Data models

We believe that it is important to keep the databases mostly unaware in which format the data was originally modelled and stored. A reference to this format will only be stored for specific purposes involving file transfers. Data models can be extended and changed to improve database performance, these changes are transparent for users.

Data models for CellBase data have been designed and implemented by using Java, ProtocolBuffer3 and/or Avro IDLs. They explicitly specify the most commonly used fields, and at the same time provide mechanisms for preserving all the information of a certain format. For instance, the fields specified for a variant would be (among others) chromosome, position, reference and alternatives; if a VCF file is being stored, then columns such as INFO are also saved in a key-value data structure.

Implementation

CellBase data models are stored in a related project **Biodata**, this guarantees that all OpenCB projects talk the same language whether if the use CellBase or not. Please, for detailed specification of the data models have a look at:

https://github.com/opencb/biodata/tree/develop/biodata-models/src/main/java/org/opencb/biodata/models
https://github.com/opencb/biodata/tree/develop/biodata-models/src/main/resources/protobuf
https://github.com/opencb/biodata/tree/develop/biodata-models/src/main/resources/avro