В этой статье было проведено сравнение биоинформатических конвейеров по их специфичности и чувствительности анализа секвенированных микробных 16S рРНК-ампликонов. В это сравнение были включены шесть различных конвейеров: QIIME (v.1.9.1), MOTHUR (v.1.39.5), DADA2 (1.7.0), Qiime2 (v.2017.6.0)-Deblur, USEARCH (v.10.0.240)-UPARSE и USEARCH (v.10.0.240)-UNOISE3. Конвейеры сравнивались с использованием макета образца, неоднократно секвенированный в течение нескольких циклов секвенирования, а также с большим (N = 2170 особей) набора данных фекальных образцов из многоэтнического исследования «Здоровая жизнь в городских условиях» (HELIUS).

**Макет сообщества.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Species** | **Strain** | **NCBI\_reference** | **16S Copy Number** | **16S variants in V4 region** |
| Acinetobacter baumannii | ATCC 17978 | NC\_009085 | 5 | 1 |
| Actinomyces odontolyticus | ATCC 17982 | NZ\_AAYI02000000 | 2 | 1 |
| Bacillus cereus | ATCC 10987 | NC\_003909 | 12 | 1 |
| Bacteroides vulgatus\* | ATCC 8482 | NC\_009614 | 7 | 3 (5:1:1 ration) |
| Clostridium beijerinckii\* | NCIMB 8052 | NC\_009617 | 14 | 2 (13:1 ration) |
| Deinococcus radiodurans | R1 (smooth) | NC\_001263 and NC\_001264 | 3 | 1 |
| Enterococcus faecalis | OG1RF | NC\_17316 | 4 | 1 |
| Escherichia coli | K12 substrain MG1655 | NC\_000913 | 7 | 1 |
| Helicobacter pylori | 26695 | NC\_000915 | 2 | 1 |
| Lactobacillus gasseri | 63 AM | NC\_008530 | 6 | 1 |
| Listeria monocytogenes | EGDe | NC\_003210 | 6 | 1 |
| Neisseria meningitidis | MC58 | NC\_003112 | 4 | 1 |
| Propionibacterium acnes | KPA171202 | NC\_006085 | 3 | 1 |
| Pseudomonas aeruginosa | PAO1-LAC | NC\_002516 | 4 | 1 |
| Rhodobacter sphaeroides | ATH 2.4.1 | NC\_007493 and NC\_007494 | 3 | 1 |
| Staphylococcus aureus# | TCH1516 | NC\_010079 | 5 | 1 (V4 identical to S. epidermidis) |
| Staphylococcus epidermidis# | FDA strain PCI 1200 | NC\_004461 | 5 | 1 (V4 identical to S. aureus) |
| Streptococcus agalactiae | 2603 V/R | NC\_004116 | 7 | 1 |
| Streptococcus mutans | UA159 | NC\_004350 | 5 | 1 |
| Streptococcus pneumoniae | TIGR4 | NC\_003028 | 4 | 1 |

Геномная ДНК из микробного имитационного сообщества был секвенирован в трех отдельных прогонах. Макет содержит ДНК из 20 бактериальных штаммов в эквимолярном (равном) количестве рибосомальной РНК оперона (100000 копий на организм на мкл). Два штамма (Bacteriodes vulgatus и Clostridium beijerinckii) имеют несколько вариантов последовательности в области V4 гена 16S рРНК. В. Vulgatus имеет три варианта (в соотношении 5:1:1), тогда как C. Beijerinckii имеет два варианта (в соотношении 13:1). Последовательности 16S рРНК *Staphylococcus aureus* и *Staphylococcus epidermidis* идентичны в области V4. Таким образом, макет содержит в общей сложности 22 варианта (ASV) гена 16S в области V4. Макет сообщества был секвенирован три раза в разных последовательных запусках.

Три пробных прогона образцов имеют 36464, 84054 и 146653 парных считываний соответственно.

Макет необработанных данных последовательности является общедоступным (<https://github.com/andreiprodan/mock-sequences>).

#### Набор данных образцов фекалий HELIUS.

В общей сложности было секвенировано 2170 образцов фекалий, полученных от взрослых особей из шести этнических групп в Амстердаме, Нидерланды (исследование HELIUS). Набор данных образцов фекалий HELIUS содержал 177,08 миллиона парных считываний, полученных из 17 отдельных запусков секвенирования. Все необработанные данные секвенирования из этого набора данных доступны в Европейском архиве генома-феномена (присоединение no. EGAD00001004106).

