Санкт-Петербургский
политехнический университет
Петра Великого

Анализ мутационного ландшафта пациентов с хроническим миелолейкозом

Выполнила студентка гр. 63601/4: Терских А.В.
Руководитель: Самсонова М.Г., проф., д.б.н.

15.06.15

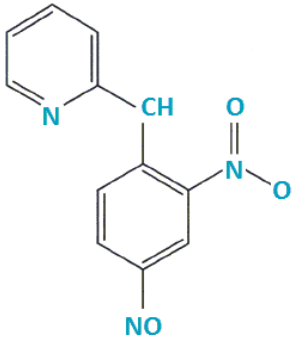
Хронический миелолейкоз: характеристика

4 случая на 100 000 человек в год

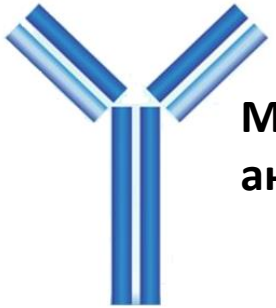
Средний возраст постановки диагноза – 72 года

заболевают в 1.5 – 2 раза чаще чем

Основные методы лечения:



Химиотерапия, радиотерапия

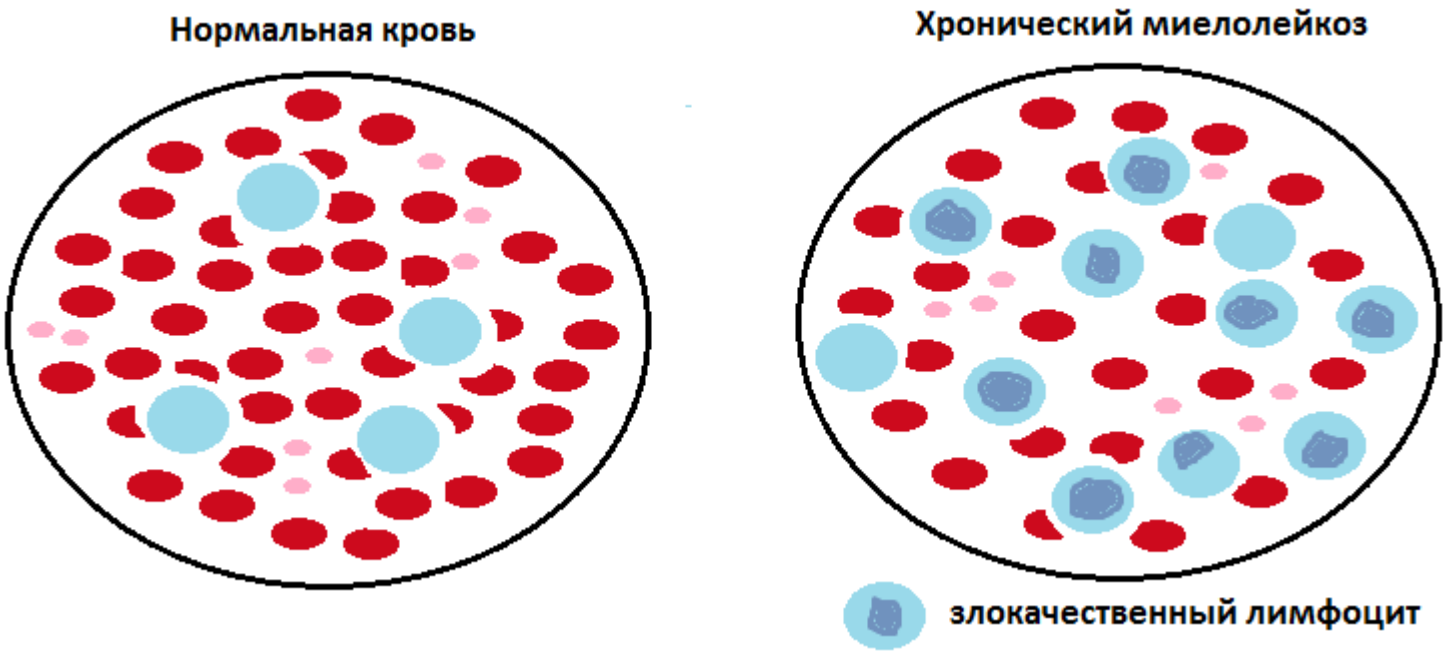


Моноклональные антитела

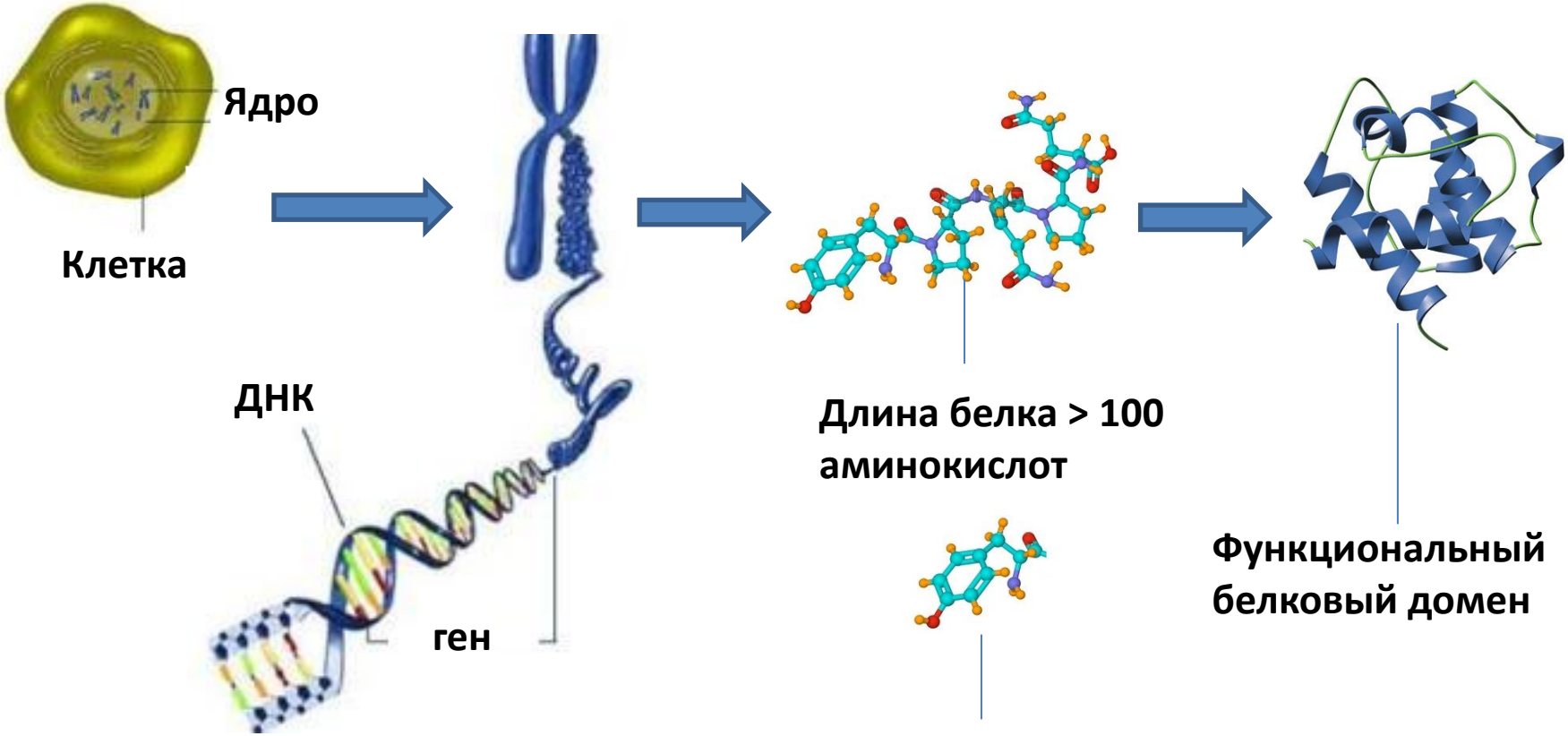
Трансплантация костного мозга и стволовых клеток

Хронический миелолейкоз: причина

Развитие клеток крови



Используемые термины



Цели и задачи анализа мутационного ландшафта

Цель

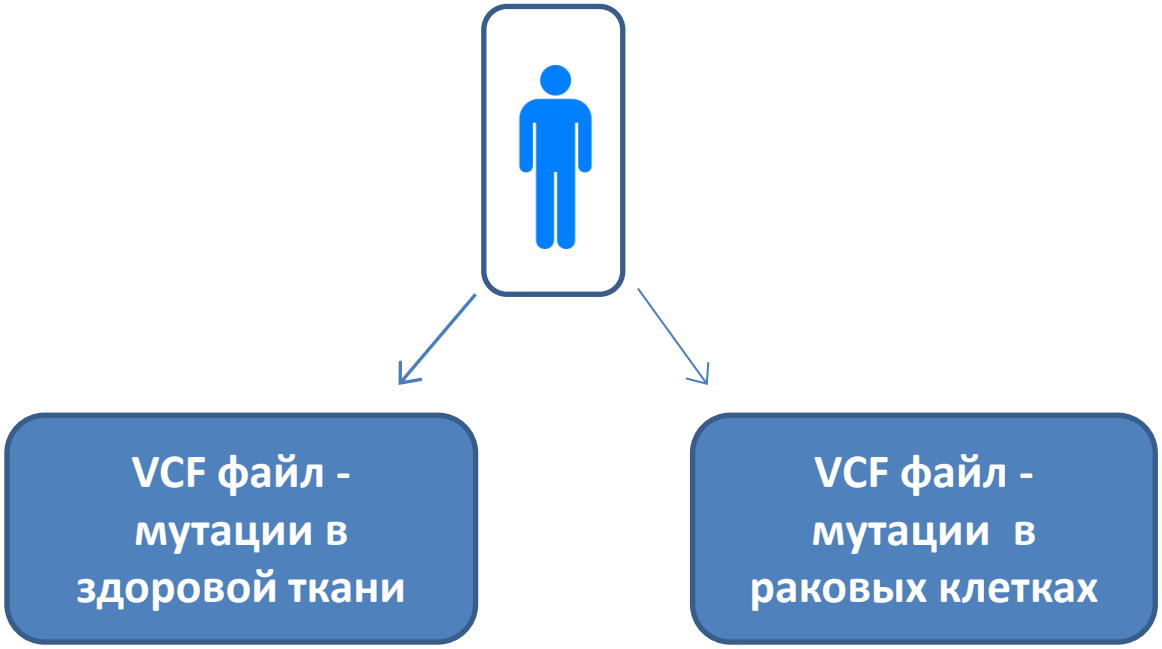
- **Определение белковых доменов, генов и клеточных сигнальных путей и процессов, которые влияют на ответ пациентов на лечение**

Задачи

- **Анализ мутаций в кодирующих областях**
  - **мутации в функциональных белковых доменах (PFAM)**
  - **мутации в коротких пептидах**

Исходные данные

Общее количество пациентов - 39

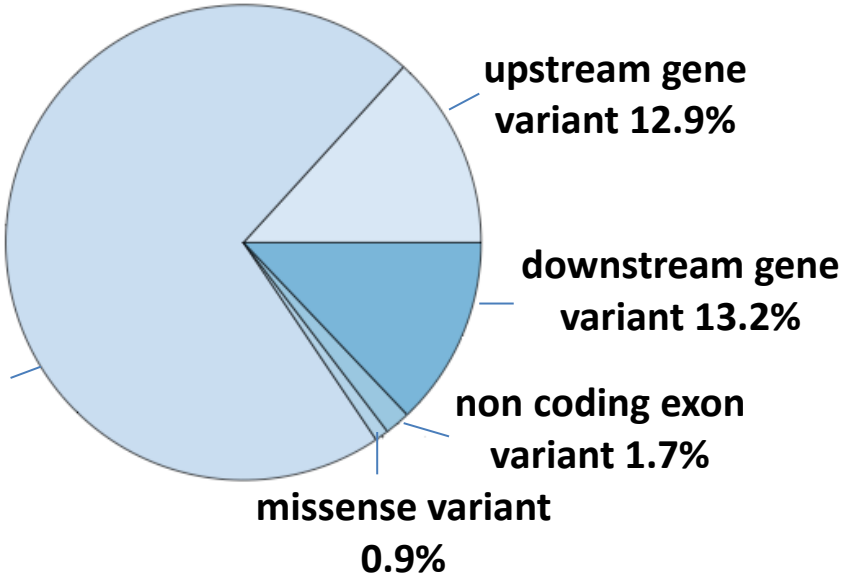


Данные предоставлены группой Dr. Anna Schuh, Оксфордский Университет, Великобритания

