ИНСТИТУТ
БИОИНФОРМАТИКИ

Анализ данных геномных секвенирований

Научный руководитель:
Геннадий Захаров
EPAM Systems, Lifesciencs department

Студенты:
Скитченко Ростислав
Черенкова Ксения
Федотова Евгения

Санкт-Петербург
2017

Цель проекта:

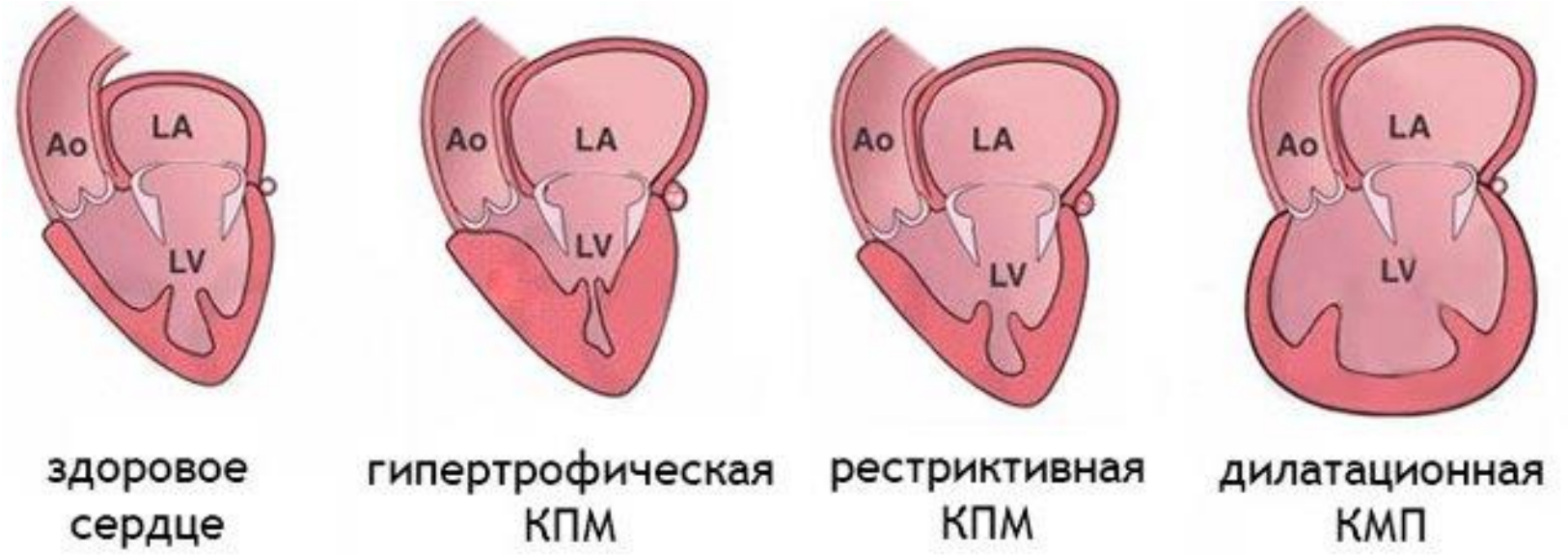
Сконструировать пайплайн для анализа данных, который помогал бы медикам находить клинически значимые вариации.

Задачи проекта:

- Обработка ДНК-сиквенсов, нахождение вариаций
- Фильтрация и ранжирование вариаций по значимости
- Разработка переносимой системы, которая делает полную обработку данных автоматически от начала и до конца

Исходные данные

- Результаты направленного сиквенса генов членов трёх семей – кардиоассоциированные гены. Наборы хранятся в формате FASTQ.

