Обзор литературы по ассембелрам

Дмитрий Яковлев

EPAM Systems

17 октября 2016 г.

План

1. Ввведение
2. Модель ошибки на данных BioNano
3. Ассемблеры
   - TWIN
   - OPTIMA
   - MAligner
   - OMBlast

Ввведение

Модель ошибки на данных BioNano

Ассемблеры

TWIN

OPTIMA

MAligner

OMBlast

OMBlast: Общие сведения

OMBlast: Алгоритм

Этапы выравнивания:

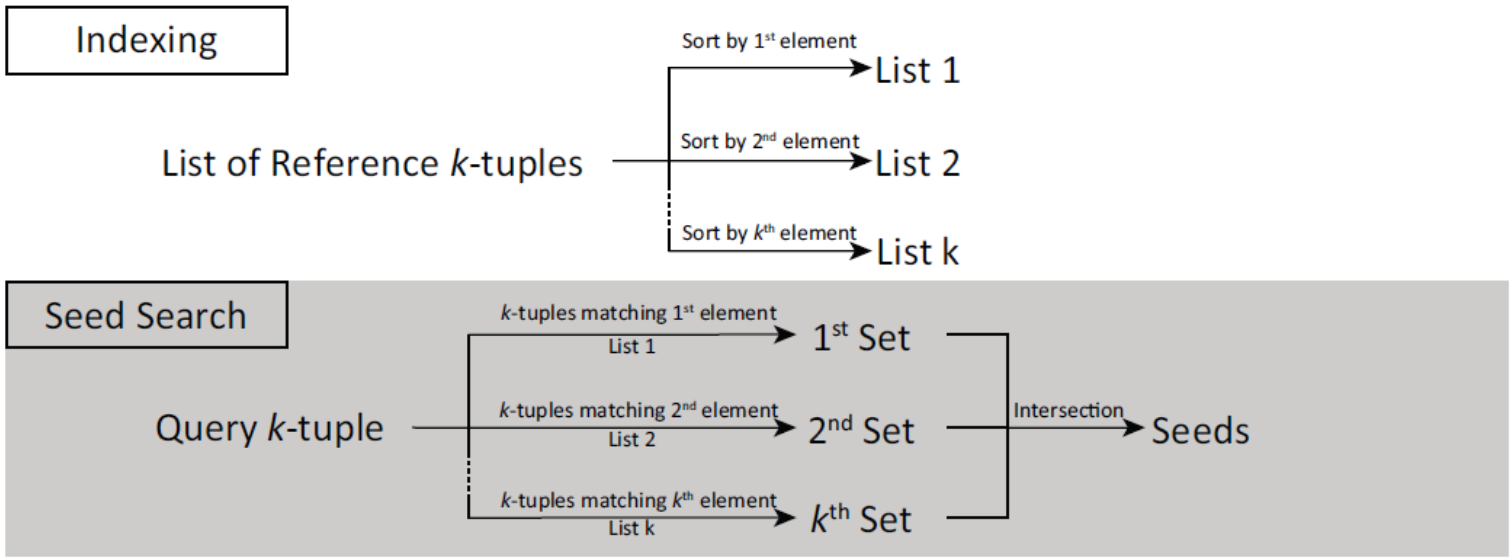
- Поиск стартовых мест (сидов) для начала выравнивания
- Расширение сидов
- Объединение пересекающих выравниваний
- Построение итогового выравнивания

