

Gut feeling: AIMing to learn microbiome signatures of disease

Group 1 · Amina Ismagulova, Maria Markidonova & Group 2 · Daniel Korkin, Ksenia Golubtsova

INTRODUCTION

- The gut microbiome is associated with human health, yet its composition **varies** substantially between individuals.
- Microbiome data are **high-dimensional** and **sparse**, while disease-specific patterns are often **unknown** in advance.

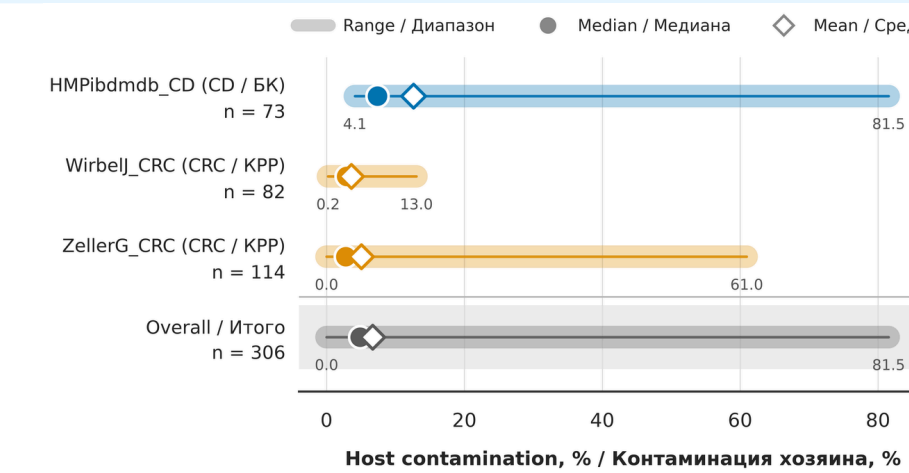
- Individual variability**, **outliers**, and **overlapping features** between diseases **make it difficult** to identify robust disease specific patterns.
- Machine learning** enables us to work with **high-dimensional** data and identify complex patterns that may be difficult to detect using traditional approaches.
- In this project, we investigate **colorectal cancer (CRC)** and **Crohn's disease (CD)** using data from multiple independent studies.

Dataset description

Study	Disease (amount)	Control (amount)	Total (amount)
HMPibdbmdb_CD	42	23	65
Wirbelj_CRC	20	54	74
ZellerG_CRC	48	55	103
Total	110	132	242

