Integrating and Connecting databases

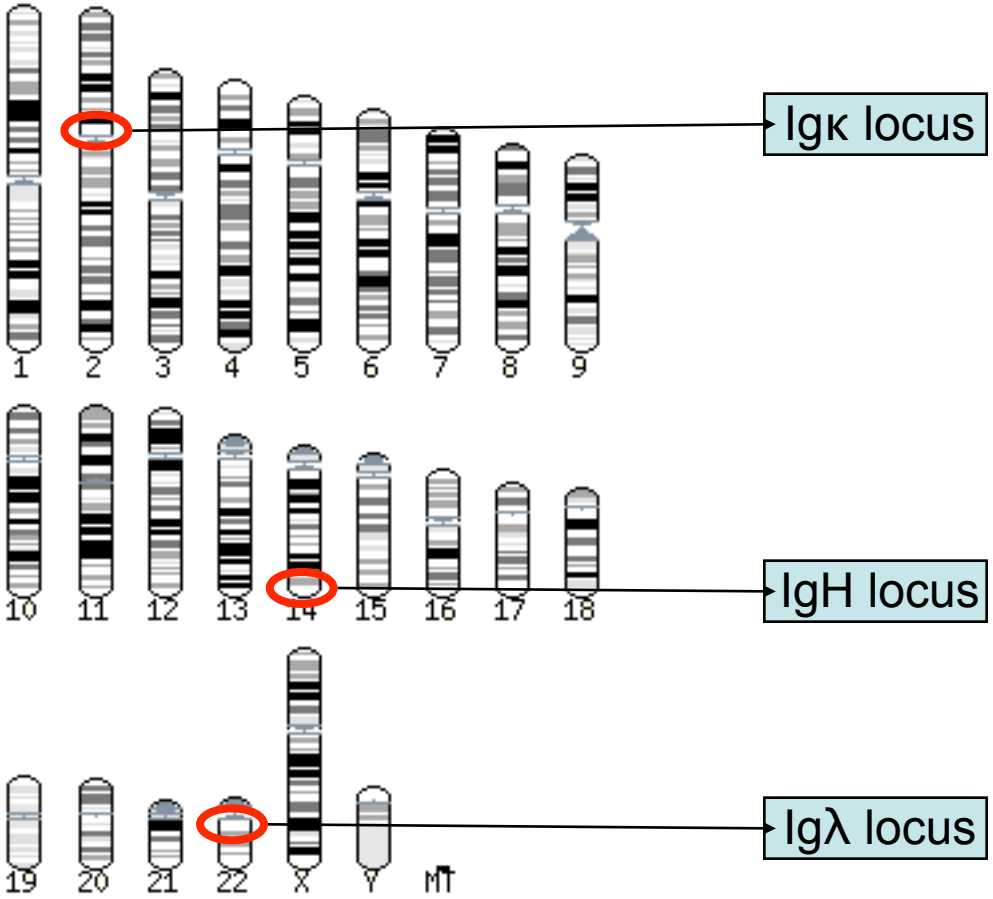
exemplified by the

VBASE 2 Database

http://www.vbase2.org

HELMHOLTZ
ZENTRUM FÜR
INFEKTIONSFORSCHUNG

immunoglobulin loci in the genome



→ DNA rearrangements in B cells

History of VBASE2

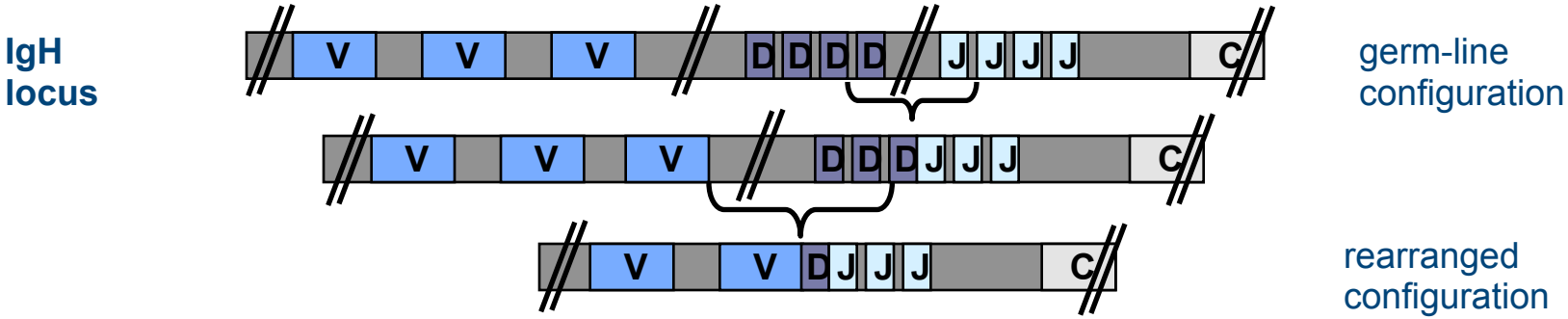
- In the beginning of the 90th of the last century, a group of Scientist thought it would be a good idea to add „expert“ databases to the EMBL Sequence database.
- The first example of such a database is the IMGT database
- The IMGT database picks up every EMBL entry and adds „expert“ annotation.
- This is very expensive and requires expert annotators

History of VBASE2

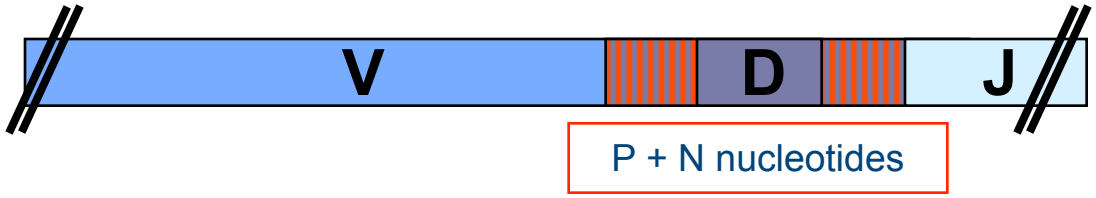
- Also in the middle of the 90th of the last century, Ian Tomlison developed the first manually collected database of human germline Immunoglobulin V genes, called VBASE.
- In this century, we developed a computer program that automatically generates a similar database of mouse and human Immunoglobulin V genes, called VBASE2.

generation of diversity

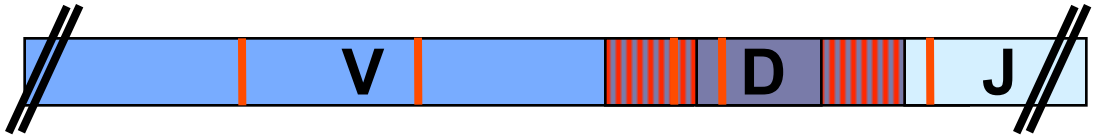
- combination of multiple gene segments



- diversity of the junctions



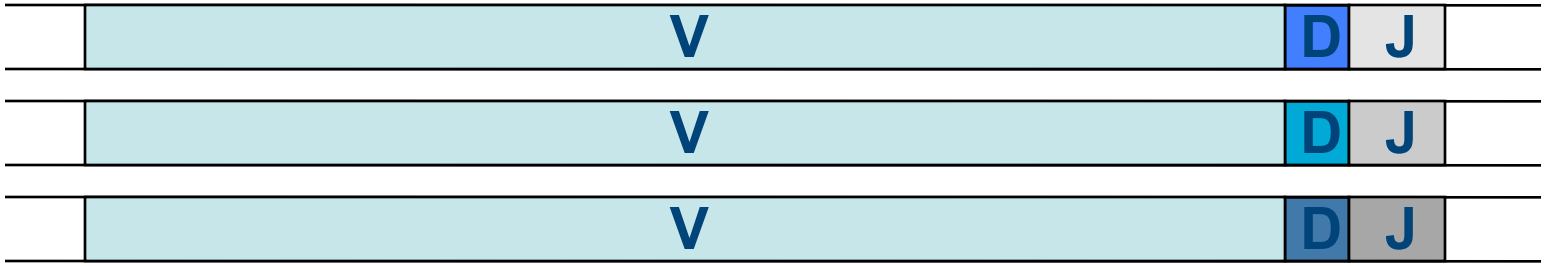
- somatic mutations



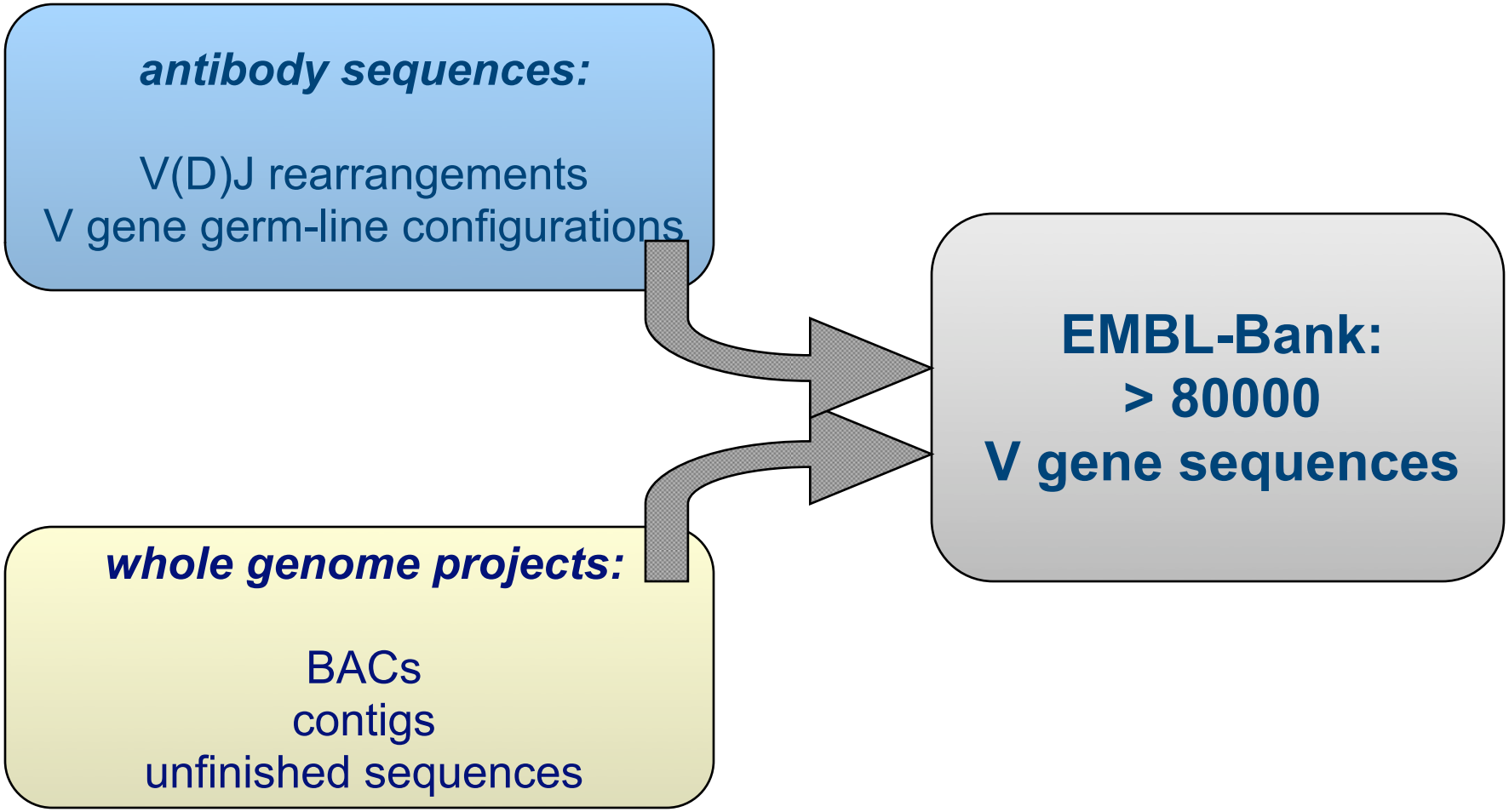
How to detect the germ-line V genes?

germ-line configuration

rearranged configuration



V gene sequences in EMBL-Bank

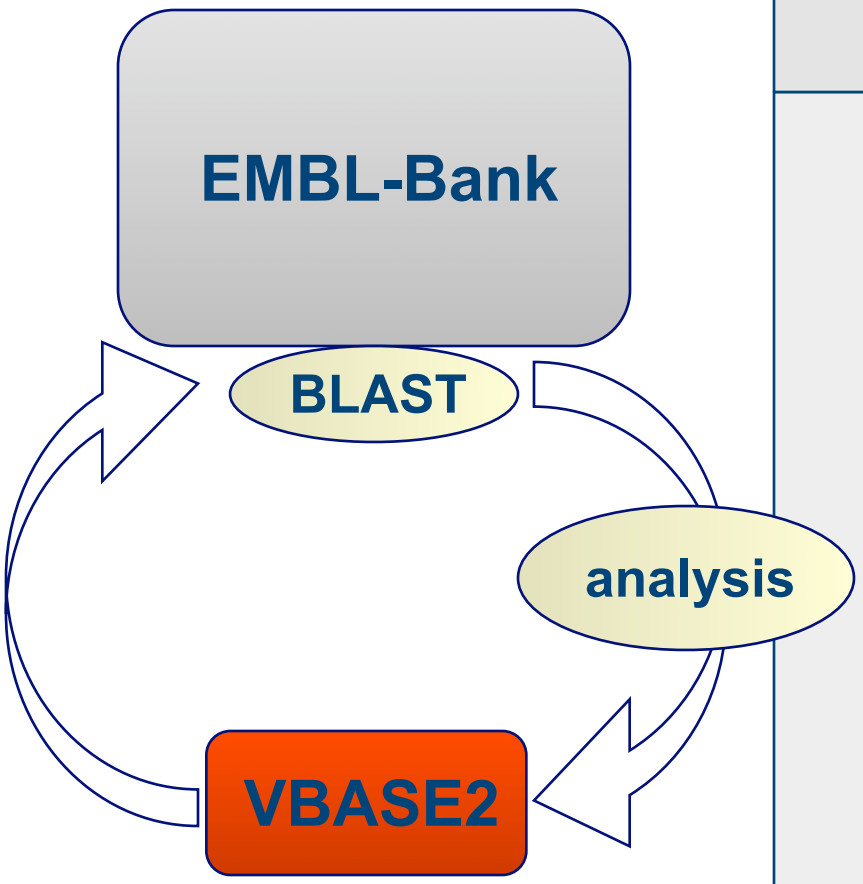


development of VBASE2

- germ-line V gene database
- derived from EMBL-Bank

- ➢ remove redundancy
- ➢ evaluate reliability
- ➢ automatically generated

VBASE2 generation procedure



