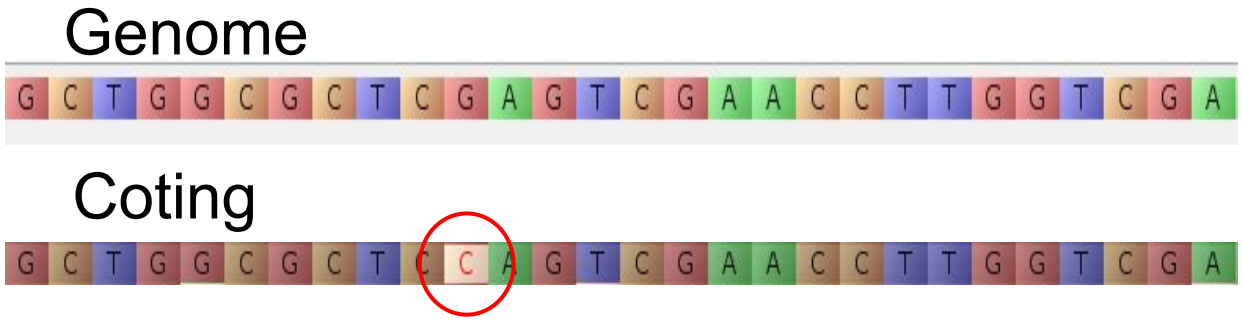
Corrector 2.1

Максим Винниченко

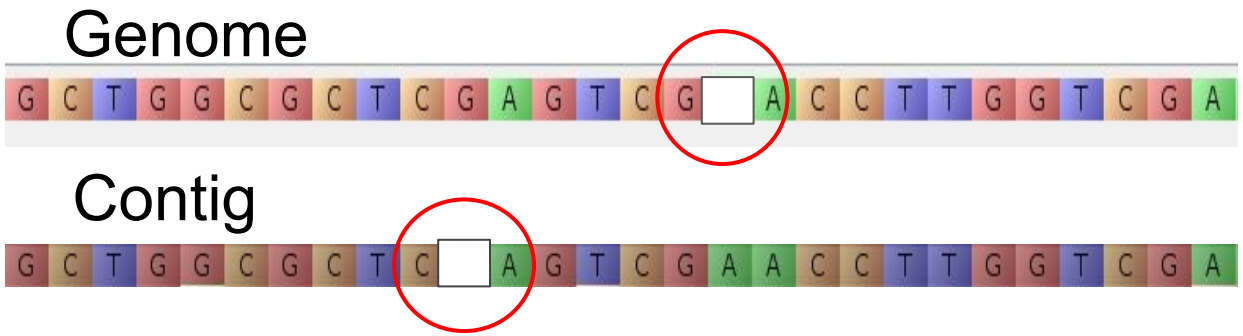