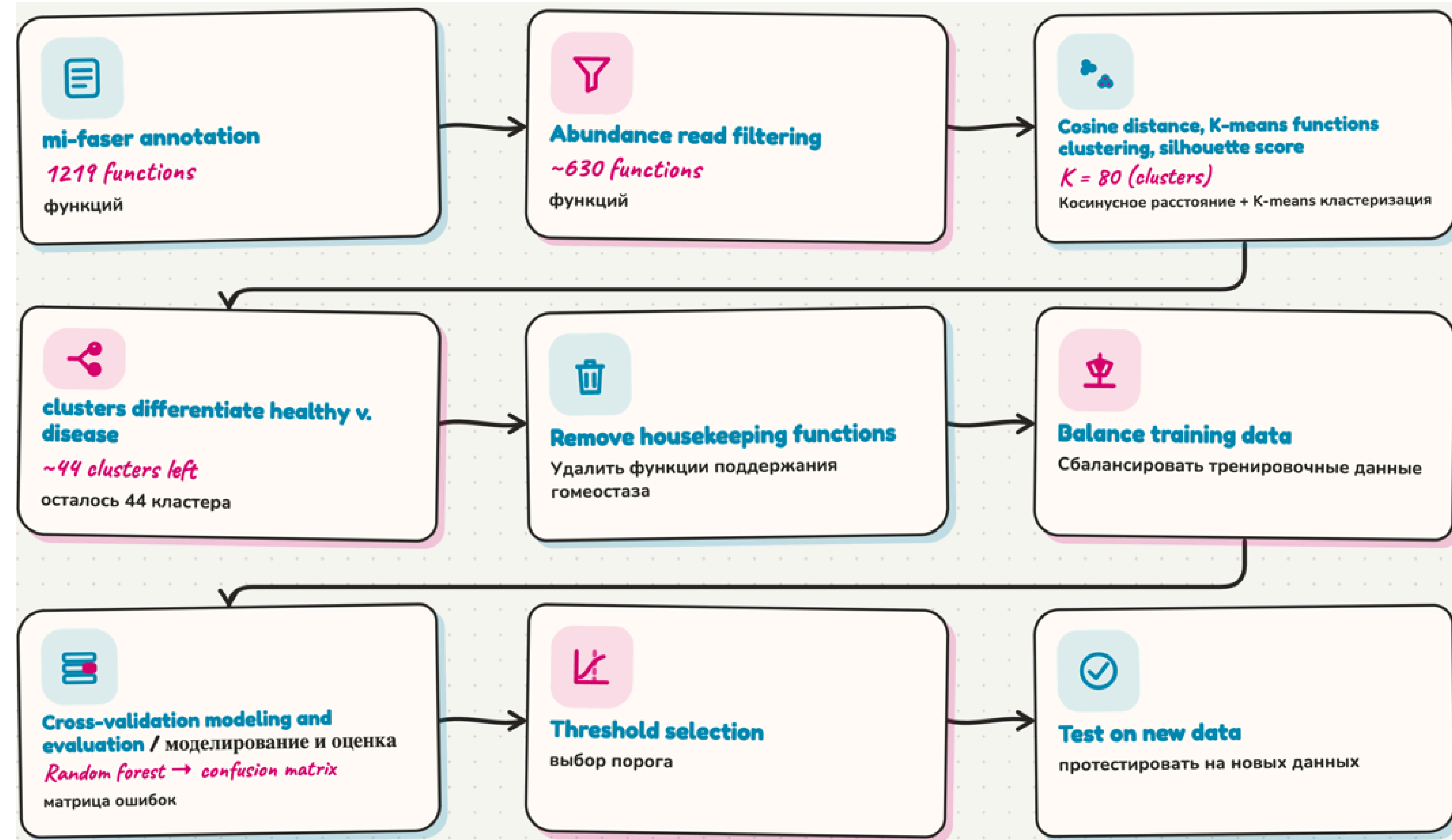
Data preparation + Quality Control

- Low-quality sections/reads → filtered by Q15 ✓
- Adapters → removed ✓
- Host reads → removed ✓

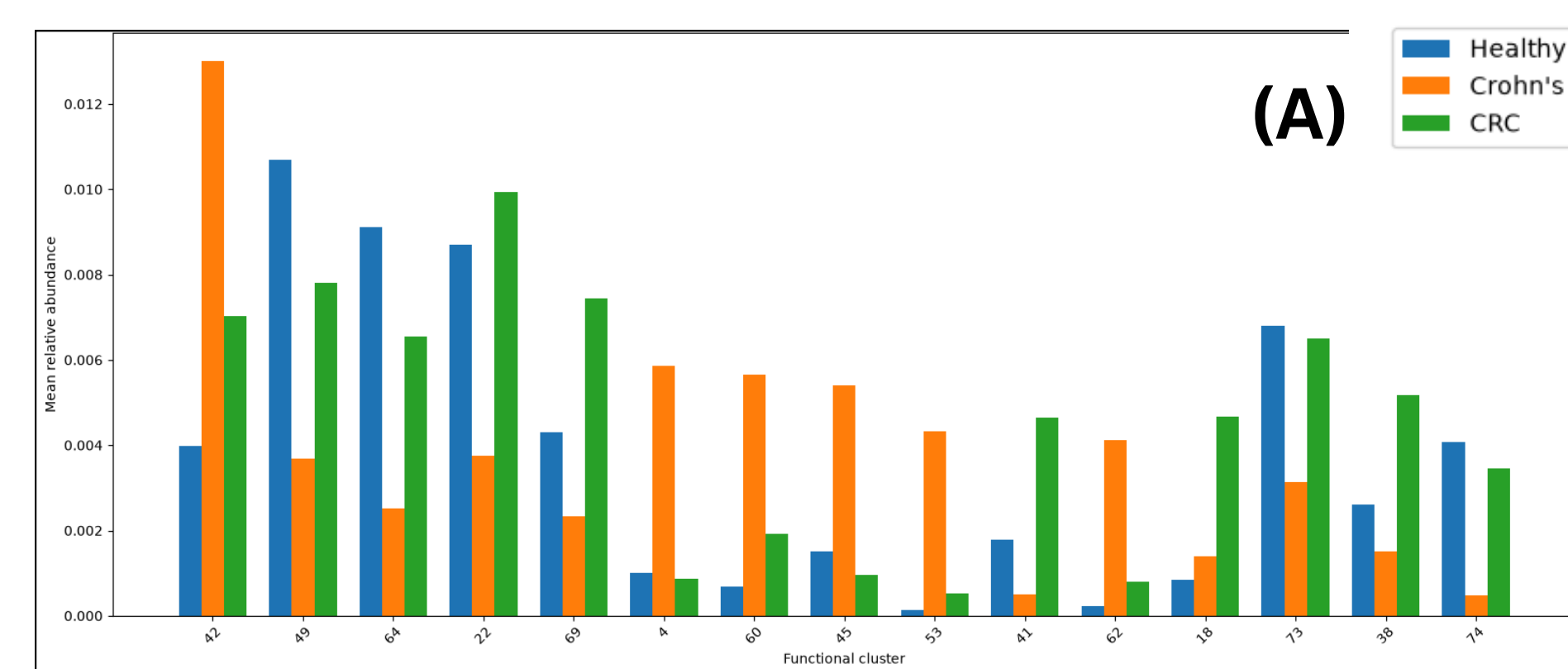


Project A/Проект А - mi-FASER (alignment based)

