Поиск новых мутаций и неаннотированных V, D, J генных сегментов в репертуаре антител

Караваева Валерия

Институт биоинформатики

Руководитель: Сафонова Яна,
ЦАБ СПБГУ

Санкт-Петербург
2016 г.

Мотивация

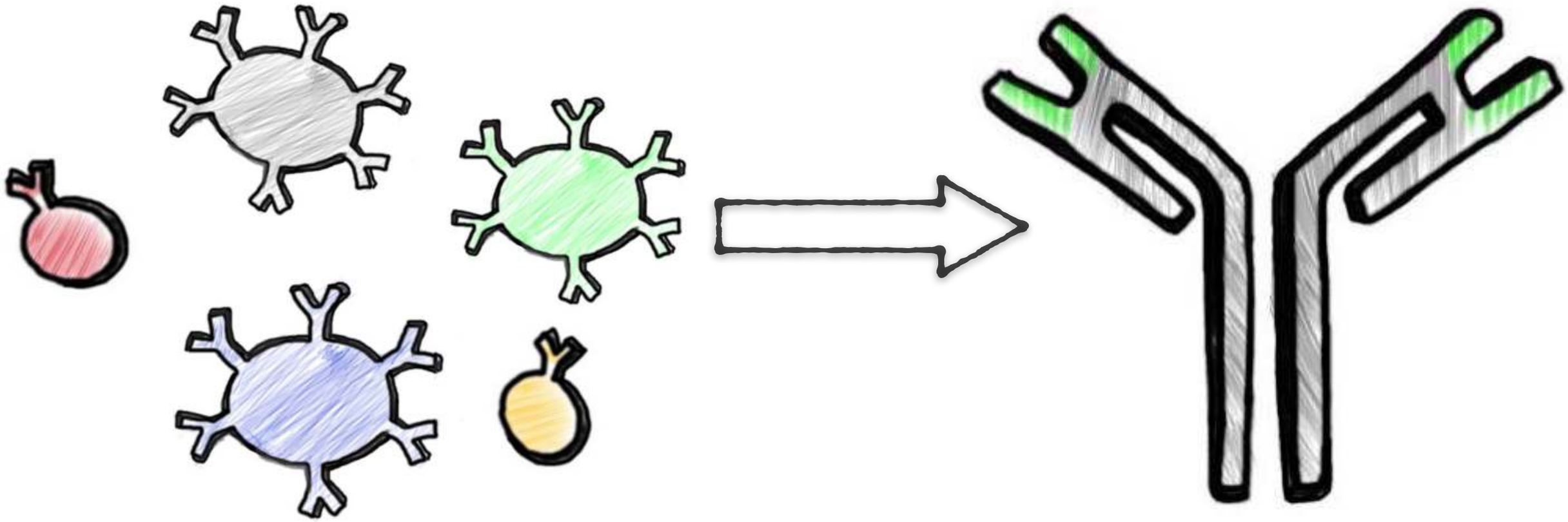
- **B-клетки** - тип клеток, играющих важную роль в обеспечении гуморального иммунитета.

- **Антитела** - специальные белки, вырабатываемые B-клетками и служащие для распознавания потенциально вредоносных объектов (антигенов).

Мотивация

- **B-клетки** - тип клеток, играющих важную роль в обеспечении гуморального иммунитета.

- **Антитела** - специальные белки, вырабатываемые B-клетками и служащие для распознавания потенциально вредоносных объектов (антигенов).



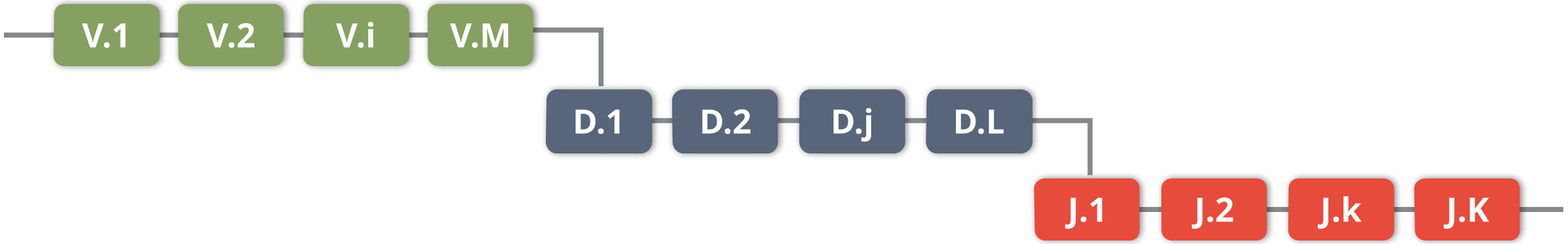
B-клетки **Антитело**

Мотивация

- **V(D)J рекомбинация** - механизм соматической рекомбинации ДНК, приводящий к формированию антиген-распознающих участков антител или Т-клеточного рецептора.

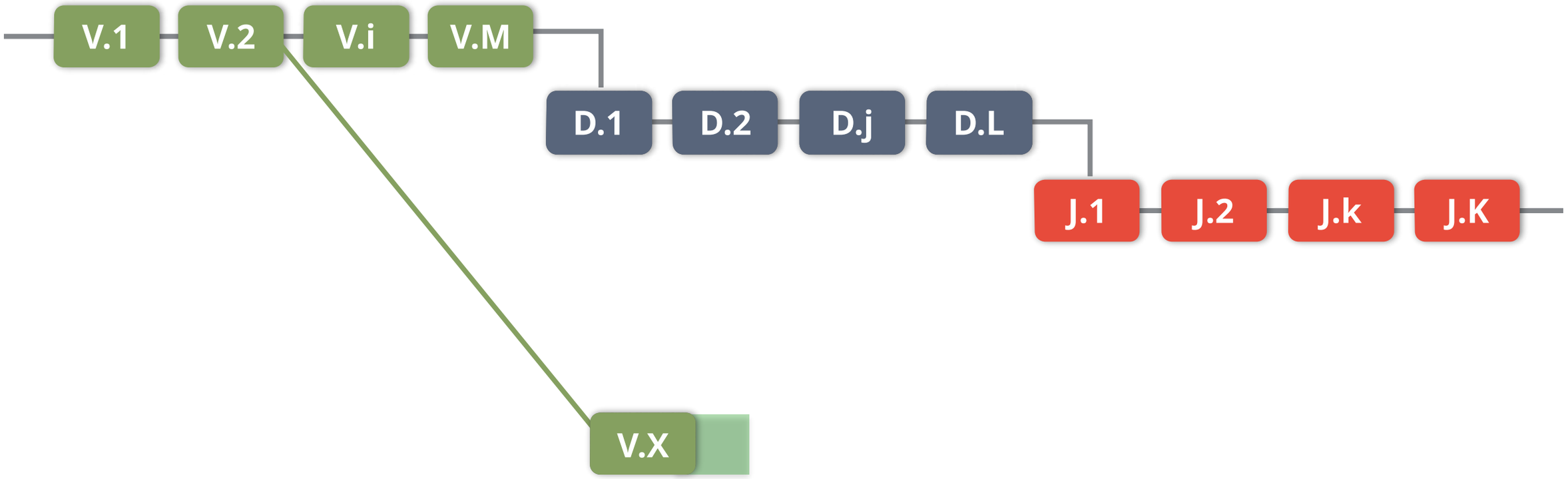
Мотивация

- **V(D)J рекомбинация** - механизм соматической рекомбинации ДНК, приводящий к формированию антиген-распознающих участков антител или Т-клеточного рецептора.