Все таблицы OTU/ASV, полученные конвейерами, были преобразованы в объекты phyloseq с использованием пакета "phyloseq"(v.1.24.2). Меры альфа-разнообразия (богатство, Chao1, индекс Шеннона, обратный индекс Симпсона) были рассчитаны с использованием функции «estimate\_richness» от «phyloseq». OTU/ASV были классифицированы как «точные» (идеальное соответствие истинной последовательности в макетном сообществе), «одноразовые» (на расстоянии 1 Хэмминга от истинной последовательности) или «другие» (на расстоянии более 1 Хэмминга от истинной последовательности). Расстояния между последовательностями Хэмминга были рассчитаны с использованием пакета R "stringdist" (v.0.9.5.1). «One-off» и «Other» были вместе помечены как «Spurious».

#### Чувствительность и специфичность.

Для анализа был взят макет сообщества с 20 бактериальными штаммами. Был проведен обзор точных, одноразовых и ложных ОТУ/ASV, создаваемых различными конвейерами с использованием считывания из трех пробных запусков секвенирования.

DADA2 показал лучшую чувствительность, обнаружив все 22 истинных ASV, присутствующих в макете, и был единственным конвейером, способным дифференцировать последовательности с одноосным разрешением даже при высоких коэффициентах изобилия (например, соотношение 13: 1 между двумя C. Beijenrickii варианты). Несмотря на лучшую чувствительность, потоки DADA2 произвели некоторые ложные ASV.

USEARCH-UNOISE3 и Qiime2-Deblur были единственными двумя конвейерами, которые показали идеальную специфичность в данных секвенирования макетов образцов, не производя ложных OTU / ASV.

**Анализ с набором фекальных образцов HELIUS**

В анализе набора данных фекальных образцов HELIUS наблюдалась 3,5-кратная разница между наибольшим количеством ASV, произведенных трубопроводом (около 25000, в DADA2) и наименьшим числом (более 7500, в USEARCH-UNOISE3). Qiime2-Deblur произвел около 11000 ASV. И снова DADA2 показал как лучшую чувствительность, так и самую высокую склонность к ложным ASV среди трех конвейеров уровня ASV.

Помимо этого, оценивалось и алфа-разнообразие. Наблюдалось значительное различие между конвейерами по альфу-разнообразию. Обычно инструменты, использующие кластеризацию ASV, имеют более высокое альфа-разнообразие, чем использующие кластеризацию OTU.

Тем не менее, два типа ошибок могут искажать воспринимаемое альфа-разнообразие. Во-первых, огромное количество ложных ОТУ может значительно раздувать воспринимаемое альфа-разнообразие (как для Qiiime-clust). Во-вторых, как наблюдалось для QIIME2-Deblur, конвейер уровня ASV может не различать очень тесно связанные истинные биологические последовательности и объединять их вместе в единый ASV. Это искусственно уменьшит воспринимаемое альфа-разнообразие по сравнению с более чувствительными трубопроводами уровня ASV, и является причиной того, что Qiime2-Deblur дал более низкие значения альфа-разнообразия по сравнению с DADA2 и USEARCH-UNOISE3.

**Заключение**

Между различными трубопроводами наблюдались большие различия в чувствительности и специфичности. DADA2 показал лучшую чувствительность и разрешение (за ним следует USEARCH-UNOISE3) при стоимости производства большего количества ложных ASV по сравнению с USEARCH-UNOISE3 и Qiime2-Deblur. USEARCH-UPARSE и MOTHUR произвели одинаковое количество ОТУ. Рабочие процессы QIIME-uclust произвели огромное количество ложных ОТУ, а также завышенные показатели альфа-разнообразия, независимо от параметров фильтрации качества.

Рабочие процессы на уровне ASV предлагают превосходное разрешение по сравнению с уровнем OTU, и в этом исследовании была показана лучшая специфичность и более низкие скорости ложных последовательностей. Кроме того, конвейеры на уровне ASV позволяют легче интегрировать биологические особенности между исследованиями, поскольку ASV имеют внутреннее биологическое значение, независимо от справочной базы данных или контекста исследования.

В статье автор считает, что DADA2 является лучшим выбором для исследований, требующих максимально возможного биологического разрешения (например, исследований, ориентированных на дифференциацию близкородственных штаммов). А USEARCH-UNOISE3 имеет лучшую общую производительность, сочетая высокую чувствительность с отличной специфичностью.