Project design and methods

The STRSeq "BioProject" serves to organize these records within the GenBank repository and is divided into categories (as specified in Gettings 2015): commonly used autosomal STRs, alternate autosomal STRs, Y-chromosomal STRs, and X-chromosomal STRs. Each of these categories is divided further into locus-specific BioProjects. This BioProject hierarchy allows easier access to the GenBank records by browsing, BLAST searching, or ftp download. The sequence records in GenBank are flat files of specified format, such that they can be downloaded and parsed en masse as shown in Fig. 3, or explored via an interactive graphic display (Fig. 4).

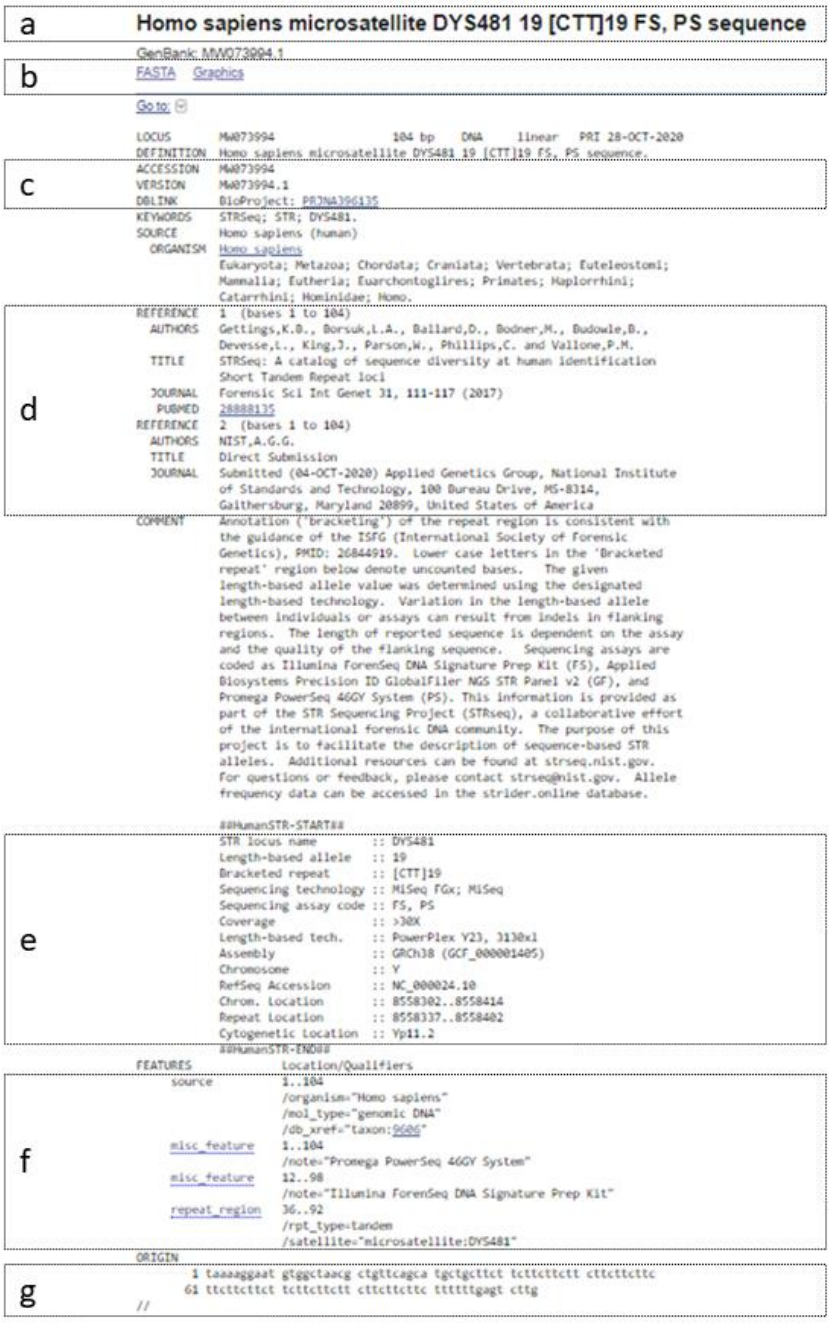

Figure 3. An example DYS481 STRSeq BioProject GenBank record. a) The Definition Line uniquely identifies each allele with components of the record. b) Hyperlinks to FASTA sequence and Graphics view (see Figure 4). c) GenBank sequence identifiers and link to the parent BioProject (e.g. DYS481), records in which the sequence string is updated branch off as e.g. MW073994.2, coexisting with MW073994.1. d) References describe the BioProject and identify the submitting laboratory. e) Field-based information relevant to STRSeq records and orienting the sequence to the GRCh38 reference sequence. f) Position of the repeat region within the sequence, position and dbSNP rs number of variations in the flanking regions (when applicable), and subset of sequence observed with different commercial assays (when applicable). Selecting a feature highlights corresponding region in sequence string. Rs numbers are hyperlinked to dbSNP. g) Full sequence string reported by the submitting laboratory. Length of reported sequence is assay- and quality-dependent, generally consistent with assay-specific configuration files published in (Woerner 2017).