OMBlast: Поиск стартовых сидов - индексация

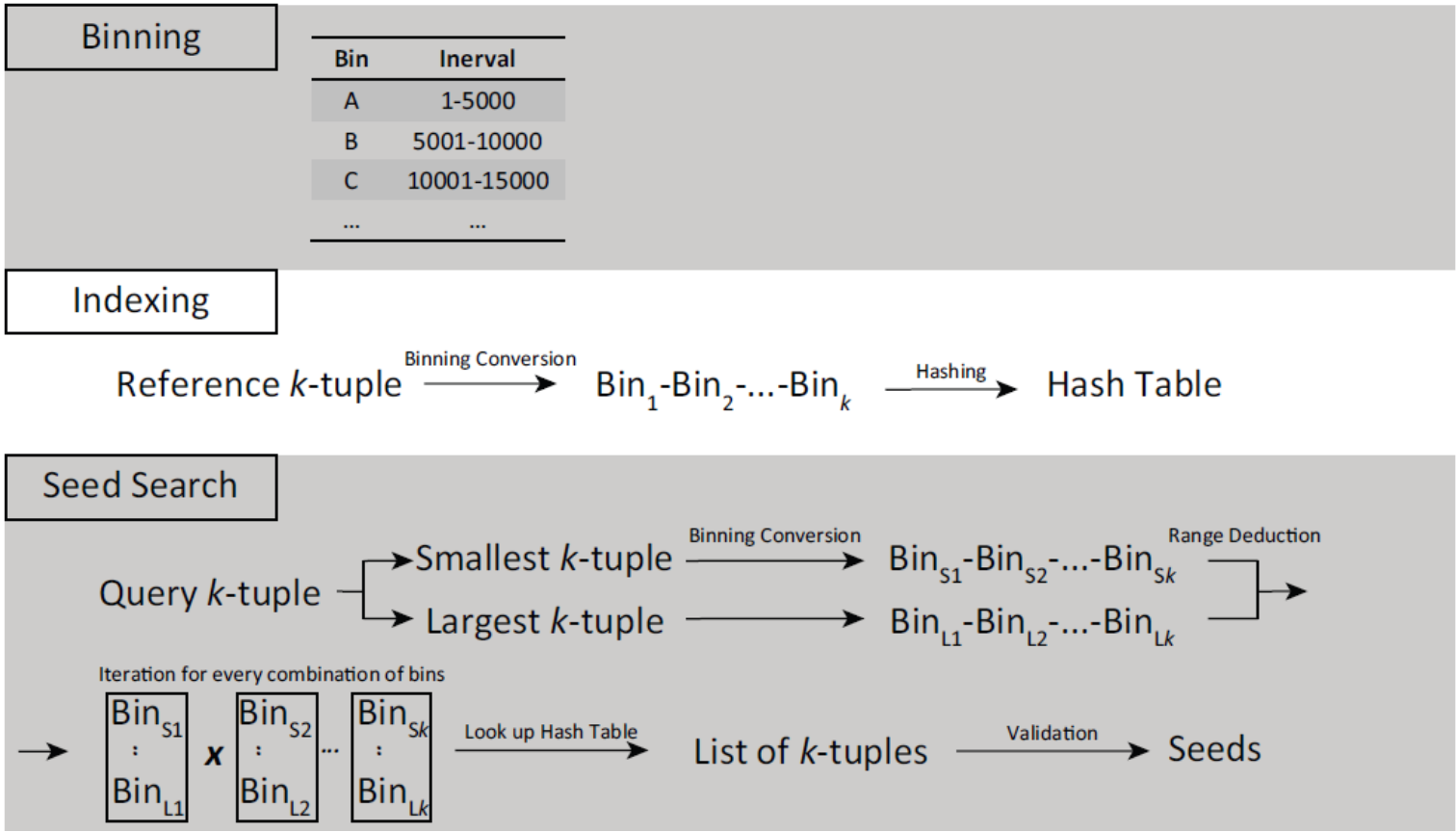
Фрагмент q на карте совпадает с фрагментом r на референсе:

$$r(1 - T_s) - T_m \leq q \leq r(1 + T_s) + T_m$$

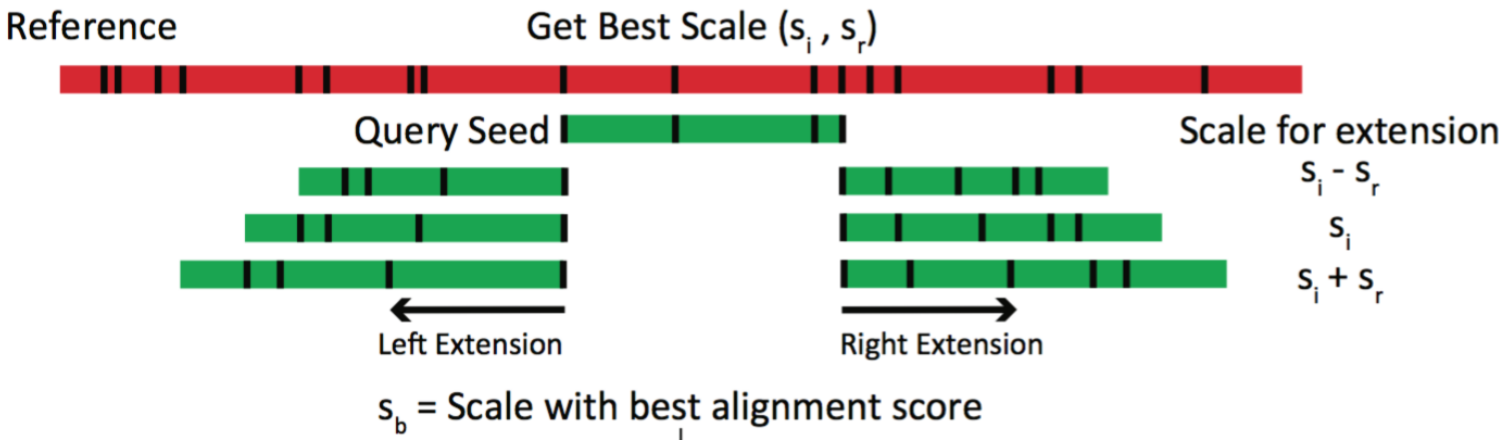
T_s - ошибка масштабирования
T_m - ошибка измерений

