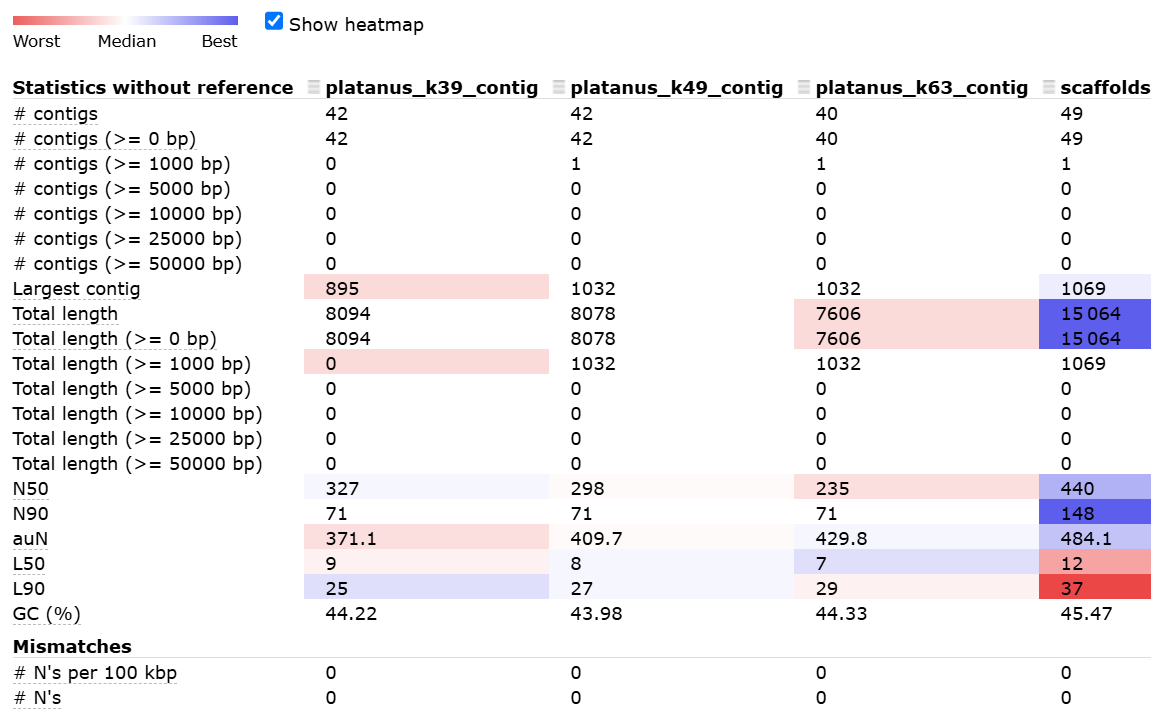
Задание 11

Трещева Мария

**На семинаре взяли данные секвенирования вируса гриппа** из /projects/mipt\_dbmp\_biotechnology/genome\_de\_novo/.

Собрали геном с разными k-мерами. С помощью Quast оценили качество сборки.



Ппервые три колонки – результаты сборки SPAdes, последняя – Platanus. У SPAdes результаты лучше, поэтому оптимизируем эту сборку. Возьмем еще несколько более длинных контигов с тем же шагом 22. Длинные контиги хороши тем, что они уменьшают риск разрывов в сборке + при наличии повторов покрывают лучше покрывают их. Проблемы длинных контигов могут быть в пропусках редких/мало представленных среди прочтений вариантов. Делаем:

python3 /projects/mipt\_dbmp\_biotechnology/soft/SPAdes-4.1.0-Linux/bin/spades.py --careful **-k 21,33,55,77,99,111**

