FAANG metadata - analysis specification

This document describes the specification for all analysis metadata. You can find an overview of our metadata and archival plans in the overview document. The experiment and sample documents are also in this git repo.

Raw data produced by each experiment will be analysed, producing *analysis results*. For example :

1. The ChIP-seq experiment will generate reads.
2. The reads are aligned to the genome.
3. Normalized signal plots and QC metrics are produced from the alignments.
4. ChIP-seq & ChIP input alignments are used for peak calling.

Steps 2-4 produce analysis results. For each of these analysis results we should record which data, reference data and protocol were used to produce them.

Analysis metadata requirements

Requirements are laid out like this:

- `attribute name` (*data type*) a brief description

The data types will be described later in this document. The metadata & data sharing (M&DS) group will seek guidance from the bioinformtics and data analysis (B&DA) group on what needs to be recorded here for each analysis type.

Process attributes

Analysis metadata needs to contain the following process attributes

1. Input data - a list of files used as input and references to the experiment records in a data archive
2. Reference data - genome assembly, gene set, etc
3. Analysis protocol - a precise description of the analysis protocol, including the following information:
    - URLs and version numbers for all software used (including in-house scripts)
    - Full command line used to run the analysis
    - Link to any VM or containers used, if applicable

The analysis must be reproducible based on the protocol document, as such we strongly recommend any inhouse scripts which are using through the process are made publicly available through github or a similar code repository. FAANG is happy to host people's code under the FAANG github repository if that is needed.

QC attributes

The analysis metadata will also contain QC attributes. These will vary based on the experiment type, but for sequencing work should always include mapping statistics as a bare minimum.

Required:

- `total reads` (*number*) the number of reads used in mapping
- `mapped reads` (*number*) the number of reads that can be mapped. Care should be taken that reads with multiple mappings are only counted once

File naming

Each file should be uniquely identifiable with a human readable name, giving sufficient information to understand what it contains. We expect analysis to be repeated at intervals, as reference data and protocols are updated, so a data freeze date is included.

Short names based on the following, separated with a dot ('.'):

- species / assembly version
- sample name
- sample description (tissue or cell type)
- assay type
- experiment target
- experiment ID
- analysis protocol name
- results type (e.g. genotypes vs. sites for 1000genomes vcf files)
- data freeze date
- file format

e.g

OAR3_1.OA_Roslin001.liver.H3K27ac.ERX053278.FAANGUK_chipv3.peaks.20150617.bb

So this hypothetical example represents a liver H3K27ac ChIP-Seq experiment for the Roslin's first Sheep sample using the FAANG v3 peak calling pipeline on the Sheep Assembly OAR_v3.1