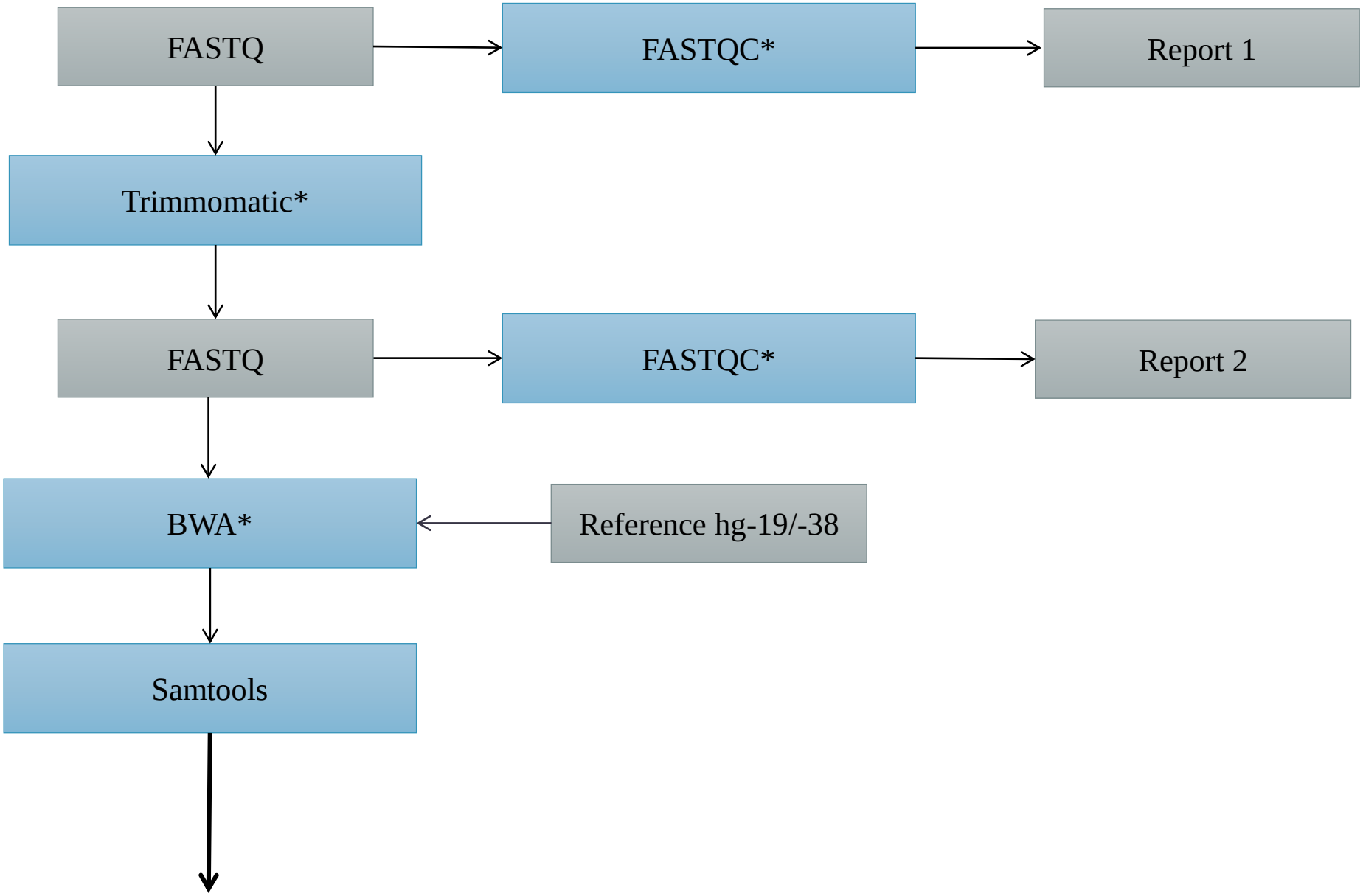


Используемые программы и ход работы

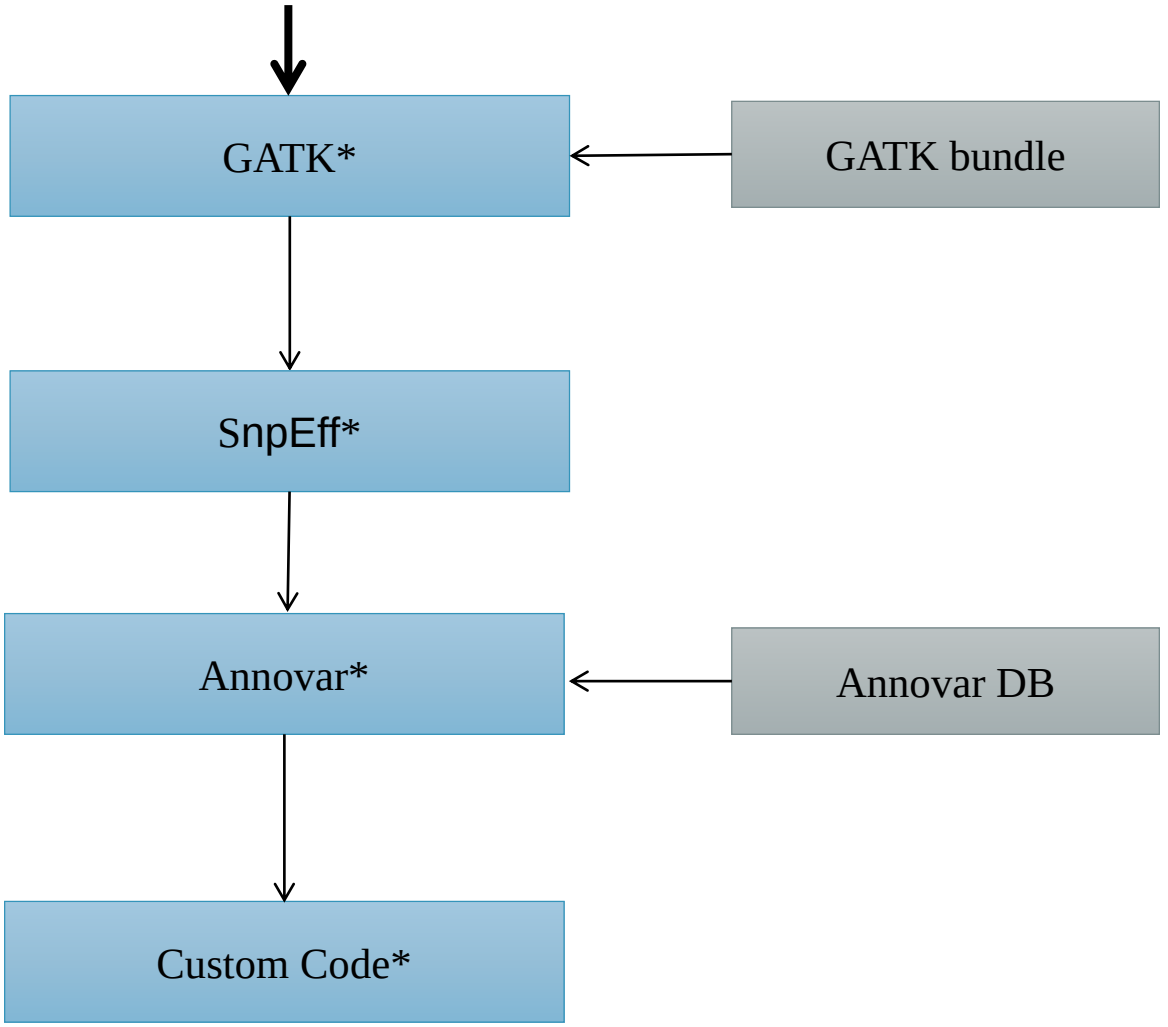
1. Fastqc (fastq)
2. Trimmomatic (fastq)
3. BWA (sam)
4. Samtools (bam)
5. Picard-tools mark duplicates (bam)
6. GATK (VCF)
7. SnpEff (VCF)
8. Annovar (VCF)

Custom code (txt)