Этапы анализа

Анализ исходных данных

Набор программ на языке программирования R



База данных коротких пептидов. Общее количество коротких пептидов - 678

оценка PROVEAN ≤ - 2.5 повреждающая мутация

оценка PROVEAN > - 2.5 нейтральная мутация

Исходные данные для логистической модели

	Домен 1 (PF00038)	Домен 2 (PF00047)	…	Домен 183 (PF00091)	Ответ на лечение
Пациент 1					0
Пациент 2					1
Пациент 3					1
…					
…					
…					
Пациент 26					0

Частоты появления мутаций в каждом белковом домене

Ответ пациента на лечение

0 → Улучшение состояния

1 → Улучшения нет

Модель логистической регрессии с регуляризацией

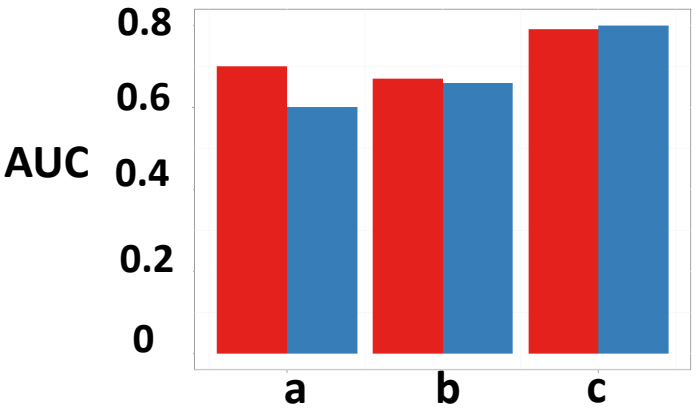
метод Лассо (a)

гребневая регрессия (регуляризатор Тихонова) (b)

эластичная сеть (c)

$$\min_{(\beta_0,\beta)\in\mathbb{R}^{n+1}} -\left[\frac{1}{m}\sum_{i=1}^{m} y_i\cdot\left(\beta_0 + x_i^T\beta\right) - \log\left(1 + e^{\left(\beta_0 + x_i^T\beta\right)}\right)\right] + \lambda\left(\frac{1-\alpha}{2}\|\beta\|_2^2 + \alpha\|\beta\|_1\right) \quad (1)$$

где λ – параметр регуляризации, α - параметр ∈ [0,1], β – вектор коэффициентов, *X* – матрица независимых переменных, *Y* – вектор зависимых переменных, n – количество независимых переменных (183), m – количество зависимых переменных (26).



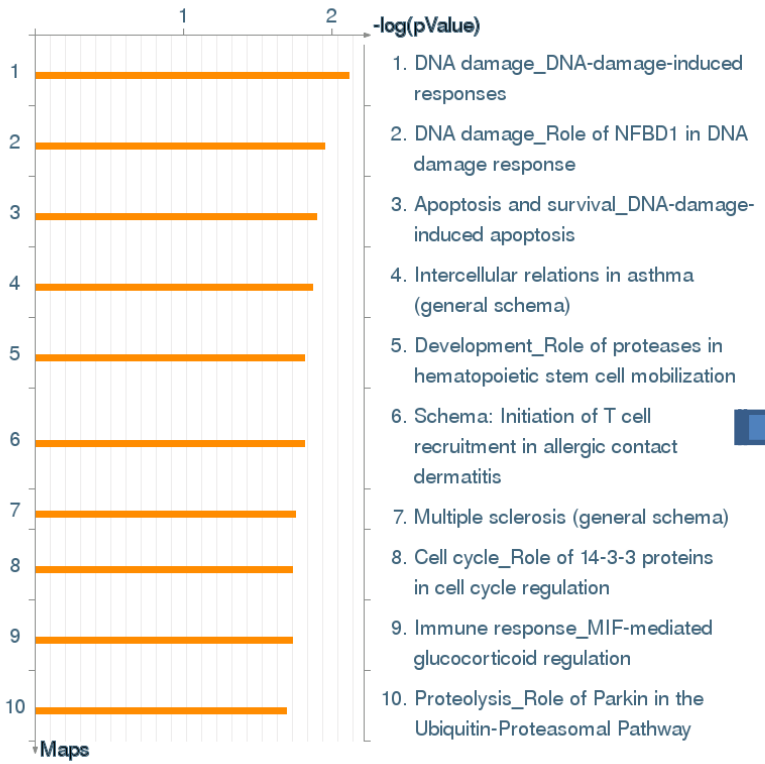
Кросс-валидация(CV)