METHOD / PIPELINE - МЕТОД / ПАЙПЛАЙН

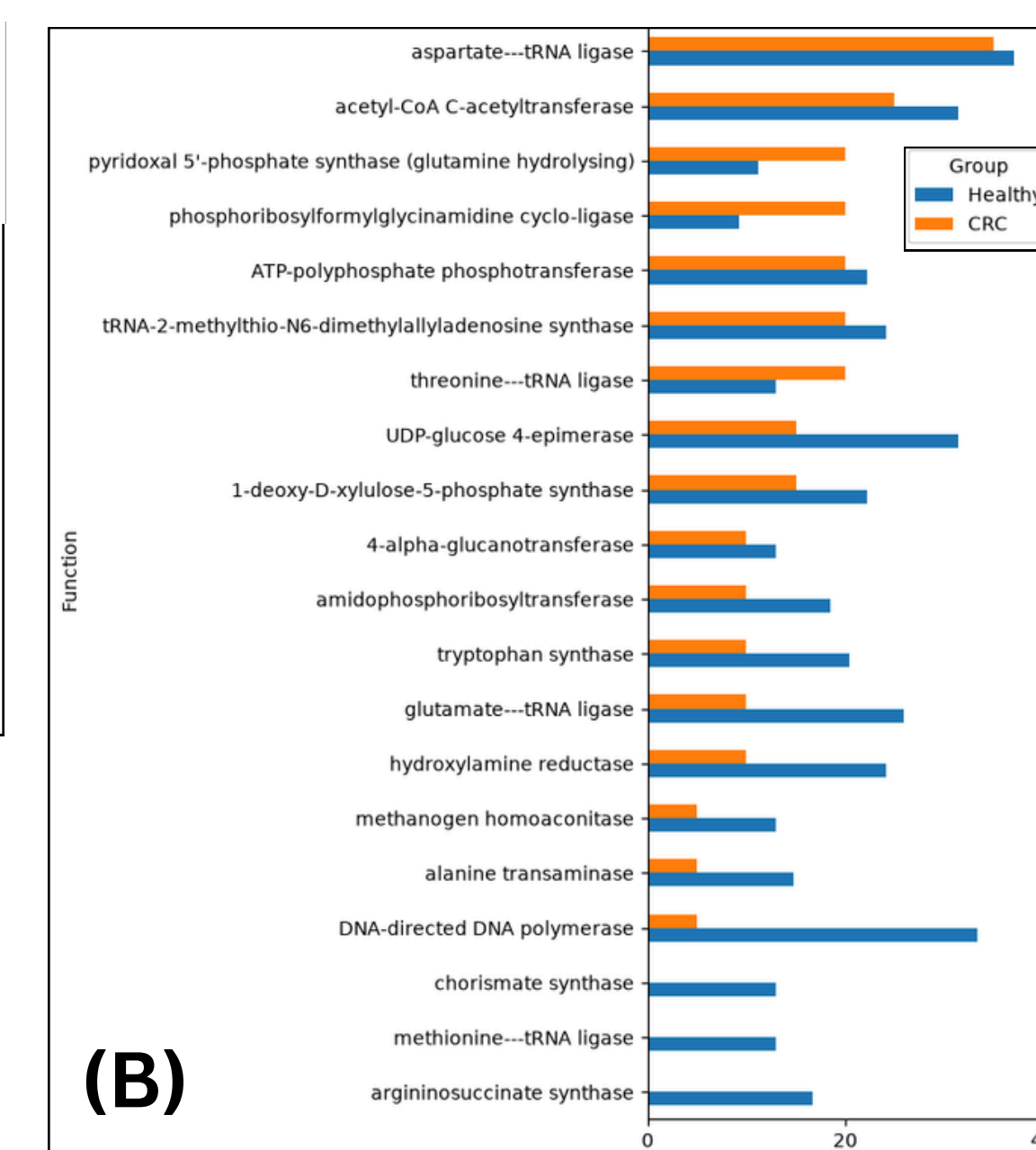


2 RESULT / METRIC - РЕЗУЛЬТАТЫ / МЕТРИКИ



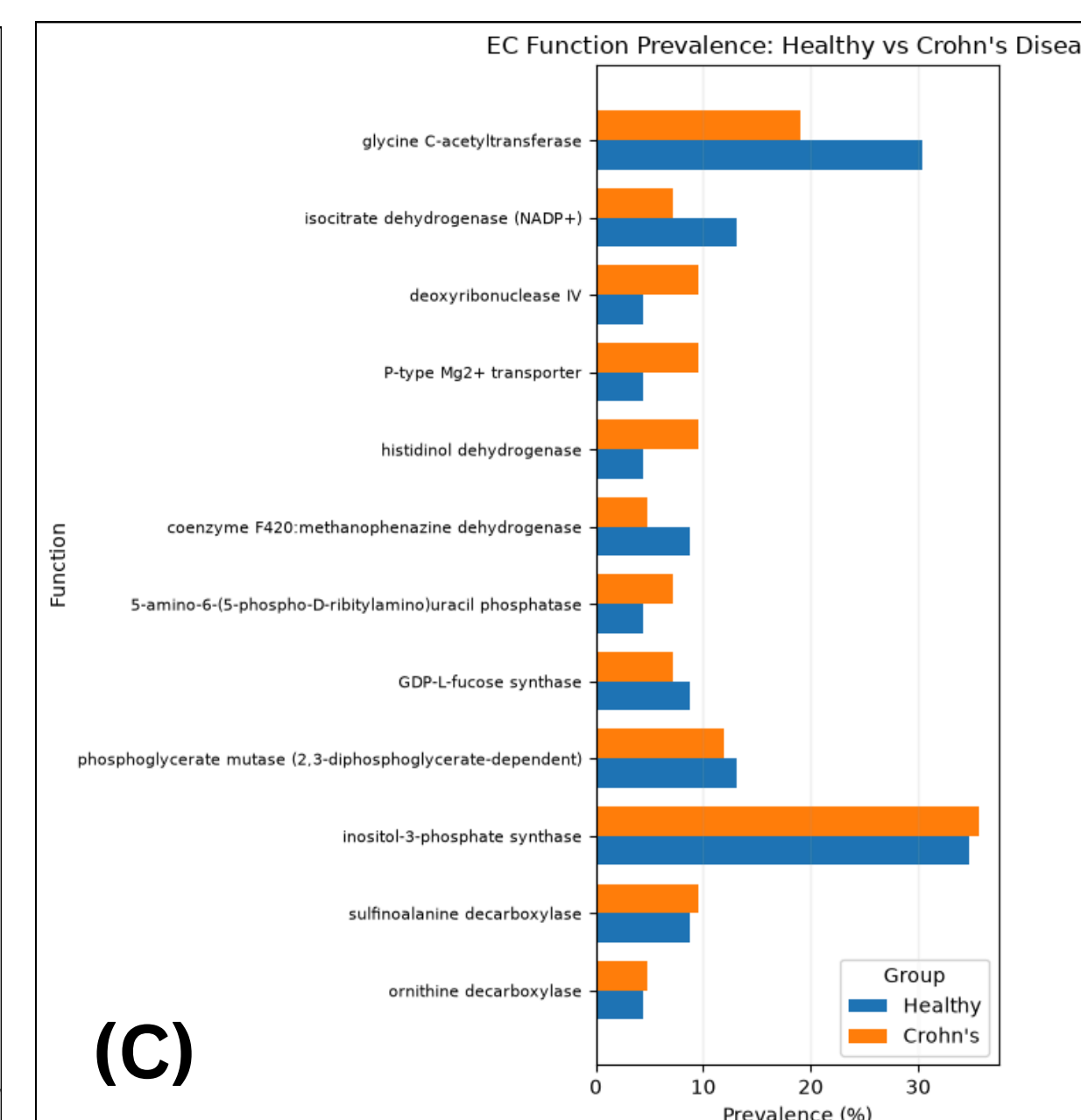
(A) Functional clusters show distinct abundance patterns across healthy, Crohn's, and CRC groups → disease-associated functional differences are detectable.

Функциональные кластеры различаются по численности между Healthy, Crohn's и CRC → выявляются связанные с заболеванием функциональные различия.



(B) Selected EC functions differ in prevalence between healthy and CRC → CRC-associated functional signatures are observed.

Распространённость выбранных ЕС-функций различается между Healthy и CRC → наблюдаются функциональные признаки, связанные с CRC.



(C) Selected EC functions differ between healthy and Crohn's, but substantial overlap remains → the functional signal is weaker than for CRC.