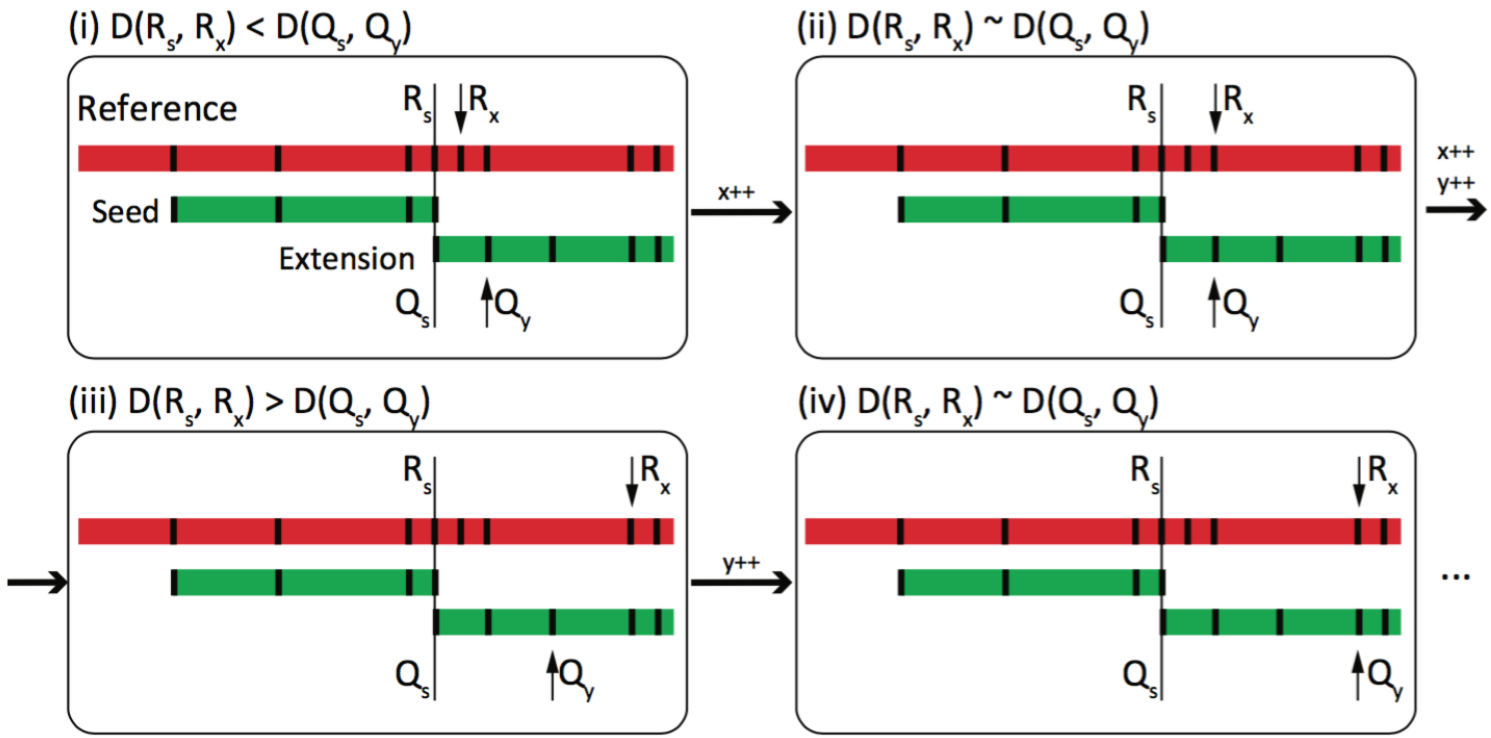


OMBlast: Поиск стартовых сидов - бины



OMBlast: Расширение сидов

