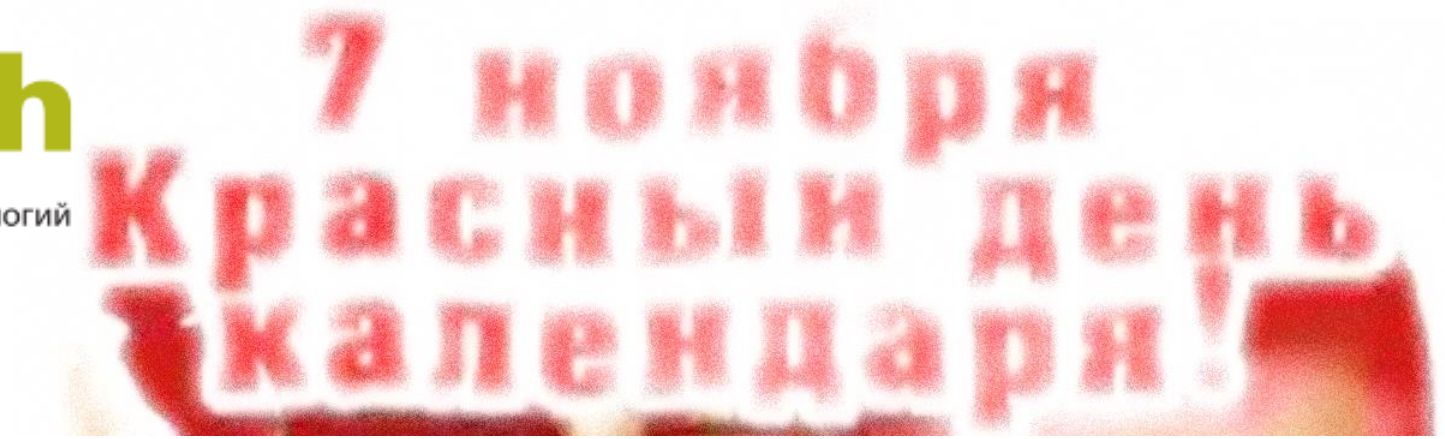
Skoltech

Сколковский институт науки и технологий



Типичные задачи биоинформатики

Михаил Гельфанд

PiterPy. 7 XI 2023

серёжа исаев
Кстати, а почему люди до сих пор работают в R, а не в Python? Если убрать фактор привычки

Установка пакетов в R (особенно каких-то нетривиальных) — это какое-то минное поле всегда у меня, игра “угадай, что пойдёт не так во время компиляции зависимостей”

1 16:45

Pavel Mazin Reply
ну по моему опыту для половины бионформатики R просто удобнее - тот же код только в два раза короче

16:46

zaira ✨
оч многие пакеты есть только в R том же, так что приходится. например, для cnv calling

16:46

Pavel Mazin
про установку пакетов не согласен - это вечное "попробовал установить что-то и сломал всю конду" бывает только с питоном

Sci Key
> **серёжа исаев**
> Кстати, а почему люди до сих пор работают в R, а не в Python...

Хз, почти все, кого я знаю, перекатились на питон в основном. А р юзают чисто по привычке и реже в 2 раза. 16:47

Pavel Mazin Reply
кроме объективных вещей (см выше) есть более индивидуальные - ноутбук - это каша из кода, картинок и аутпута, имхо, это ОЧЕНЬ сложно читать 17:02

серёжа исаев
ты через год сможешь хоть что то из скрипта понять?

1 17:02

я вот не могу, из своего же 17:02

Pavel Mazin
> **серёжа исаев**
> ты через год сможешь хоть что то из скрипта понять?

если это надо запомнить, то это надо записать в том или ином виде 17:03

Казанов Марат
еще имхо debug в R-studio гораздо проще, чем в любой python-овской оболочке. Выбираю R, но если честно не люблю оба языка :)

3 17:03

Fed Taratorkin
Думаю если ты крутой программист, то фактор привычки может быть более весомым, чем все остальное вместе взятое, если говнокодер, то большой разницы между R и питоном нет

17:01

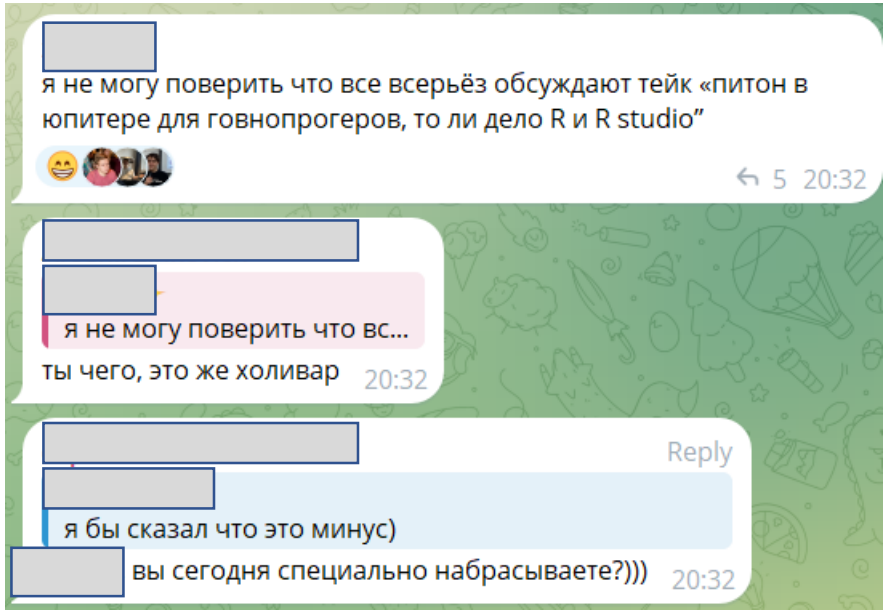
Pavel Mazin
вечер пятницы хорошее время для срача! 16:55

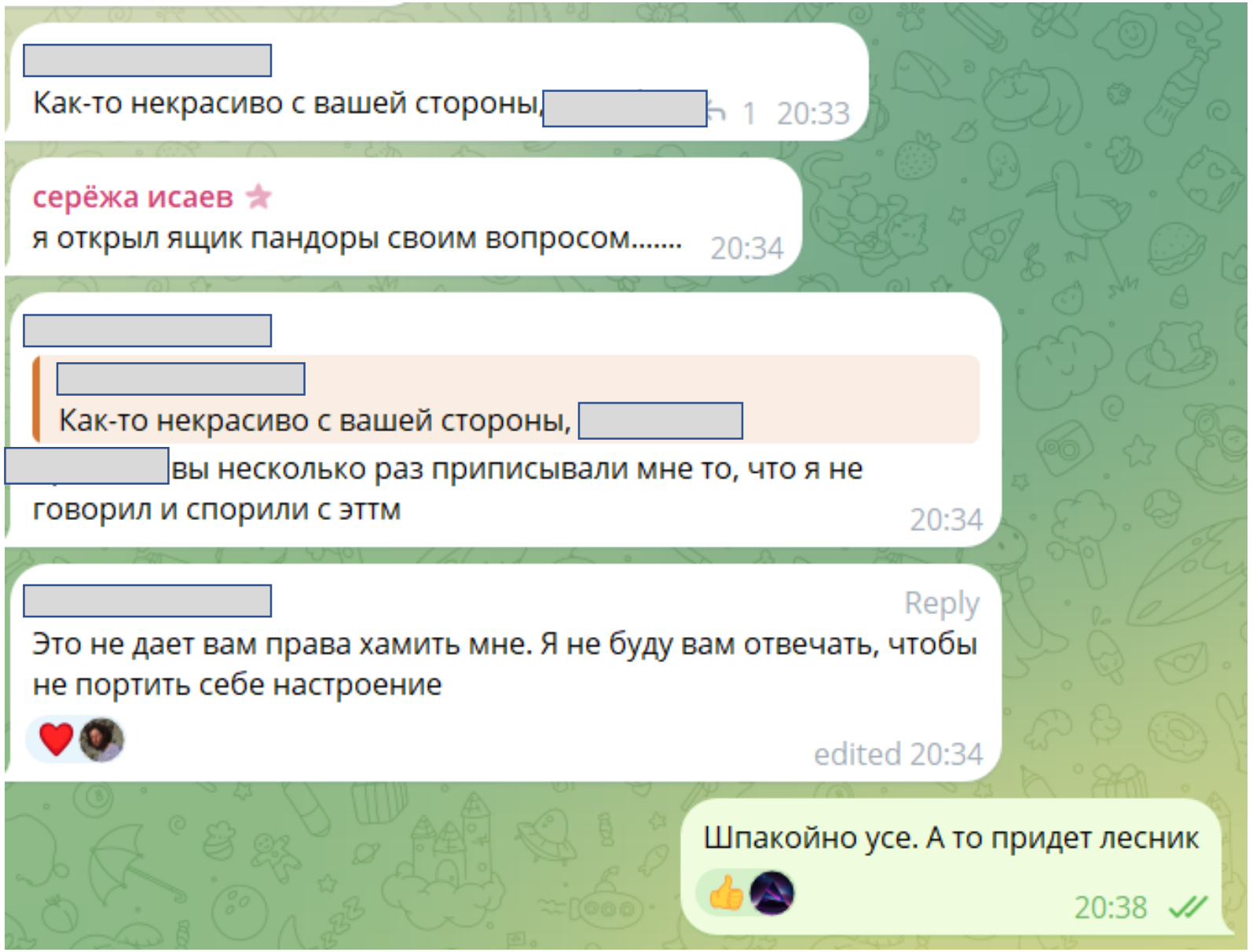
серёжа исаев
Когда я смотрю в чей-то ноутбук я хотя бы могу понять, что и в какой последовательности они делали

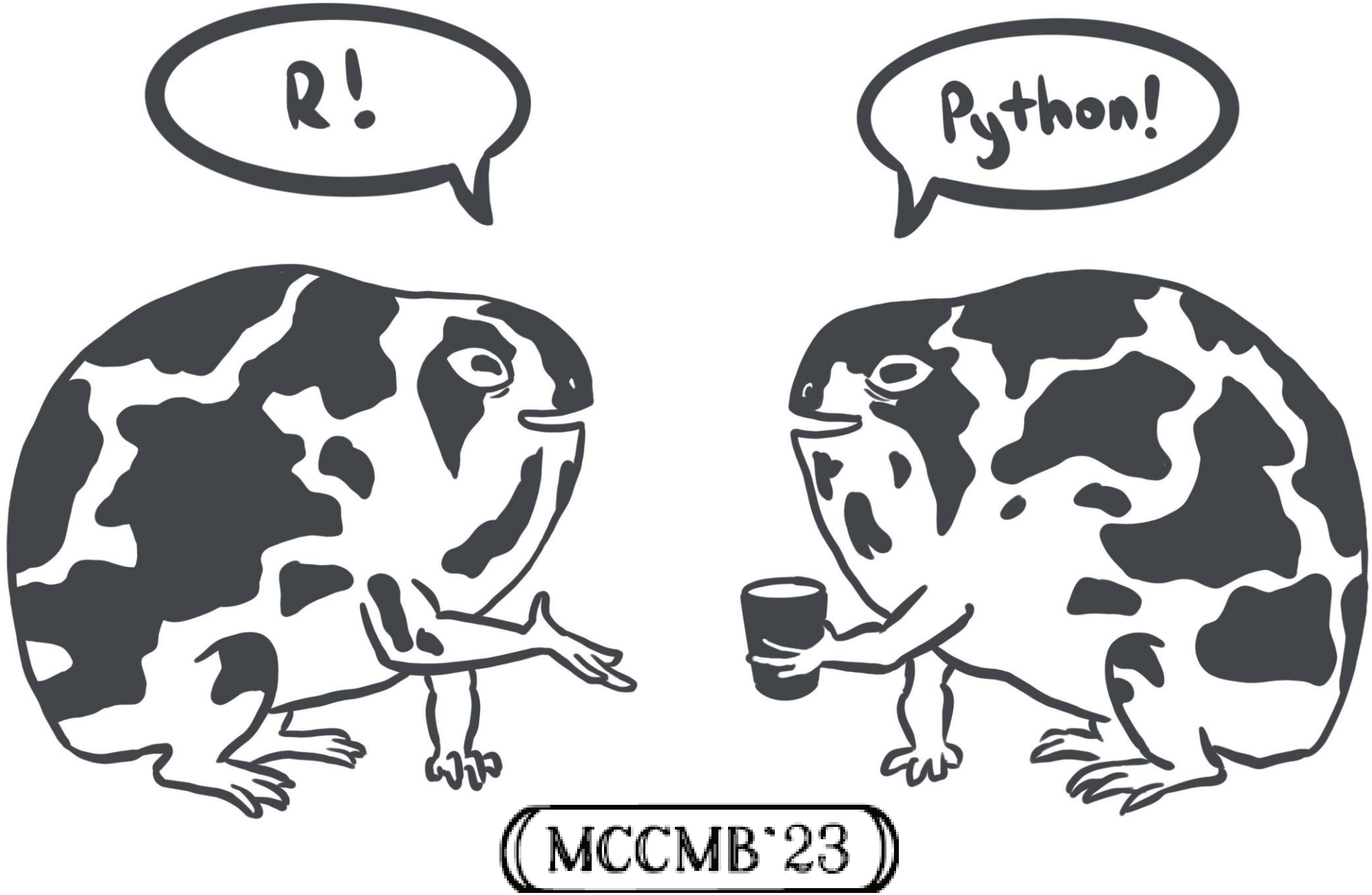
Когда я смотрю в чей-то R Script это обычно a huge mess, логика вообще не понятна

2 16:56

Прошло четыре часа...