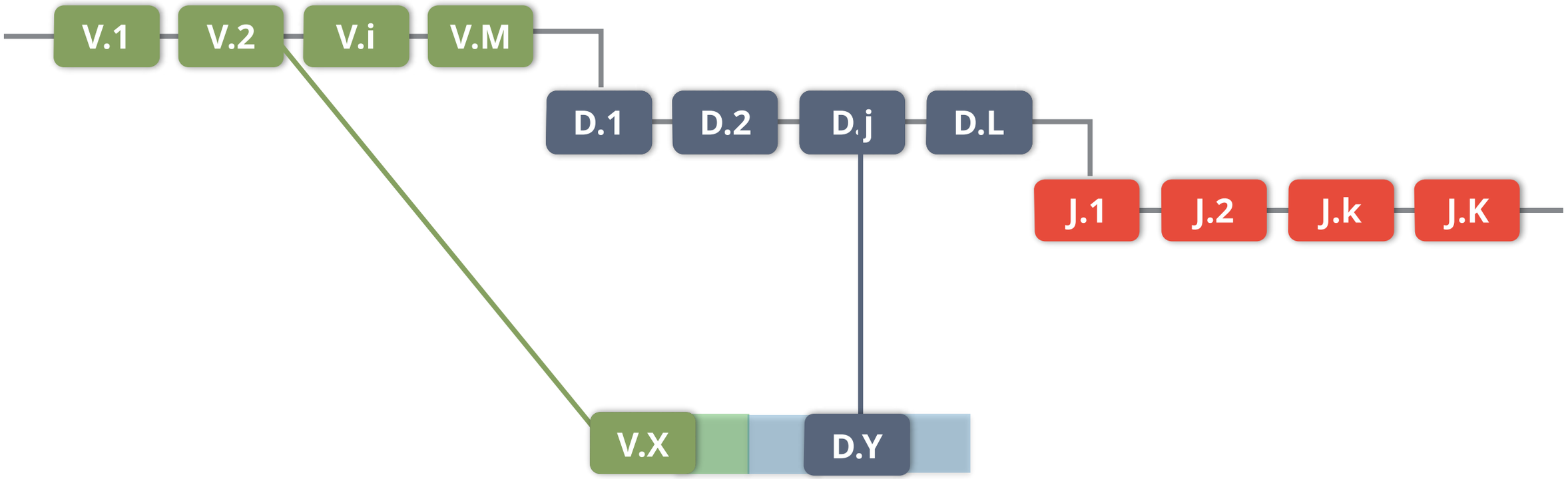
Мотивация

- **V(D)J рекомбинация** - механизм соматической рекомбинации ДНК, приводящий к формированию антиген-распознающих участков антител или Т-клеточного рецептора.



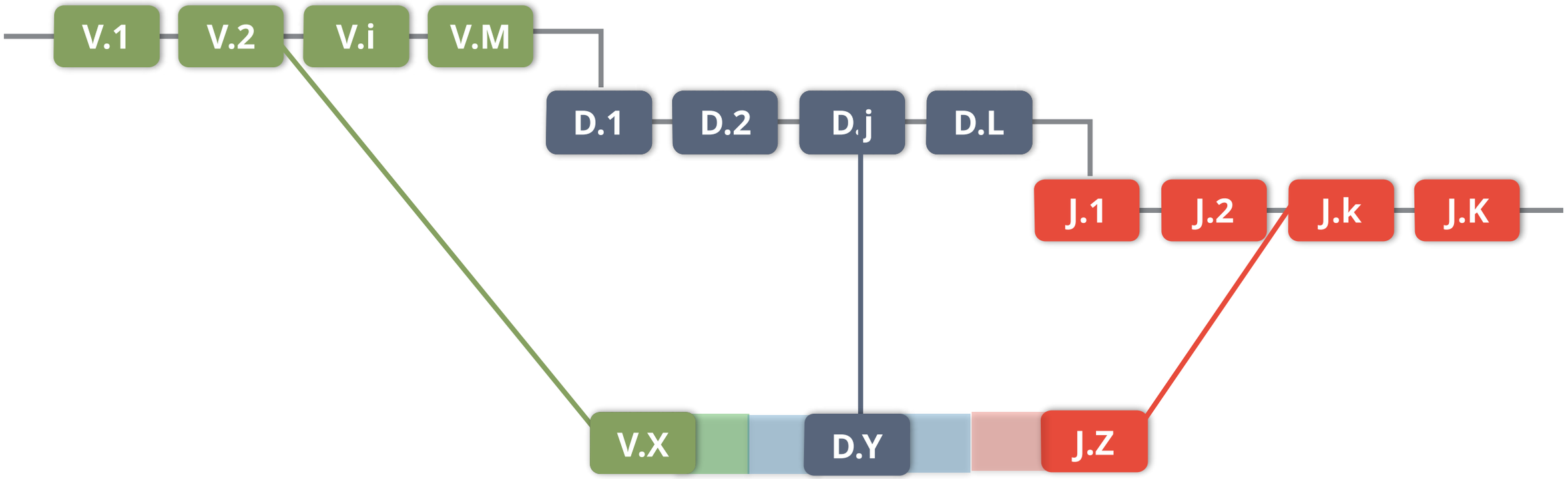
Мотивация

- **V(D)J рекомбинация** - механизм соматической рекомбинации ДНК, приводящий к формированию антиген-распознающих участков антител или Т-клеточного рецептора.



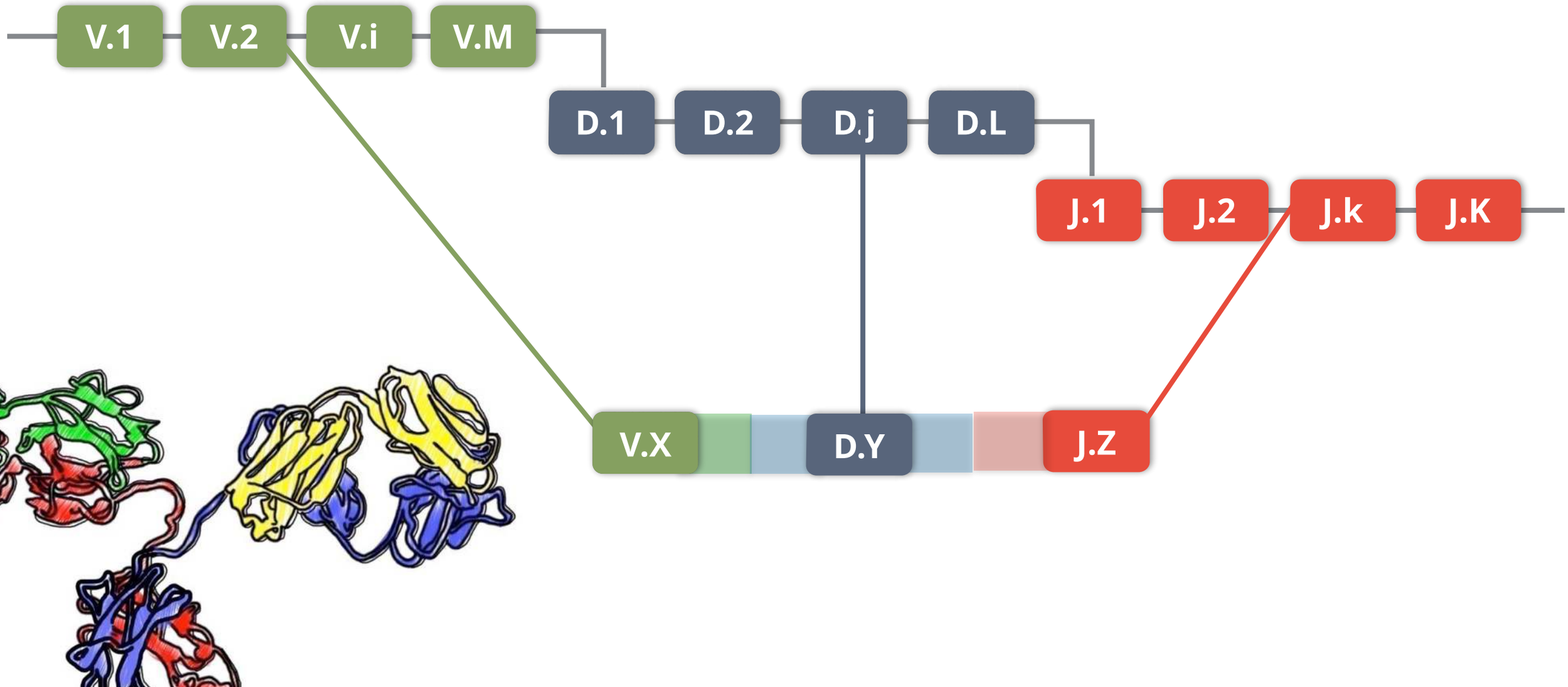
Мотивация

- **V(D)J рекомбинация** - механизм соматической рекомбинации ДНК, приводящий к формированию антиген-распознающих участков антител или Т-клеточного рецептора.