руководитель: Дмитрий Антипов

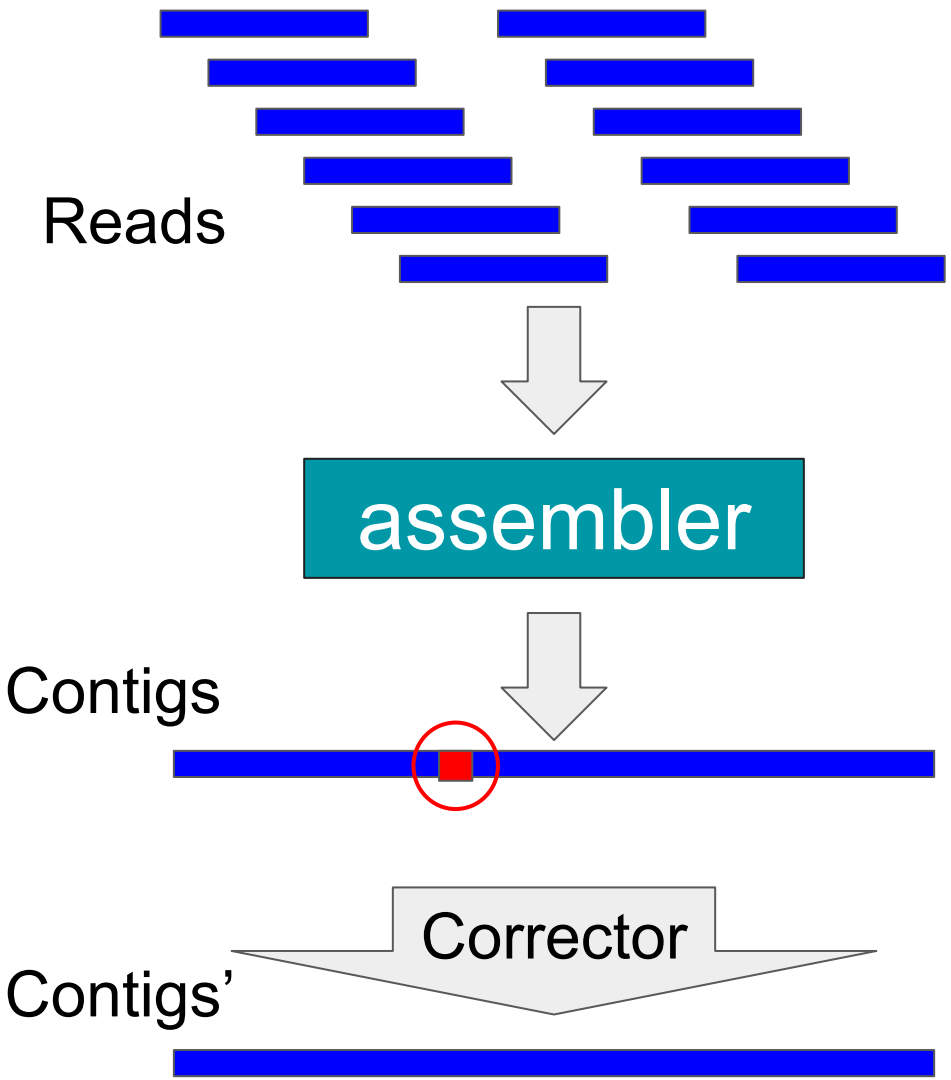
Ошибки в контигах

mismatch (substitution)