Выбранные ЕС-функции различаются между Healthy и Crohn's, но сохраняется значительное перекрытие → функциональный сигнал слабее, чем для CRC.

ЗАКЛЮЧЕНИЕ

- Такие заболевания, как болезнь Крона и колоректальный рак, ассоциированы с изменениями микробиома кишечника
- Глубокое обучение повышает точность выявления этих изменений, поскольку не ограничено нашими текущими знаниями о функциях генов/микробов
- Эти изменения микробиома также можно выявить с помощью классического машинного обучения

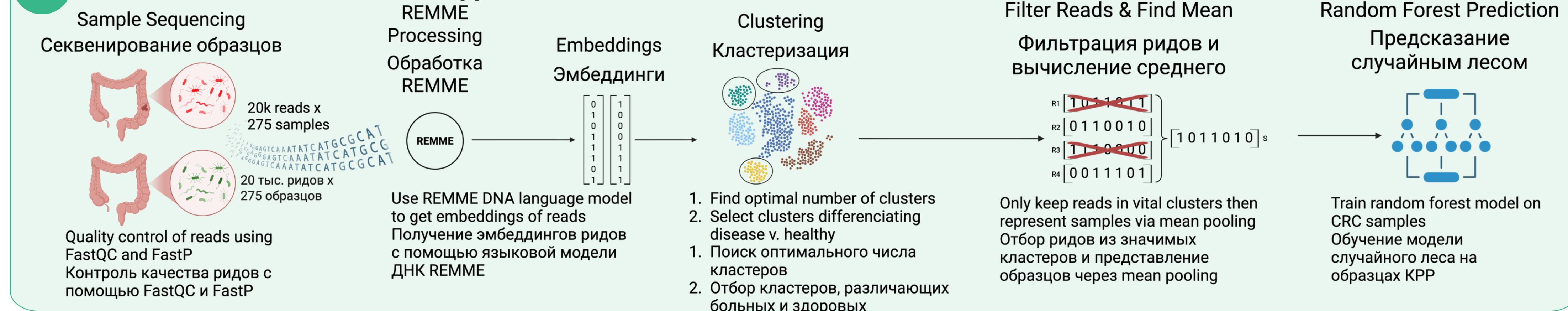
ВСТУПЛЕНИЕ

- Микробиом кишечника тесно связан со здоровьем человека, однако его состав значительно **различается** между людьми.
- Данные микробиома **высокоразмерны** и **разрежены**, а закономерности, связанные с конкретными заболеваниями, заранее **неизвестны**.