Кросс-валидация с исключением (LOOCV)

Общее количество доменов: 13

Общее количество генов: 19

Биологические процессы, ассоциированные с мутировавшими доменами



Диаграмма, полученная с помощью ПО MetaCore

- повреждение ДНК
- апоптоз
- клеточный цикл
- имунный ответ
- нейрофизиологические процессы
- транспорт марганца
- транскрипция
- протеолиз
- клеточное взаимодействие лимфоцитов
- развитие хрящевой ткани
- цитоскелет

Гены, содержащие домены с мутациями

Гены, НЕ ассоциированные ранее с хроническим миелолейкозом (17)

Гены, ассоциированные с хроническим миелолейкозом (2)

ATM
NR4A1

KRT32
CD163L1
SLC22A9
SLC2A14
ARHGEF38
PLEKHG4B
TMC2

нет информации о заболеваниях (8)

VCAM1
DMBT1
ESRRA
THRA
ABCC10
MGAM
TUBB8

другие типы рака (6)

DSCAM
NELL1
PDE4DIP

другие заболевания (3)

Заключение

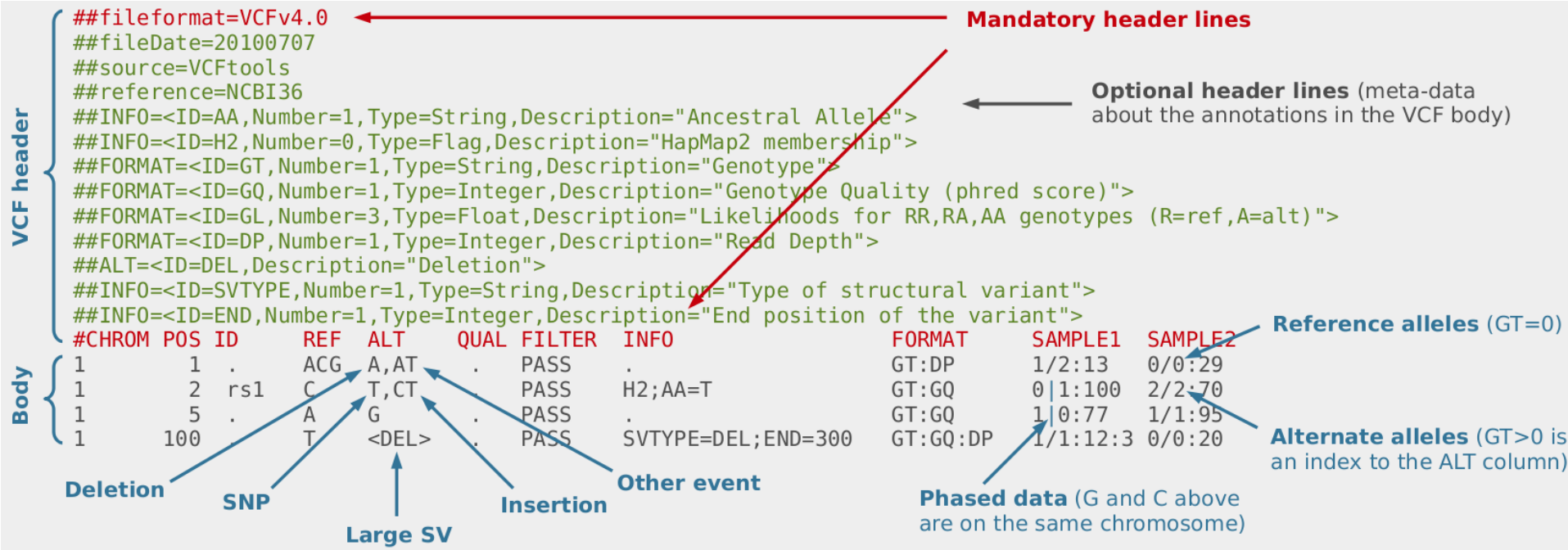
- разработана методика обработки геномных данных и создан набор программ на языке R
- определены домены, мутации в которых связаны с ответом на терапию
- определены сигнальные и метаболические пути, ассоциированные с этими доменами
- определены 17 новых генов, ранее не ассоциирующихся с хроническим миелолейкозом
- создана база данных коротких пептидов человека
- проведен первичный анализ последствий мутаций в коротких пептидах

