**Инструкция по использованию программы Alignment analysis**

**Общие определения и цели программы**

Программа анализирует нуклеотидные пангеномы, которые получаются при запуске программы npge и исправляет в ней типичные ошибки аннотации. Для этого все гены разбиваются на группы – множества взаимопересекающихся генов с перекрытием выше заданного, а после анализирует, есть ли в данной группе те или иные ошибки аннотации и исправляет саму аннотацию.

С типичным пангеномом программа работает в течение нескольких минут, однако это время зависит от размера пангенома.

Текущая версия программы 1.1.0

**Требования программы**

Для запуска программы на компьютер должен быть установлен питон 3, а также такая библиотека как pandas.

**Входные данные для программы**

В качестве входных данных используется выдача программы npg\_explorer, а именно файлы “genes/partition-ungrouped.tsv” и “pangenome/pangenome.bs”.

**Выходные данные программы**

В выходной директории программы (которая называется “alan\_result”) содержатся следующие файлы и папки: “align\_split” – вспомогательная директория с файлами разбиения пангенома, не является выходным файлом, нужна для работы программы, “general\_result.txt” – общие результаты работы программы (число групп, число групп с ошибкой в старте, число групп с местами, требующими внимания), “group\_result.xlsx” – таблица групп и указаний, в каких из них присутствует ошибка в старте, а также, где требуется внимание, “partition\_remastered.xlsx” – переделанная таблица “partition-ungrouped.tsv” с добавленной туда информацией о принадлежности каждого гена группе.

**Синтаксис и параметры программы**

Запускается командой “python alignment\_analysis.py [набор параметров]”. Пример – “python alignment\_analysis.py -i directory1/directory2 -l 100”.

Программа содержит только необязательные параметры. Для типичного запуска без параметров, программу требуется запускать из папки с выбранным пангеномом (выдачей программы npge).

-h – написать сообщение с помощью и завершить программу

-i – указать входную директорию (если требуется запустить программу не из папки с пангеномом, указать требуется как раз саму папку с пангеномом, по умолчанию подаётся текущая директория)

-o – указать выходную директорию (в ней будет создана папка “alan\_result” с результатами работы, по умолчанию текущая папка)

-l – указать требуемое минимальное перекрытие генов для образования группы (по умолчанию 90 нуклеотидов)

-a – указать, если разбиение пангенома уже было произведено ранее, ускоряет работу программы

**Известные проблемы и ошибки**

Программа находится в разработке, поэтому пока что не выполняет всех выдвинутых к ней требований. Список известных проблем:

1. Программа не учитывает возможность пересечения генов суммарно в нескольких блоках, однако гены из разных блоков объединяет вместе
2. Есть проблема, когда два разных гена в выдаче программы npg\_explorer названы одинаково, тогда данная программа их не обрабатывает
3. У программы временное имя, требуется придумать постоянное
4. Программа не анализирует последовательности в блоках, которые не имеют на данном месте аннотированного гена
5. Сделать вывод количества ошибок программы