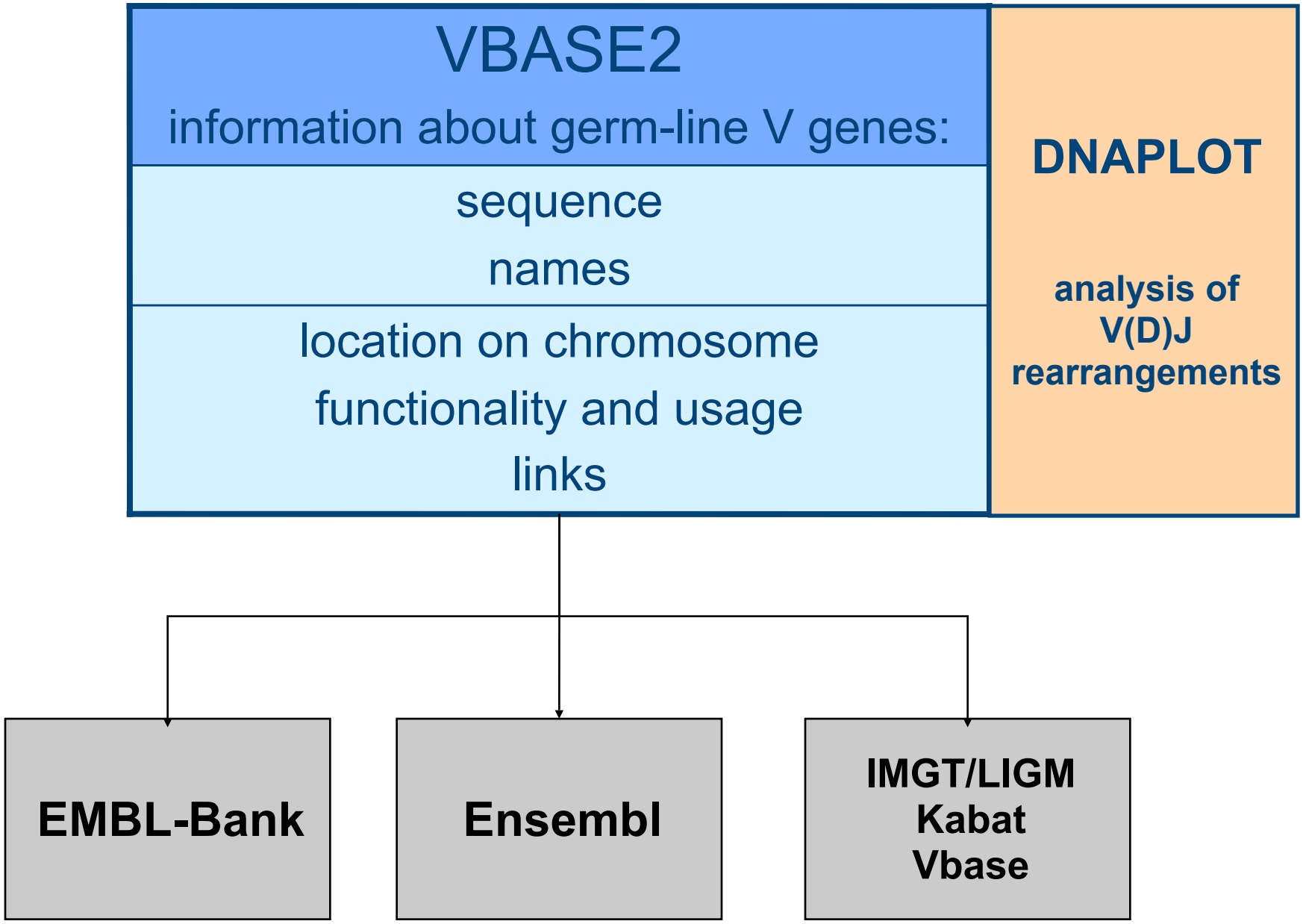
DNAPLOT / Perl

1. recognition
   - RSS elements / germ-line V genes
   - V pseudo-genes
   - J segments
2. sorting
   → assignment into 3 classes of sequence reliability
3. annotation
   → family- / name assignment, database cross references