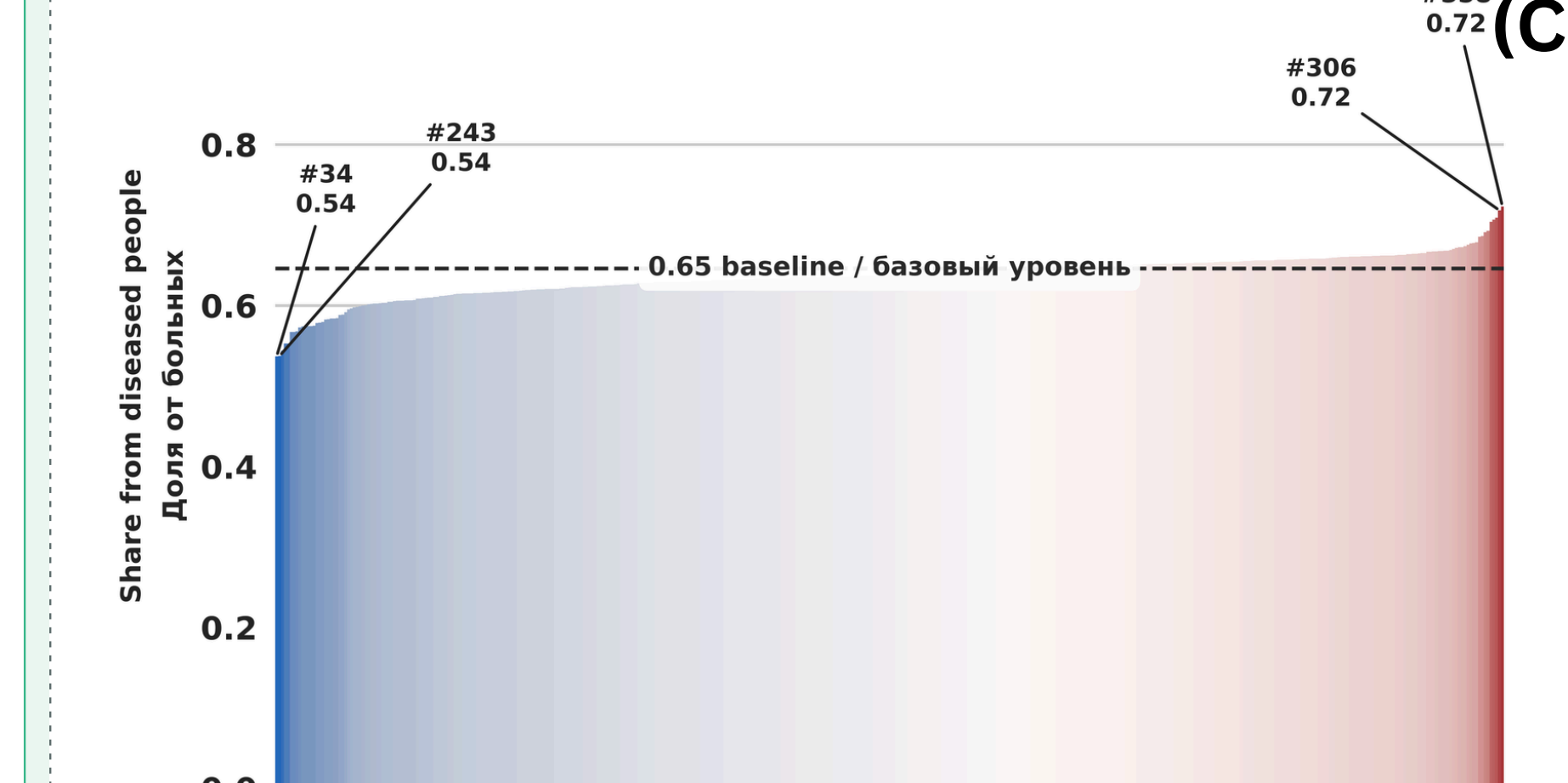
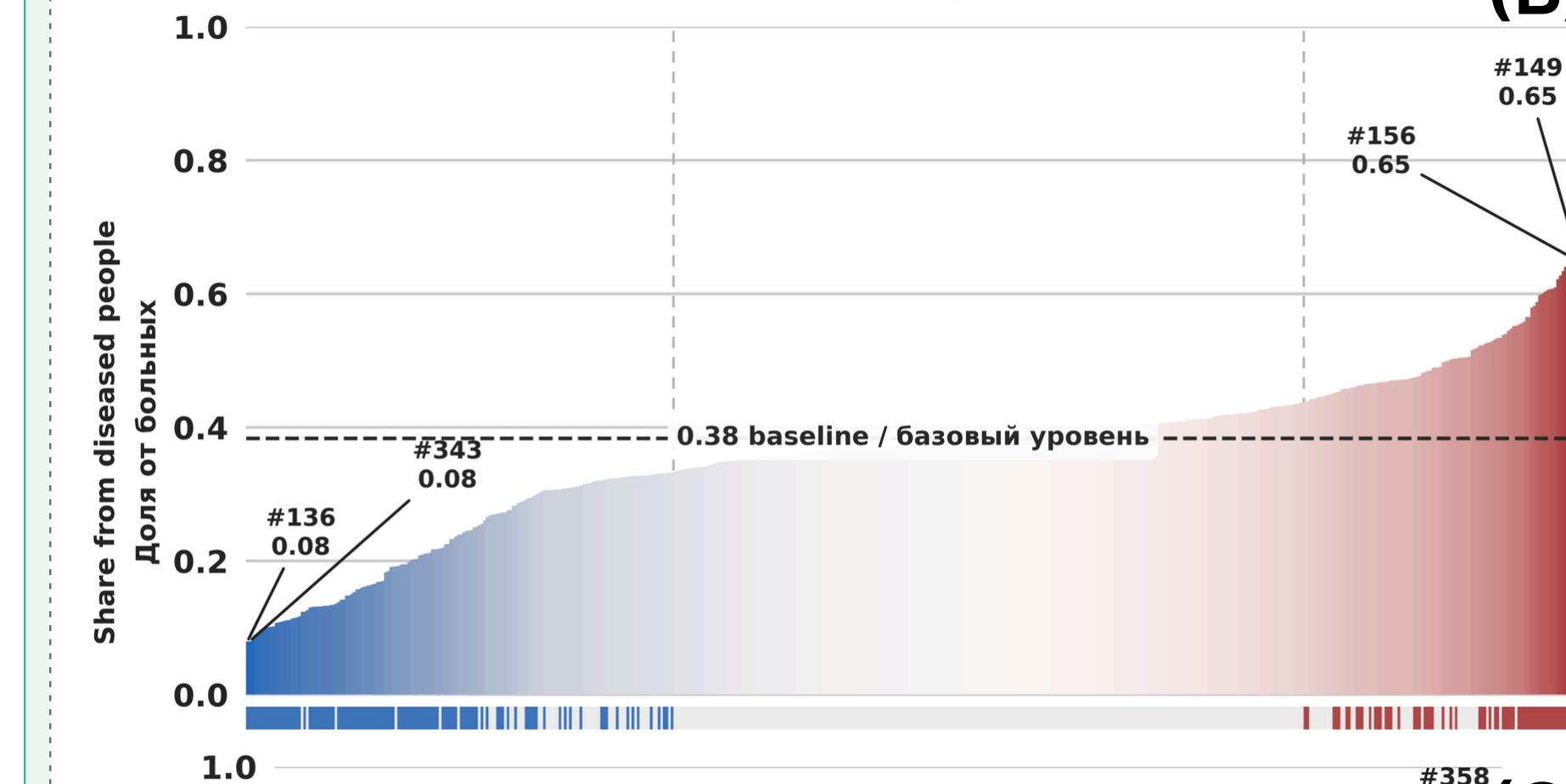
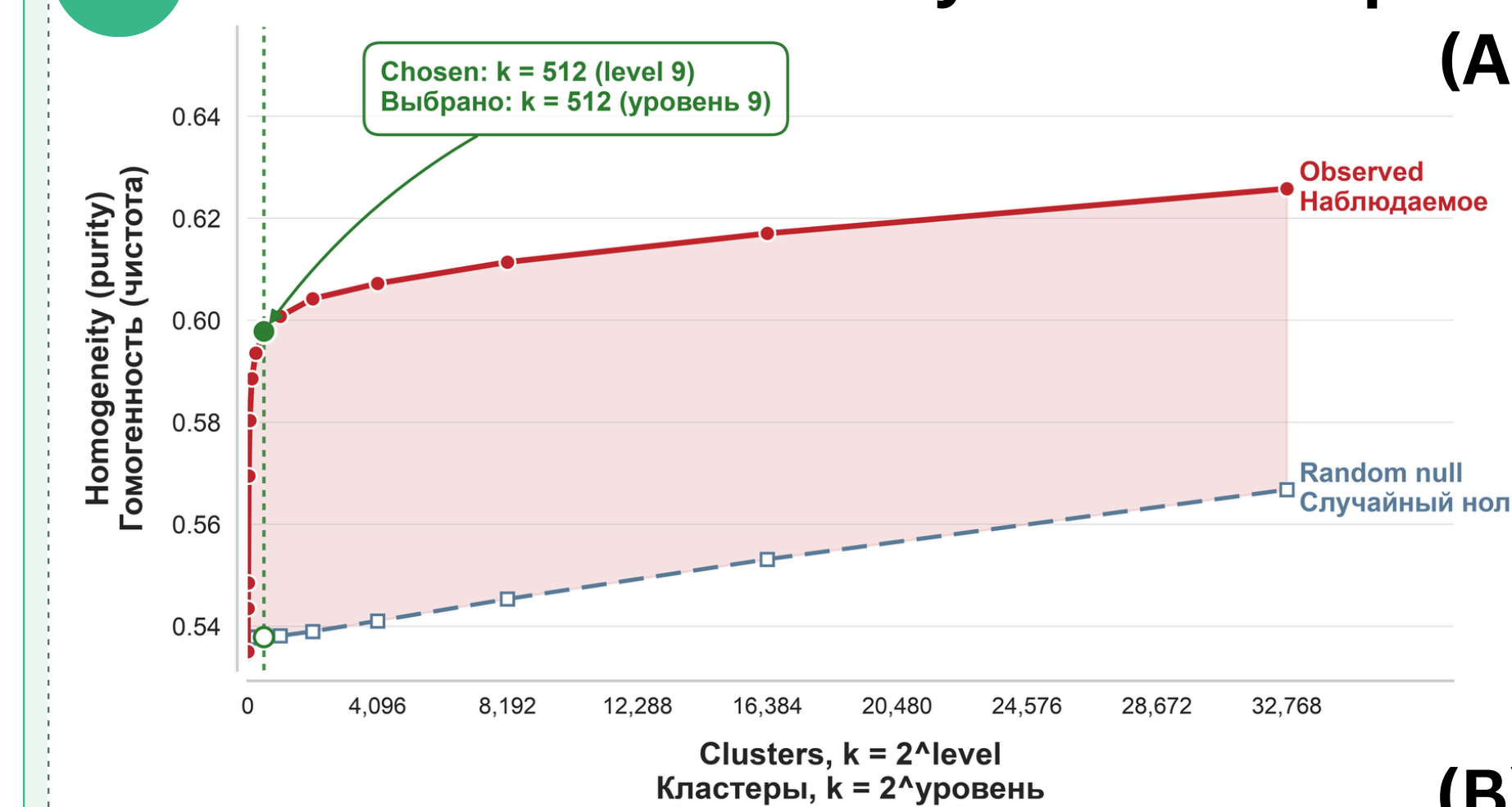


