Germline Sets	Purpose and Content	Usage	Manifestation
	A set of genes, representative of the species. There may be multiple sets for one organism, e.g. high confidence; exhaustive; subsets for subgroups.  Contains links to the Gene Descriptions and a subset of their content. May contain additional annotation, e.g. additional field delineations.	Curators publish and maintain standard sets, mediated through OGRDB.  The format is intended to be directly usable by tools such as annotation software.  Most users will use one or more standard sets. Some users may tailor standard sets, or define their own, for particular purposes.	A single file or set of files per germline set, containing the information likely to be needed by tools.  Tool providers are encouraged to support the format.  File format is human readable and modifiable.
Gene Descriptions	A comprehensive summary of the evidence supporting existence of a single gene. References source records and publications. Names the gene, according to one or more naming schemes. Provides all available information on genomic location and context. Contains evaluations of confidence of the gene's existence. Paralogs are identified via references to other Gene Descriptions.	Curators maintain the descriptions and keep them up to date as new information becomes available.  Some users may refer to Gene Descriptions to obtain the full background to genes of interest  Some tools may access Gene Descriptions to provide enhanced functionality or reporting.	A single file per gene.  File format is human readable and modifiable, but sufficiently structured for automatic maintenance.  The structured data contained in the Descriptions may allow standard Germline Sets to be maintained automatically.
Source Records	References to primary or secondary evidence on which the inference of one or more genes is based.  Notes and evaluations of confidence  Exact nucleotide position of the sequence of interest	Curators of Gene Descriptions create a Source Record for each item of primary or secondary evidence. They provide a confidence score for each item.  Curators link each item to the Gene Description(s) for which it provides evidence.	Tables for each type of Source Record: for example genome sequence, RNA sequence, publication.  For ease of use, Source Records may also be denormalised into Gene Descriptions, to provide a convenient file that does not require cross-referencing.
External Records	Primary evidence: experimentally observed sequences (genome or RNA)  Secondary evidence: gene sequences inferred from Ab repertoire studies  Publications	Sequences are contributed by their observers — the experimental group, in the case of experimentally observed sequences, or the group conducting the analysis in the case of inferred sequences.  Curators of Gene Descriptions will review contributed source records and act on them as appropriate.	This primary evidence will be deposited in a public database of record such as Genbank. Discussion re. inferred records is in progress.  Submitters will adhere to the public database's guidelines. Additional guidelines defined by AIRR may apply for sequences to be considered as evidence of germline genes.

# Example Records

### **Source Record**

Observed Sequence : O100 id version : 23 : mRNA type : Genbank db ref : GI:BB234567

position : 50-250

cnf\_score : 1

sequence : acgu...

: e8f612a77ac1 md5 sum

Reference Genome

: R100 id version : 83

: GRCh38 assembly

patch : p7

: CHR\_HSCHR14\_3\_CTG1 gff\_seqid

: 106739363 gff\_start gff\_end : 106739862

gff\_strand :+ cnf\_score sequence : agct...

: e8f612a77ac1 md5\_sum

**Publication** 

id : P100 version : 93

PubMed id : 25501908

doi : 10.1038/nm.3743

Citation : In-depth determination...

cnf\_score : 1

# **Gene Description**

id : D100 : 9 version authors

species : Mus musculus subspecific\_name : domesticus additional info : BALB/C

: IGHV2-6-7\*01 canonical name Alt-name-1 : musIGHV134

Alt-name-2

inference class : 2

cnf score

: D101 paralog\_of

delineation:

delin\_name : IMGT FR1 : 1-24 CDR1 : 25-30

delin name : Chothia FR1 :1-30 CDR1: : 31-36

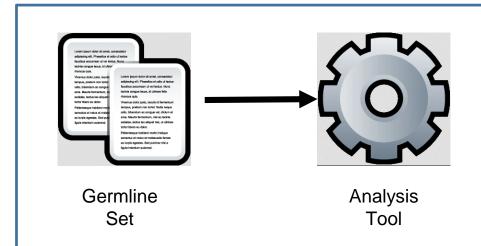
observed\_sequences : O500,O503,... reference\_genome\_links : R456, R837,... publications : P398, P8902, ...

#### **Germline Set**

: S100 id version : 48 : "High title confidence B-cell repertoire of the BALB/C mouse" authors reported\_in : P302, P109,... species : Mus musculus : domesticus subspecific\_name additional info : BALB/C : D100, D101,... genes

> Only illustrative fields are shown. For the full list of fields, consult the schema definition.

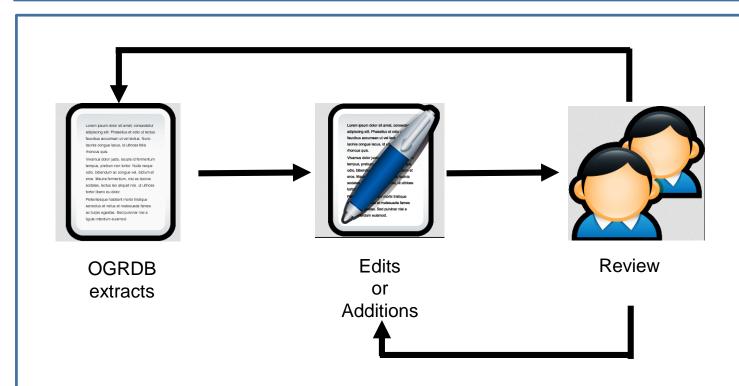
# **OGRDB** In Use



OGRDB will generate germline sets for use by repertoire parsers and other tools.

The format of the germline set will be convenient for tools to import.

Some users may wish to modify the germline sets, or produce their own sets in the same format: we hope to establish a standard for the interchange of germline sets.



Curators will download records from OGRDB for review and modification.

Downloaded records will be provided in CSV format for convenience.

OGRDB will support a review and acceptance process.

Proposed modifications will be checked by OGRDB.

Some fields will be auto-populated (e.g, if a PubMed ID is provided, OGRDB will populate the citation and doi fields).

All records have a version number which is generated by OGRDB. This will be used to resolve conflicting updates to the same record by different editors.

# **OGRDB** File Formats

## Source Record

All source records of a particular type, or an extract

id	version	
O001	23	
O002	54	

### Germline Set Definition

# The contents of a Germline Set Record

id : S100 version : 48 title : "High

confidence B-cell repertoire of the BALB/C mouse"

authors

reported\_in : P302, P109,... species : Mus musculus subspecific name : domesticus

additional\_info : BALB/C

genes : D100, D101,...

---

# Gene Description

A Gene Description Record, plus all associated Source Records in full

id : D100 version authors species : Mus musculus subspecific name : domesticus additional info : BALB/C : IGHV2-6-7\*01 canonical name Alt-name-1 : musIGHV134 Alt-name-2 inference class : 2 cnf\_score : 2 : D101 paralog of delineation: delin name : IMGT : 1-24 FR1 CDR1 : 25-30

#### **Observed Sequences**

P045

id	version	
O045	12	

## Reference Genome Sequences

12

id	version	
R045	12	
Publications		
id	version	

Primarily for curators

Primarily for tool users

Germline Set

The contents of a Germline Set Record, plus extracts from all associated Gene Descriptions

: S100 id version : 48 : "High title confidence B-cell repertoire of the BALB/C mouse" authors reported\_in : P302, P109,... : Mus musculus species subspecific name : domesticus additional info : BALB/C

id	version	
D100	12	
D101	54	

The Gene Description and Germline Set formats are non-canononical files meant to ease curation or use in specific circumstances