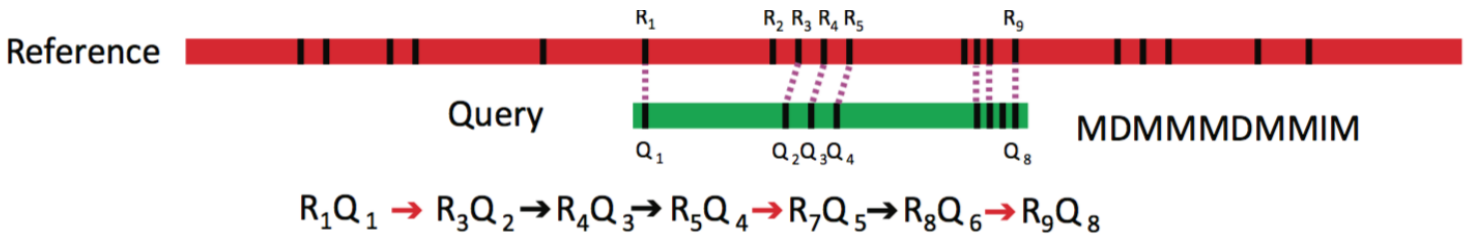




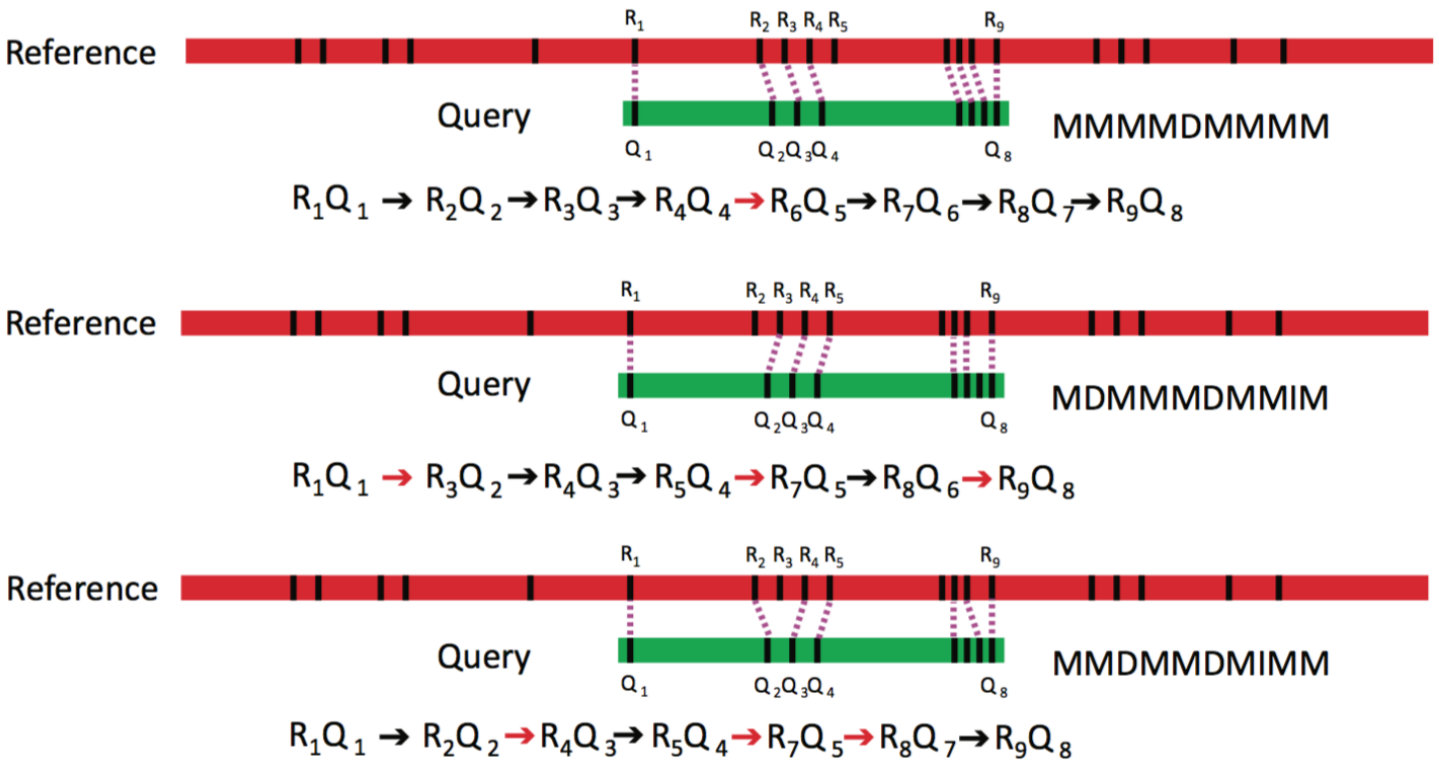
OMBlast: Объединение выравниваний (1)