Project B/Проект В - REMME (deep learning based)

1 METHOD / PIPELINE - Метод / Пайплайн



2 RESULT / METRIC - Результат / Метрика



(A) Bisecting k-means: homogeneity (y-axis) of read clusters rises sharply then plateaus. X-axis k = 512 chosen as the balance between purity and cluster size

Бисекционная k-means: гомогенность (ось Y) кластеров ридов сначала резко растёт, а затем выходит на плато. Ось X — k = 512 выбран из-за баланса между чистотой и размером кластеров

Per cluster we scored the **share of disease samples** to find reads specific to disease or control. Для каждого кластера считалась **доля больных образцов**, чтобы найти риды, специфичные для болезни или контроля.

(B) CRC: sharp tails = pure control / pure disease clusters → usable.

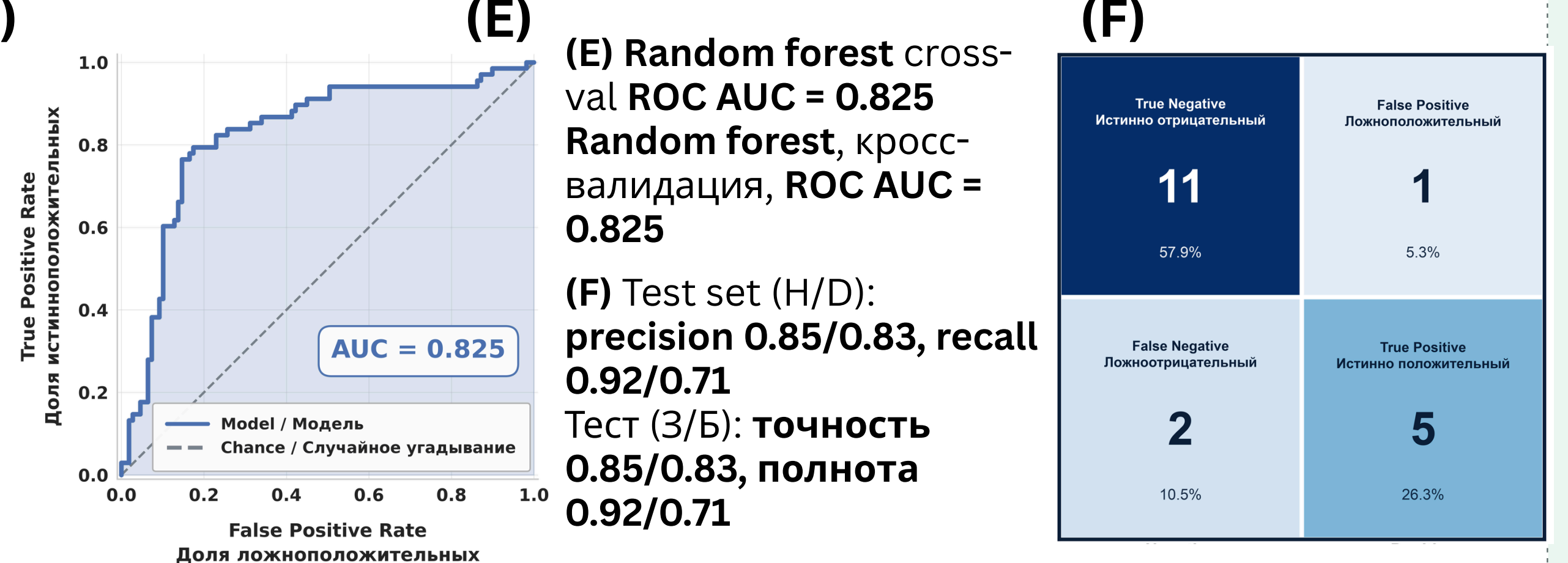
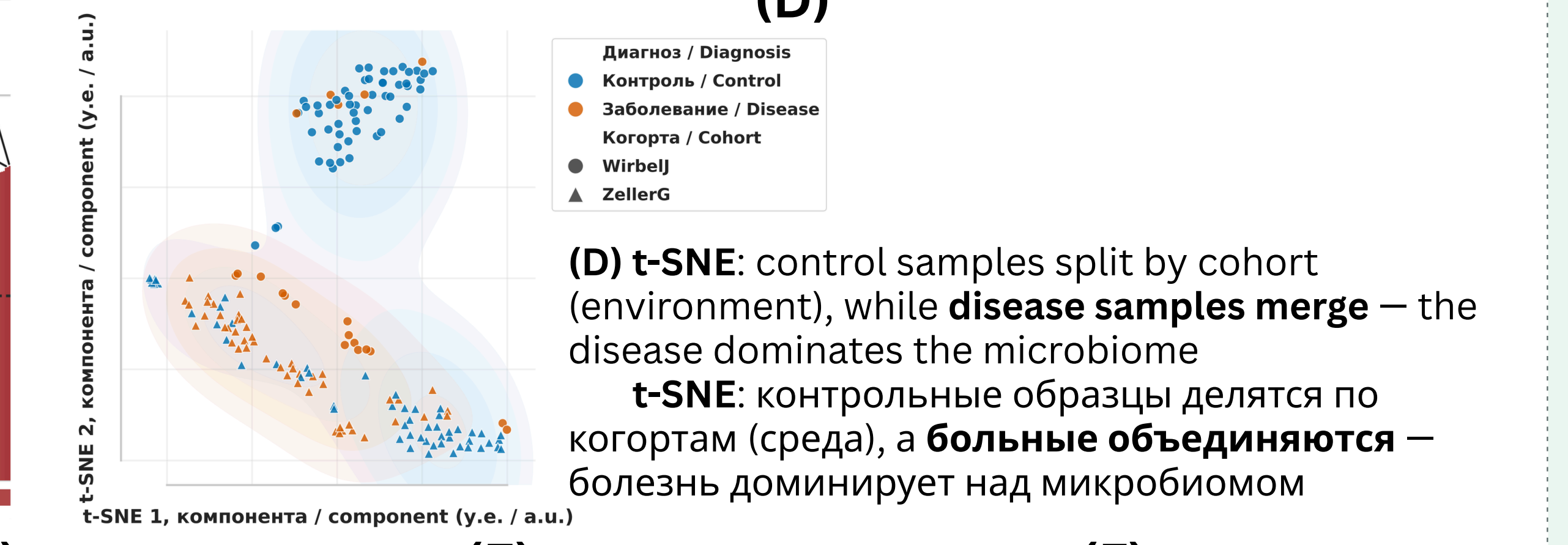
KPP: резкие хвосты = чистые кластеры контроля / болезни → пригодно.

(C) Crohn's: flat, all mixed → no signal.

Крон: плоско, всё смешано → сигнала нет.

CRC edge clusters kept and their reads **mean-pooled** → 1 embedding per sample.

Краевые кластеры KPP сохранены, их риды **усреднены** → 1 эмбединг на образец.



(F) Test set (H/D): precision 0.85/0.83, recall 0.92/0.71. Тест (3/5): точность 0.85/0.83, полнота 0.92/0.71

CONCLUSION

- Diseases such as Crohn's disease and colorectal cancer are associated with changes in the gut microbiome
- Deep learning improves detection of microbiome shifts, as it isn't limited by our current knowledge of gene/microbial function
- These microbiome shifts can also be detected using classical machine learning as well

REFERENCES

- Prabakaran R & Bromberg Y. Nucleic Acids Res 53(16):gkaf836 (2025).
- Zhu C, Miller M, Marpa S, et al. Nucleic Acids Res 46:e23 (2018).
- Wirbel J, Pyl PT, Kartal E, et al. Nat Med 25:679–689 (2019).
- Zeller G, Tap J, Voigt AY, et al. Mol Syst Biol 10:766 (2014).
- Sankaranarayanan K, Ozga A, Warinner C, et al. Curr Biol 25:3161–3169 (2015).
- Lloyd-Price J, Arze C, Ananthakrishnan AN, et al. Nature 569:655–662 (2019).
- Chen S, Zhou Y, Chen Y, Gu J. Bioinformatics 34(17):i884–i890 (2018).
- Andrews S. FastQC: A Quality Control Tool for High Throughput Sequence Data (2010).
- Li H. arXiv:1303.3997 [q-bio.GN] (2013).