Благодарности

- **University of Oxford**
  - **Dr. Alexander Kanapin**
  - **Anastasia Samsonova, Ph.D.**
  - **Dr. Christiana Kartsonaki**
  - **Dr. Anna Schuh**

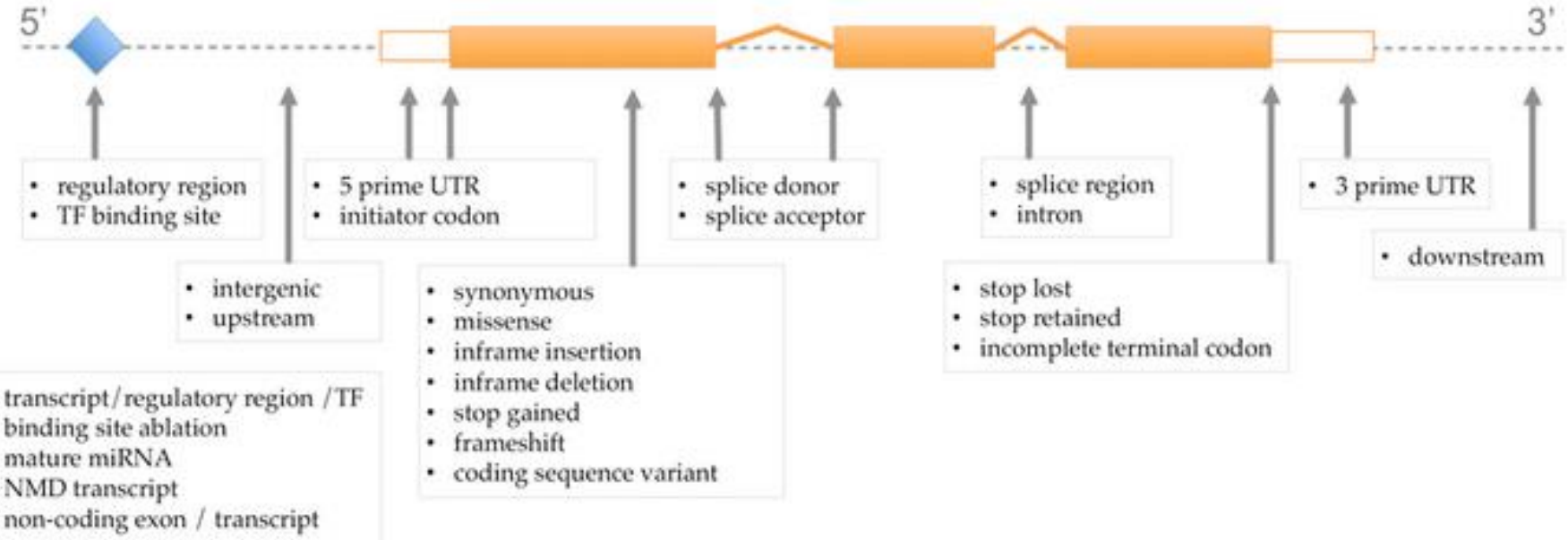
- **СПбГПУ**
  - **Самсонова М.Г., проф., д.б.н.**
  - **Кадырова Н.О., доц., к.т.н.**