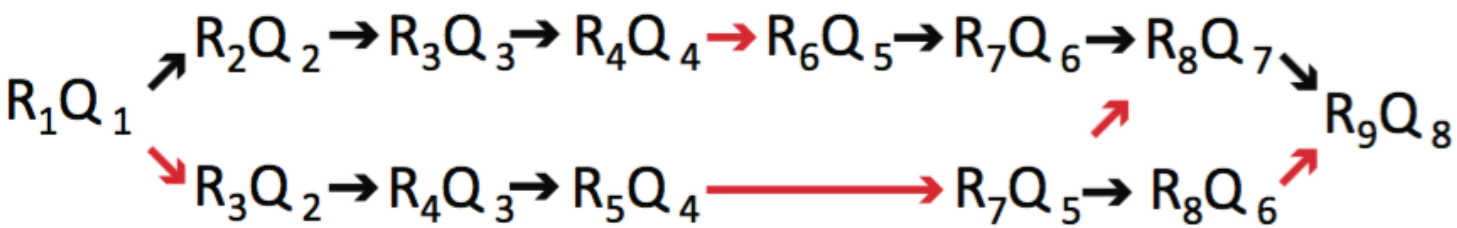
Строится взвешенный ациклический граф:

- Вершины - выравненные разрезы
- Рёбра - между двумя парами последовательно (на одной карте) выравненных разрезов
- Веса - $t_m\, u_m - t_{es}\, u_{es} - t_{ms}\, u_{ms}$
  u_m - количество совпадений
  u_{es} - количество лишних разрезов
  u_{ms} - количество пропущенных разрезов



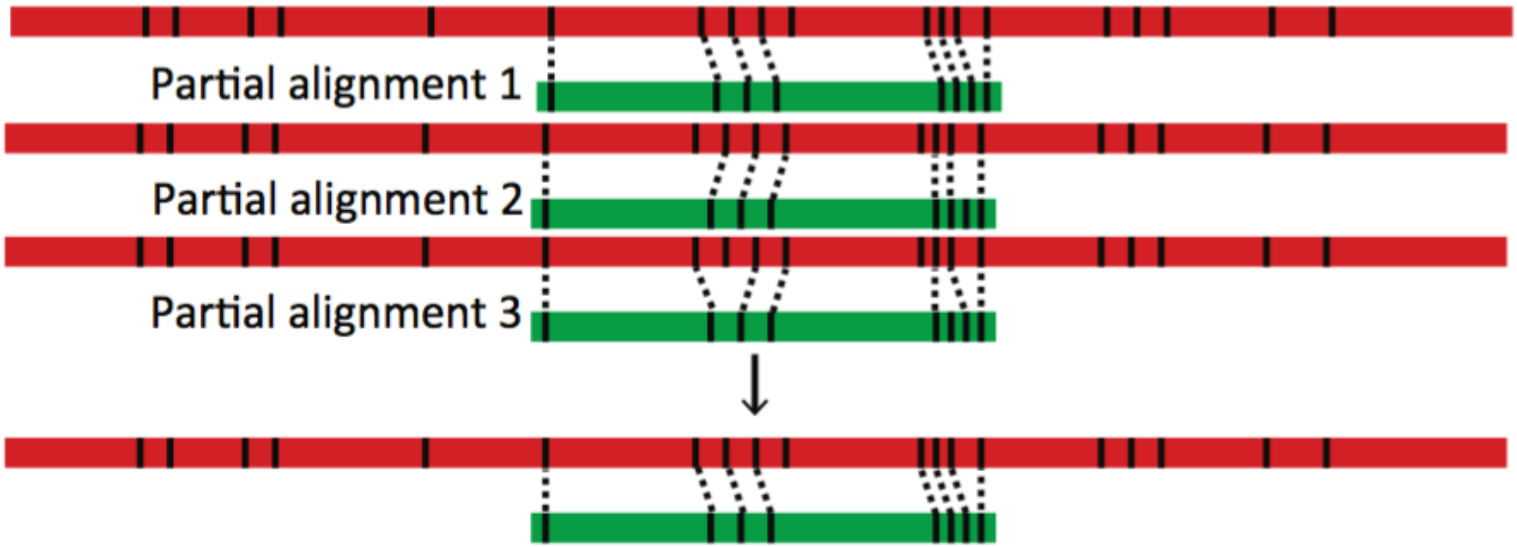
OMBlast: Объединение выравниваний (2)