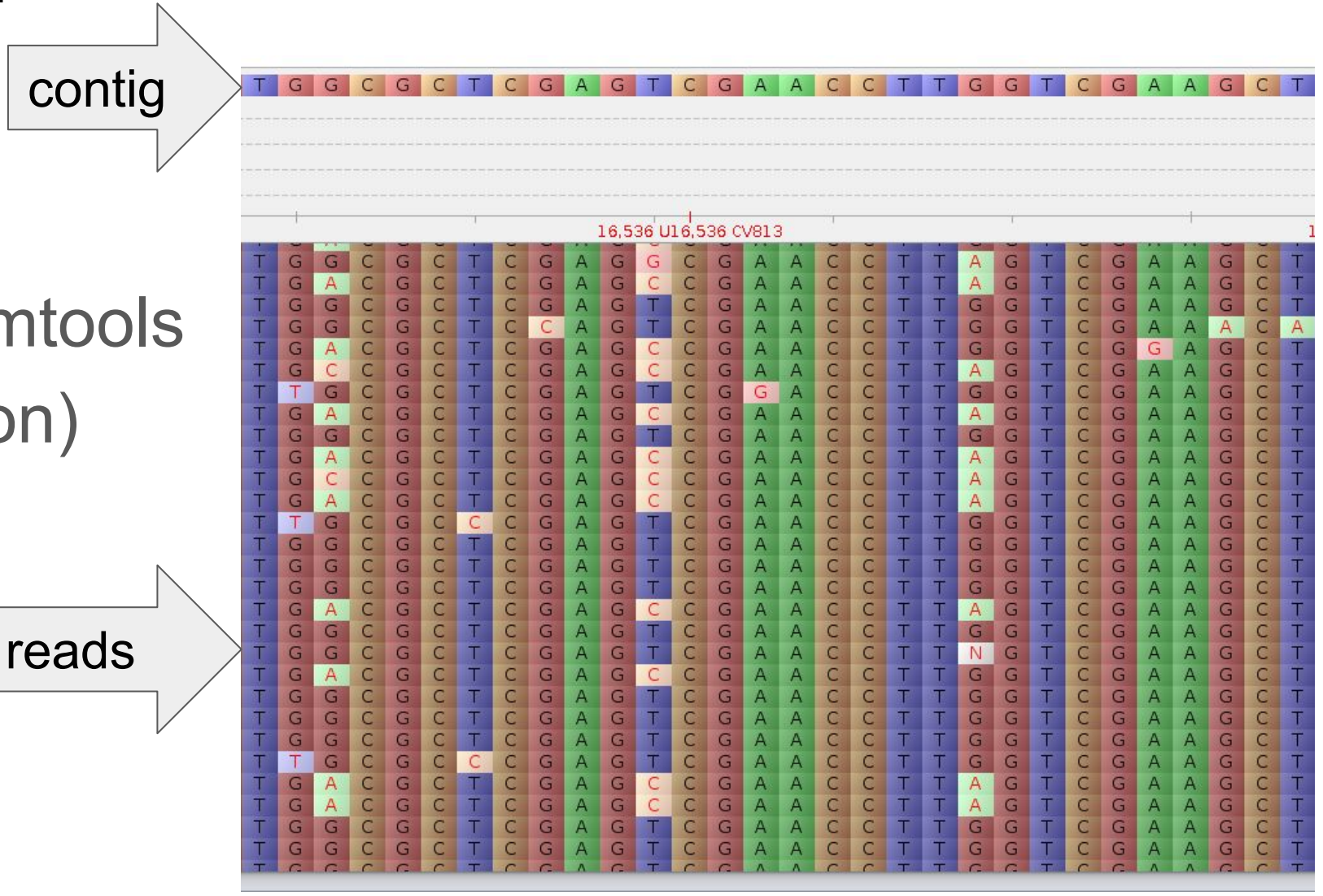
indels





Прикладывание ридов

- bwa/bowtie2 + samtools
- Tablet (visualization)



Актуальность

Pilon (релиз 2014)

- Corrector 1.0 (из Spades)

- за время стажировки была создана новая версия
  - другой алгоритм (работающий менее агрессивно)

- Часто сборки из pacbio и nanopore полируют иллюминой

Цели и Задачи

- улучшить работу на сборках из ридов Nanopore
- SAM→BAM (меньше места на диске, ускорить)
- улучшить качество с помощью использования множественных выравниваний

Эвристика для сборок из ридов Nanopore

- В сборках из ридов Nanopore много индел, даже больше чем мисматч.
- Исправляем инделы даже при маленьком покрытии
- Используем bowtie2 для выравнивания
- Добавлен выбор bwa / bowtie2

Результаты эвристики

сборка бактерии	per 100 kbp	было	pilon	старая версия	версия со стажировки	indels fix
Enterobacter cloacae	mismatches	86.54	6.44	4.18	0.08	0.12
	indels	649.51	5.78	5.18	3.24	2.18
Klebsiella pneumoniae	mismatches	41.02	3.31	1.45	0.25	0.27
	indels	661.65	5.40	4.09	5.20	3.07
Serratia marcescens	mismatches	6.38	3.56	2.12	0.98	1.29
	indels	464.70	9.73	4.47	5.14	2.51

SAM→BAM

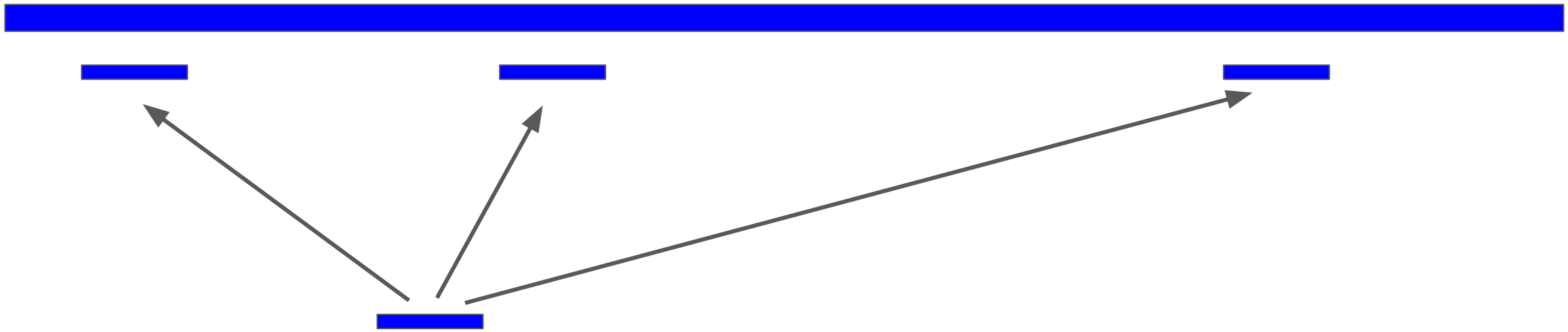
- SAM — текстовый формат для хранения выравнивания ридов на референс

- BAM — бинарная и сжатая версия

- теперь выравнивание ридов на контиги храниться в BAM
- появилась настройка, позволяющая вернуться на SAM

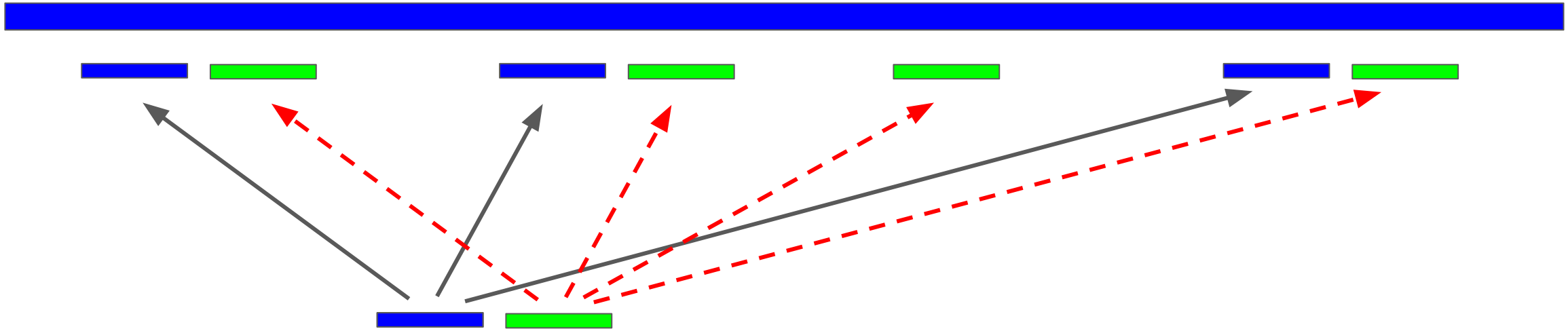
- занимает в **3.5** раза меньше места
- запись на **20%** быстрее на медленном (HDD) диске
- распределение выравниваний ридов дрожжей по контигам: 180 мин → 145 мин

Множественное выравнивание (multiple alignment)



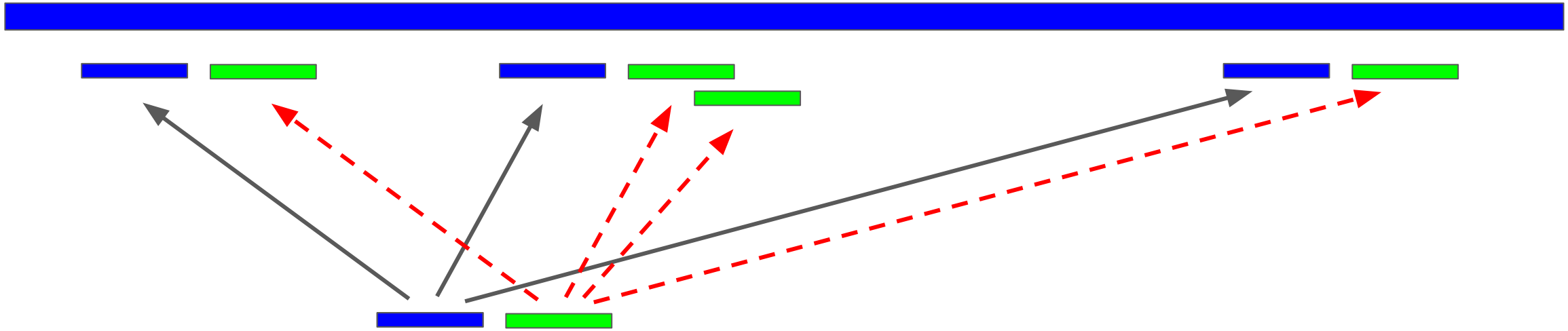
- У рида есть несколько вариантов куда приложиться
- В файле выравнивания одного рида идут подряд