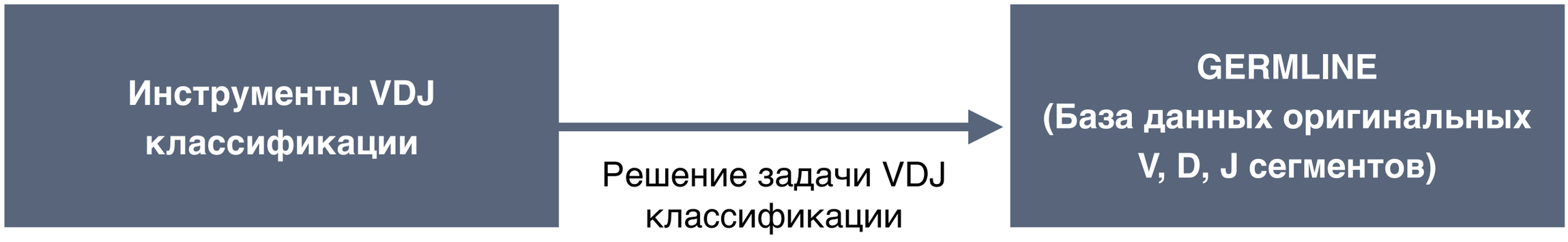
Мотивация

- **V(D)J рекомбинация** - механизм соматической рекомбинации ДНК, приводящий к формированию антиген-распознающих участков антител или Т-клеточного рецептора.



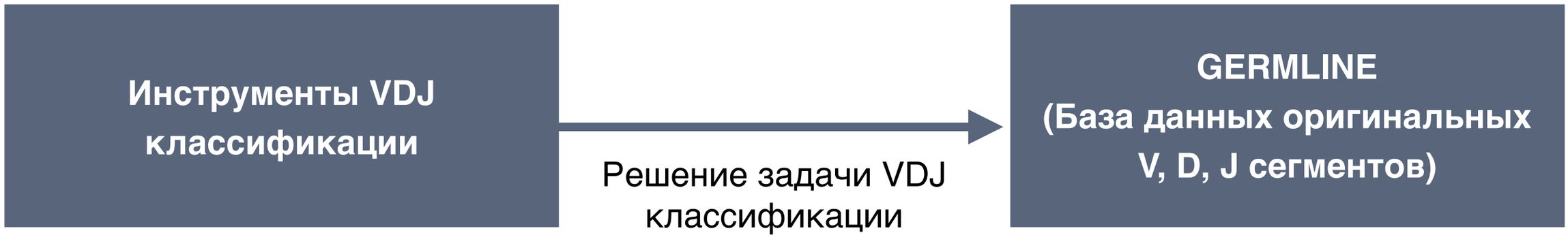
Задача

- **V(D)J классификация:** для данного рекомбинированного гена найти ближайший сегменты исходного ДНК



Задача

- **V(D)J классификация:** для данного рекомбинированного гена найти ближайший сегменты исходного ДНК



- **Проблема:** для многих рекомбинированных генов не удается найти достаточно близкие сегменты из базы данных, для некоторых организмов база неизвестна.

Задача

Мотивация: сборка локусов иммуноглобулинов требует длинных ридов (Sagner или PacBio), это значительно дороже, чем секвенирование репертуара с помощью Illumina MiSeq.

Задача

Мотивация: сборка локусов иммуноглобулинов требует длинных ридов (Sagner или PacBio), это значительно дороже, чем секвенирование репертуара с помощью Illumina MiSeq.

Задача проекта: разработка подхода для поиска новых V, D, J сегментов в репертуаре рекомбинированных антител.

Результаты

Результаты

- **Данные:** три репертуара - наивный, ASC, плазматический. Каждый определяется типом клеток, из которых он состоит.
- **Пути решения:** построение базы данных (БД) из наивного репертуара антител.
- **Причины:** а. наивный репертуар не содержит соматических мутация.

Результаты

- **Данные:** три репертуара - наивный, ASC, плазматический. Каждый определяется типом клеток, из которых он состоит.
- **Пути решения:** построение базы данных (БД) из наивного репертуара антител.
- **Причины:** a. наивный репертуар не содержит соматических мутация.
  b. простота верификации с помощью плазматического и ASC репертуаров.

Результаты

- **Данные:** три репертуара - наивный, ASC, плазматический. Каждый определяется типом клеток, из которых он состоит.
- **Пути решения:** построение базы данных (БД) из наивного репертуара антител.
- **Причины:** a. наивный репертуар не содержит соматических мутация.
  b. простота верификации с помощью плазматического и ASC репертуаров.

Стандартная БД мышей

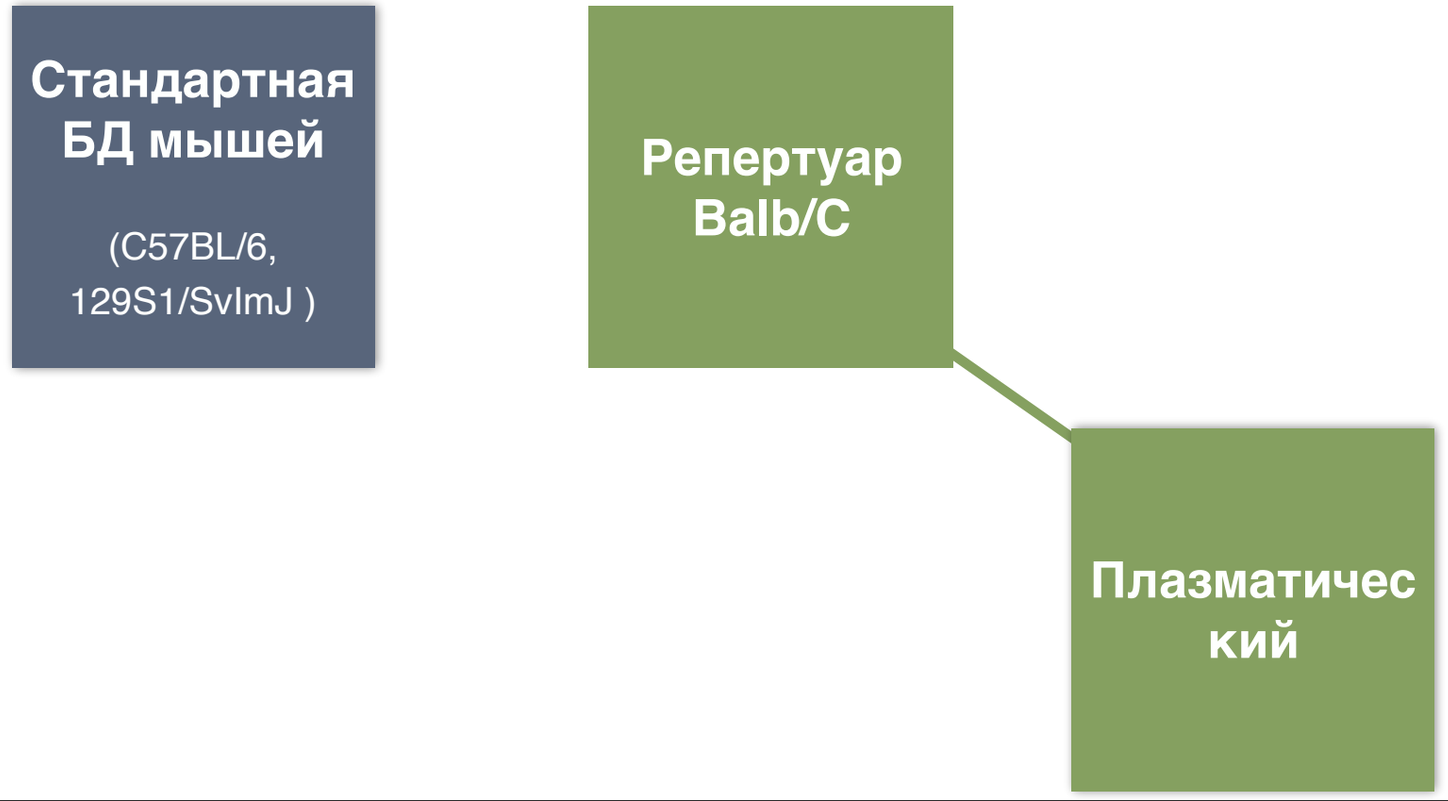
(C57BL/6, 129S1/SvImJ)

Результаты

- **Данные:** три репертуара - наивный, ASC, плазматический. Каждый определяется типом клеток, из которых он состоит.
- **Пути решения:** построение базы данных (БД) из наивного репертуара антител.
- **Причины:** a. наивный репертуар не содержит соматических мутация.
  b. простота верификации с помощью плазматического и ASC репертуаров.