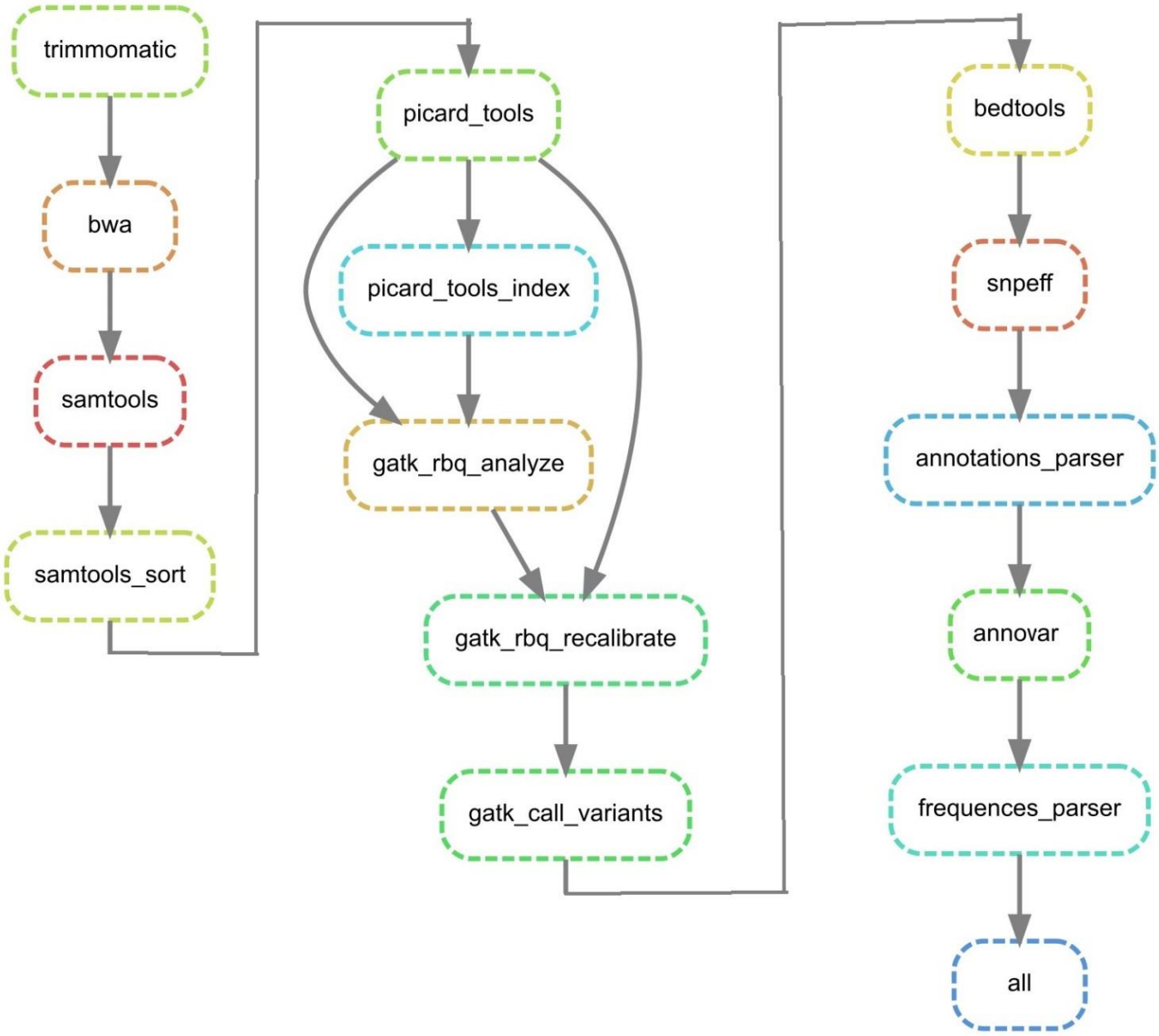
Общая схема анализа (Часть 1):



Общая схема анализа (Часть 2):