R!
Python!
MCCMB`23


Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS


biopython

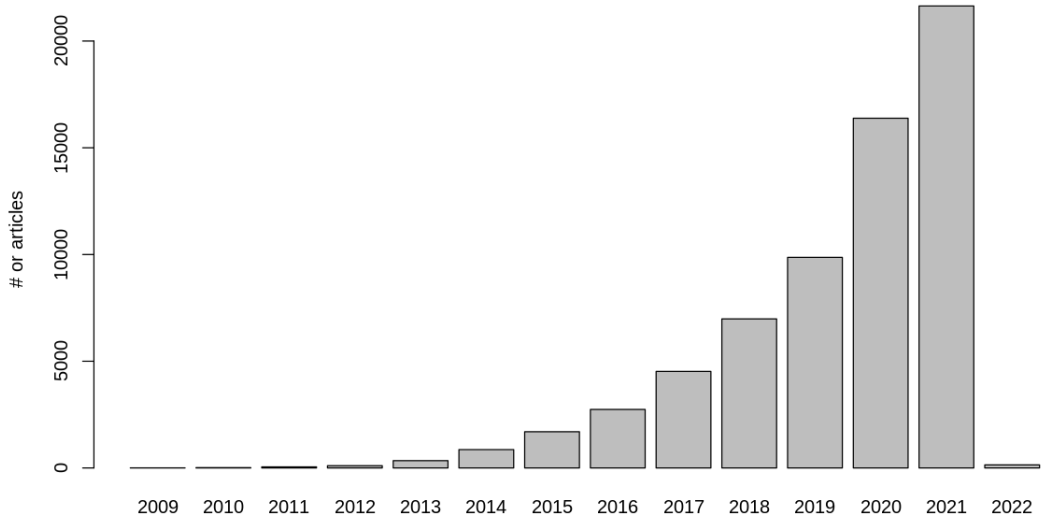
Один диссовет, 10 лет

		Python		
		Да	Нет	Всего
R	Да	3	9	12
	Нет	2	8*+14**	24
	Всего	5	31	36

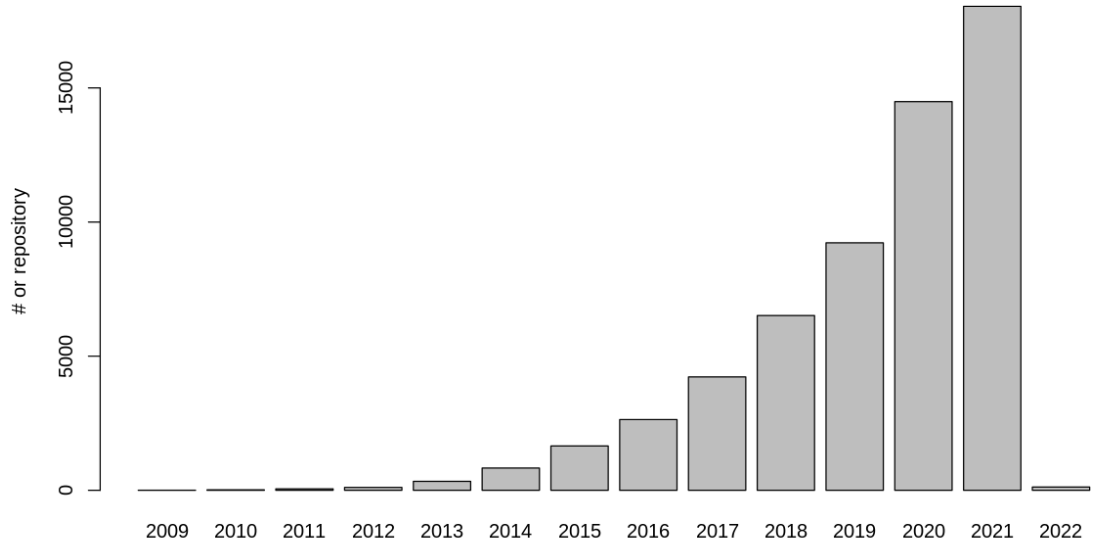
* – Java, Perl, bash, C++, Matlab.

** – Программы упомянуты, языки нет.

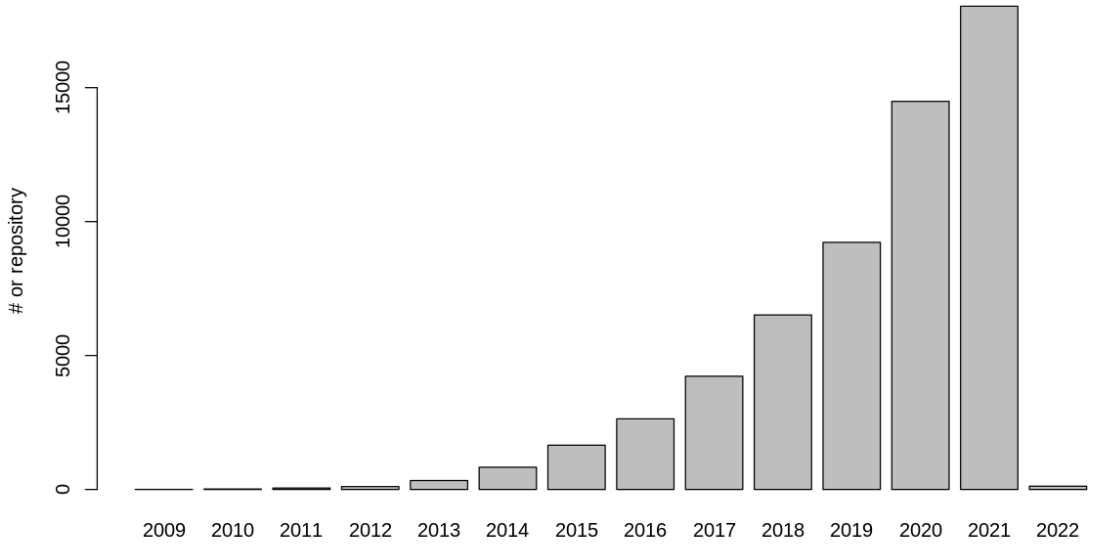
Github, связанный со статьями в PMC



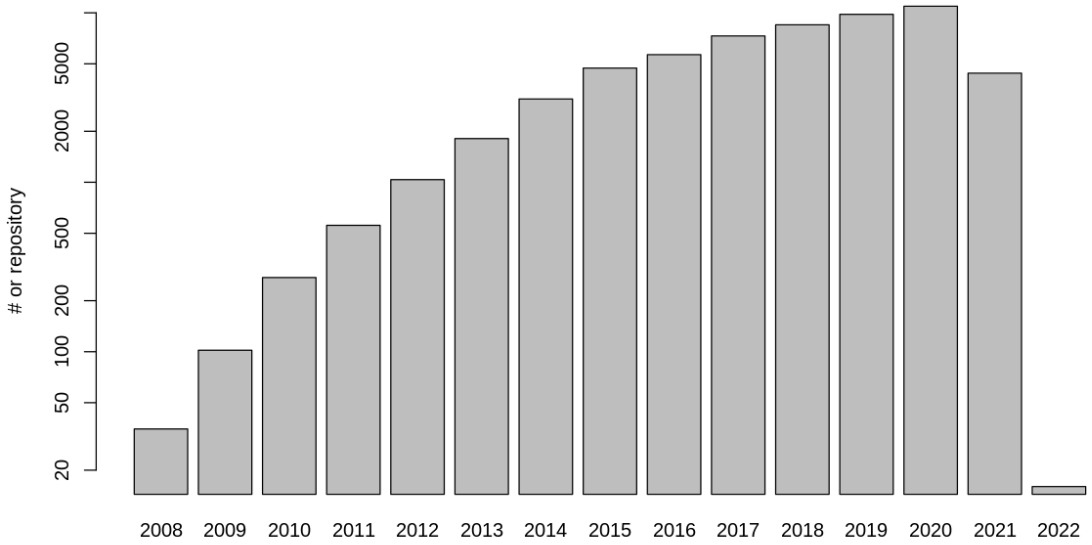
статей



депозиториев

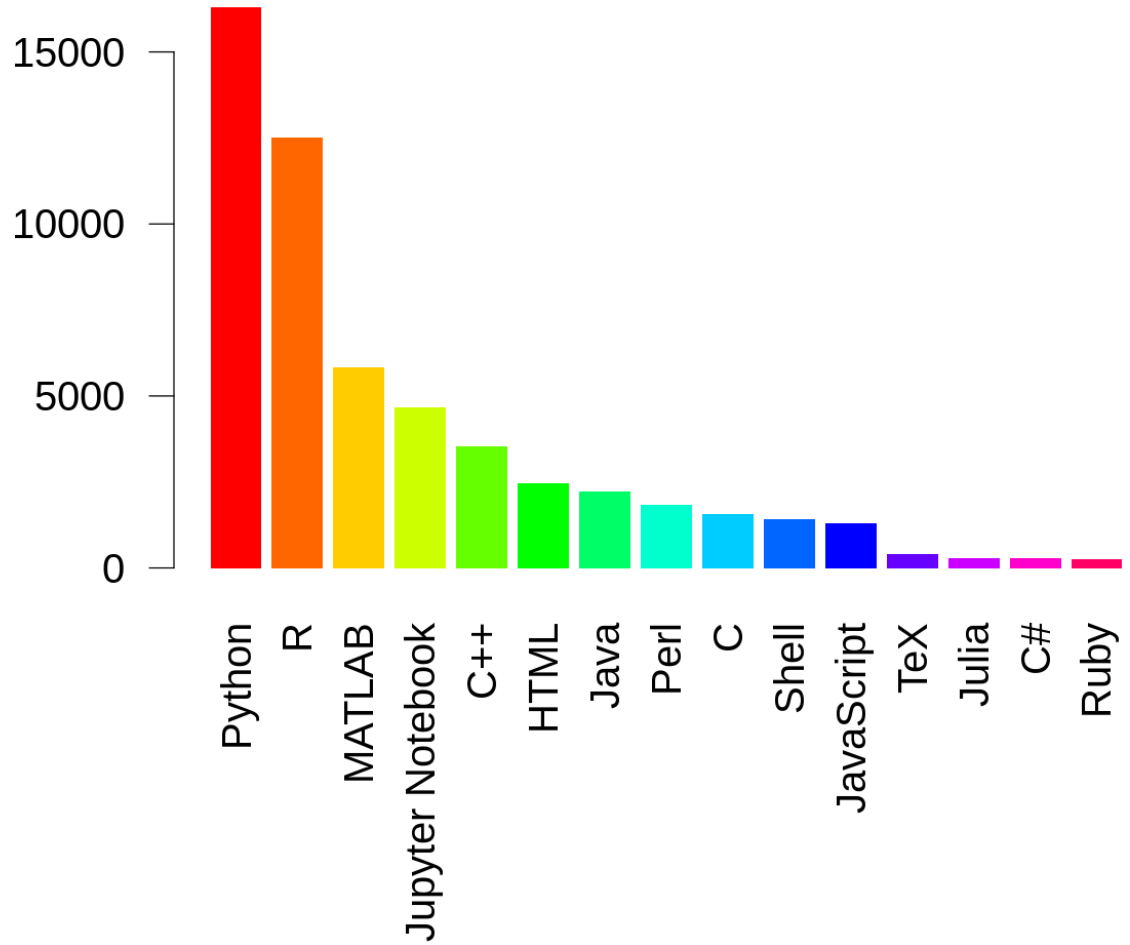


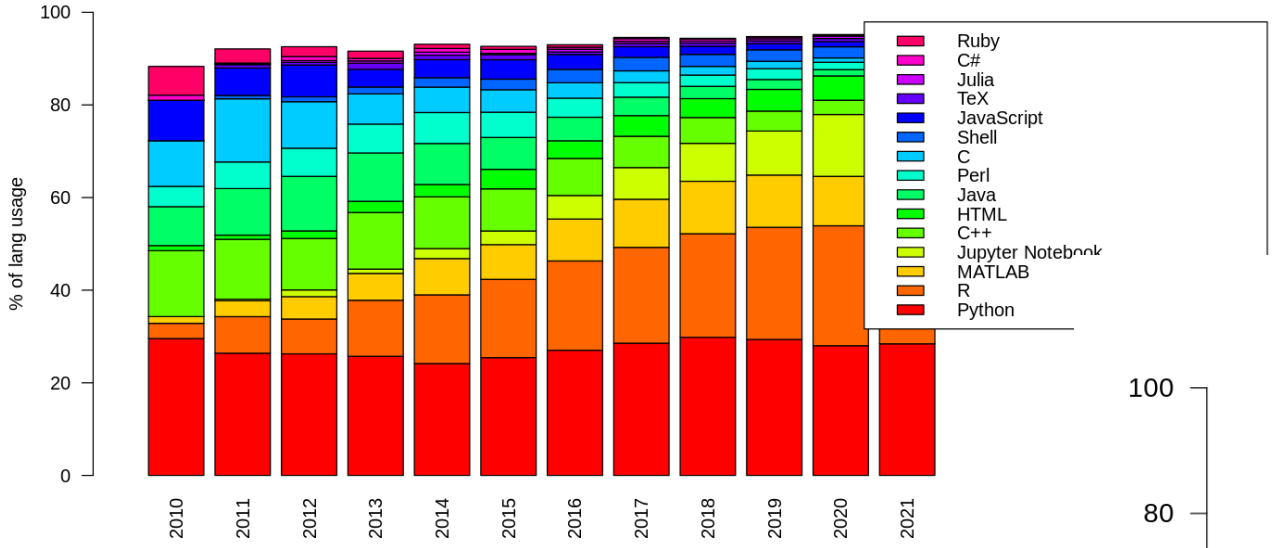
Депозиториев по дате публикации

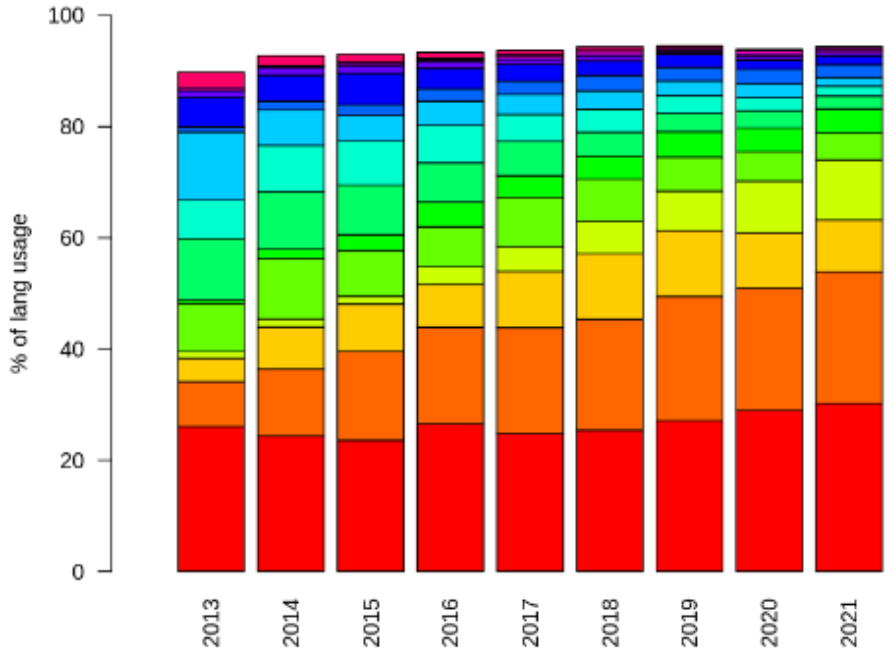


Депозиториев по дате создания

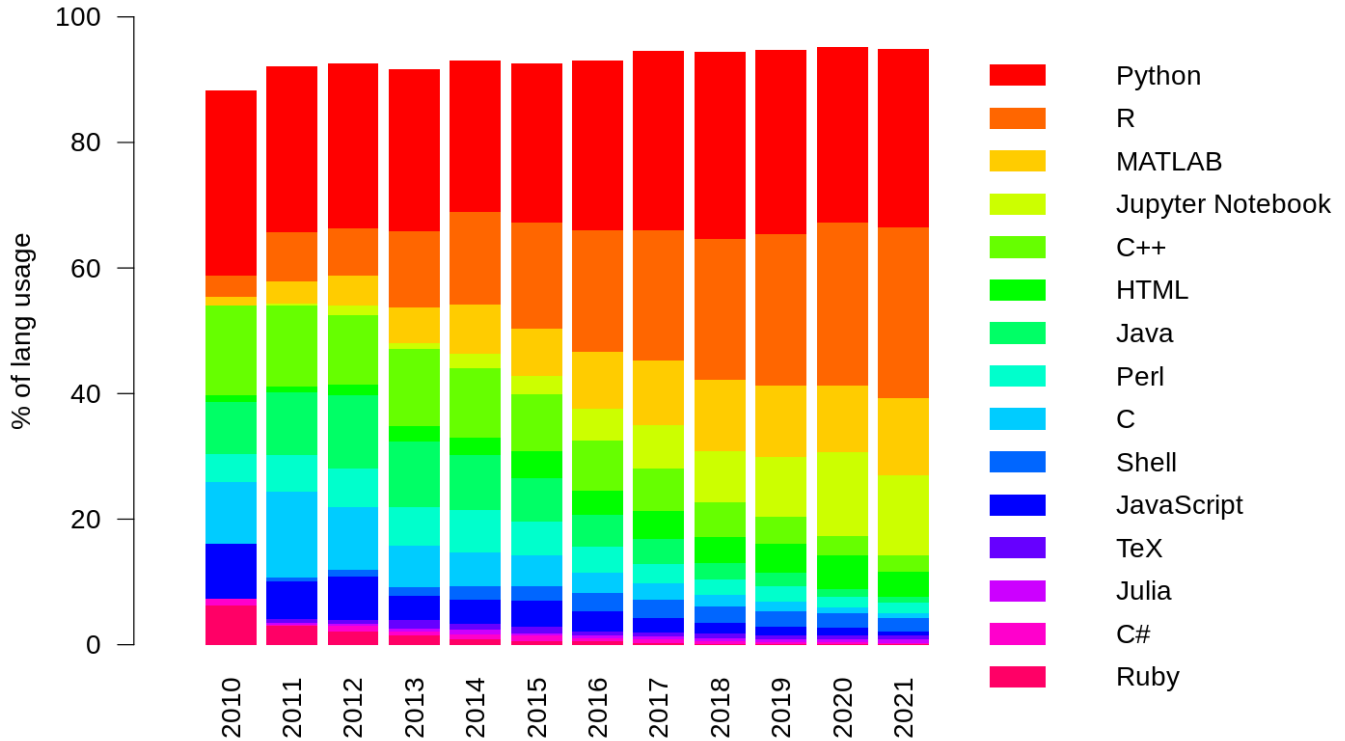
2008-2021



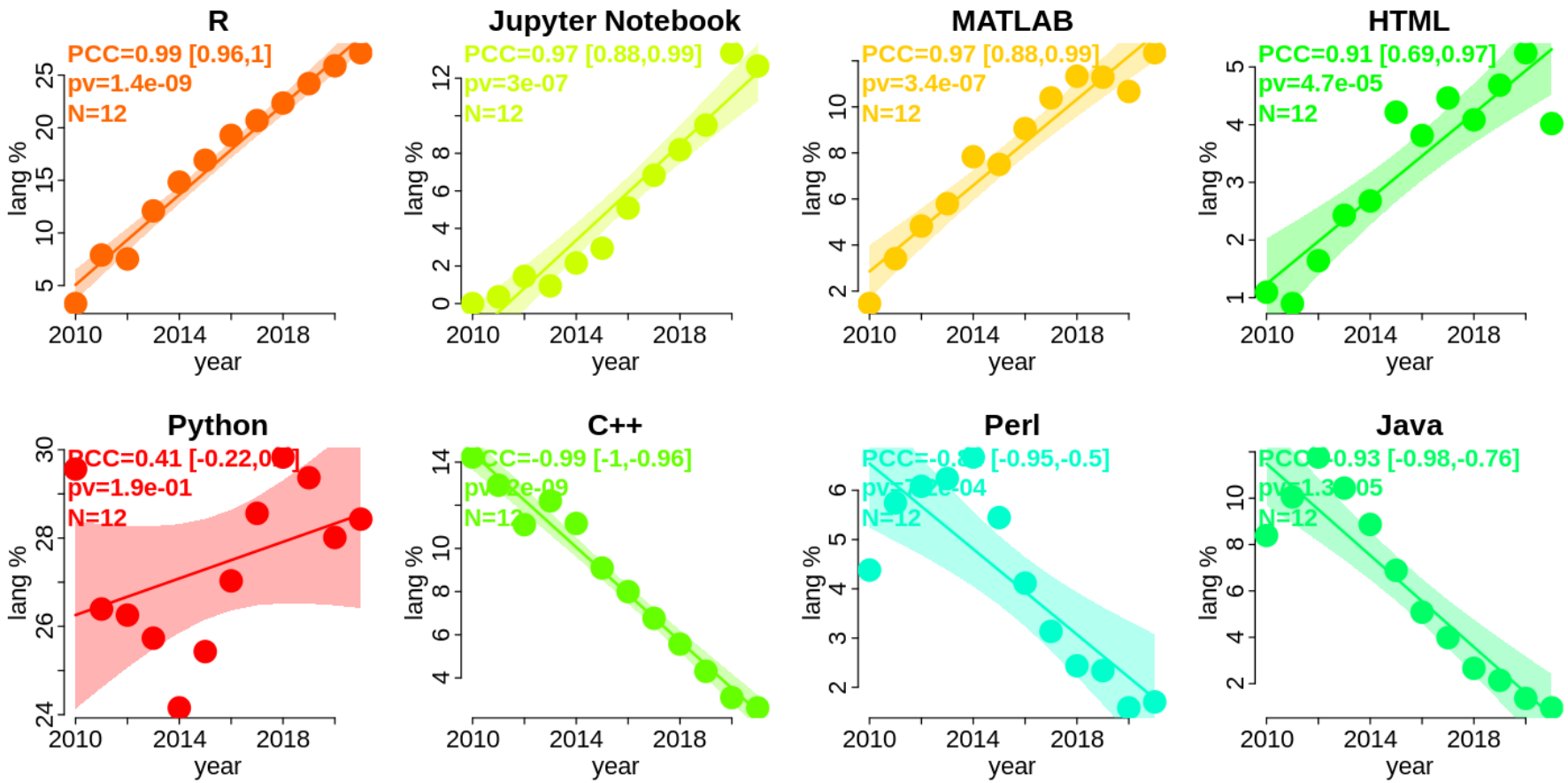




По дате создания



(c) Павел Мазин, Сколтех => EBI



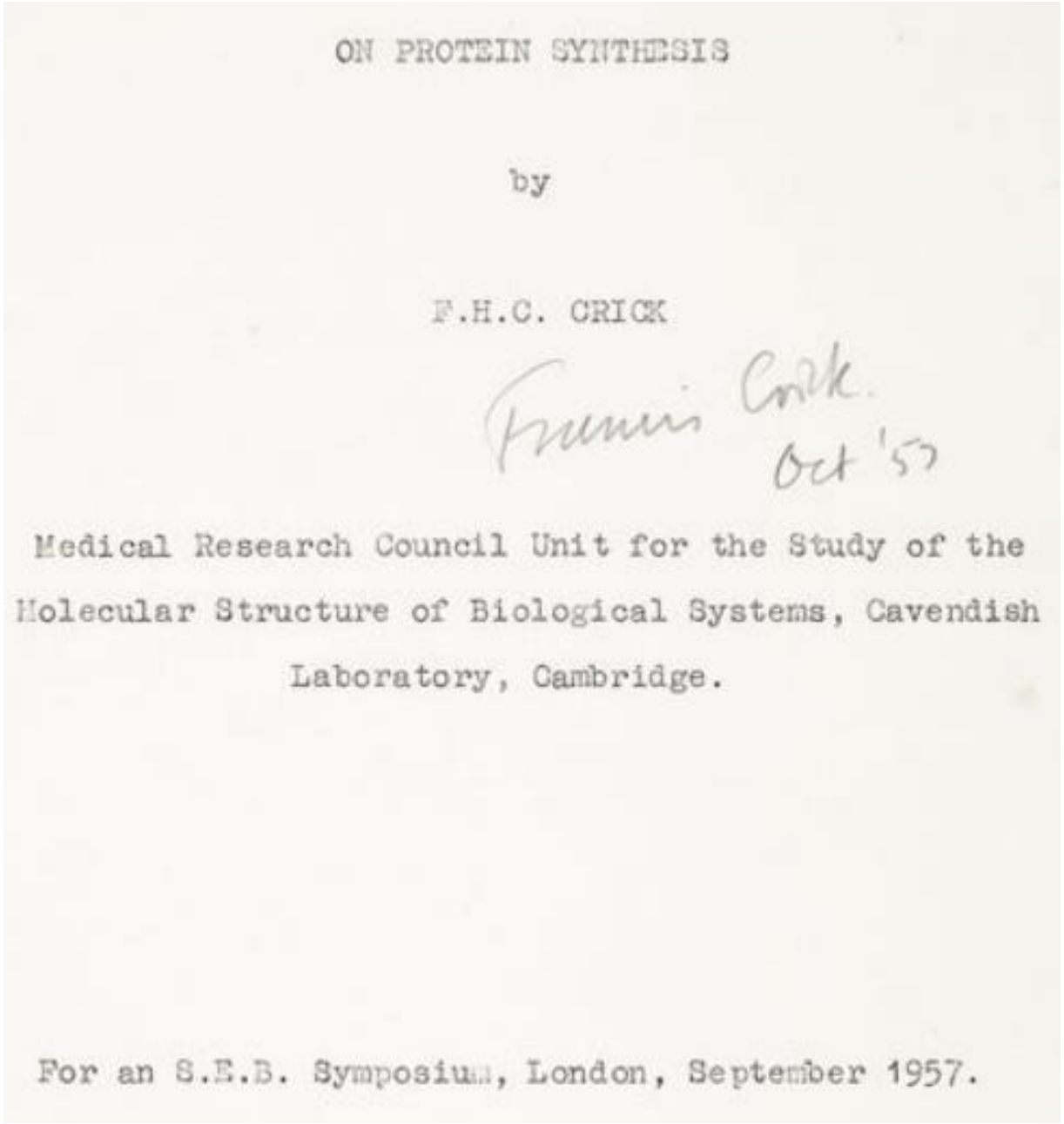
(c) Павел Мазин, Сколтех => EBI

Central dogma

The sequence hypothesis: “Specificity of a piece of nucleic acid is expressed solely by the sequence of its bases, and this sequence is a (simple) code for the amino acid sequence of a particular protein”.

The Central Dogma: “Once ‘information’ has passed into protein it cannot get out again. The transfer of information from nucleic acid to nucleic acid, or from nucleic acid to protein may be possible, but transfer from protein to protein, or from protein to nucleic acid is impossible”.

(F. Crick, 1957)



ON PROTEIN SYNTHESIS

by

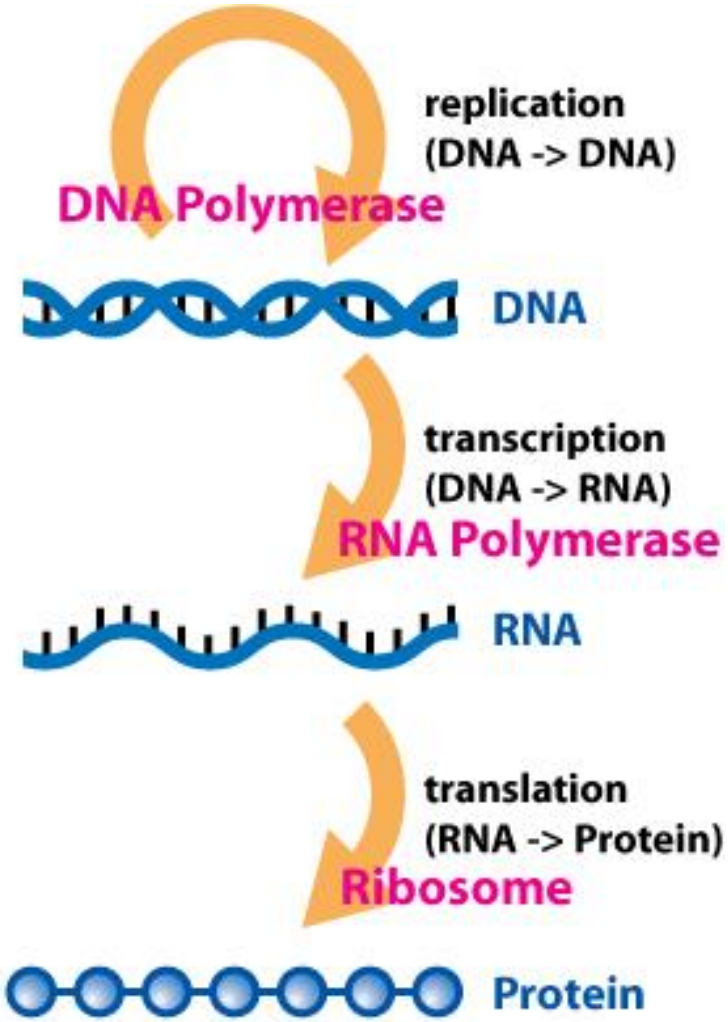
F.H.C. CRICK

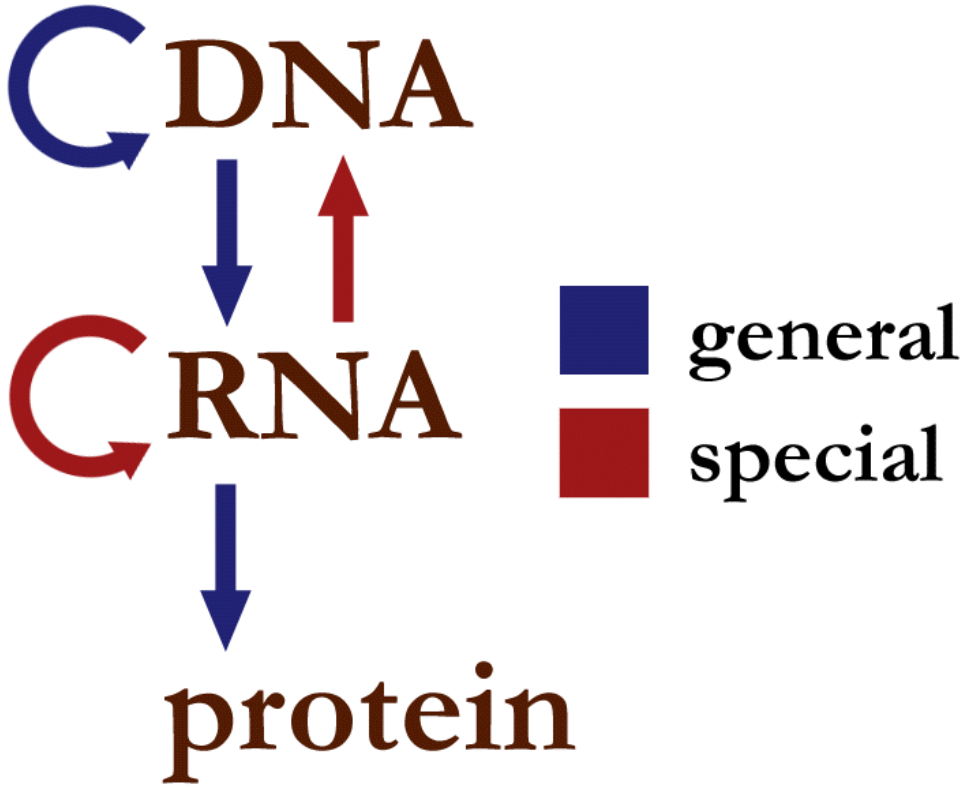
Francis Crick Oct '57

Medical Research Council Unit for the Study of the Molecular Structure of Biological Systems, Cavendish Laboratory, Cambridge.

For an S.E.B. Symposium, London, September 1957.

Центральная догма





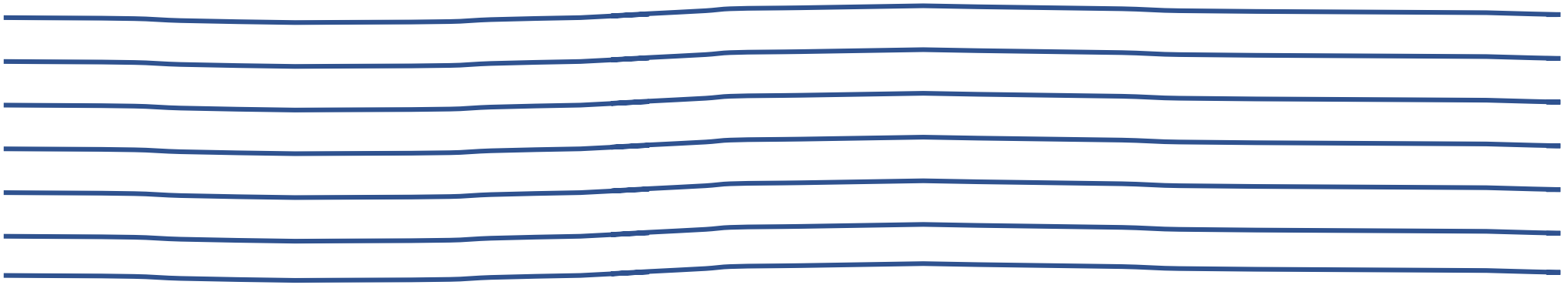
Википедия

Типичные задачи биоинформатики

- Алгоритмы на строках
  - Сборка
  - Картирование
- Технические скрипты
  - Очистка, фильтрация, баркодирование
  - Выбор данных по условиям
- Анализ
  - Дифференциальная экспрессия
  - Peak calling

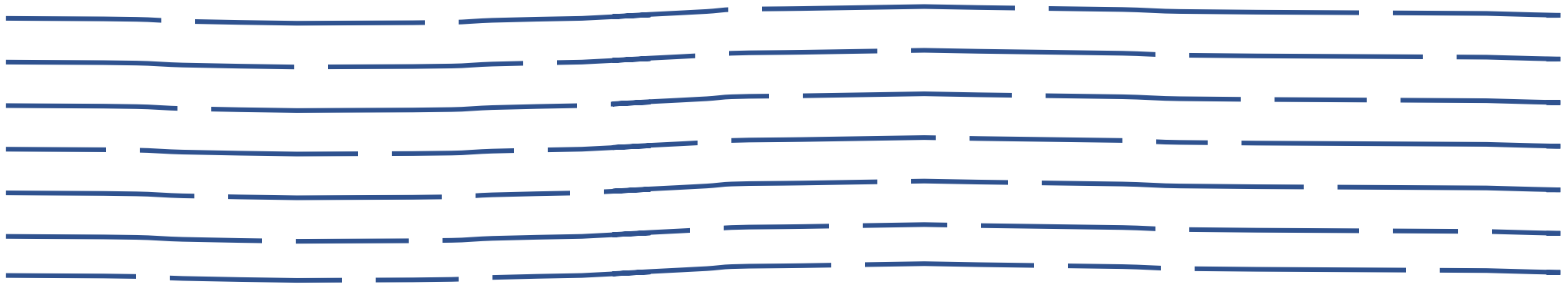
Секвенирование и сборка геномов

1. Много идентичных геномов



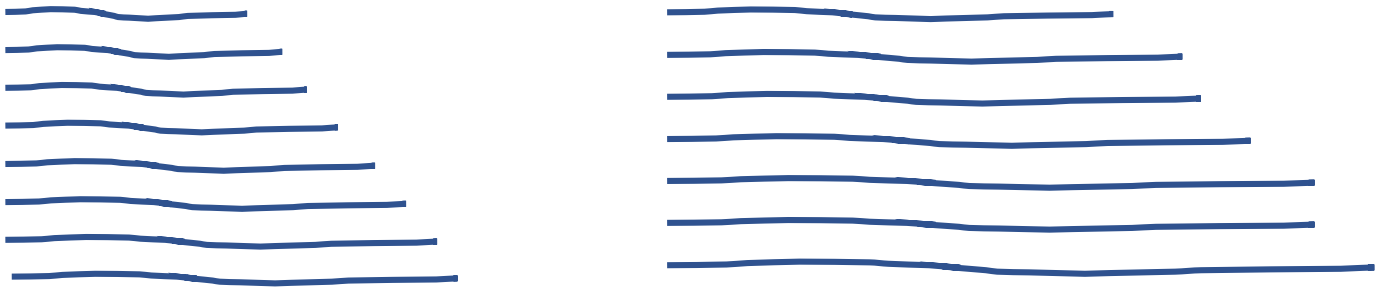
Секвенирование и сборка геномов

2. Режем в случайных местах



Секвенирование и сборка геномов

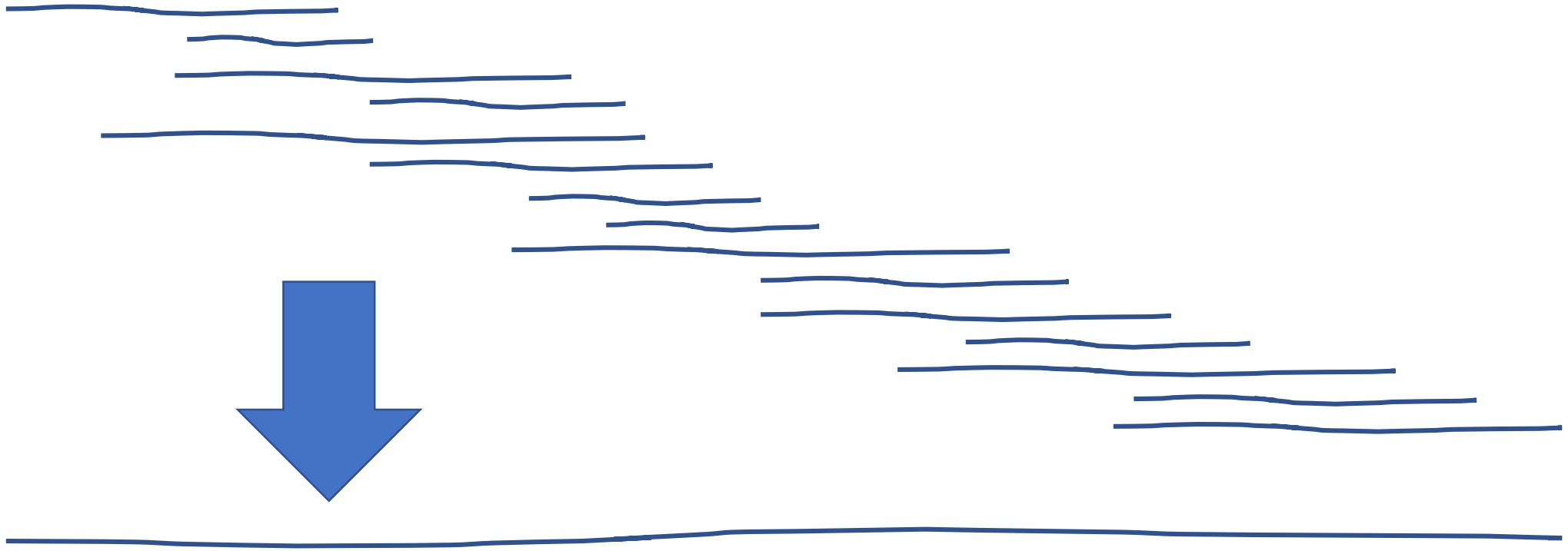
3. Определяем последовательности фрагментов (секвенируем)



И т.д.

Секвенирование и сборка геномов

4. Фильтруем (+ другая предобработка)
5. Собираем по перекрываниям



Секвенирование и сборка геномов

Проблемы:

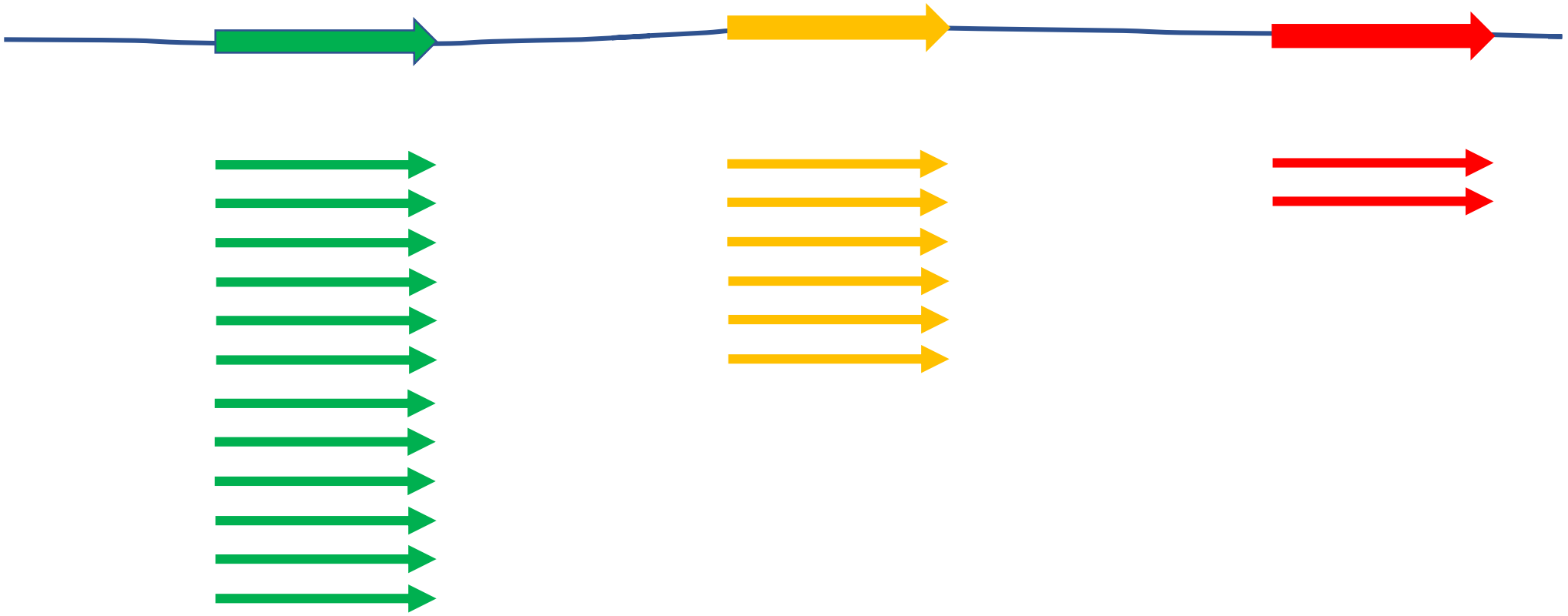
- Большой объем данных
- Контаминация
- Ошибки секвенирования
- Повторы в геноме
- …

Картирование

Вхождение короткой строки в длинную:

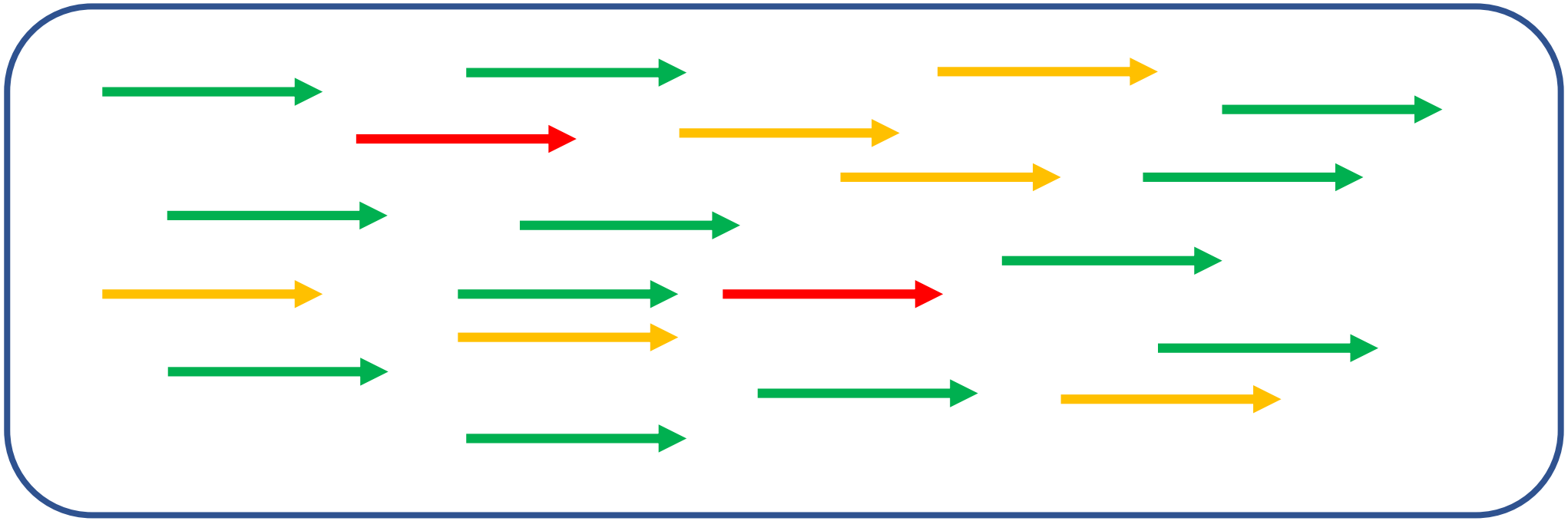
- Пересеквенирование и картирование на референсный геном
  - Точечные несовпадения (отличать от ошибок секвенирования!)
  - Большие вставки / делеции (отличать от артефактов эксперимента!)
- Измерение экспрессии генов (транскриптомика)

Транскриптомика



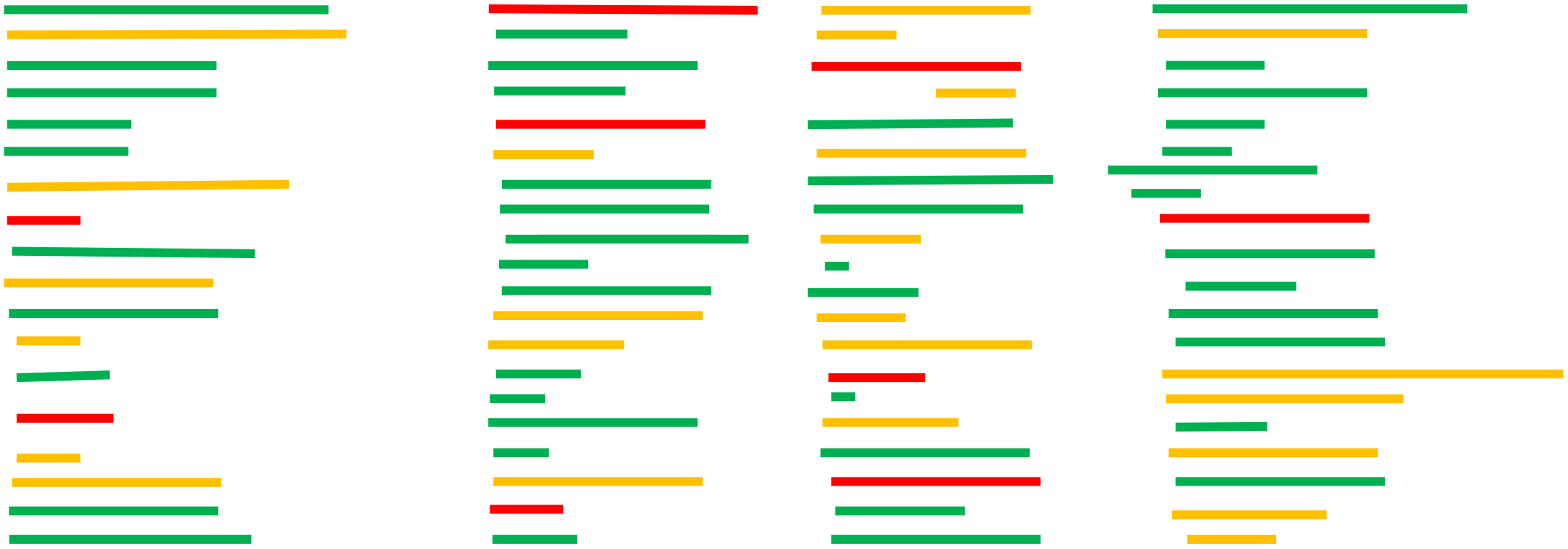
Транскриптомика

1. Выделяем мРНК



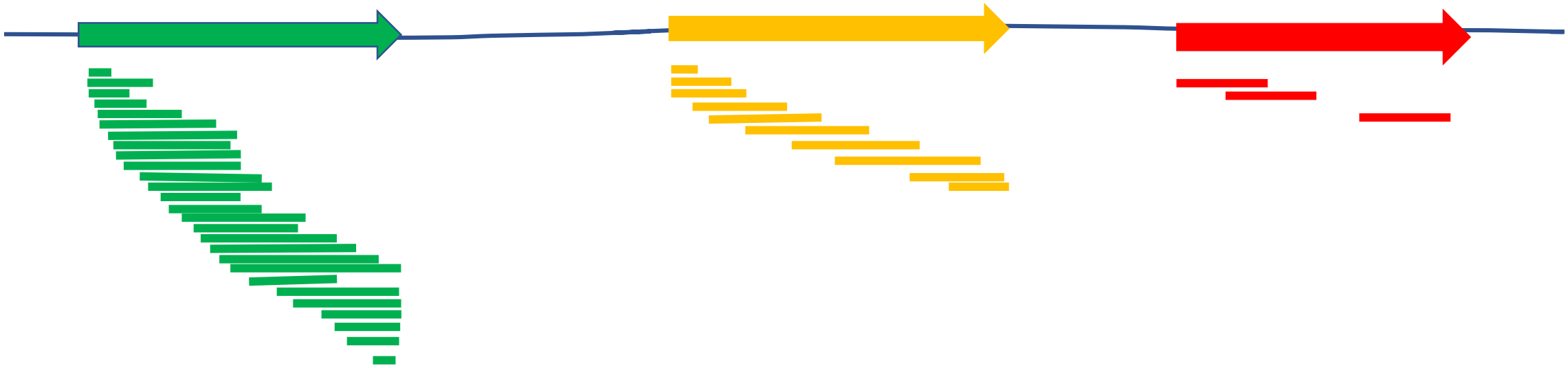
Транскриптомика

2. Режем и секвенируем

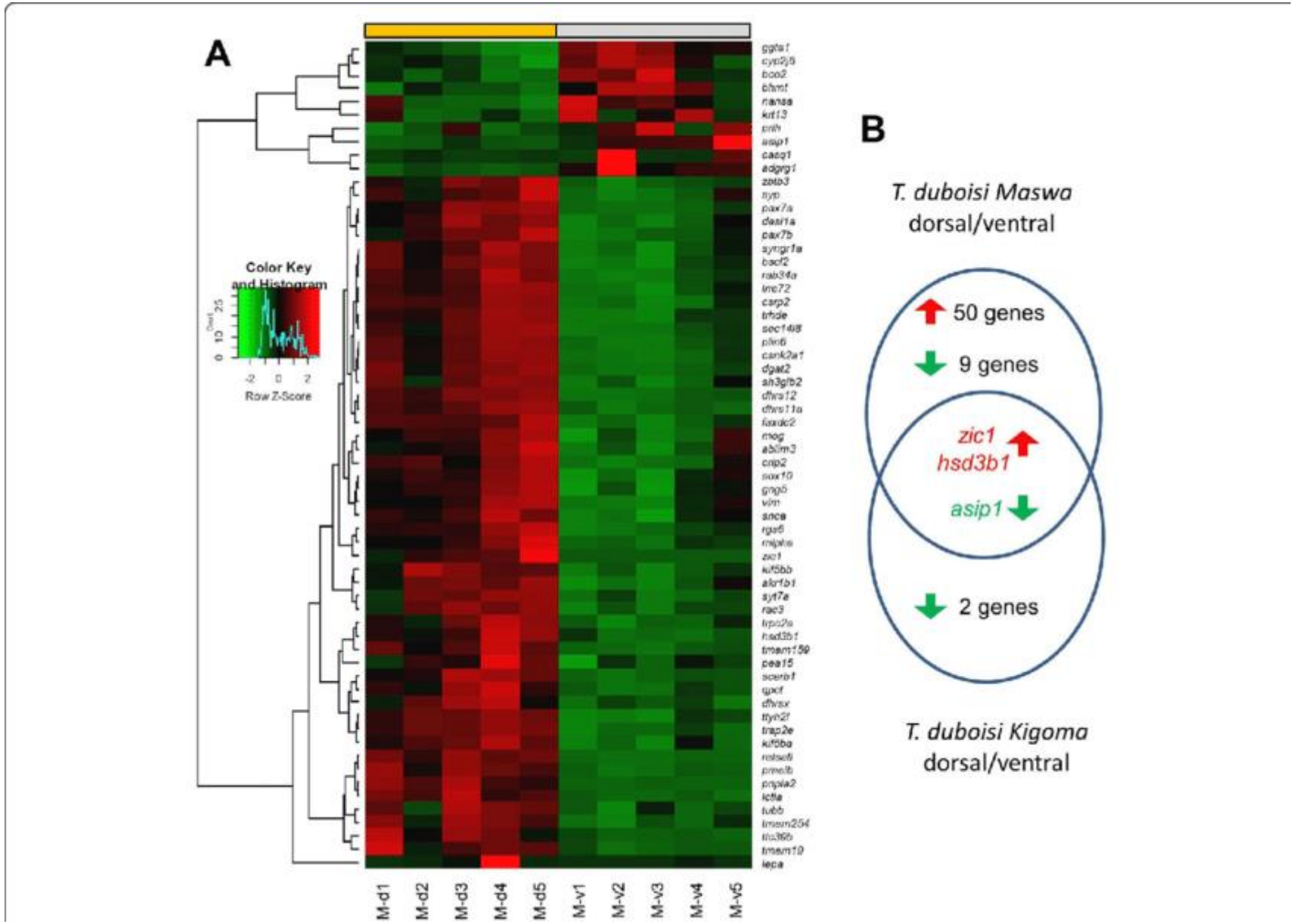


Транскриптомика

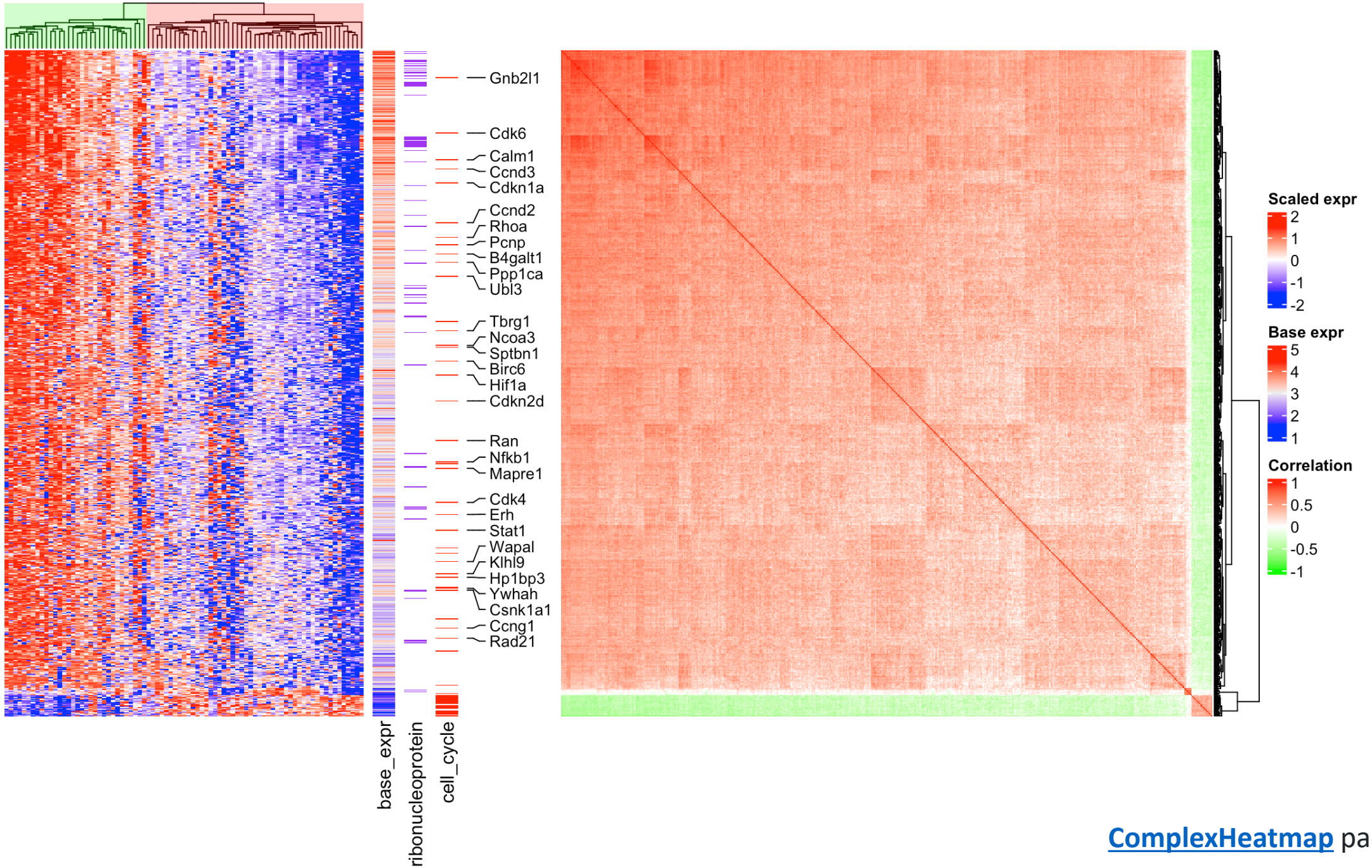
3. Фильтруем и т.п.
4. Картируем на геном



Тепловые карты



Comparative transcriptomics reveals candidate carotenoid color genes in an East African cichlid fish

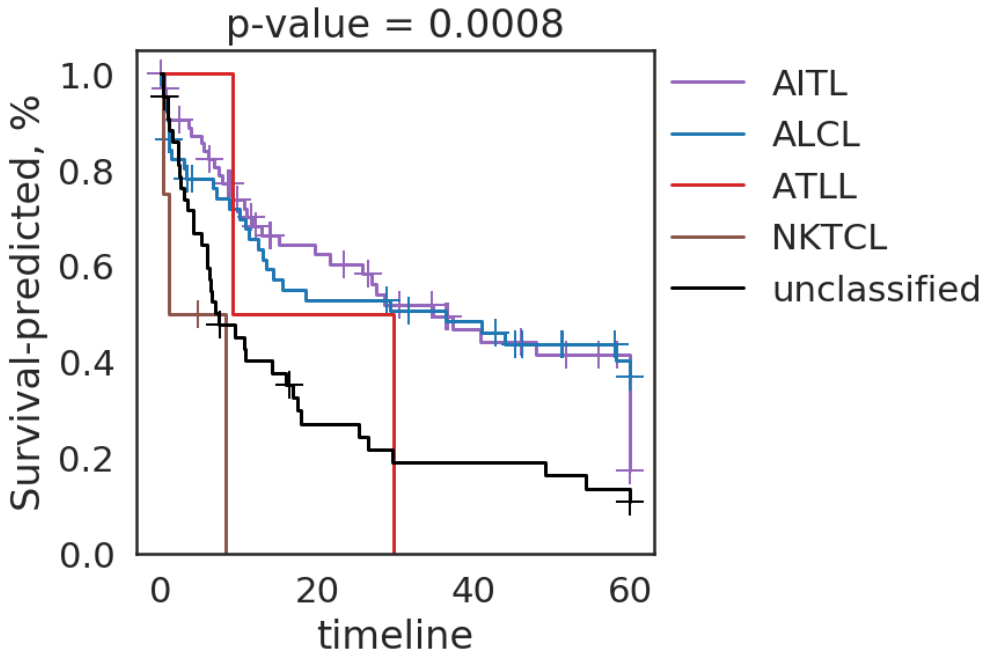


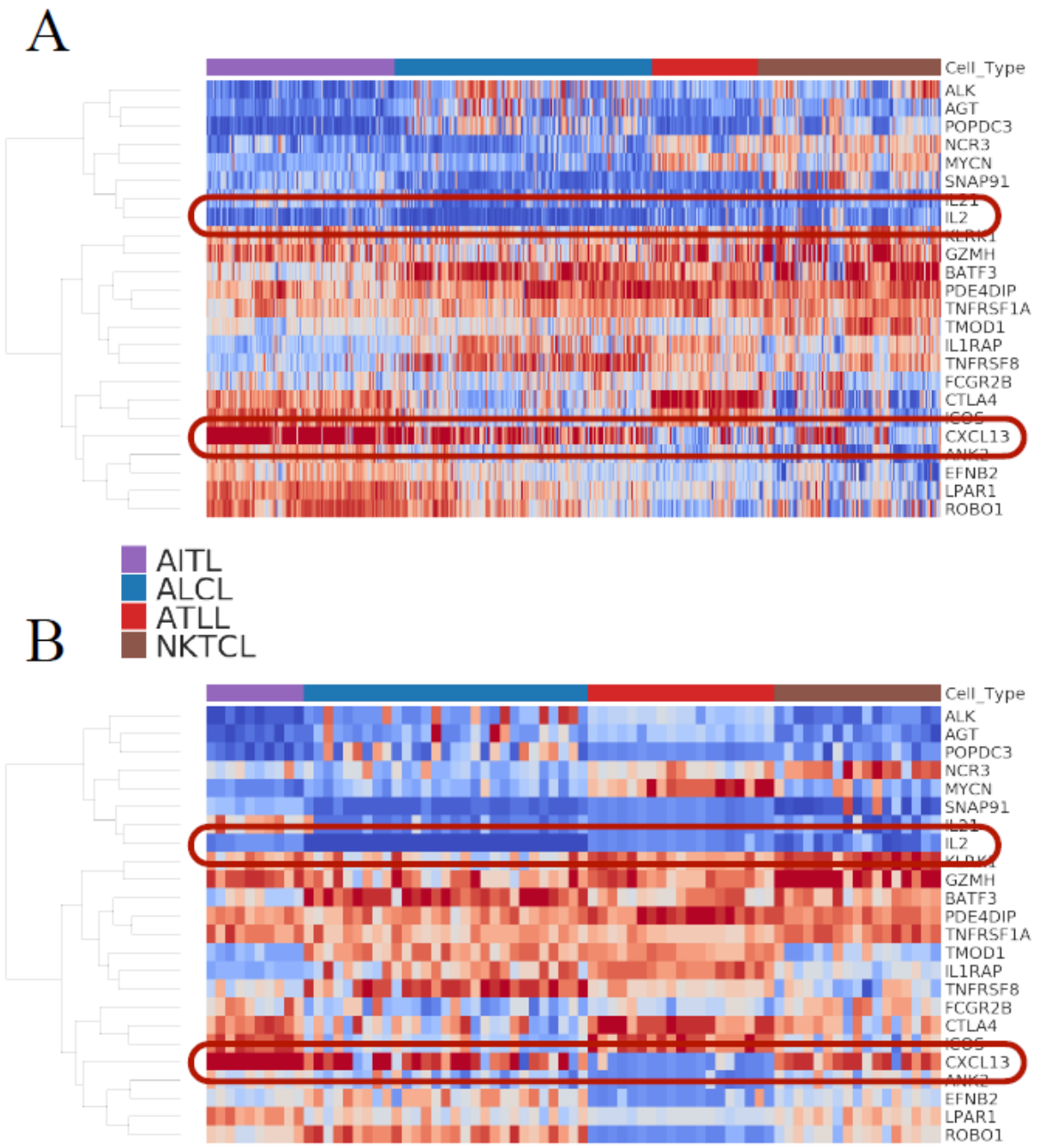
ComplexHeatmap package

Типичные задачи

- Двойная кластеризация
- Классификация образцов
  - Транскриптомные подписи – диагностика
- Ко-экспрессируемые гены
- ВременнЫе профили
- Типы клеток (single cell), дифференцировка

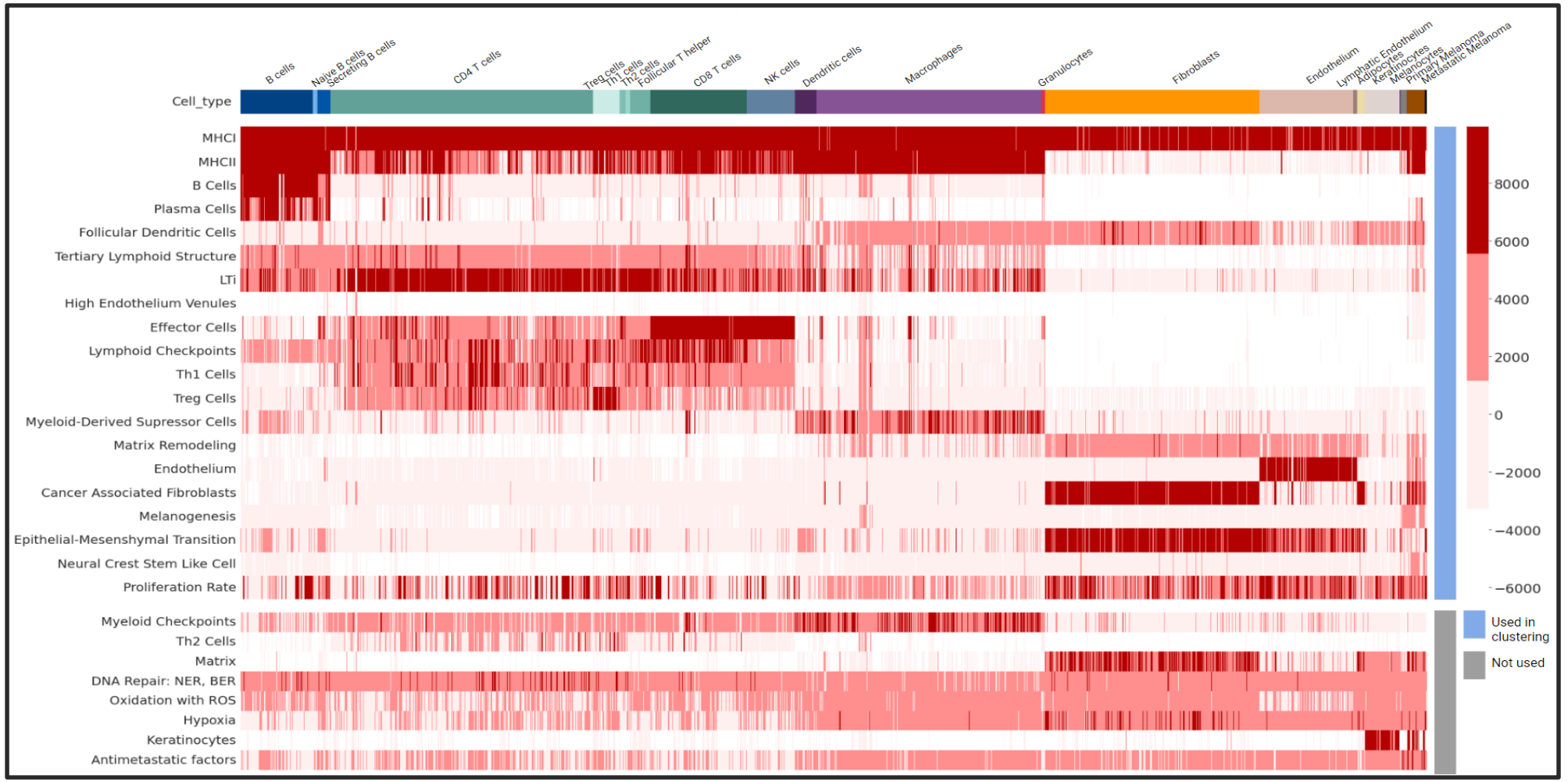
Классификация – транскриптомные подписи





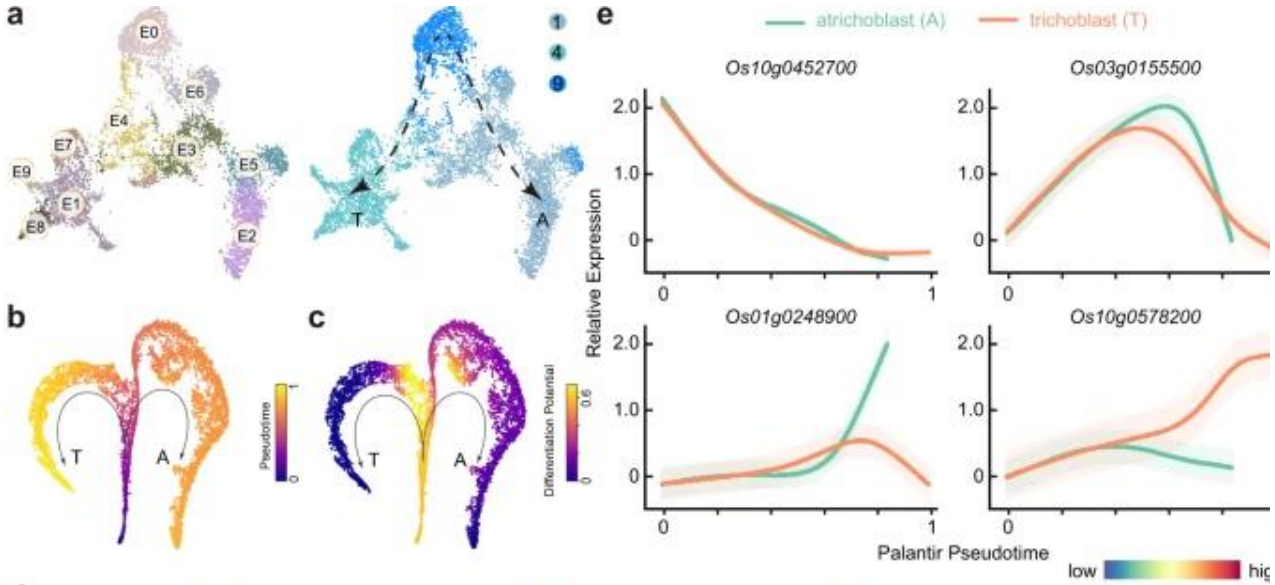
Analysis of PTCL heterogeneity and RNA-based classification of subtypes (A.Karelin, MSc thesis, Skoltech, 2020)

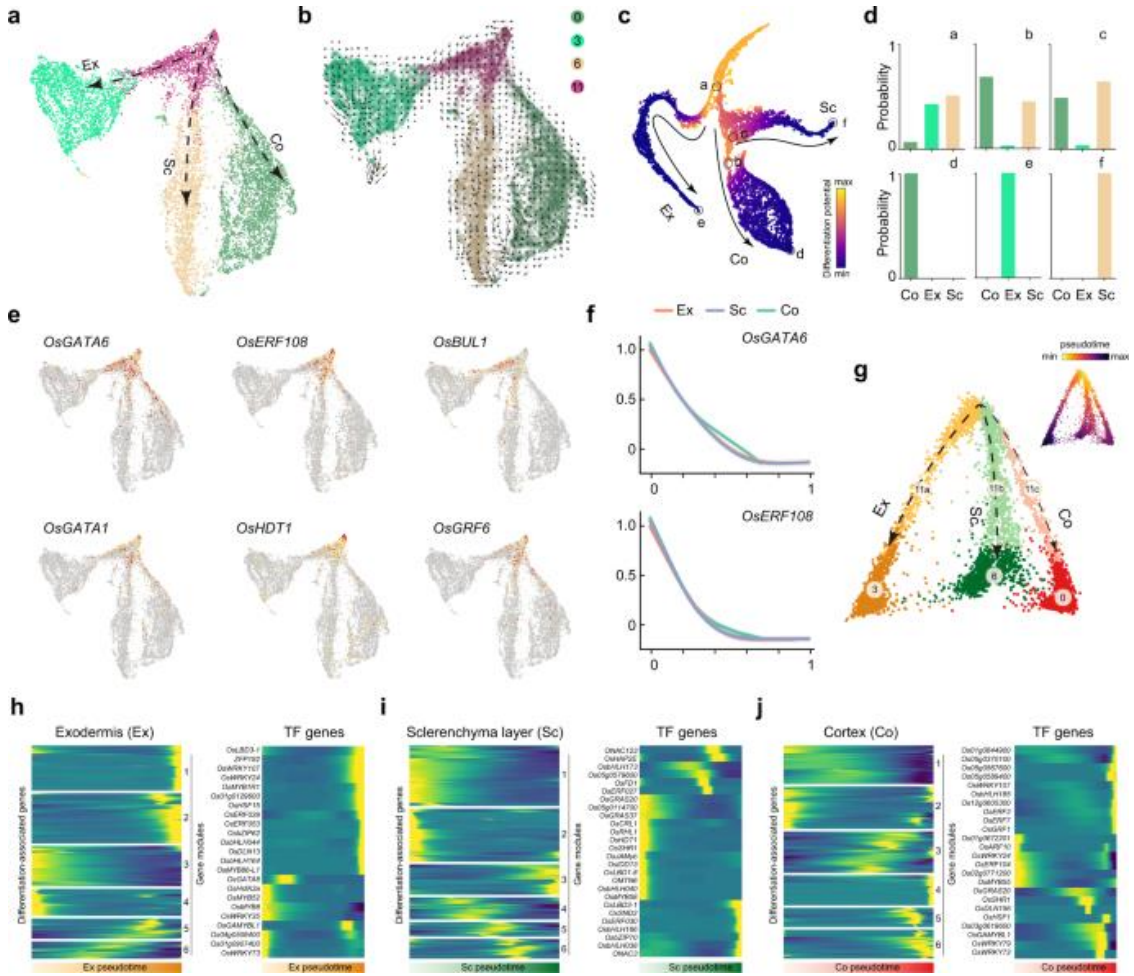
Транскриптомные подписи – типы клеток



Analysis of Tumor-Associated Tertiary Lymphoid Structures Based on Transcriptomics Data of Cutaneous Melanoma (N. Lukashevich, MSc thesis, Skoltech, 2022)

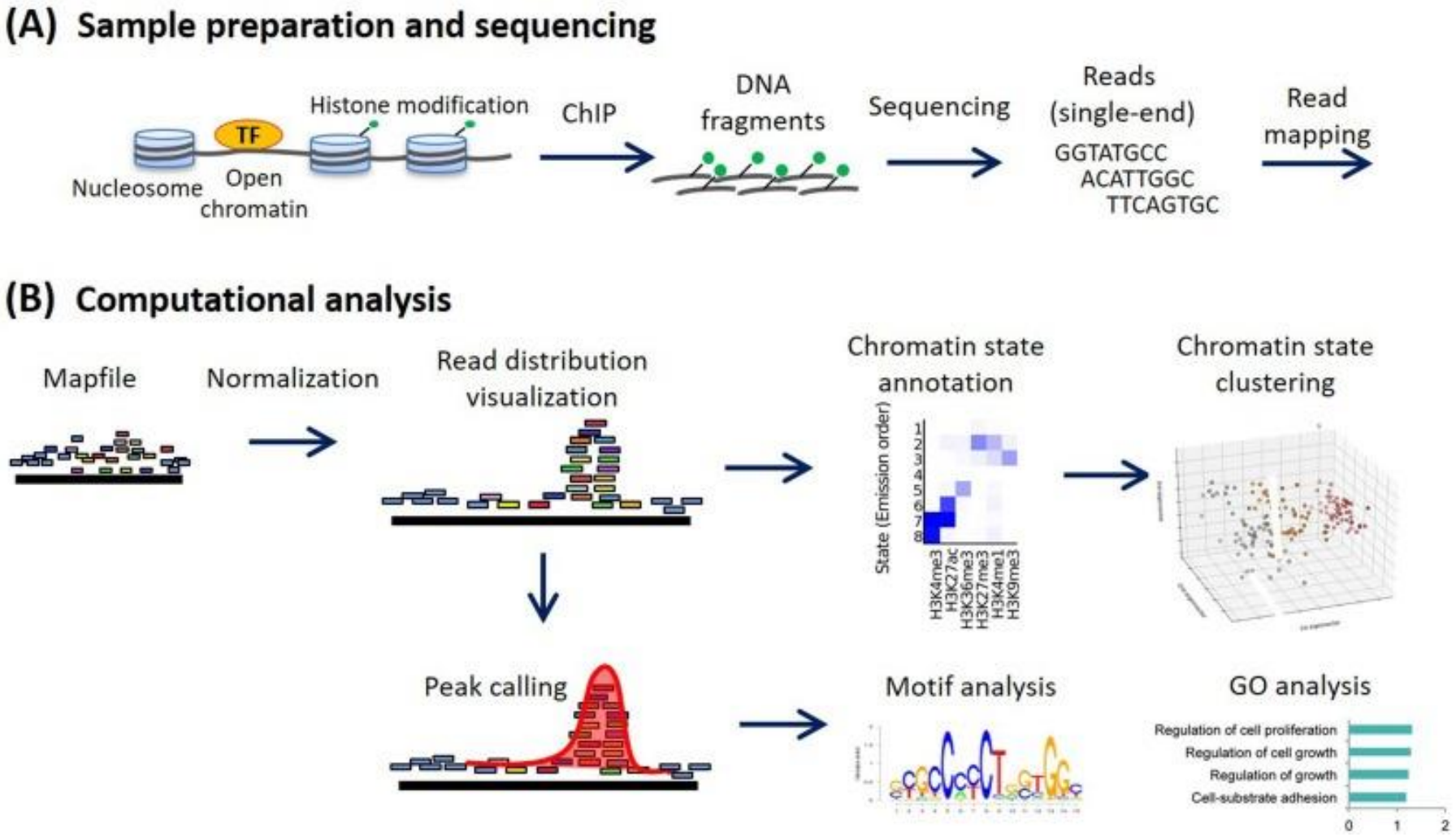
Транскриптомика единичных клеток





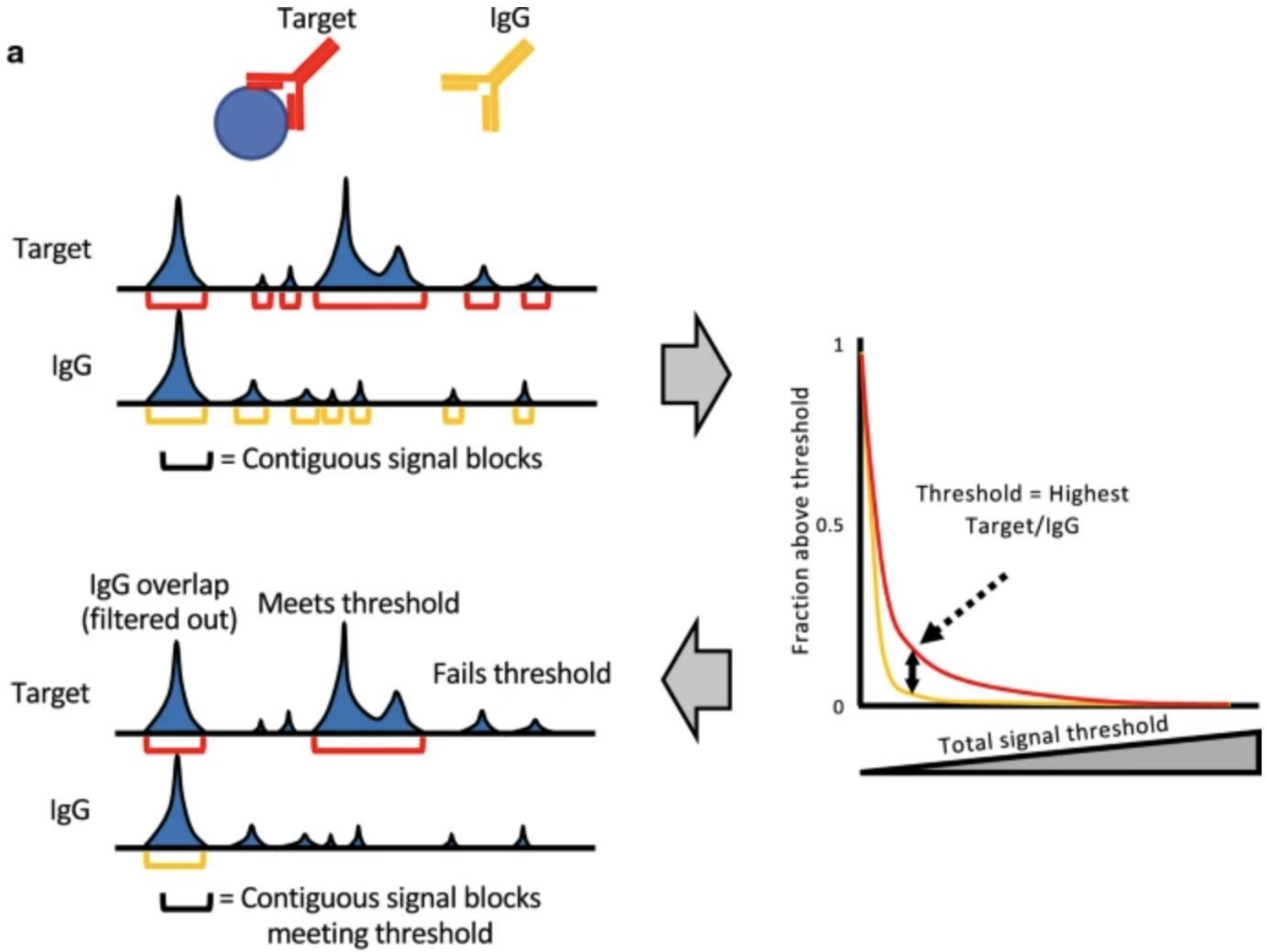
Single-cell transcriptome atlas and chromatin accessibility landscape reveal differentiation trajectories in the rice root

ChIP-Seq и peak calling

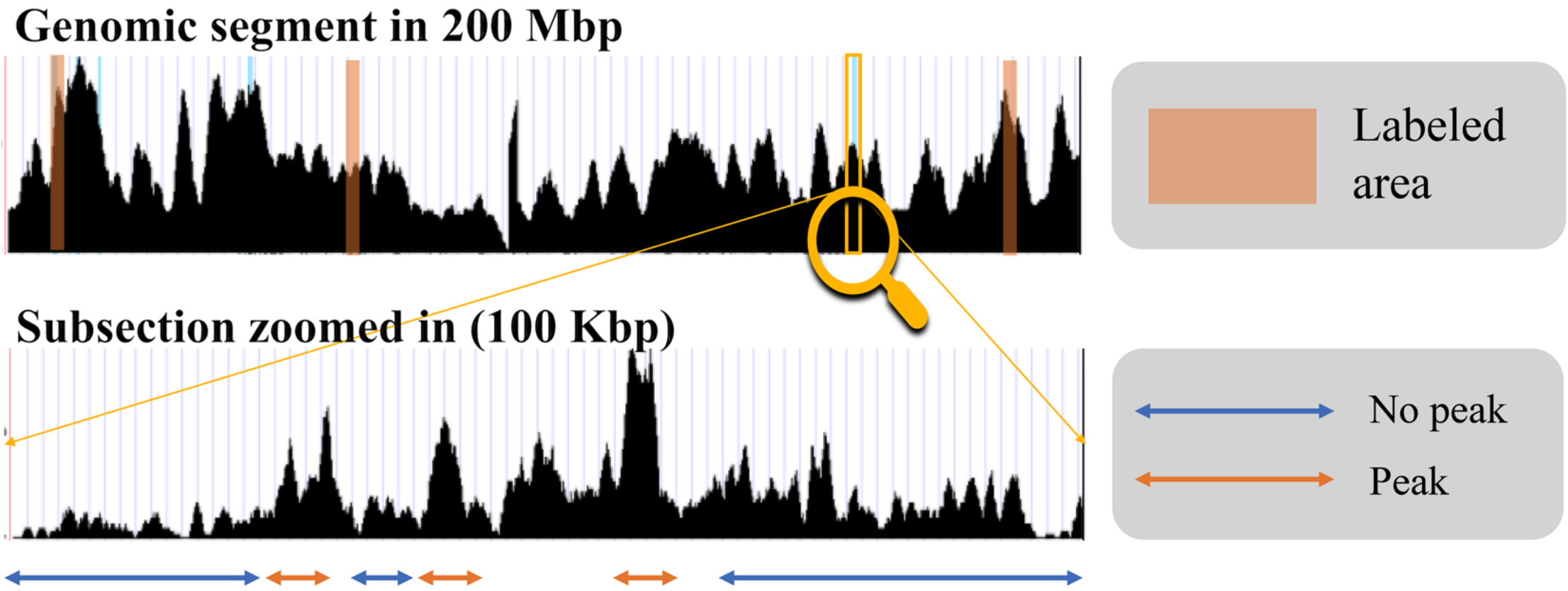


Methods for ChIP-seq analysis: A practical workflow and advanced applications

Fig. 1

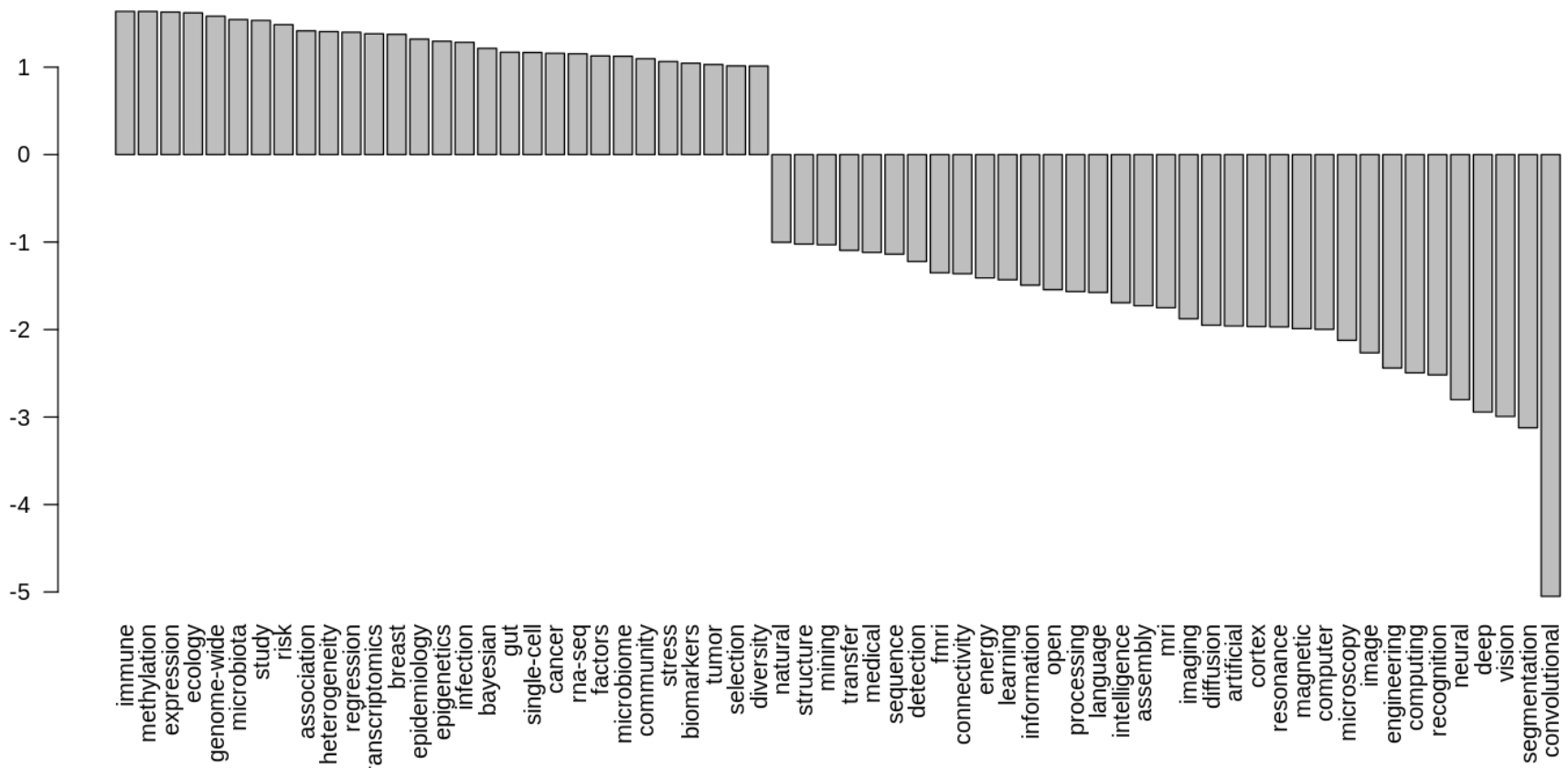


Peak calling by Sparse Enrichment Analysis for CUT&RUN chromatin profiling

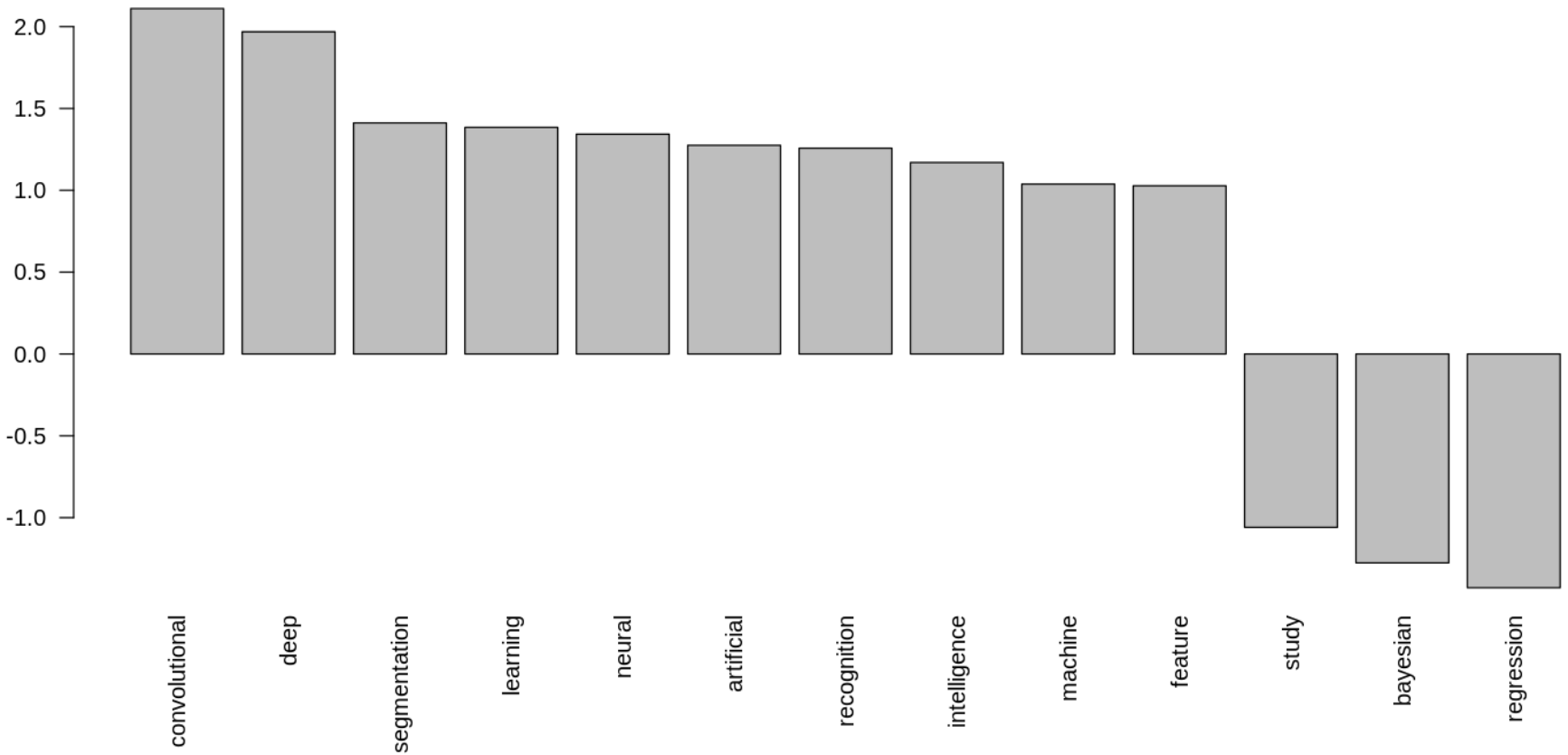


CNN-Peaks package

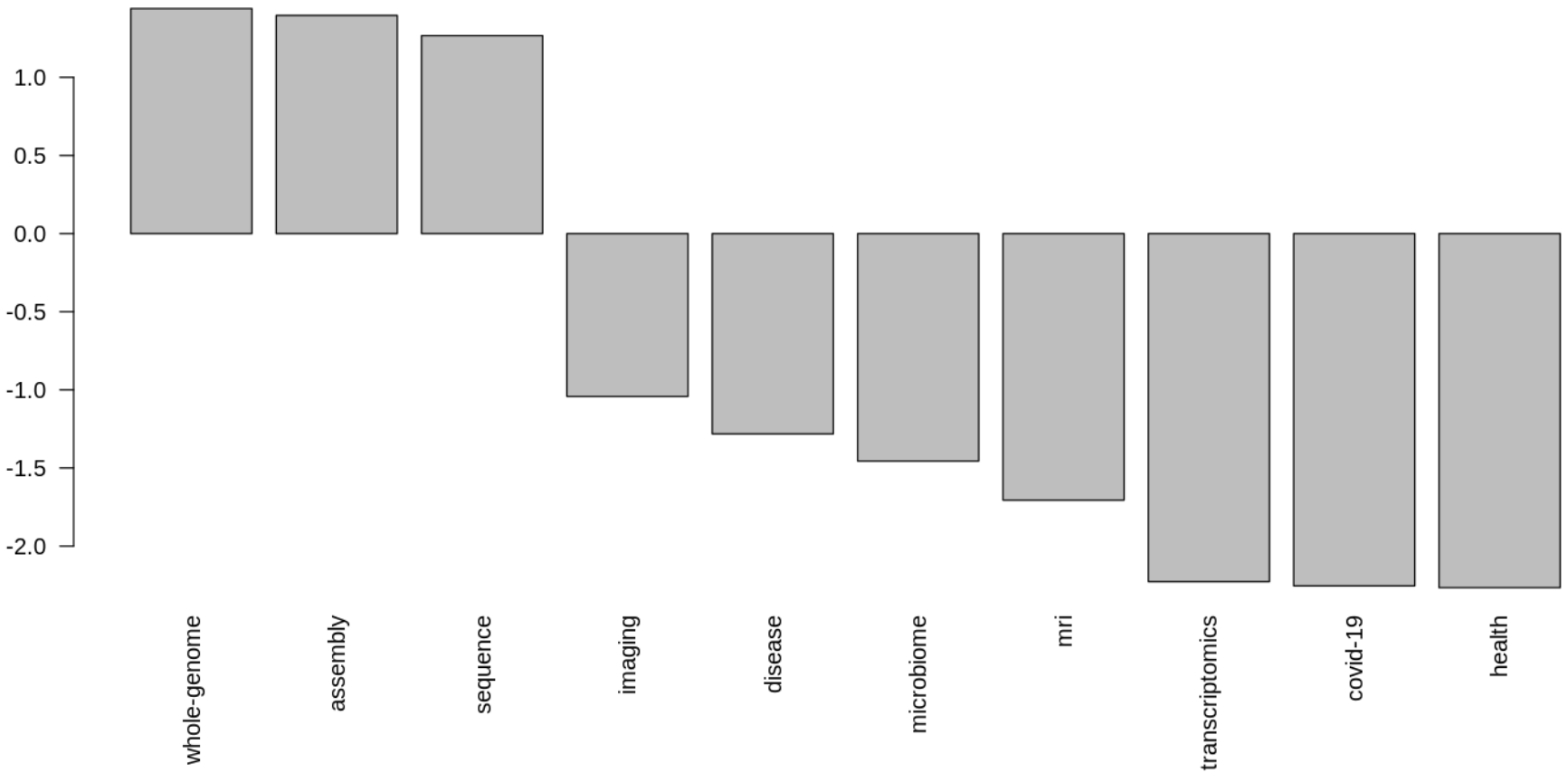
Ключевые слова – R



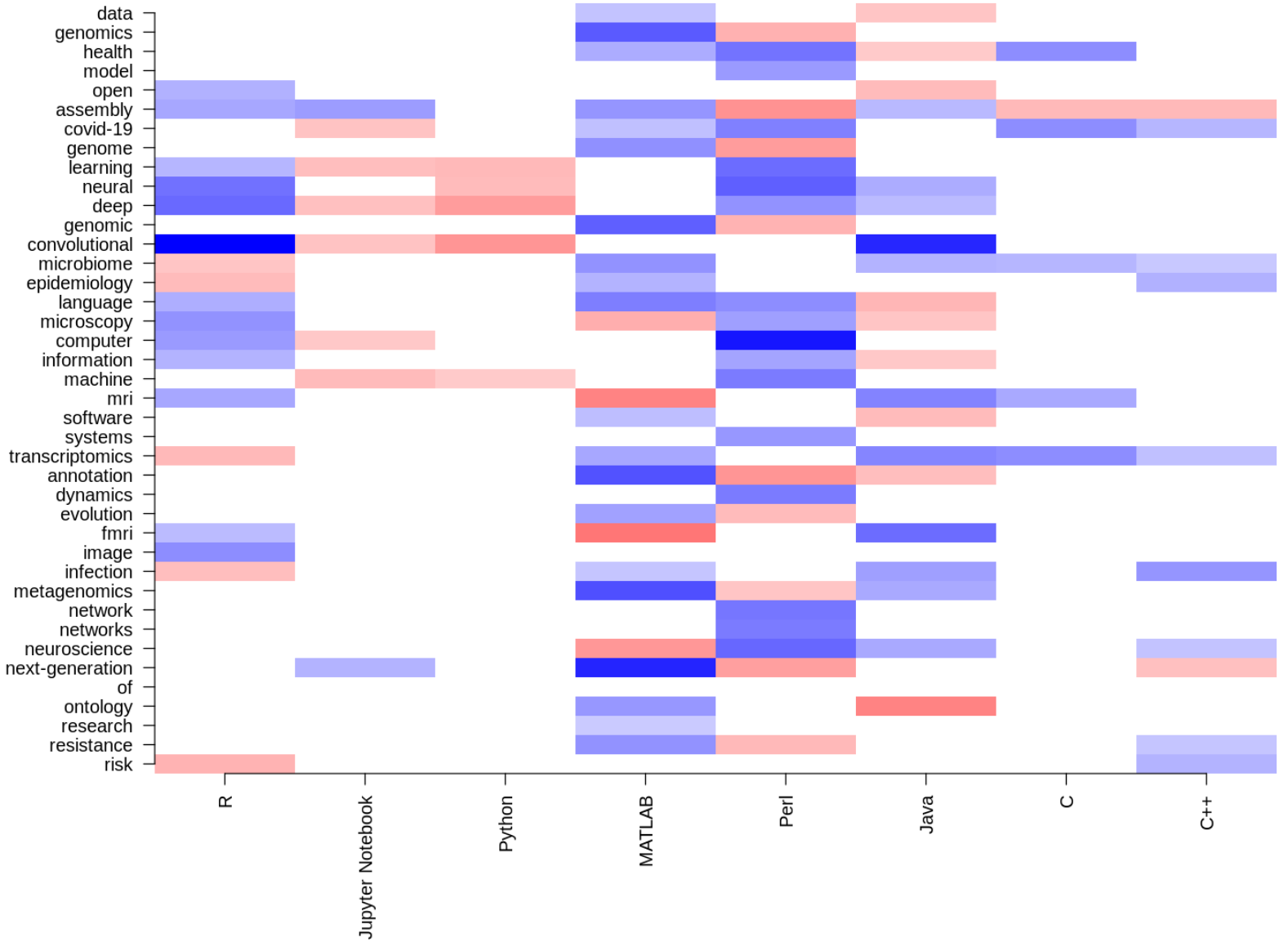
Ключевые слова – Python



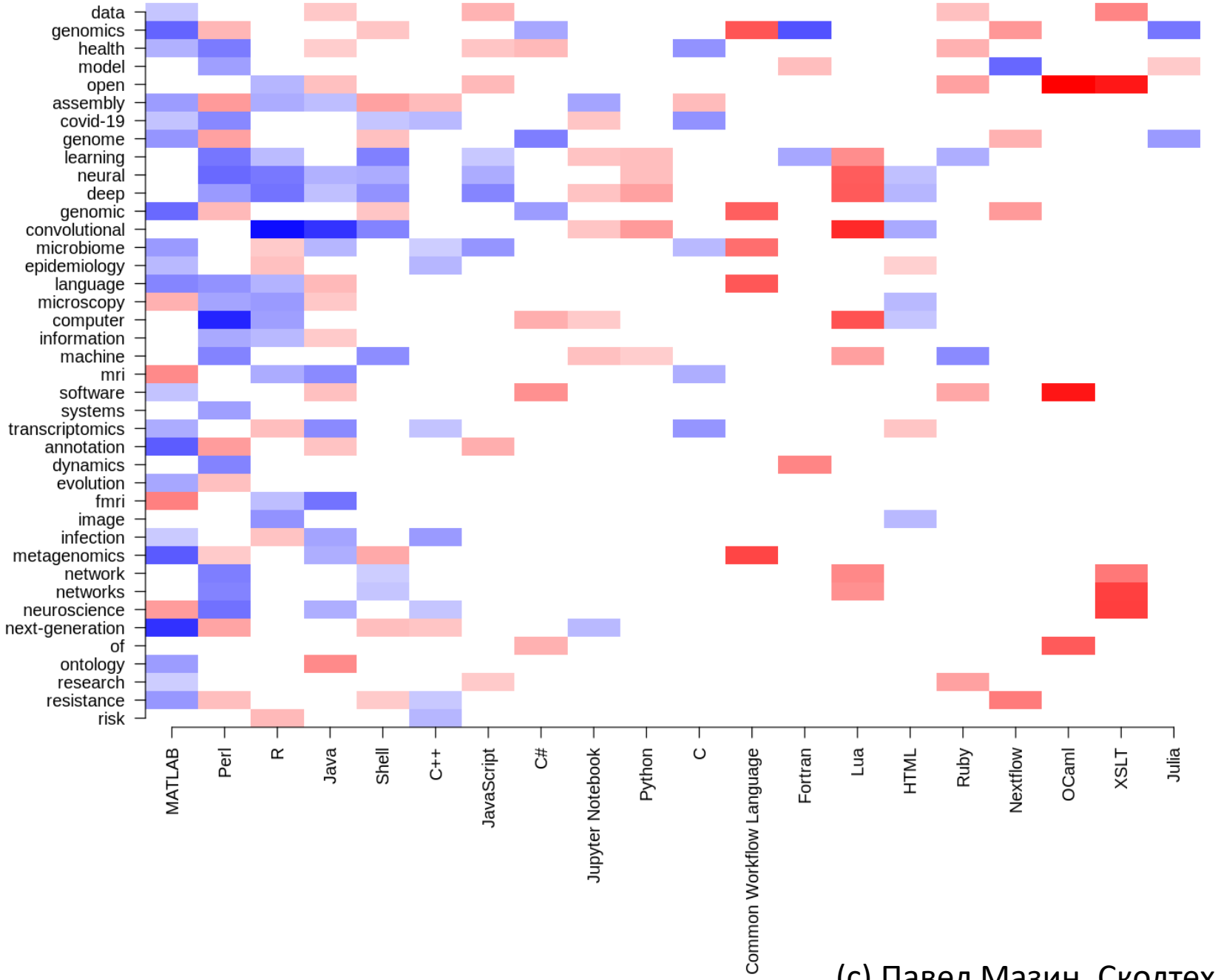
Ключевые слова – C



Ключевые слова



Ключевые слова



Семантическая аура

- R – омиксы
- Python – глубокое обучение
- C – вычислительно интенсивные алгоритмы на строках (сборка)
- Perl – эволюционные / биологические задачи
- Matlab – нейробиология

Тут были еще два слайда…

… но организаторы просили меня их убрать.

Вместо этого – фотография Павла Мазина, рассчетами которого я пользовался. Он предпочитает R, поэтому я выбрал фотографию в шлеме.