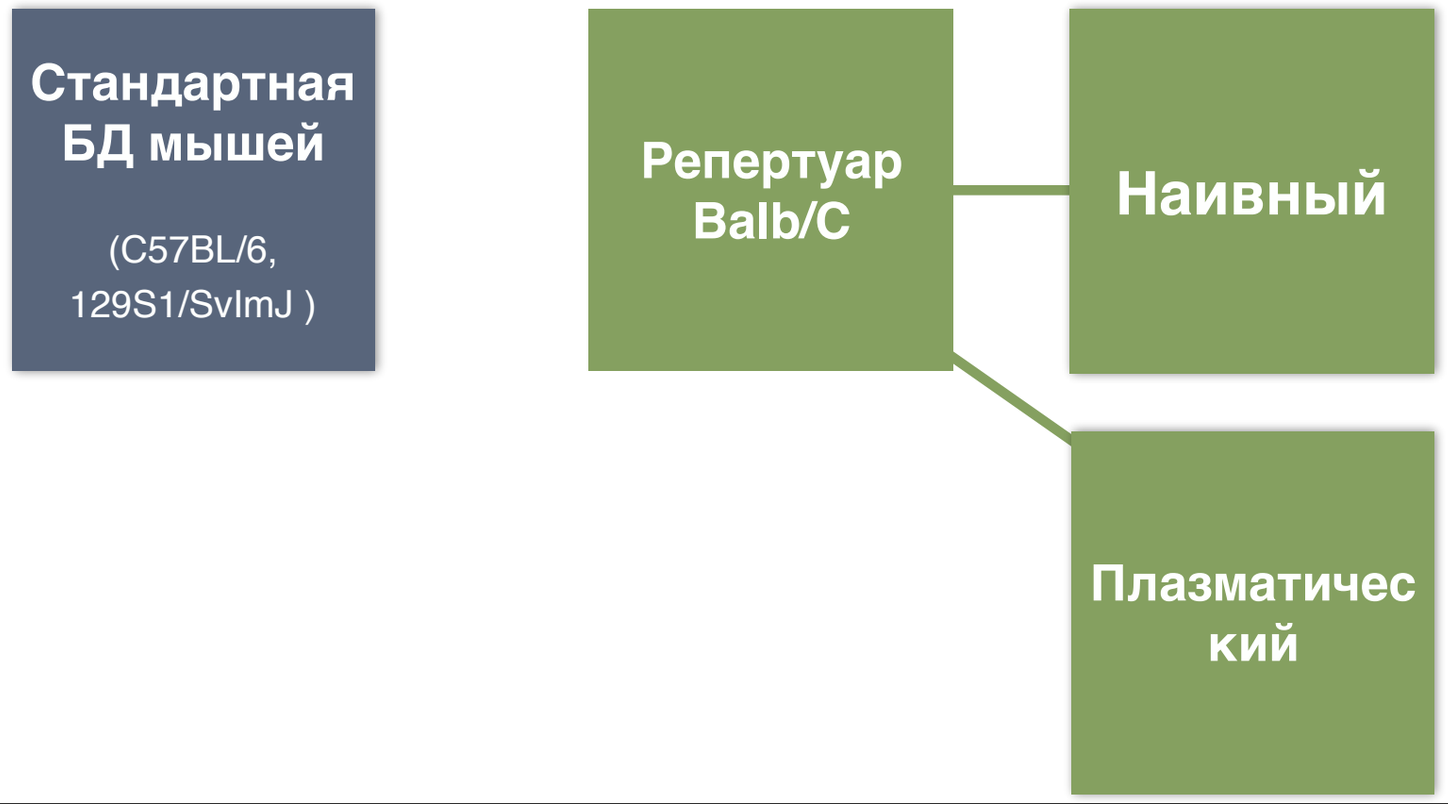
Результаты

- **Данные:** три репертуара - наивный, ASC, плазматический. Каждый определяется типом клеток, из которых он состоит.
- **Пути решения:** построение базы данных (БД) из наивного репертуара антител.
- **Причины:** a. наивный репертуар не содержит соматических мутация.
  b. простота верификации с помощью плазматического и ASC репертуаров.



Результаты

- **Данные:** три репертуара - наивный, ASC, плазматический. Каждый определяется типом клеток, из которых он состоит.
- **Пути решения:** построение базы данных (БД) из наивного репертуара антител.
- **Причины:** a. наивный репертуар не содержит соматических мутация.
  b. простота верификации с помощью плазматического и ASC репертуаров.

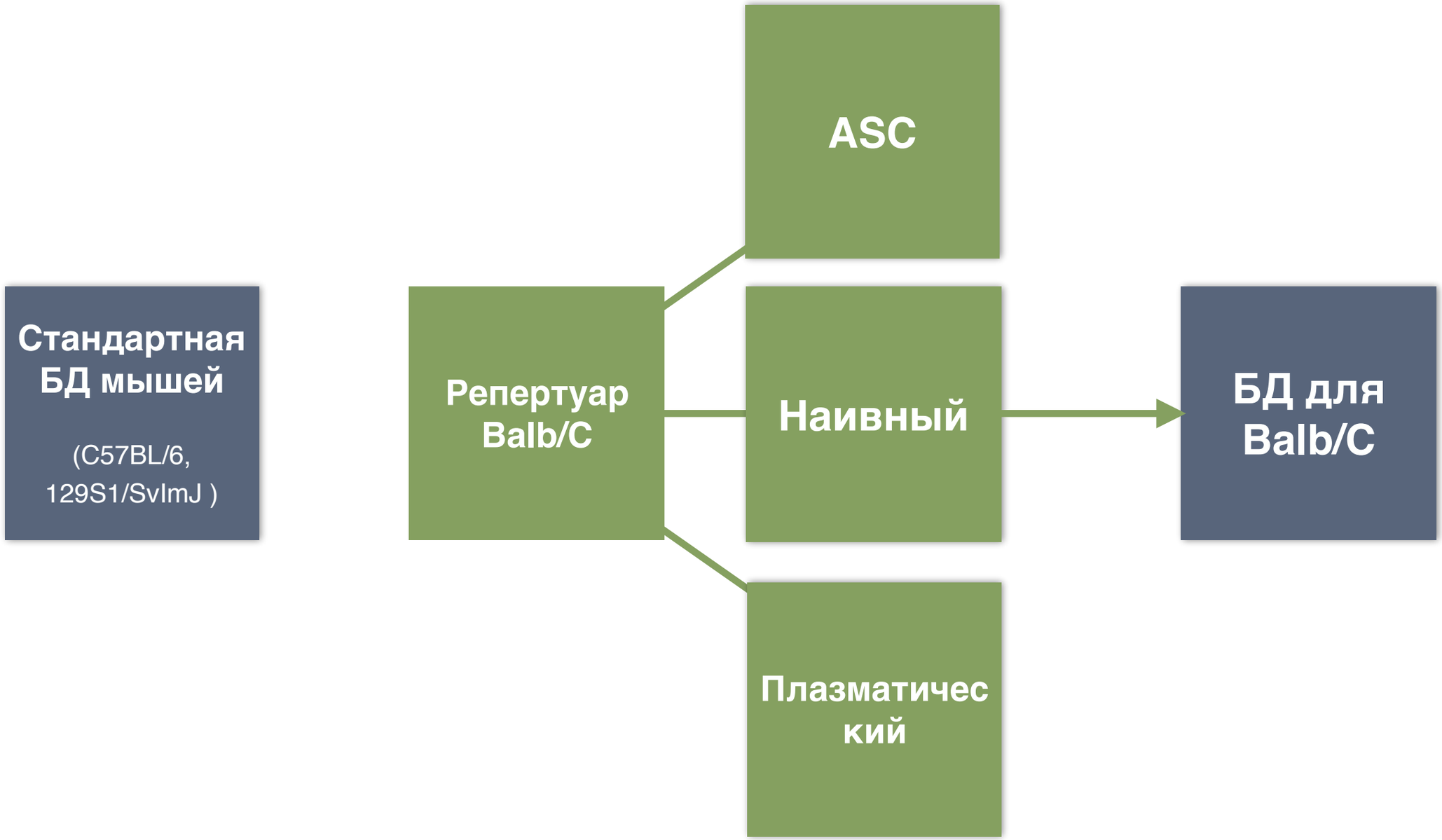


Результаты

- **Данные:** три репертуара - наивный, ASC, плазматический. Каждый определяется типом клеток, из которых он состоит.
- **Пути решения:** построение базы данных (БД) из наивного репертуара антител.
- **Причины:** a. наивный репертуар не содержит соматических мутация.
  b. простота верификации с помощью плазматического и ASC репертуаров.