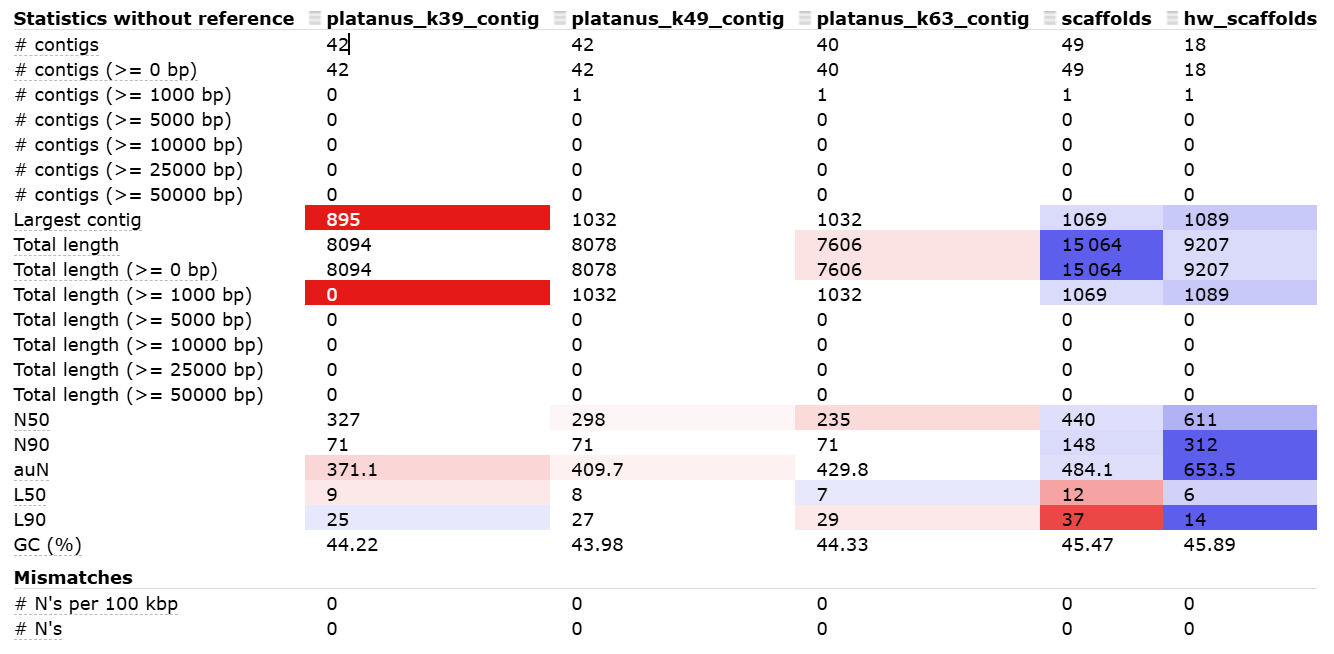
-1 /projects/mipt\_dbmp\_biotechnology/genome\_de\_novo/7\_S4\_L001\_R1\_001.fastq

-2 /projects/mipt\_dbmp\_biotechnology/genome\_de\_novo/7\_S4\_L001\_R2\_001.fastq

-o ~/homeworks/hw\_11/genome\_assembly\_results/spades

Используемые k-меры те же: 39, 49, 63. Для проверки качества сборки:

~/soft/bin/quast.py -o ~/homeworks/hw\_11/assembly\_analysis -m 0 --threads 1 ~/classes/class\_11/genome\_assembly\_results/k39/platanus\_k39\_contig.fa ~/classes/class\_11/genome\_assembly\_results/k49/platanus\_k49\_contig.fa ~/classes/class\_11/genome\_assembly\_results/k63/platanus\_k63\_contig.fa ~/classes/class\_11/genome\_assembly\_results/spades/scaffolds.fasta ~/homeworks/hw\_11/genome\_assembly\_results/spades/hw\_scaffolds.fasta

**

Длина сборки стала меньше. Длина генома вируса ~13kb, так что семинарский вариант чуть «пересобрал» геном – могли перепредставить повторы в геноме, а мы «недособрали» - могли потерять часть информации.

Но мы улучшили метрики!

**N50** – самый короткий контиг из множества контигов, сумма которых образует половину длины всей сборки (генома).

**Увеличилась**, то есть контиги стали менее короткими => меньше фрагментировали геном для сборки.

**L50** – наименьшее число контигов, сумма длин которых составляет половину размера генома.

**Уменьшилась**, то есть контиги стали длиннее.

Та же тенденция с **N90, L90.**

**auN** – интегральная метрика по всем N1,…,100. Мы тоже радуемся, что она выросла.