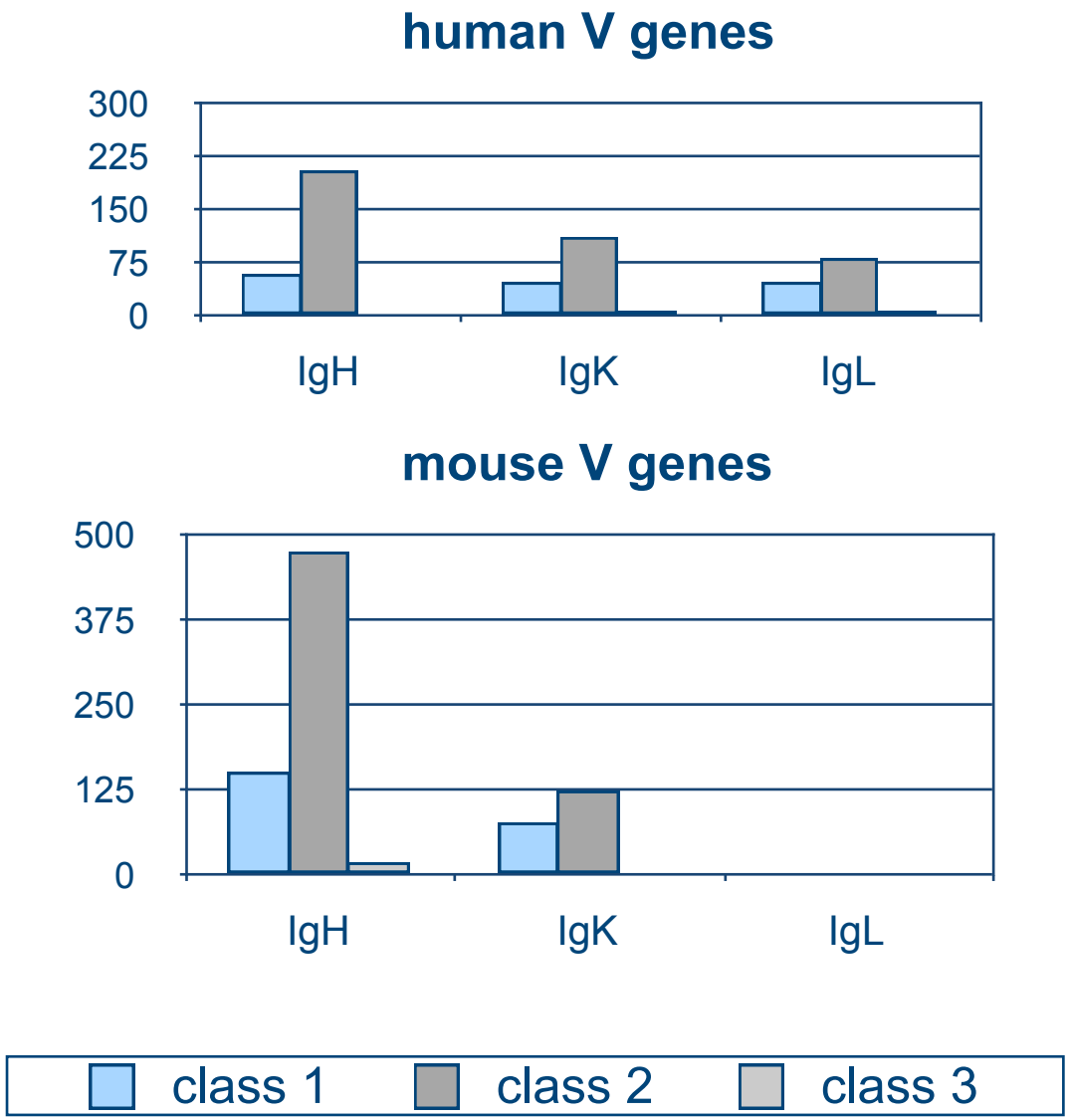
VBASE2 classes

evaluation of the available sequence information

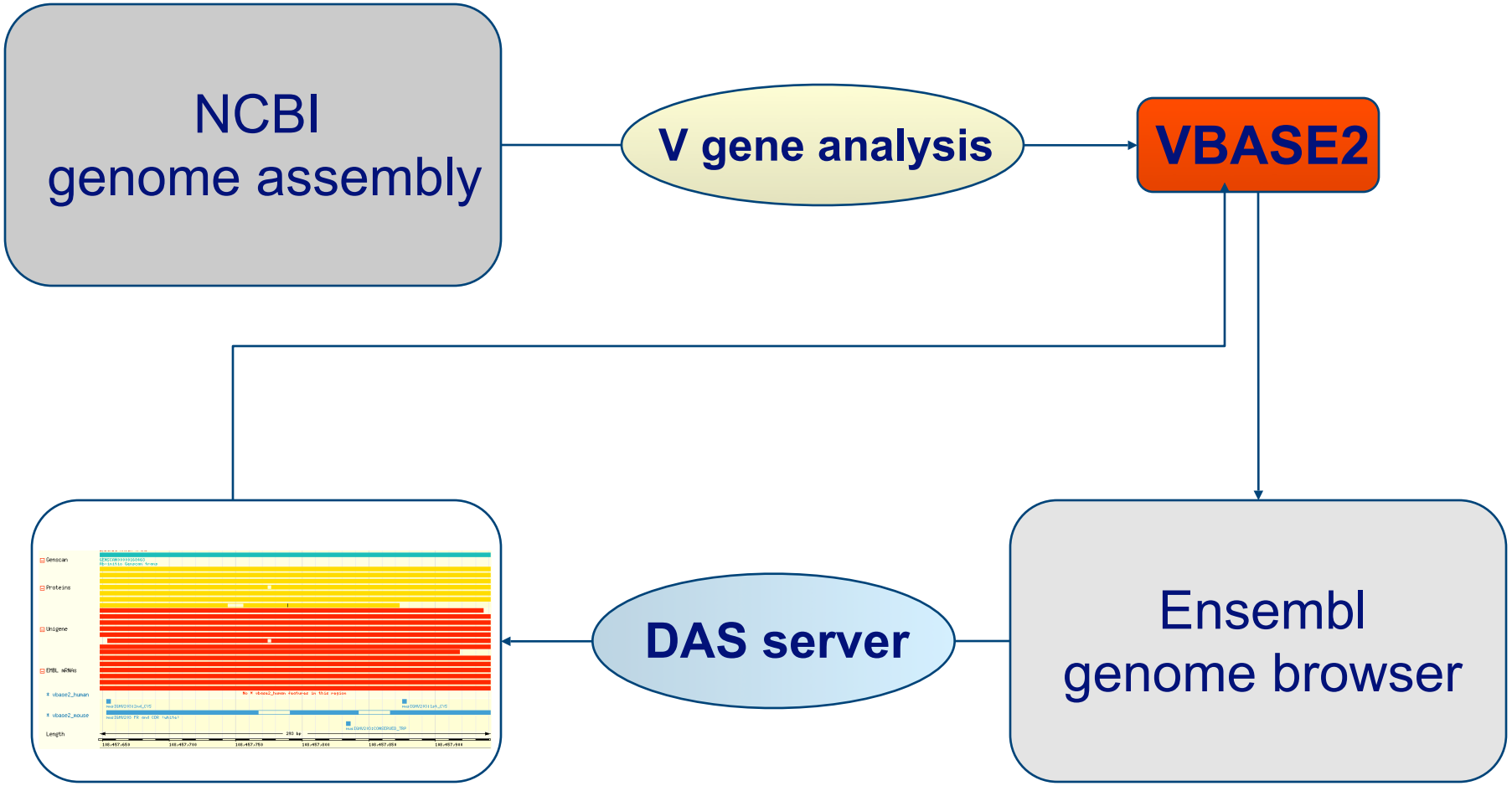
	sequence information	***resulting quality information***
class 1	V(D)J rearrangement + germ-line configuration	germ-line V genes, functional
class 2	germ-line configuration	germ-line V genes: pseudo genes or functionality unknown
class 3	V(D)J rearrangement	potential germ-line V genes


VBASE2
information about germ-line V genes:
sequence
names
location on chromosome
functionality and usage
links
DNAPLOT
analysis of
V(D)J
rearrangements
EMBL-Bank
Ensembl
IMGT/LIGM
Kabat
Vbase