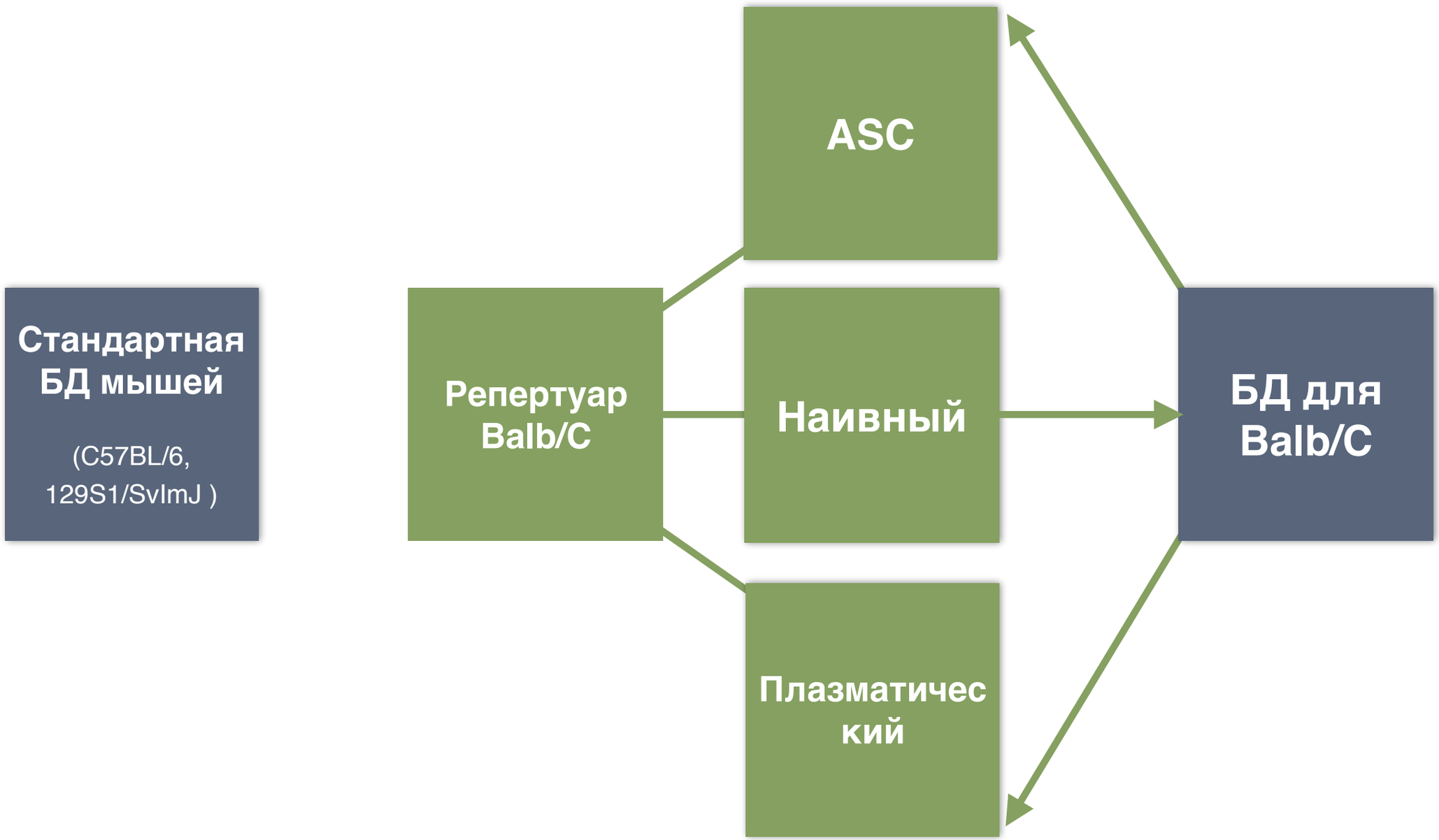
Результаты

- **Данные:** три репертуара - наивный, ASC, плазматический. Каждый определяется типом клеток, из которых он состоит.
- **Пути решения:** построение базы данных (БД) из наивного репертуара антител.
- **Причины:** a. наивный репертуар не содержит соматических мутация.
  b. простота верификации с помощью плазматического и ASC репертуаров.



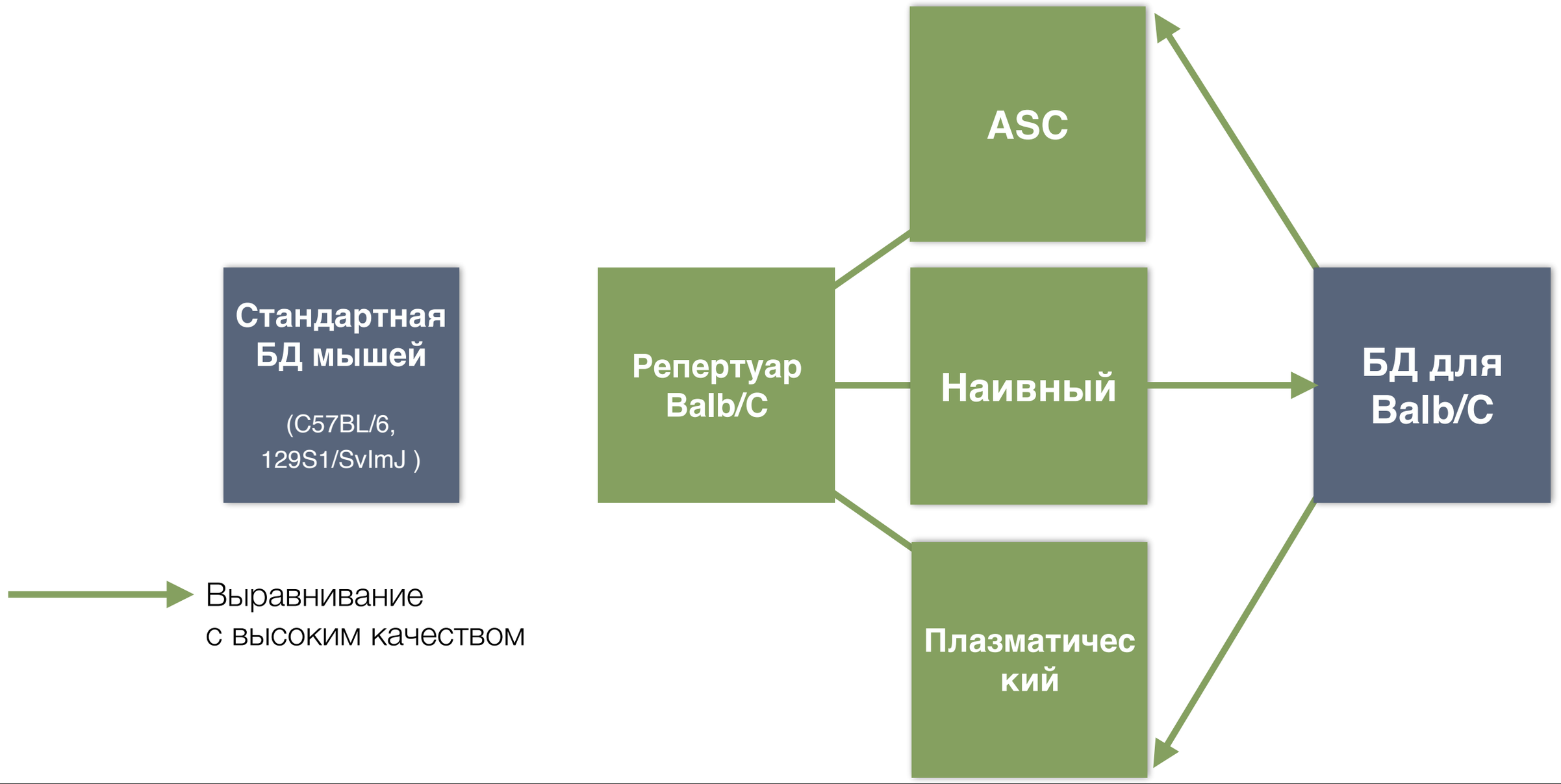
Результаты

- **Данные:** три репертуара - наивный, ASC, плазматический. Каждый определяется типом клеток, из которых он состоит.
- **Пути решения:** построение базы данных (БД) из наивного репертуара антител.
- **Причины:** a. наивный репертуар не содержит соматических мутация.
  b. простота верификации с помощью плазматического и ASC репертуаров.



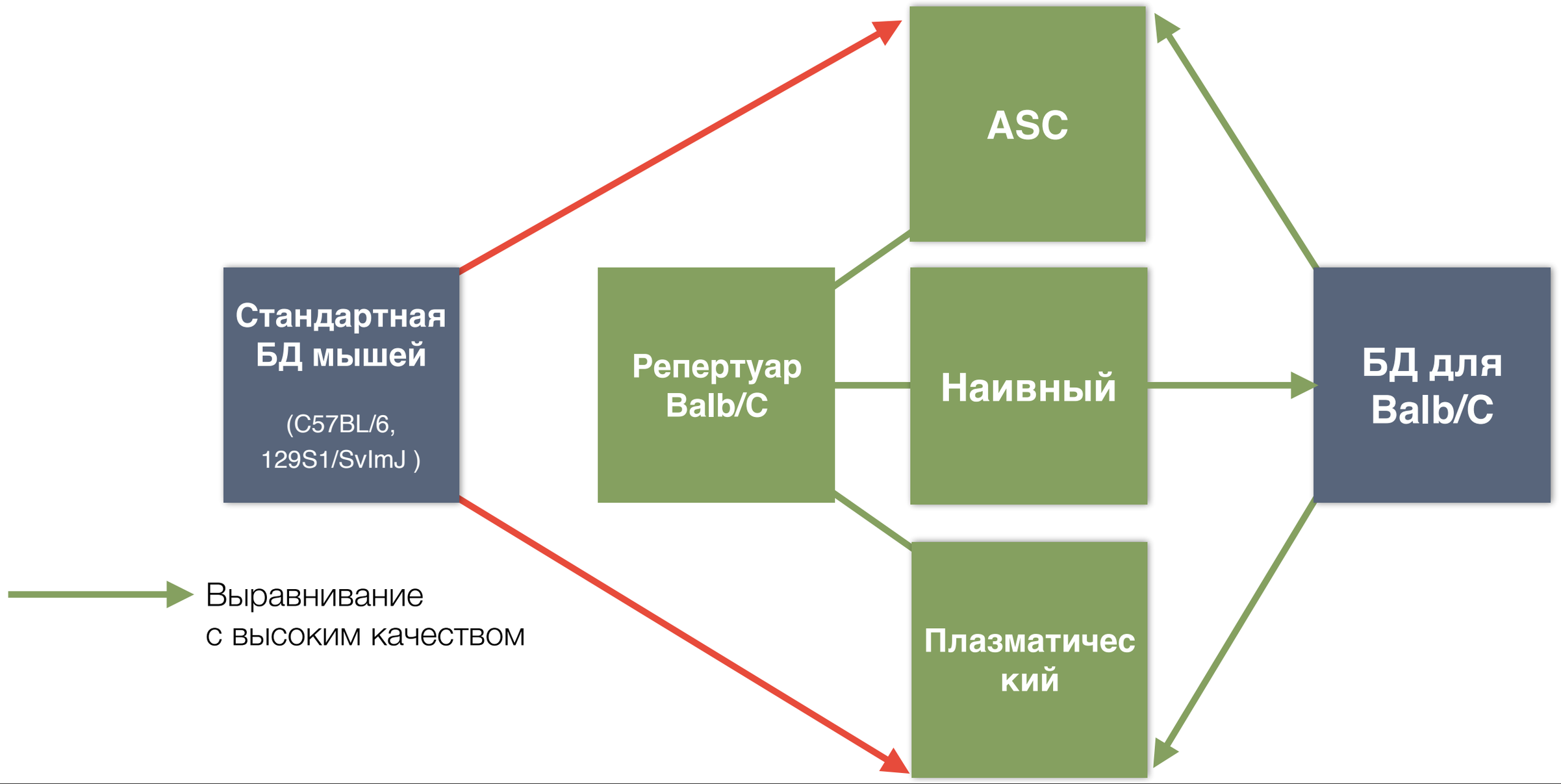
Результаты

- **Данные:** три репертуара - наивный, ASC, плазматический. Каждый определяется типом клеток, из которых он состоит.
- **Пути решения:** построение базы данных (БД) из наивного репертуара антител.
- **Причины:** a. наивный репертуар не содержит соматических мутация.
  b. простота верификации с помощью плазматического и ASC репертуаров.