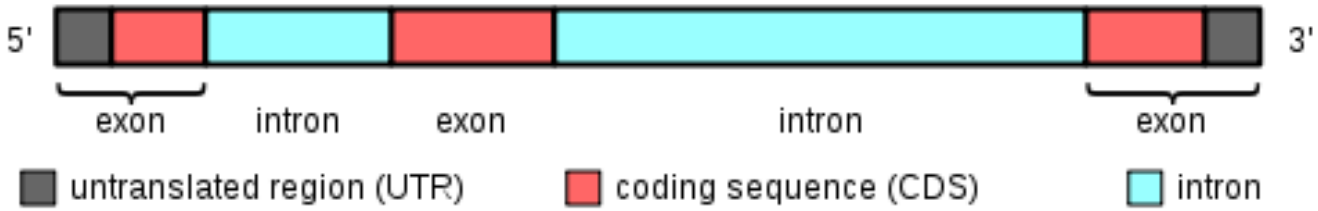
Variant Calling Format



Transcript structure

See below a diagram showing the location of each display term relative to the transcript structure:



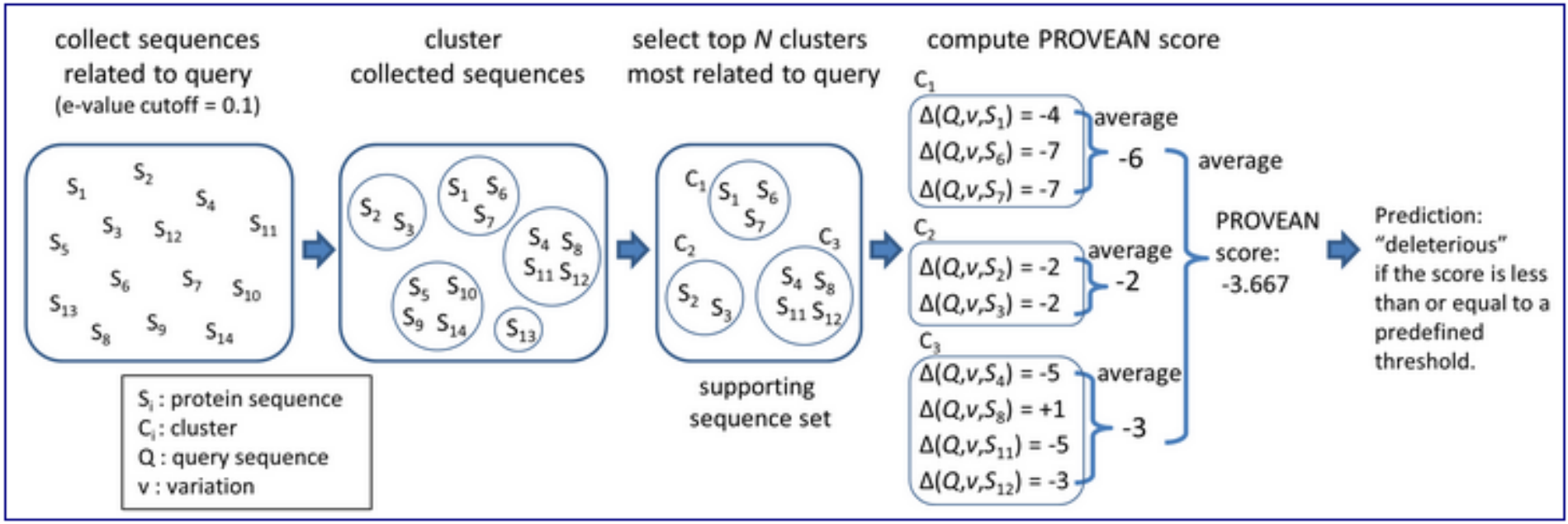


PROVEAN score

An overview of the PROVEAN procedure is shown in Figure 1. Briefly, clustering of BLAST hits is performed by CD-HIT with a parameter of 75% global sequence identity. The top 30 **clusters** of closely related sequences form the **supporting sequence set**, which will be used to generate the prediction.

A **delta alignment score** is computed for each supporting sequence. The scores are then averaged within and across clusters to generate the final PROVEAN score. If the PROVEAN score is equal to or below a predefined threshold (e.g. -2.5), the protein variant is predicted to have a "deleterious" effect. If the PROVEAN score is above the threshold, the variant is predicted to have a "neutral" effect.

Figure 1. Computing the PROVEAN score.