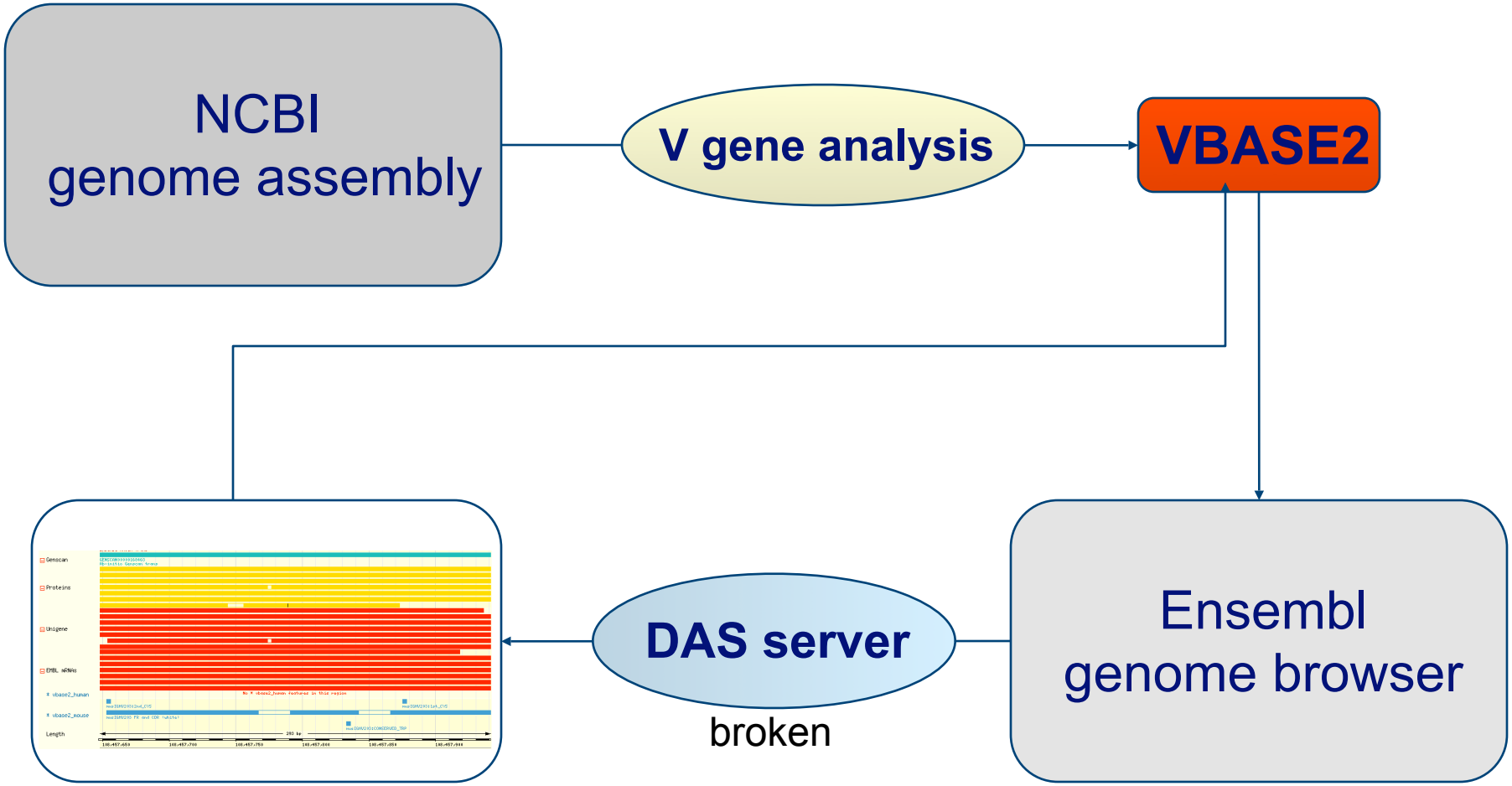
VBASE2 statistics



VBASE2 connection to Ensembl



VBASE2 connection to Ensembl



Problems when interconnecting databases

- Amount of Data to download and to process
- Defining the best time to update the database
- Keeping the DAS Server up and running

Problems when interconnecting databases

- Amount of Data to download and to process
  - Extensive Computing is required to build the database
  - Remote Processing of the data not practical
  - Web Services would generate to high overhead

Problems when interconnecting databases

- Defining the best time to update the database
  - Ideally the databases used for VBASE 2 would tell the VBASE 2 database automatically, when a new major update is due
  - The generation of a new release of the VBASE database could start automatically
  - RSS feed ?

Problems when interconnecting databases

- Keeping the DAS Server up and running
  - The current version does not support postgres databases
  - Ideally
    - ENSEML would create an open DAS Server
    - The user would provide the name of the Service and a pointer to a datafile to be uploaded by ENSEMBL with an indication to link it to a certain ENSEMBL version of a particular species

Thanks!

Helmholtz Centre for Infection Research,
Department of Experimental Immunology:

Ida Retter
Miguel Nunes
Svetlana Mollova

TU Braunschweig, Institute for Microbiology:
Richard Münch

HELMHOLTZ
ZENTRUM FÜR
INFEKTIONSFORSCHUNG

Bundesministerium
für Bildung
und Forschung

VBASE2 server is provided by EU-gene