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 itself) – e.g. the MIAME (**M**inimal **I**nformation About a **M**icroarray **E**xperiment) 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

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

Policies for Data Sharing and Data Access

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.

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/	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.

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

	help/hgTrackHubHelp.html).		
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 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.	None that directly take submissions.	Contact Infravec2 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.