Инструменты для пайплайна: язык Snakemake

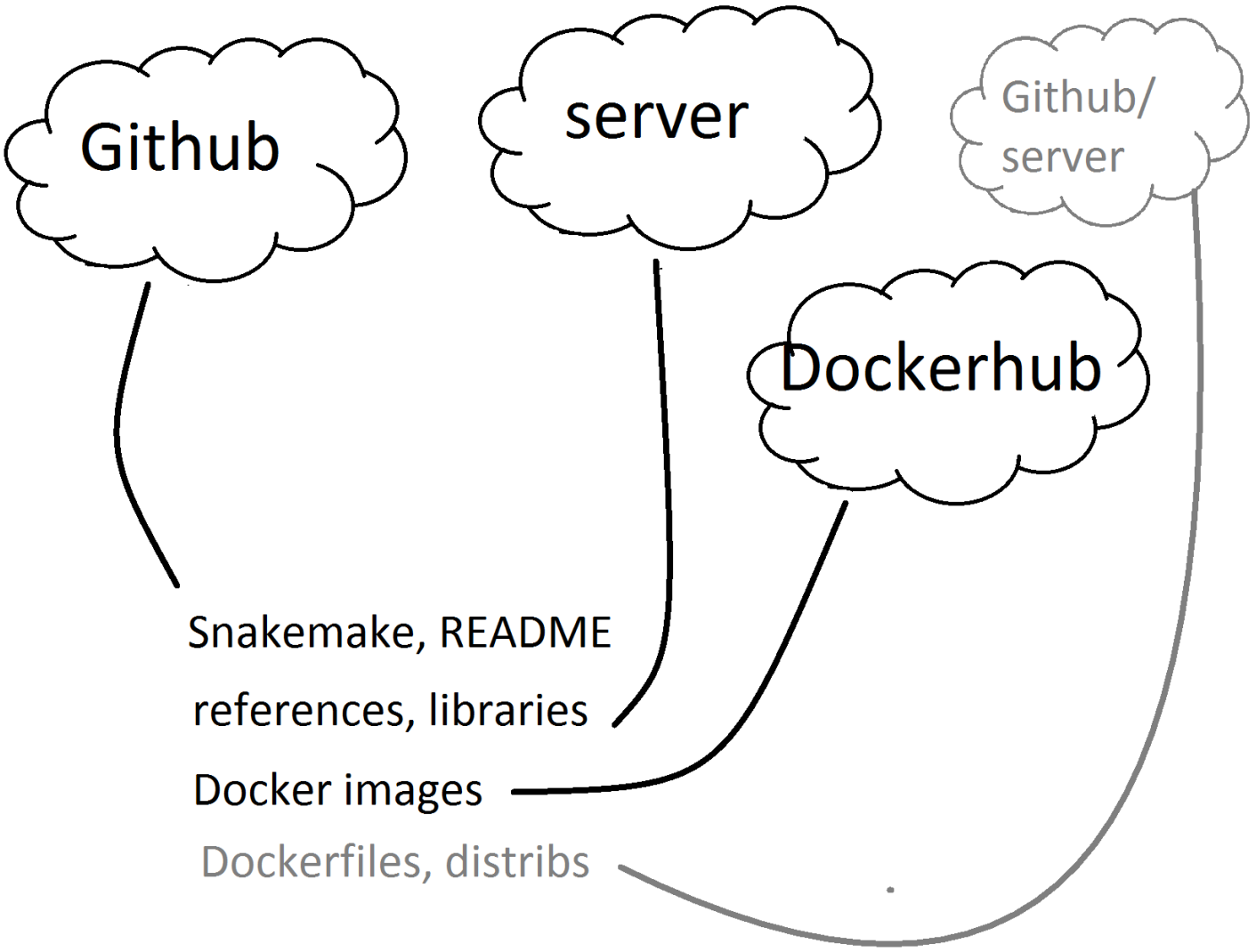