OMBlast: Объединение выравниваний (3)

С помощью динамического программирования определяется путь в графе с наибольшим весом



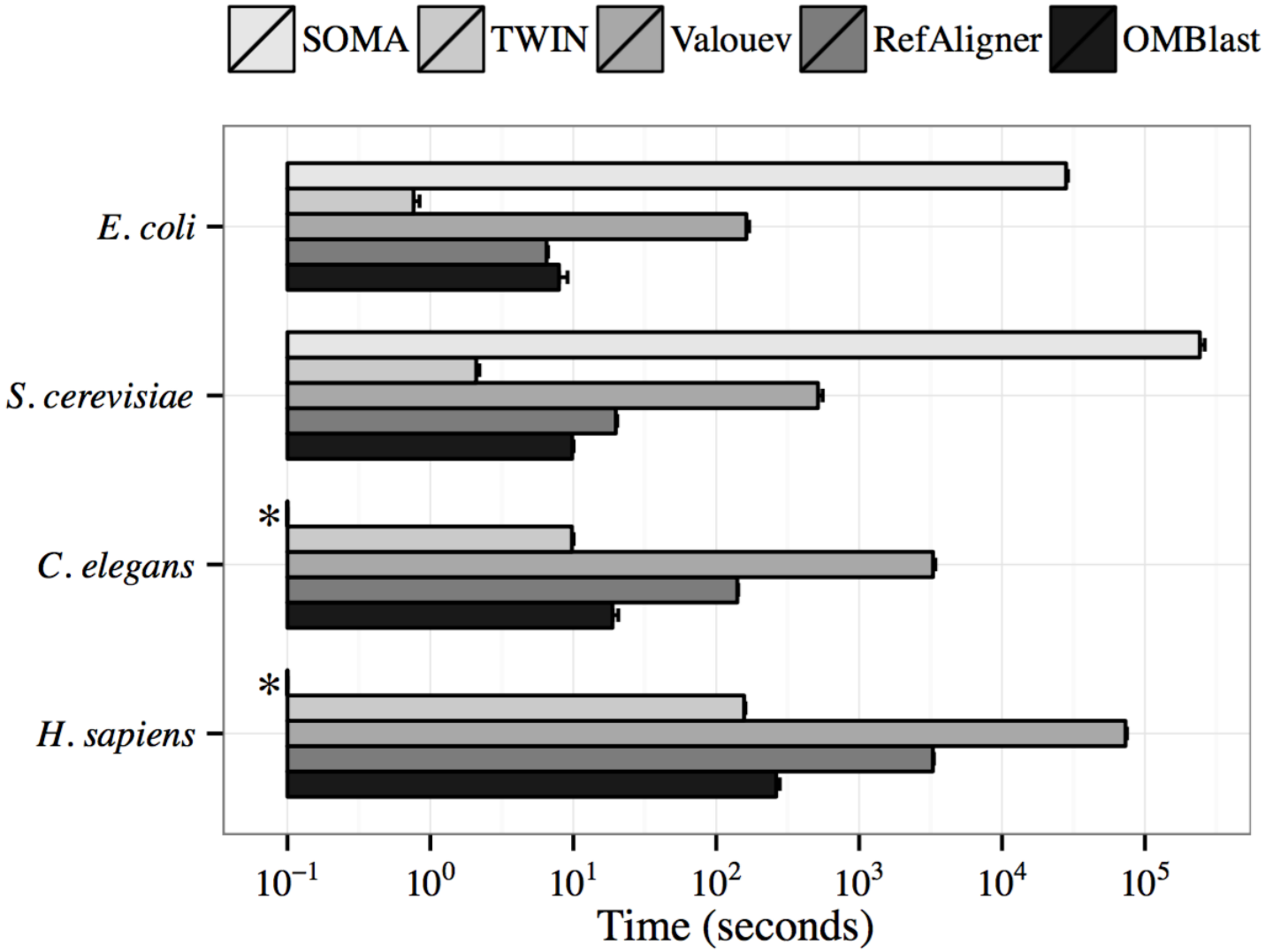
OMBlast: Построение итогового выравнивания