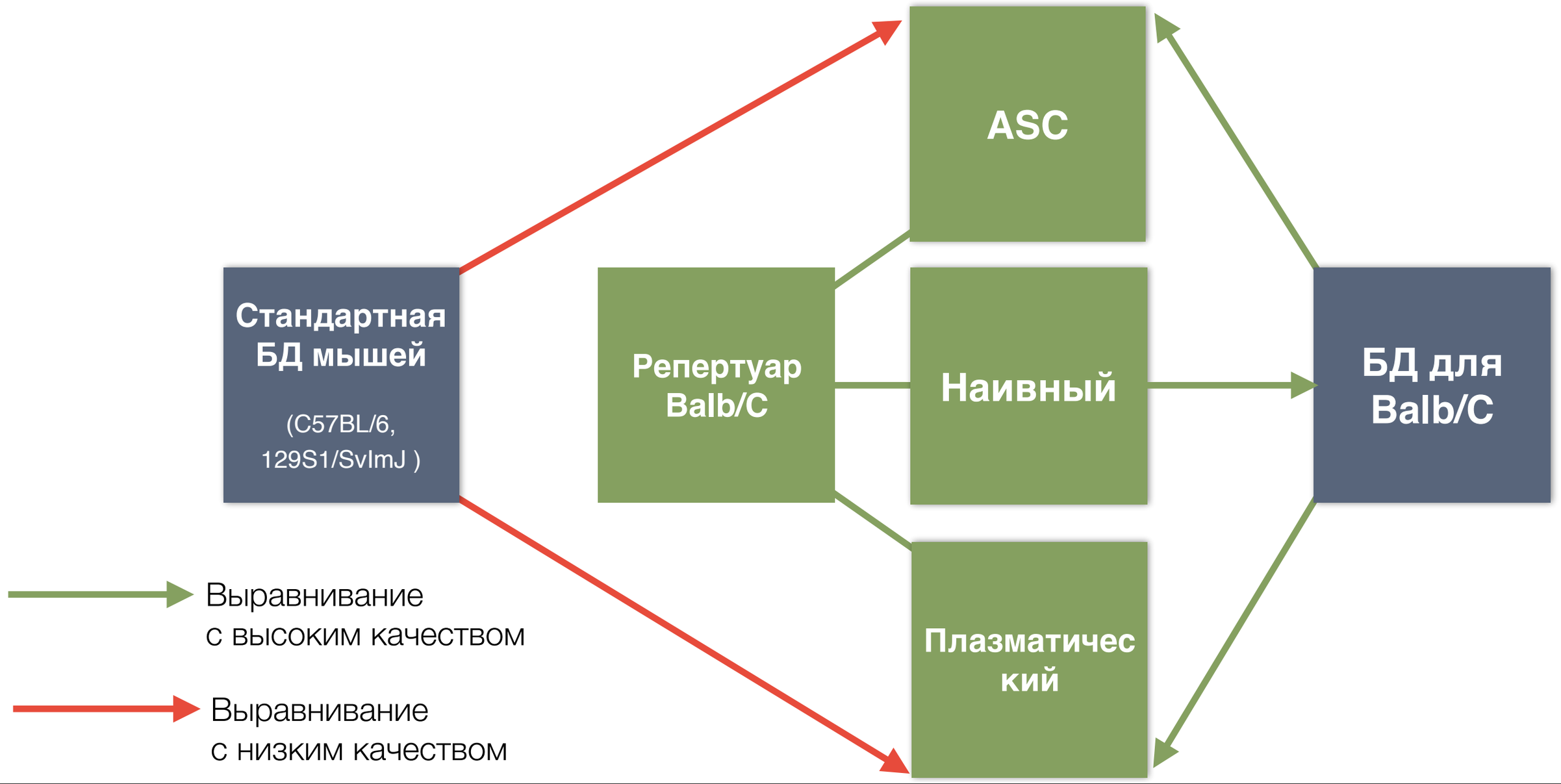
Результаты

- **Данные:** три репертуара - наивный, ASC, плазматический. Каждый определяется типом клеток, из которых он состоит.
- **Пути решения:** построение базы данных (БД) из наивного репертуара антител.
- **Причины:** a. наивный репертуар не содержит соматических мутация.
  b. простота верификации с помощью плазматического и ASC репертуаров.



Результаты

- **Данные:** три репертуара - наивный, ASC, плазматический. Каждый определяется типом клеток, из которых он состоит.
- **Пути решения:** построение базы данных (БД) из наивного репертуара антител.
- **Причины:** a. наивный репертуар не содержит соматических мутация.
  b. простота верификации с помощью плазматического и ASC репертуаров.



Результаты. Поиск новых V сегментов

GERMLINE

DB1

(Стандартная БД для мышей C57BL/6, 129S1/SvImJ)

DB2

- На основе наивного репертуара Balb/C
- Идея: наивная база данных по многократно повторяющимся V сегментам

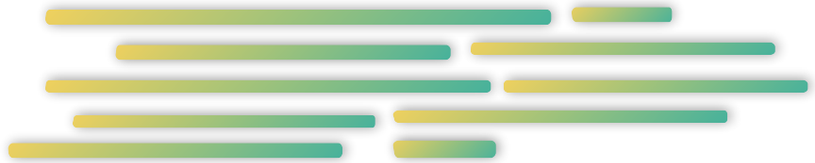
DB3

- На основе наивного репертуара Balb/C
- Идея: предварительная кластеризация для исправления ошибок секвенирования и ПЦР

Результаты. Поиск новых V сегментов

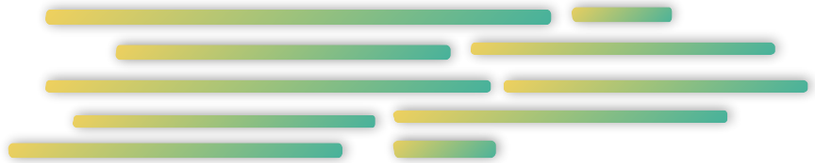
Результаты. Поиск новых V сегментов

1. *VJ Finder:* выравнивание

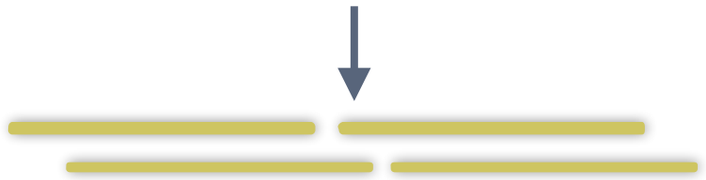


Результаты. Поиск новых V сегментов

1. *VJ Finder:* выравнивание



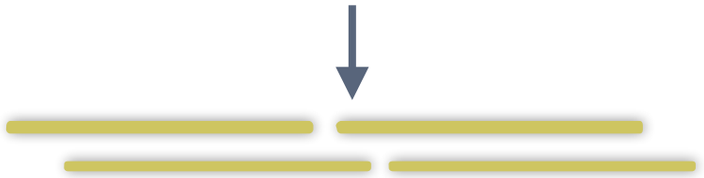
2. Вырезание V сегментов



Результаты. Поиск новых V сегментов

1. *VJ Finder:* выравнивание

2. Вырезание V сегментов



3. Построение БД

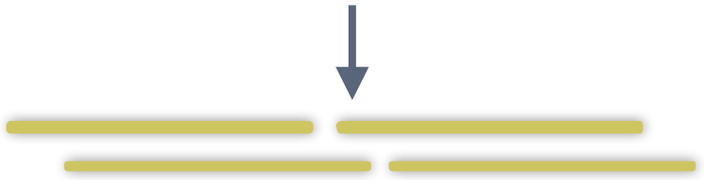
DB2

DB3

Результаты. Поиск новых V сегментов

1. *VJ Finder:* выравнивание

2. Вырезание V сегментов



3. Построение БД

DB2

a. *Tree compressor:* сжатие ридов

CTGGTAAG

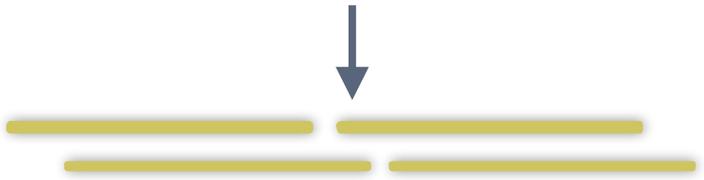
CTGGTACG → CTGGTAAG

DB3

Результаты. Поиск новых V сегментов

1. *VJ Finder:* выравнивание

2. Вырезание V сегментов



3. Построение БД

DB2

a. *Tree compressor:* сжатие ридов

CTGGTAAG

CTGGTACG → CTGGTAAG

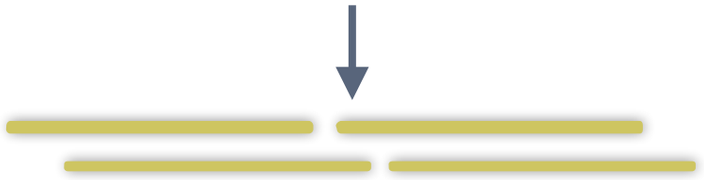
b. Формирование DB2. Abundance > t

DB3

Результаты. Поиск новых V сегментов

1. *VJ Finder:* выравнивание

2. Вырезание V сегментов



3. Построение БД

DB2

a. *Tree compressor:* сжатие ридов

CTGGTAAG

CTGGTACG → CTGGTAAG

b. Формирование DB2. Abundance > t

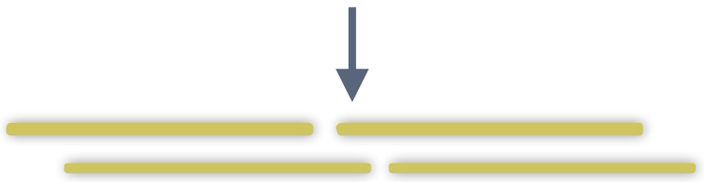
DB3

a. Построение Хэмминг графа
b. Поиск плотных подграфов
c. *ClustalW:* множественное выравнивание V сегментов в подграфе
d. *Single linkage:* кластеризация ридов в одном подграфе
e. Формирование DB3. Построение для каждого кластера консенсуса