To increase the specificity of detection (i.e. find deleterious variants with higher confidence), a lower score threshold can also be used (e.g. -4.1). To increase the sensitivity of detection (i.e. find more deleterious variants by including those with lower confidence), a higher score threshold can also be used (e.g. -1.3). A summary of specificity and sensitivity obtained using different score cutoffs is shown in Figure 3.

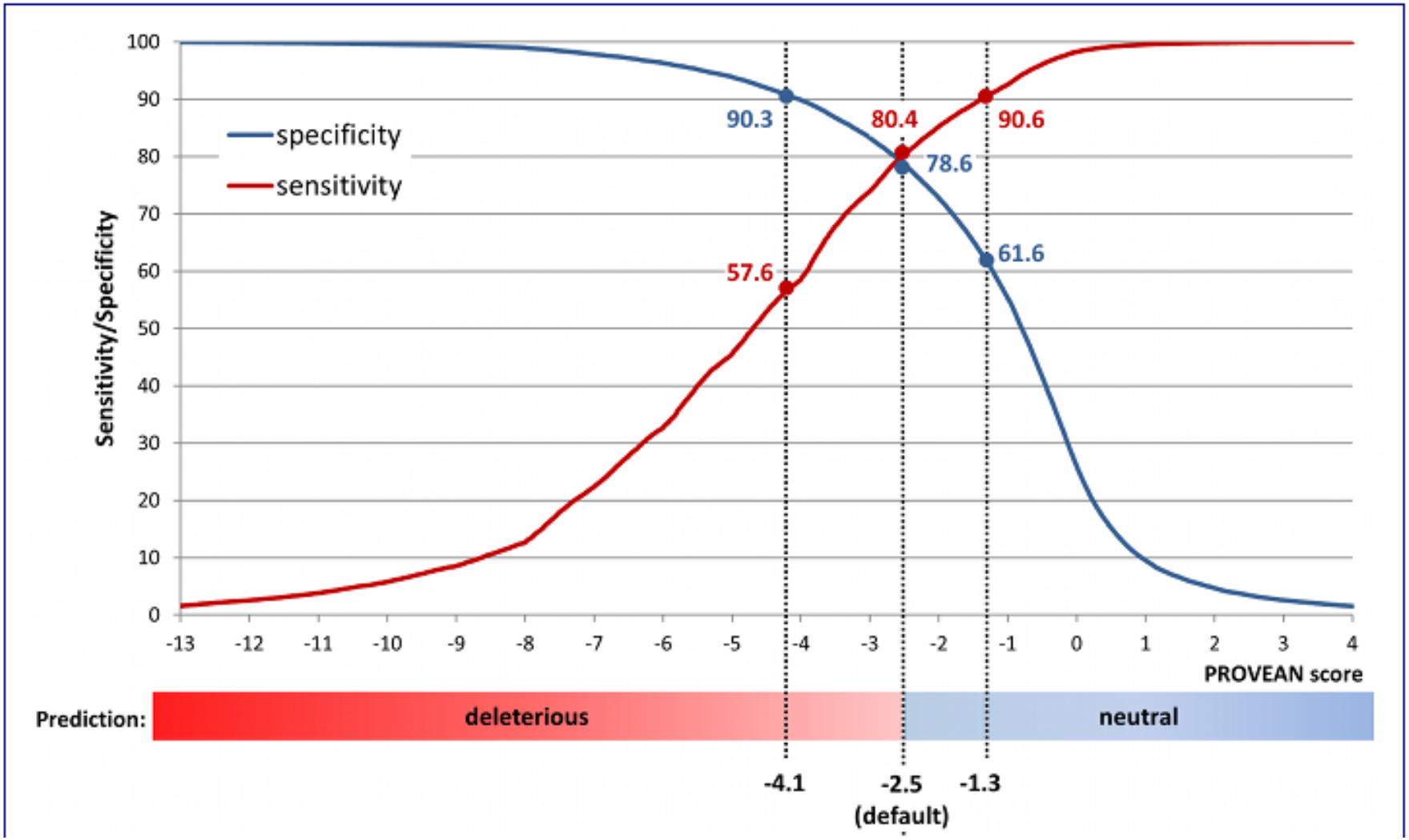



Figure 3. Sensitivity and specificity at different PROVEAN score cutoffs.