README.md

awesome-bioinformatics-formats

Curated list of bioinformatics formats and publications. Not every format here is "awesome" per se, but if you are thinking about creating a new format this could be your first place to look at potential pre-existing formats. We also include formats not specific to bioinformatics, but should be considered for bioinformatics applications.

Please feel free to contribute.

EDAM

> EDAM is a comprehensive ontology of well-established, familiar concepts that are prevalent within bioinformatics and computational biology, including types of data and data identifiers, data formats, operations and topics. EDAM provides a set of concepts with preferred terms and synonyms, definitions, and some additional information - organised into a simple and intuitive hierarchy for convenient use.

EDAM is a more exhaustive and established ontology for bioinformatics data including formats. This is not intended to be a replacement or contain as much information as EDAM, please refer to their great resources including this explorable ontology for more information. We ask that where possible you link to the EDAM ontology for any formats your contribute. If your format is not available, then it is a great opportunity to contribute to EDAM as well.

Table of Contents

Formats

General

Formats not specific to bioinformatics that should be considered.

- HDF5 - [edam:format_3590] HDF5 is a data model, library, and file format for storing and managing data.
- NetCDF - [edam:format_3650] Network Common Data Form is a set of interfaces for array-oriented data access and a freely distributed collection of data access libraries for C, Fortran, C++, Java, and other languages.
- SQLite - [edam:format_3621] SQLite is a C-language library that implements a

small, fast, self-contained, high-reliability, full-featured, SQL database engine.
- tiledb - TileDB manages massive dense and sparse multi-dimensional array data that frequently arise in important scientific applications.

Dense Genomic Data

Formats associated with storing dense functional genomics data.

- bigWig - [edam:format_3006] The bigWig format is useful for dense, continuous data and is a binary form of wiggle.
- Genomedata - a format for efficient storage of multiple tracks of numeric data anchored to a genome.
- GenomicsDB - GenomicsDB is an open sourced library and tools with a focus on optimizing sparse array storage specifically for genomic data.
- wiggle - [edam:format_3005] ASCII format for dense, continuous data.

Genomic Intervals

Formats associated with storing genomic intervals (e.g., contig, start, stop, strand).

- BED - [edam:format_3003] Browser Extensible Data format provides a flexible way to define the data lines that are displayed in an annotation track.
- bedGraph - [edam:format_3583] The bedGraph format allows display of continuous-valued data in track format.
- bigBed - [edam:format_3004] Binary and indexed form of BED.
- interval list - The intervals are given in the form `<chr> <start> <stop> + <target_name>`, with fields separated by tabs, and the coordinates are 1-based (first position in the genome is position 1, not position 0).
- narrowPeak - [edam:format_3613] This format is used to provide called peaks of signal enrichment based on pooled, normalized (interpreted) data. It is a BED6+4 format.
- segmentation file - A tab-delimited text file that lists loci and associated numeric values associated with copy number.
- tabix - [edam:format_3616] An *index* file format for genomic intervals (can be used on bed, gtf, vcf, etc).

Genomic Features

Formats for describing genomic features (e.g., gene models, etc.).

- genePred - [edam:format_3011] a table format commonly used for gene prediction tracks.
- GFF2 - [edam:format_1974] The general feature format is a file format used for describing genes and other features of DNA, RNA and protein sequences version 2.
- GFF3 - [edam:format_1975] The general feature format is a file format used for describing genes and other features of DNA, RNA and protein sequences version 3.
- GTF - [edam:format_2306] GTF stands for Gene transfer format. It borrows from GFF2, but has additional structure that warrants a separate definition and format name.

Genotype Data

Formats associated with genotype data.

- BCF - [edam:format_3020] Binary and compressed VCF format.
- GDS - Genomic Data Structure is a storage format for bioinformatics data similar to NetCDF.
- GVF - [edam:format_3019] The Genome Variation Format (GVF) is a very simple file format for describing sequence_alteration features at nucleotide resolution relative to a reference genome.
- MAF - A Mutation Annotation Format (MAF) file (.maf) is a tab-delimited text file that lists mutations.
- oxford-bgen - Binary version of the native Oxford gen format. Operations on bgen files are generally faster and more descriptive than on plain gen files, and the file size of bgen files is also smaller -- UK Biobank genotypes are in bgen format. Latest bgen version is 1.3.
- oxford-gen - [edam:format_3812] Native text genotype file format for Oxford statistical genetics tools, such as IMPUTE2 and SNPTEST.
- pileup - [edam:format_3015] Describes the base-pair information at each chromosomal position. This format facilitates SNP/indel calling and brief alignment viewing by eyes.
- plink-bed - PLINK binary biallelic genotype table.
- plink-ped - [edam:format_3288] PLINK plain-text genotype format. Mostly has been replaced by bed/bim/fam, but is useful if someone wants to actually look at the SNPs in plain-text since Plink bed is in binary, and also wants to retain more information than from a VCF (eg. additional individual information).

- plink2-pgen - PLINK2 binary genotype table capable of representing mixed-phase, multiallelic, and mixed-hardcall/dosage/missing genotype data.
- VCF - [edam:format_3016] Variant Call Format.

Unaligned Sequencing Data

Formats associated with storing unaligned sequencing data.

- FASTA - [edam:format_1929] ASCII format for storing nucleotide/amino acid sequences.
- FASTQ - [edam:format_1930] Designed as a replacement for FASTA, combining the sequence and quality information in the same file.

Aligned Sequencing Data

Formats associated with storing aligned sequencing data.

- BAM - [edam:format_2572] Binary SAM format. *note: it is becoming more common to store unaligned reads in a BAM called a uBAM*
- CALF - The Compact ALignment Format records the base qualities and mapping qualities of the aligned reads, and unaligned reads data.
- CRAM - [edam:format_3462] Further lossless compression of SAM format.
- MAF - [edam:format_3008] The multiple alignment format stores a series of multiple alignments in a format that is easy to parse and relatively easy to read.
- SAM - [edam:format_2573] The Sequence Alignment/Map format.

Molecular Structural Data

- CTfile - The CTfile Formats document fully describes the formats for CTfiles (chemical table files) including: Molfiles, RGfiles, Rxnfiles, SDfiles, RDfiles, XDfiles.
- mmCIF - [edam:format_1477] Another format for PDB molecular structures.
- PDB - [edam:format_1476] The Protein Data Bank (PDB) format provides a standard representation for macromolecular structure data derived from X-ray diffraction and NMR studies. This representation was created in the 1970's and a large amount of software using it has been written.

Medical Imaging Data

Formats associated with medical imaging data.

- DICOM - [edam:format_3548] Medical image format corresponding to the Digital Imaging and Communications in Medicine (DICOM) standard.
- NifTI - [edam:format_3549] Medical image and metadata format of the Neuroimaging Informatics Technology Initiative.

Miscellaneous

Formats that currently don't have a good section to place them yet (but still very relevant).

- BIOM - [edam:format_3746] The Biological Observation Matrix (BIOM) file format (canonically pronounced biome) is designed to be a general-use format for representing biological sample by observation contingency tables.
- Pairs - The specification for the text contact list file format for chromosome conformation experiments (e.g., Hi-C).
- SBML - [edam:format_2585] Systems Biology Markup Language (SBML), the standard XML format for models of biological processes such as for example metabolism, cell signaling, and gene regulation.

Review Papers and Blogs

This section contains links to relevant review papers and blog posts about bioinformatics formats.

- hts-specs - Specs from HTS

License

Releases

No releases published

Packages

No packages published

Contributors 2

kmhernan Kyle Hernandez

jzluo Jon Luo