Инструменты для пайплайна: ПО Docker

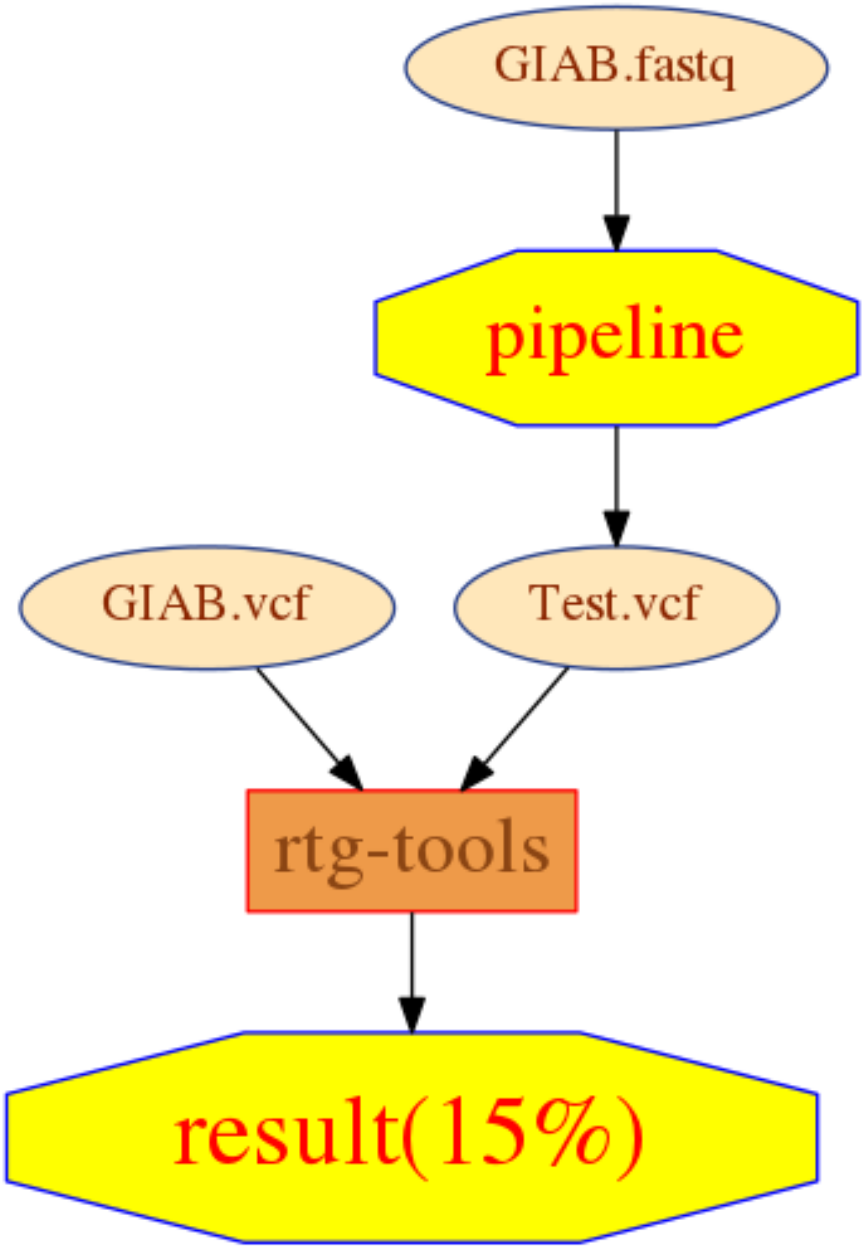
Сложности при работе с пайплайном:

