Benchmarking tools for discovery of novel mice immunoglobulin V-gene alleles

Sofiya Asadchaya, Anton Katsuba, Cynthia Ronkowski

Introduction / Введение

There are a few tools for discovery of new immunoglobulin alleles, but they generally assume that the reference allele databases are mostly complete. We are interested in finding out whether they would perform well if supplied with incomplete databases for underrepresented organisms. This tests the situation with less-researched animals, in particular some mice strains. The focus of this work was a tool called TIgGER.

Существует несколько инструментов для обнаружения новых аллелей иммуноглобулинов, но они обычно предполагают, что эталонные базы данных аллелей в основном полные. Нам интересно выяснить, будут ли они хорошо работать, если их снабдить неполными базами с данными о малопредставленных организмах. Это имитирует ситуацию с менее изученными организмами, в частности, с некоторыми линиями мышей. В центре внимания этой работы был инструмент под названием TIgGER.

V, D and J alleles sequences for a certain mice strains were taken from the IMGT database.

Последовательности V, D и J аллелей определенных линий мышей были взяты с базы данных IMGT.

Original V(D)J database

Custom Python script was developed to randomly delete some percentage of all the sequences, specified by the user.

Был написан кастомный код на Питоне, позволяющий случайно удалить заданный пользователем процент от всех последоовательностей.

Random deletion of some VDJ alleles

ImmuneSIM is a tool which simulates V(D)J recombination and produces recombined sequences given a database of existing germline genes.

ImmuneSIM - это инструмент, который моделирует V(D)J рекомбинацию и производит рекомбинированные последовательности, беря за основу базу данных существующих зародышевых генов.

ImmuneSIM

IgBLAST is an alignment tool tailored for work with immunoglobulin sequences. We chose to drop down to the level of pure IgBLAST since other tools which incorporated it did not provide enough control — we needed to build our own reduced reference databases for alignment.

IgBLAST - это инструмент выравнивания, адаптированный для работы с последовательностями иммуноглобулинов. Мы решили использовать IgBLAST напрямую, поскольку другие инструменты, включающие его, не обеспечивали достаточного контроля - нам нужно было создать собственные урезанные базы данных для выравнивания.

TIgGER analyzes mutation patterns in sequences considered to align to each germline allele in order to determine which positions might be polymorphic. It produces a list of putative novel alleles as its output.

TIgGER анализирует паттерны мутаций в последовательностях, которые считаются соответствующими каждому зародышевому аллелю, чтобы определить, какие позиции могут быть полиморфными. TIgGER выдаёт список предполагаемых новых аллелей в качестве своего результата.

TIgGER

ART is an NCBI-developed software for NGS sequencing read simulation. It allows to produce reads with qualities characteristic to specific platforms. In our case, we chose to simulate Illumina-family sequencer output.

ART - это разработанное NCBI программное обеспечение для моделирования ридов, получаемых при NGS-секвенировании. Оно позволяет получать риды с качествами, характерными для конкретных платформ. В нашем случае мы решили симулировать результаты, получаемые от секвенаторов семейства Illumina.

ART

IgBLAST

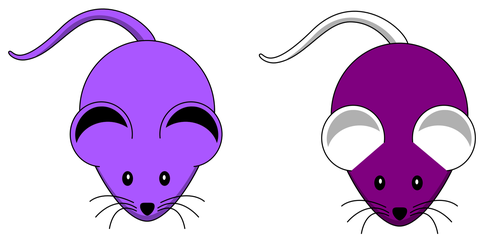
Performance analysis consists of calculating the precision and sensitivity of the tool. The number of true positive findings would have to be calculated.

Анализ эффективности заключается в расчете точности и чувствительности программы. Необходимо будет подсчитать количество истинно положительных результатов.

IGV (Integrative Genomics Viewer) is a tool which allows us to validate that read simulation has the desired properties: that they align well to the original sequences, have good coverage, etc.

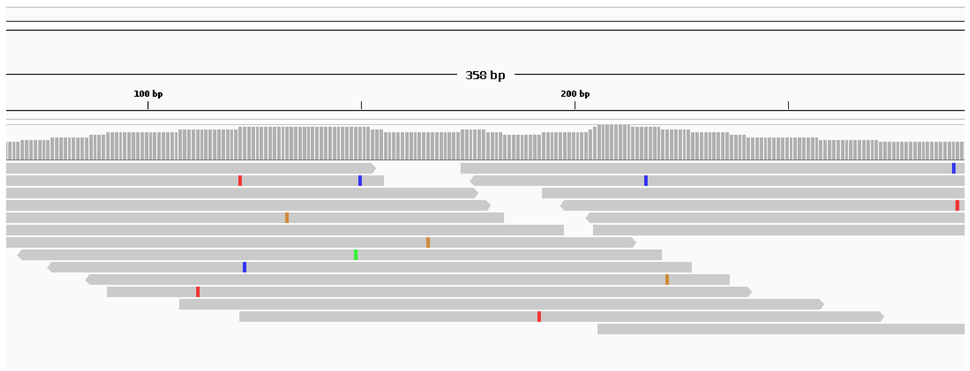
IGV (Integrative Genomics Viewer) - это инструмент, который позволяет нам проверить, что симулированные риды имеют желаемые свойства: что они хорошо выравниваются на исходные последовательности, имеют хорошее покрытие и т.д.

IGV

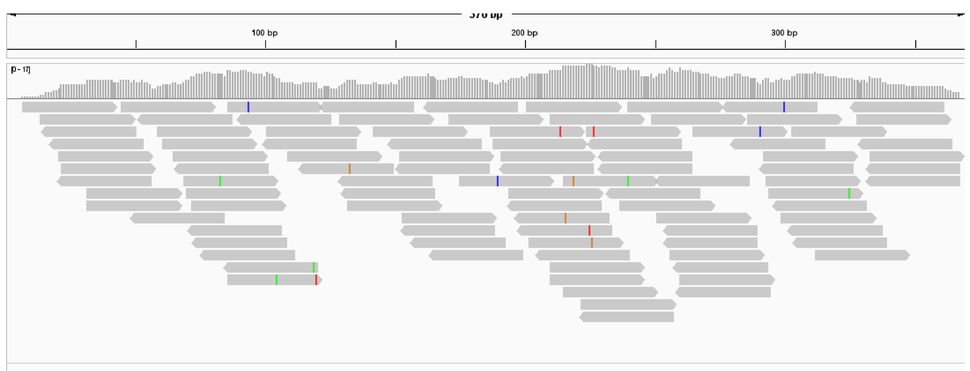


Performance analysis

IGV visualisation examples



Mouse strain BALB/c



Mouse strain A/J

Results/ Результаты

We assembled a pipeline for testing the performance of TIgGER. However, we could not retrieve any meaningful output from the tool: it seems that there was not enough different J genes in the database, and TigGER internally tests the sequences for sufficient J diversity. Further investigation is required. Meanwhile, another tool like IgDiscover could be examined.

Мы создали пайплайн для тестирования работы TIgGER. Однако мы не смогли получить от этого инструмента никаких значимых результатов: похоже, что в базе данных недостаточно различных J-генов, а алгоритм TigGER проверяет последовательности на достаточное J-разнообразие. Требуется дальнейшее исследование. Тем временем можно рассмотреть другой инструмент, например, IgDiscover.

ZIMIN FOUNDATION

SMTB

hhmi Howard Hughes Medical Institute

TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632