Множественное выравнивание (парные риды)



- У рида каждого рида *из пары* есть несколько вариантов куда приложиться
- Даёт больше информации
  - о повторности участков
  - о плохо покрытых участках

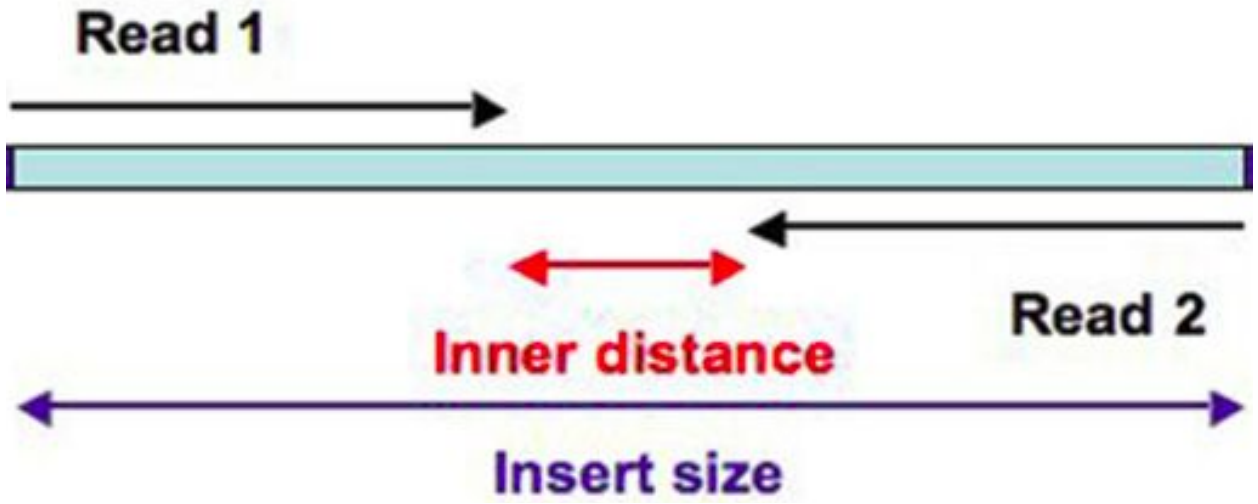
Множественное выравнивание (парные риды)



- Сложнее разбить риды на пары
- Больше шума

Множественное выравнивание

- Определяем медианный **insert size** по ридам, которые выровнялись однозначно, оцениваем его разброс



- Обработка *парного рида*, который приложился неоднозначно
  - по прикладыванию *Read 1* прикладывание *Read 2* нашлось однозначно по разбросу insert size
    ⇒ обрабатываем пару
  - Иначе игнорируем
- Вносим изменения в настройки алгоритма коррекции

Результаты

сборка бактерии	per 100 kbp	было	версия со стажировки	indels fix	multiple alignment
Enterobacter cloacae	mismatches	86.54	0.08	0.12	0.82
	indels	649.51	3.24	2.18	0.52
Klebsiella pneumoniae	mismatches	41.02	0.25	0.27	0.34
	indels	661.65	5.20	3.07	0.43
Serratia marcescens	mismatches	6.38	0.98	1.29	2.97
	indels	464.70	5.14	2.51	1.17

Вопросы