OMBlast: Результаты - входные данные

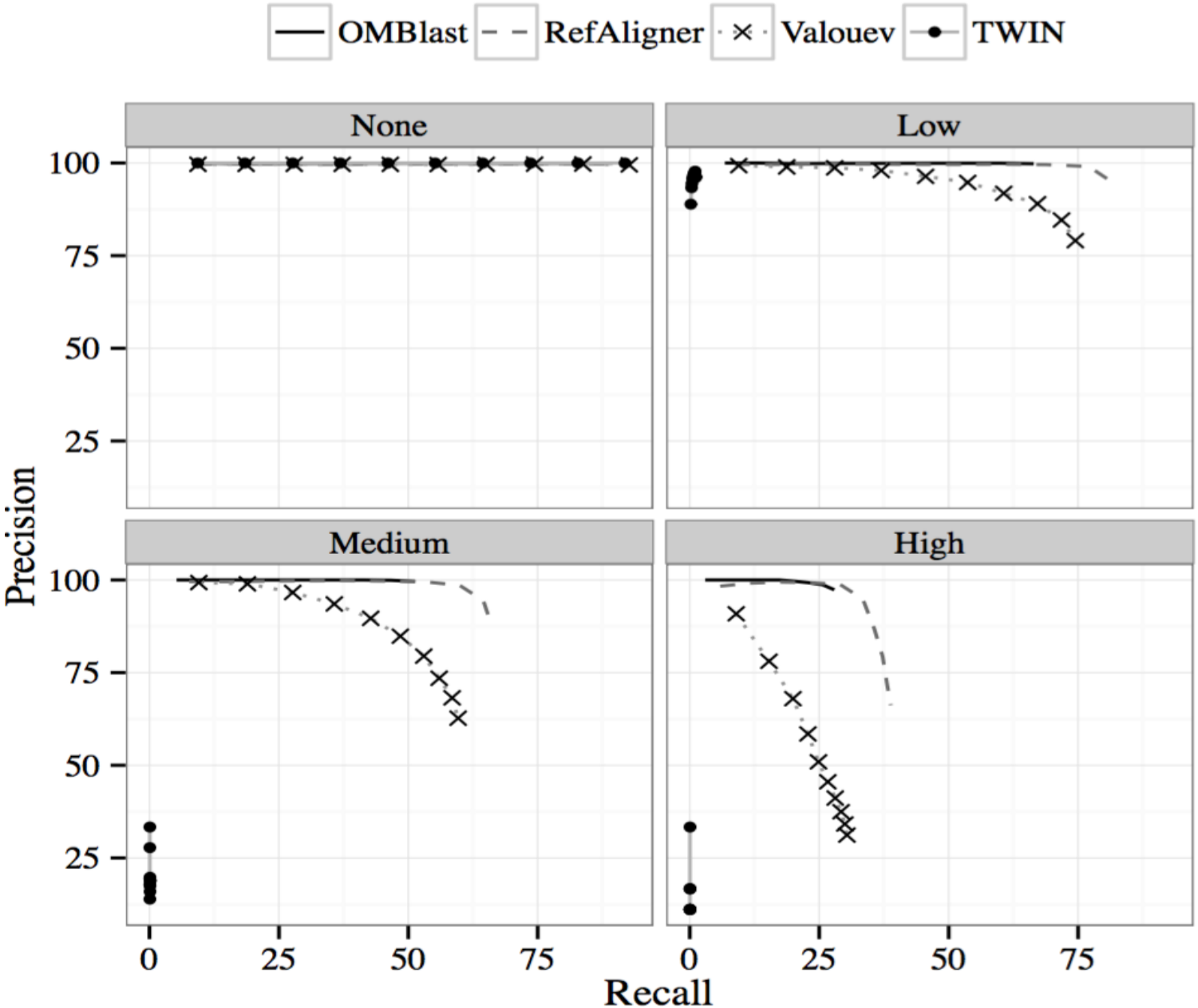
Organism	Genome Size (Mbp)	Total Signals	Average Bases Between Signals (kbp)
E. coli	4.6	683	6.8 ± 7.3
S. cerevisiae	12.1	1953	6.2 ± 6.7
C. elegans	100.3	14837	6.8 ± 8.0
H. sapiens	3088.3	377143	8.2 ± 83.2