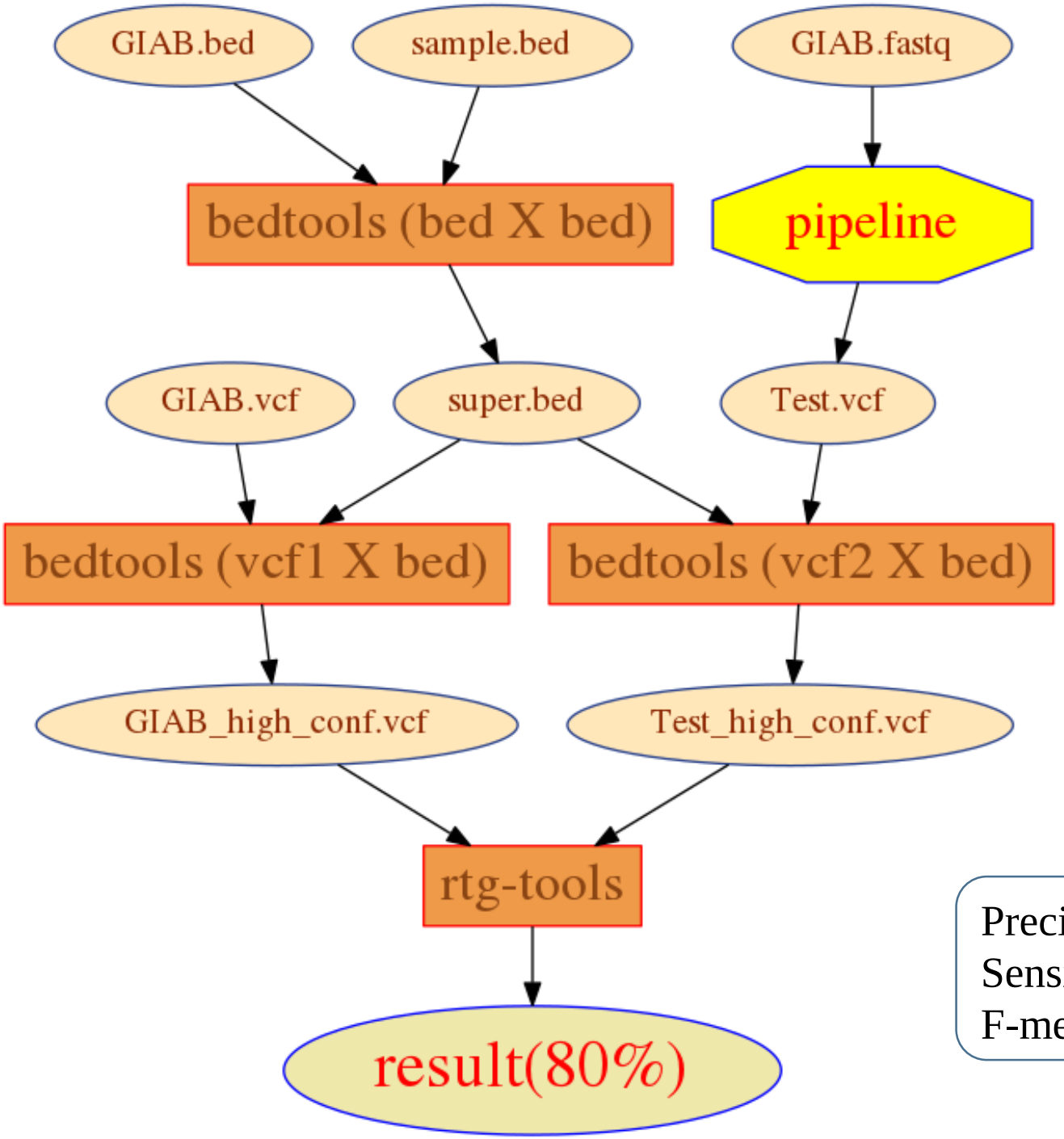
- Надо учитывать версии используемых программ;
- Надо учитывать окружение и зависимости используемых программ;
- Сложность перенесения и развертывания пайплайна на других компьютерах.

