Reference-assisted chromosome assembly

Kim J, Larkin DM, Cai Q, Asan, Zhang Y, Ge RL, Auvil L, Capitanu B, Zhang G, 2Lewin HA, Ma J.

PNAS USA 2013 Jan 29

Докладчик: Илья Минкин

СПбАУ РАН

27 апреля 2013

План

- Мотивация
- Общий обзор метода
- Детали
- Результаты
- Заключение

Мотивация

- Секвенируется все больше и больше геномов
- Не так сложно получить контиги
- Как собрать контиги в геном?
- Необходимо геном картировать
- Трудоемко и дорого
- Боль

Сборка по референсу

Что предлагалось ранее:

- Люди пытались прикладывать риды/контиги к референсу
- Склеивая контиги и находя возможные неправильные сочленения
- Некоторые даже используют филогенетику
- Тем не менее, референс при этом один
- Либо используются попарные сравнения

Общий обзор метода

- Сравнительная геномика нам поможет
- Возьмем референс
- Найдем synteny-блоки между двумя геномами
- Посмотрим на блоки на концах контигов
- Беда — могут быть нелинейные перестройки
- Давайте возьмем еще *внешние* геномы
- И попробуем оценить *вероятность* следования фрагментов

Общий обзор метода

- Находим synteny-фрагменты между референсом и собираемым геномом
- Отслеживаем те же фрагменты во внешних геномах
- Оцениваем вероятность следования одного фрагмента после другого
- Прикручиваем парную информацию
- Строим взвешенный граф из соединений между блоками
- Вес ребра это взвешенная сумма = вероятность + парная информация
- Будем жадно склеивать контиги

Граф

- У каждого блока есть голова b^h и хвост b^t
- Строим граф, где $V = \{b^h, b^t | b \in B\}$
- Каждый блок имеет номер со знаком
- Номера можно получить, если обойти граф
- Каждое ребро это пара (i, j), где i и j это номера блоков

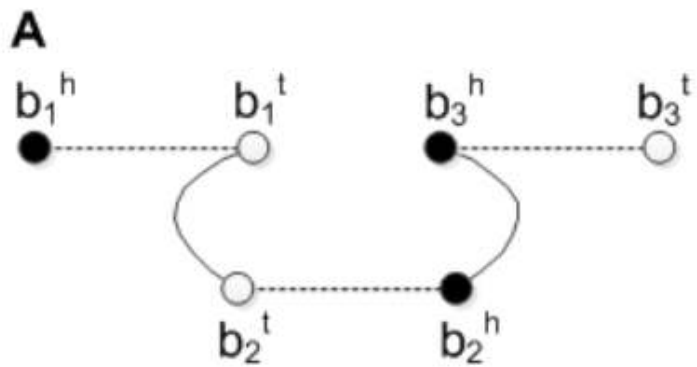


Рис. 1: Пример графа из трех блоков. Номера блоков можно прочитать как $(b_1, -b_2, b_3)$ либо как $(-b_3, b_2, -b_1)$

Ребра

- Веса ребер определяются как:

$$w(i,j) = \begin{cases} 1 & i = -j \\ \alpha Prob(i,j) + (1-\alpha) Link(i,j) & \text{иначе} \end{cases}$$

- $Prob(i,j)$ это апостериорная вероятность следования блоков i и j
- $Link(i,j)$ это *score* посчитанный при помощи парных ридов
- α можно оценить исходя из реальных данных

Содержательная картинка

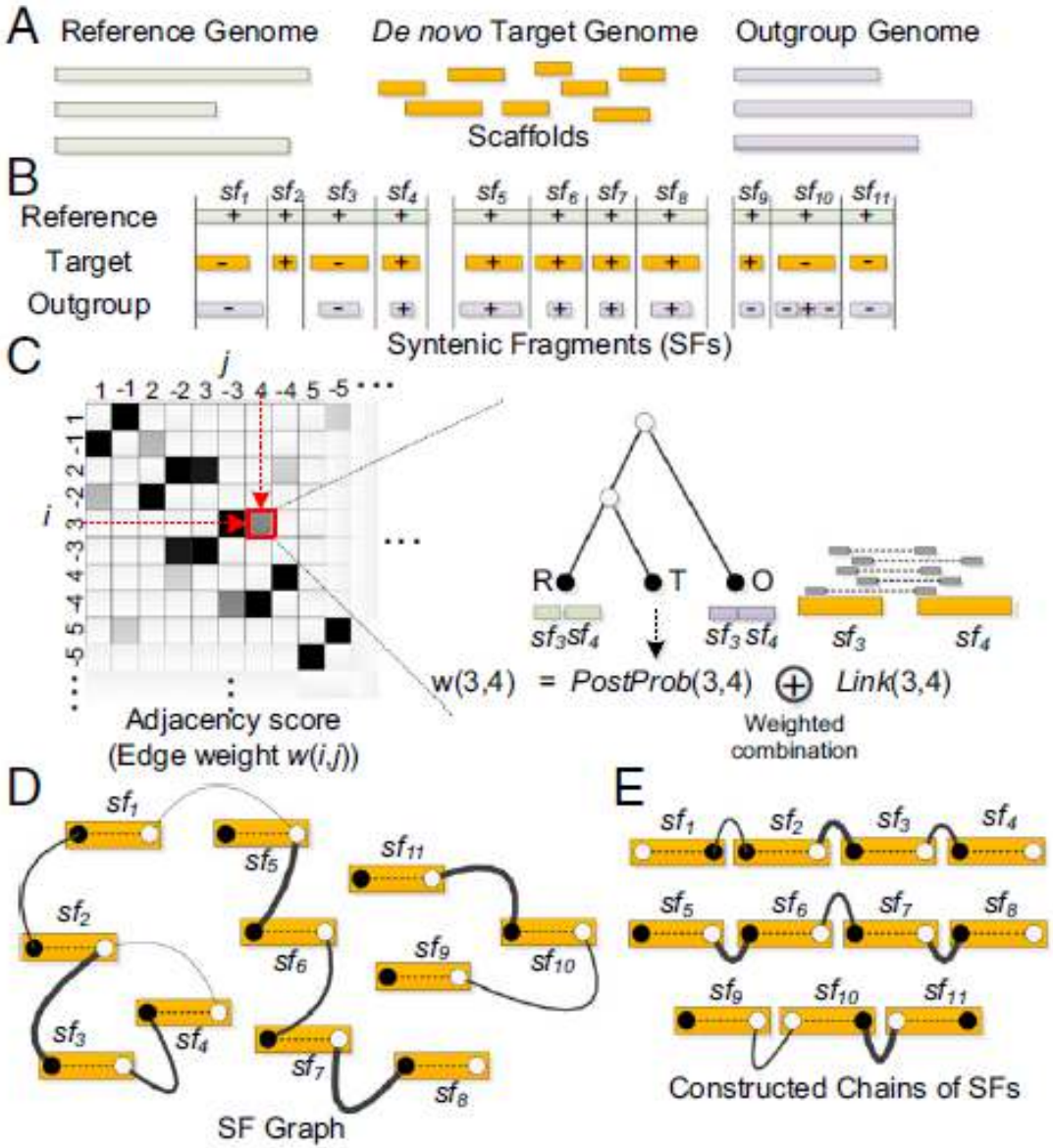


Рис. 2: Обзор метода

Как мы считаем $Prob(i, j)$

- Мы предполагаем, что у нас есть филогения
- Сначала пересаживаем дерево
- