference genome assemblies, data generated from outside the project, and relevant additional data from comparator (vector and non-vector) species. Access will be provided through interactive (web browser), programmatic (Perl and RESTful Application Programming Interfaces) and data mining (BioMart) interfaces. Through Ensembl Metazoa, data will be shared with VectorBase, a leading international resource focusing on vector data.

These detailed project requirements for data standards and data publication have been set out in a specific document, which has been incorporated in this data management plan as Appendix 1.

Open Source Software Used and Developed by the Project

Ensembl Metazoa uses the Ensembl software platform for the storage, analysis and dissemination of genomic data. Ensembl is fully open source (available under the Apache software licence, version 2.0) and the code base can be accessed via Github (https://github.com/Ensembl). Ensembl will be extended as necessary to meet any specific demands required by the Infravec2 project and the resulting code deposited in the Ensembl Github.

Genome assembly and annotation will be carried out using standard open-source Bioinformatics tools that will be selected according to the state-of-the-art at the time of annotation. There is a strong culture of open source tool development in bioinformatics and we are committed to the use of open source software wherever possible, to maintain an open audit chain for the analyses that have been performed.

Knowledge management and protection

Project-generated data

A major thrust of Infravec2 is the development of a robust data infrastructure to support researchers working on vector biology and vector control; a strong commitment to open data is integral to the proposal. The mechanisms we will employ to maximise the application of the data generated by making it available to the research community have been described above. For these data to be useful, however, high quality metadata is essential: the data needs to be annotated, organised, and made accessible through appropriate interpretative user interfaces. To ensure that data is identified and appropriately annotated, we will implement Networking activities to harmonise community data standards for quality, and conformance with open data (WP4 Task 5). There are two aspects to this task: the development, endorsement and adoption of standards, and the management, quality control and release of project data. Dedicated project resources have been assigned to develop standards for use within the project, identify new data sets as they become available, coordinate data producers, quality control i.e. validating data

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731060.

according to the standards agreed, and ensuring timely data submission into the appropriate public archives. TNA requests for units directly generating genomic or sequence data will be provided on condition that all data generated as a result of the access is published within 18 months of the completion of the access, or within 6 months of the termination of Infravec2, whichever is the sooner.

Sequence data deposition

Infravec2 has established at the European Nucleotide Archive (ENA) an 'umbrella' project record that allows it to link all of the sequencing records generated by project TNA requests back to the Infravec2 project (PRJEB32819). This is desirable both for the Infravec2 project (it allows funding bodies visibility of the results that have been generated), and also aids researchers in locating Infravec2 data (all Infravec2 sequence data can be located under a single master accession for the umbrella project). ENA/Genbank database records are however owned by whoever submits them. This means researchers need to decide whether they wish to submit and curate their own sequence data, or whether they would prefer Infravec2 to handle this process for themselves. Individual Infravec2 users can submit and maintain their own sequence records if they would like, but we request that this option is selected at the start of the TNA request so that we know the researcher has taken on the responsibility for depositing the data. Infravec2 will also request details of the subsequent sequence submission (ENA/Genbank accession), and that permission is granted for Infravec2 to link this project record to the Infravec2 umbrella study.

Project-generated publications

The Infravec2 consortium and the Coordinator, aided by the Institute Pasteur Grants Office, will assist and monitor the visibility of research output and publication of results in peer-reviewed journals. The Infravec2 Project Manager (PM) will inform and emphasize, throughout the project lifetime, the importance of Open Access. As a first step, the PM, with the help of the Media Library of Institute Pasteur, will present during the kick off meeting the Open Access principle under H2020 to ensure that all partners are aware of the practice of providing on-line access to scientific information that is free of charge to the end-user. The project manager will explain opportunities to make research articles available in open access:

- Open access publishing (‘gold’ open access): article is immediately provided in open access mode by the scientific publisher.
- Self-archiving ('green' open access): the published article or final peer-reviewed manuscript is archived by the researcher in an online repository before, after or alongside publication. Publications will also be linked in the project web site. Institute Pasteur will present to project partners the possibilities to publish in green Open access, in conjunction with the Open access infrastructure in Europe (OpenAir).

An open access good practices charter will be presented and will be signed by all partners to guarantee that each scientist in the Infravec2 consortium is aware of the Open Access principle and commits to respect this obligation. Moreover, a section of all periodic reports and the final

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 731060.

report will be dedicated to describing measures to provide open access, and enforcement of our open access strategy.

Policies for Data Sharing and Data Access

Infravec2 supports the principle of open research data, and intends that data generated by the project is widely available for use by the research community and working to establish community-wide standards for data representation and publication. This will include data generated by the project beneficiaries to contribute as "data infrastructure" to overall capacity developed by the project for the genetic control of mosquitoes and other vectors. Data will be publicly released as soon as it is capable of making a meaningful contribution to the overall infrastructure, using the formats and repositories described below.

These policies will also apply to data generated by any separate research projects which are granted access to the Infravec2 infrastructure following application through the product catalogue that will appear on the project website from August 2017. We require users of the infrastructure to agree to publish any data with which they are provided in accordance with these policies, within 18 months of the completion of the use; these conditions are specified in the Infravec2 common Material Transfer Agreement, that specifies the terms under which use of the infrastructure is allowed. This policy provides a reasonable trade-off, allowing users the first opportunity to interpret the outputs of their work, while ensuring that the public funding of the facilities used results in their availability to contribute to the public good hereafter.

Infravec2 will advise and assist its users in the formatting, annotation and publication of their data according these standards; nonetheless, responsibility for these tasks is accepted by the user as a condition of access of the infrastructure.

In short, Infravec2 requires (of both the project beneficiaries, and external users of its facilities):

- Publication of all appropriate data to a persistent, relevant repository. Appropriate data includes any data used to support a publication; but in addition, certain types of large-scale molecular data ("omics" data) whose re-use value is well proven (e.g. DNA sequence, RNA sequence, protein sequence, etc.).
- Where well-established repositories exist for a given data type (e.g. European Nucleotide Archive, GenBank and the DNA Database of Japan for nucleic acid sequence data), data should be submitted to such a repository. For data types for which no well-established dedicated repository exists, other types of repositories may be appropriate (for example, institutional repositories, publishers' repositories, or generic data storage infrastructures such as EUDAT or figshare). Repositories should assign unique identifiers by which data records can be identified (such as Digital Object Identifiers (DOIs)), and should be selected according to criteria including longevity (of repository), the absence of barriers to downstream data access (e.g. through charging or licensing restrictions), and familiarity (will the intended user community expect to find this type of data in this repository?).
- Where there is a well-established standard for the experimental meta-data (i.e. information about the experiment, the data producer/owner/publisher, or the results set

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731060.

itself) – e.g. the MIAME (Minimal Information About a Microarray Experiment) standard for microarray data, or other standards conforming to the Minimal Information Standards for Biological and Biomedical Investigations (http://www.dcc.ac.uk/resources/metadata-standards/mibbi-minimum-information-biological-and-biomedical-investigations) – published data is expected to conform.

- Where there are no established domain-specific standards for minimal meta data, or repository-specific requirements, users should describe their data with a minimal set of metadata in accordance with the standards of Dublin Core Metadata Element Set, Version 1.1 (http://dublincore.org/documents/dces/).

Data Types, Formats and Ontologies

Data types expected to be generated in the project include genome sequence and assembly, structural annotation (gene models, repeats, other functional regions) and functional annotation (protein function assignment), variation data, transcriptome data, proteomic and metabolomic and sampling, arbovirus and malaria experimental infection data, linked to archived samples; and microbiome data (Operational Taxonomic Units), including natural virome composition.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731060.

Data type	Appropriate Format(s)	Appropriate Repository	Comments
Nucleotide sequence data (short reads)	FASTQ, BAM, CRAM, and other machine-specific formats accepted by the ENA (see here http://www.ebi.ac.uk/ena/submit/read-file-formats for more detail). Meta-data should be MIxS-compliant (http://gensc.org/mixs).	European Nucleotide Archive (ENA) (http://www.ebi.ac.uk/ena); ArrayExpress (http://www.ebi.ac.uk/arrayexpress) (specifically, for RNA-seq data generated for the purposes of quantification).	Submission to a partner database of ENA (GenBank or DDBJ) is also acceptable.
Nucleotide sequence (long reads, annotated assembled sequences)	EMBL format (for sequence/annotation), AGP (for assembly description). Meta-data should be MIxS-compliant (http://gensc.org/mixs).	European Nucleotide Archive	Annotation can also be submitted as Tracks to the Ensembl Track Hub Registry provided the underlying sequence is submitted to ENA.
Annotation	• BED • Bed Graph • GFF2/GTF • GFF3 • Pairwise interactions (WashU) • PSL • WIG • BAM/CRAM • BigBed • BigWig • VCF See http://www.ensembl.org/info/website/upload/large.html#vcf-format for more details. Tracks should be packaged as Track hubs (https://genome.ucsc/goldenpath/help/hgTrackHubHelp.html).	Track Hub Registry (http://trackhubregistry.org), for subsequent incorporation in Ensembl Metazoa, VectorBase and other genome browsers.	Much annotation can be visualized as positions or spans on a genomic reference sequence (tracks). Track hubs are collections of tracks with common metadata.
Structural Annotation	Sequence features and their attributes should be described using the Sequence Ontology (http://www.sequenceontology.org/) within GFF2 or GFF3 files	Ensembl Metazoa can accept GFF based structural annotation for submissions that have a public ENA entry for the assembly.	
Functional Annotation	Depends entirely on data under annotation; use of structured controlled	None that directly take submissions.	Contact Infravec2

This project has received funding from the European Union's Horizon 2020 research innovation programme under grant agreement No 731060.

	vocabularies (such as the Gene Ontology, http://www.geneontology.org/) appropriate to the domain is recommended. GAF format (http://www.geneontology.org/page/go-annotation-file-format-20) may be appropriate for annotating other biological objects.		data management team for advice.
Variation data	Variant Call Format (VCF) v4.1 or higher (https://samtools.github.io/hts-specs/)	European Variation Archive (https://www.ebi.ac.uk/eva/)	
Microarray data	Meta-data should be MIAME-compliant (http://fged.org/projects/miame/).	ArrayExpress (http://www.ebi.ac.uk/arrayexpress)	MIAME-compliance is enforced by ArrayExpress submission interface
Proteomic data	Meta-data should be MIAPE-compliant (http://www.psidev.info/miape).	PRIDE (http://www.ebi.ac.uk/pride)	MIAPE-compliance is enforced by PRIDE submission interface
Metabolomic data	MetaboLights ISA format	MetaboLights (http://www.ebi.ac.uk/metabolights)	MetaboLights provides software support for generating compliant data
Infection data	No standard exists. Representation in a spreadsheet is currently normal.	Data should be deposited in a general purpose repository and identifiable through DOIs or similar identifiers.	
Microbiome data (from colonization experiments)	Meta-data should be MIxS-compliant (http://gensc.org/mixs)	Nucleotide sequence should be deposited in the European Nucleotide Archive (ENA) (http://www.ebi.ac.uk/ena).	Expected data types: 16S hypervariable region amplicon sequences, and taxonomic calls

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731060.