Развертывание пайплайна

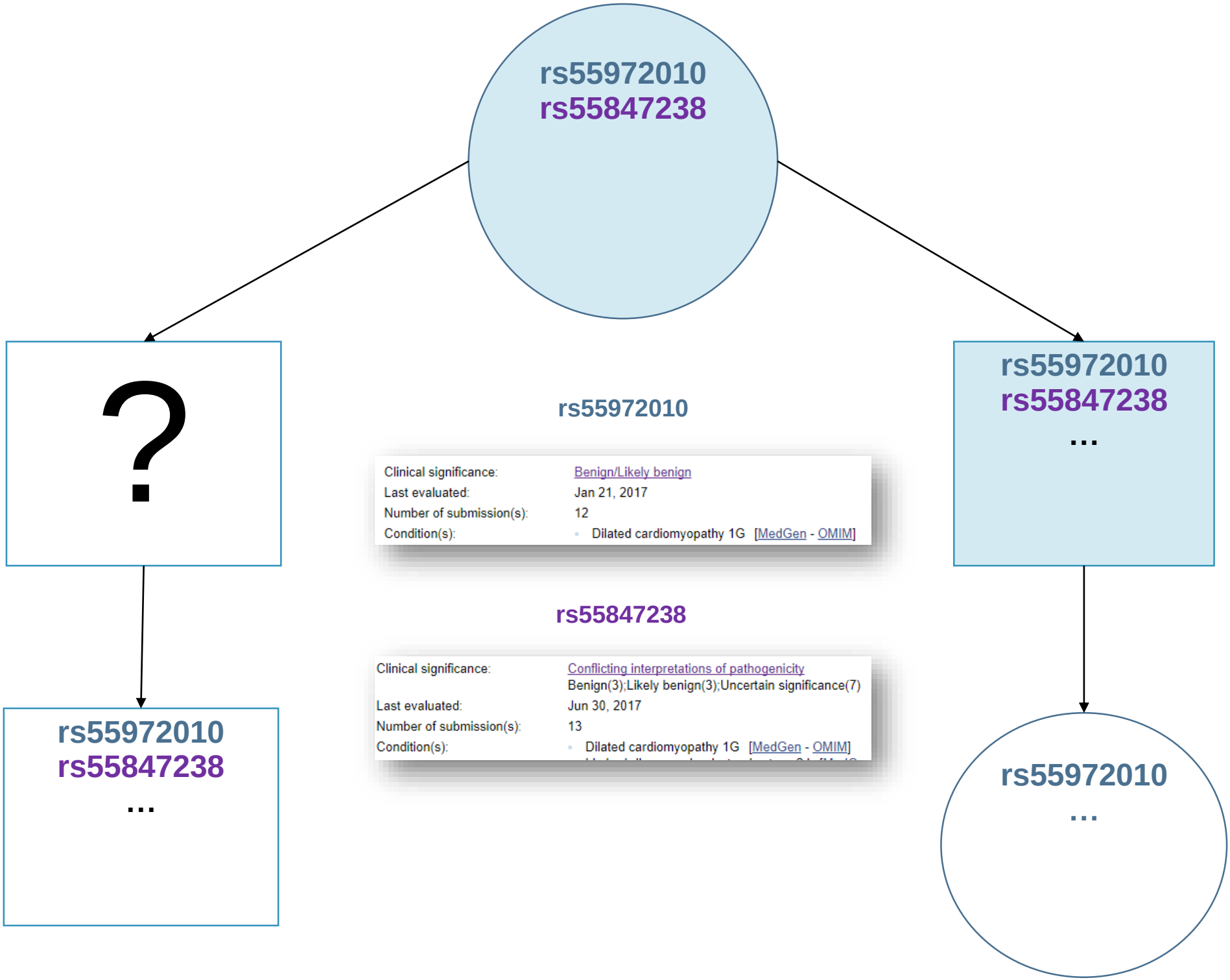


Алгоритм проверки пайплайна





Precision 0.9504
Sensitivity 0.7753
F-measure 0.8540


rs55972010
rs55847238
?
rs55972010
rs55847238
...
rs55972010
Clinical significance: Benign/Likely benign
Last evaluated: Jan 21, 2017
Number of submission(s): 12
Condition(s): Dilated cardiomyopathy 1G [MedGen - OMIM]
rs55847238
Clinical significance: Conflicting interpretations of pathogenicity
Benign(3);Likely benign(3);Uncertain significance(7)
Last evaluated: Jun 30, 2017
Number of submission(s): 13
Condition(s): Dilated cardiomyopathy 1G [MedGen - OMIM]
rs55972010
rs55847238
...
rs55972010
...

rs55972010	Dilated Cardiomyopathy, Dominant	2q31
rs55847238	Dilated Cardiomyopathy, Dominant	2q31
rs13087941	Hypertrophic cardiomyopathy	3p25
rs505058	Dilated Cardiomyopathy, Dominant	1q22
rs538089	Dilated Cardiomyopathy, Dominant	1q22

Низкая частота встречаемости в популяции
Высокая частота встречаемости в популяции