Результаты. Поиск новых V сегментов

1. *VJ Finder:* выравнивание

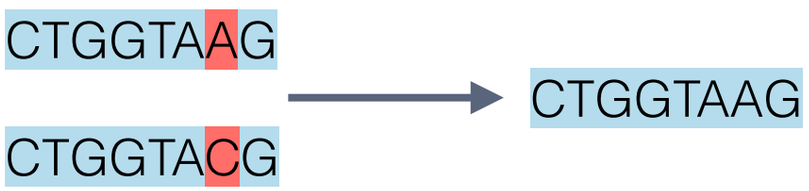
2. Вырезание V сегментов



3. Построение БД

DB2

a. *Tree compressor:* сжатие ридов



b. Формирование DB2. Abundance > t

Проблема: чувствителен к ошибкам в начале

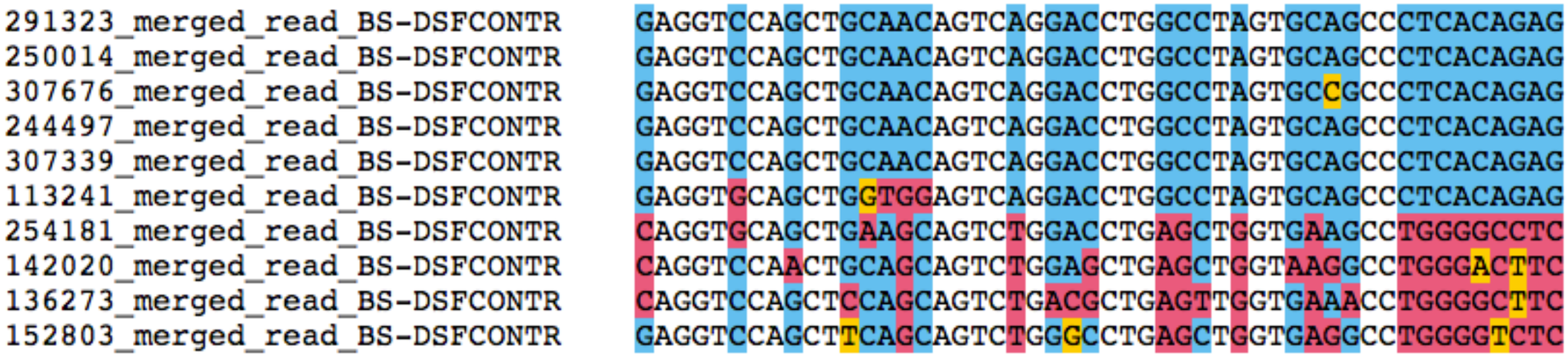
DB3

a. Построение Хэмминг графа
b. Поиск плотных подграфов
c. *ClustalW:* множественное выравнивание V сегментов в подграфе
d. *Single linkage:* кластеризация ридов в одном подграфе
e. Формирование DB3. Построение для каждого кластера консенсуса

Проблема: подбор параметров

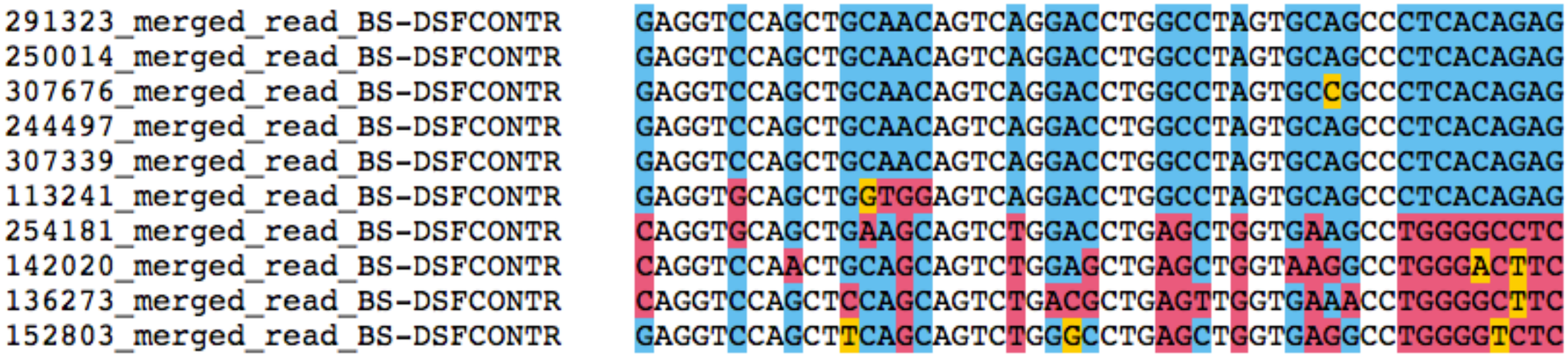
Результаты. Поиск новых V сегментов. DB3

ClustalW

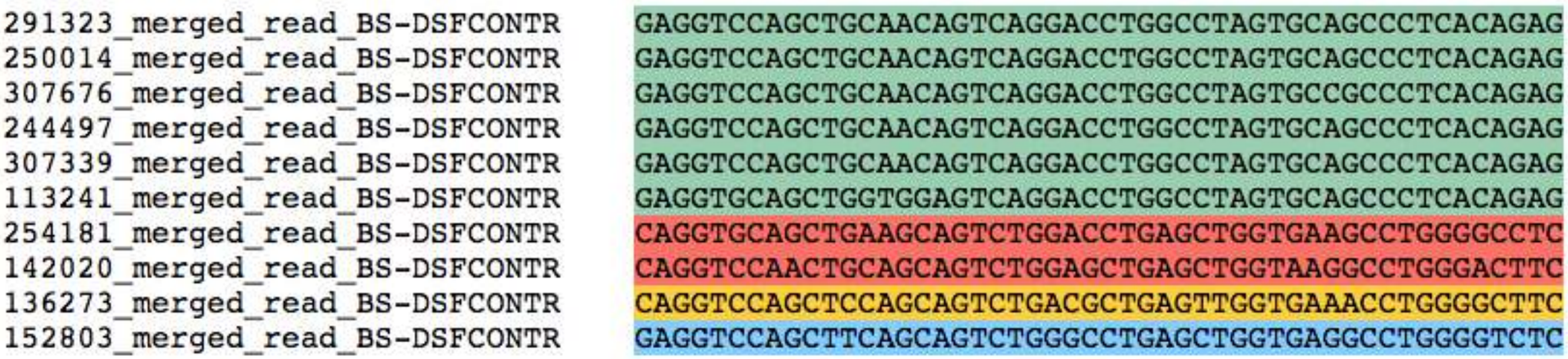


Результаты. Поиск новых V сегментов. DB3

ClustalW



Single Linkage, t = 0.2



Результаты. Сравнение БД

Результаты. Сравнение БД

ASC репертуар

	Mean	Median	Max	Min	% align
DB1	0.9381	0.95	1	0.22	98.32
DB2	0.9834	0.99	1	0.34	98.32
DB3	0.9841	0.97	1	0.31	98.32

Результаты. Сравнение БД

ASC репертуар

	Mean	Median	Max	Min	% align
DB1	0.9381	0.95	1	0.22	98.32
DB2	0.9834	0.99	1	0.34	98.32
DB3	0.9841	0.97	1	0.31	98.32

Плазматический репертуар

	Mean	Median	Max	Min	% align
DB1	0.9437	0.95	1	0.17	97.13
DB2	0.9881	0.99	1	0.19	97.15
DB3	0.9860	0.97	1	0.19	98.22

Выводы

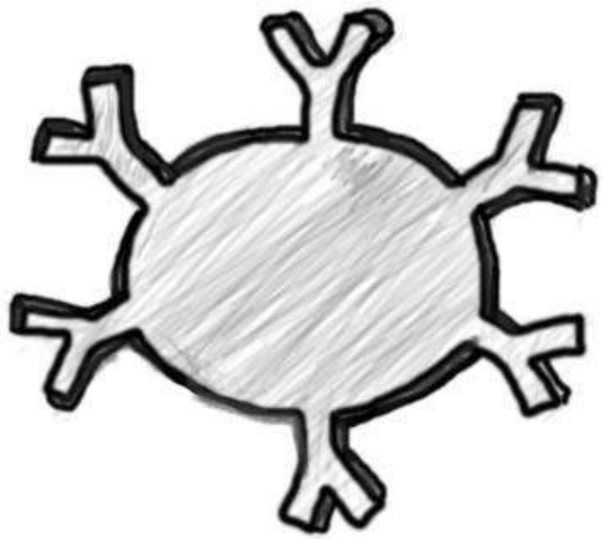
- При выравнивании ASC и плазматического репертуаров на базы данных, построенных на основе наивных сиквенсов, дают больше точности, чем выравнивания на стандартную базу данных мышей.

Выводы

- При выравнивании ASC и плазматического репертуаров на базы данных, построенных на основе наивных сиквенсов, дают больше точности, чем выравнивания на стандартную базу данных мышей.

Планы

- Оптимизация подхода построения баз данных.
- Тестирование на данных с неизвестными V, D, J сегментами.



Спасибо за внимание!