Error Rate	None	Low	Medium	High
Extra Signal Rate	0	0.000005	0.00001	0.00002
Missing Signal Rate	0	0.05	0.1	0.2
Scaling	0	0.02	0.04	0.08
Measurement (bp)	0	500	500	500
Resolution (bp)	0	1200	1200	1200

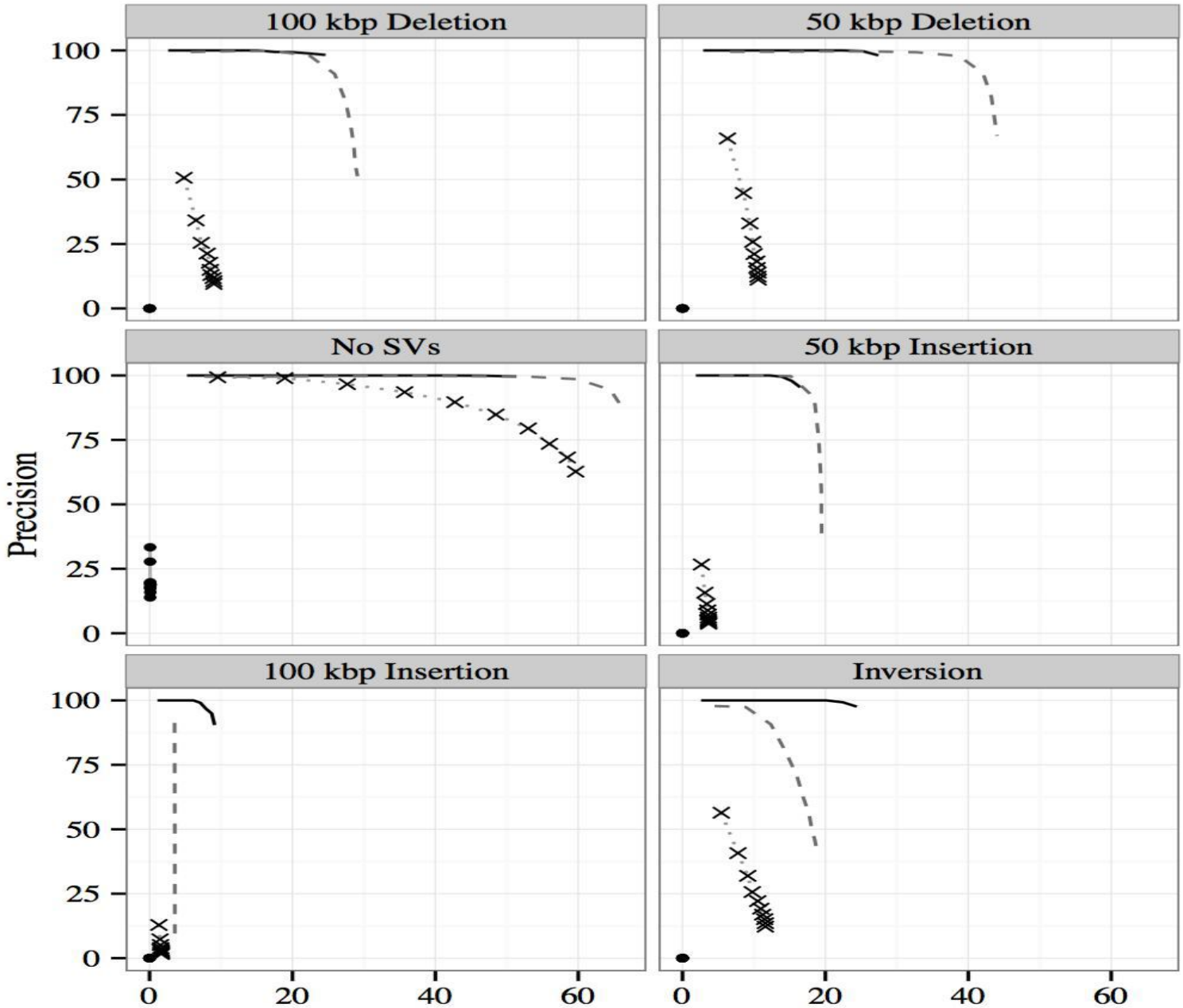
OMBlast: Результаты - время работы



OMBlast: Результаты - точность и полнота



OMBlast: Результаты - наличие SV



Спасибо за внимание!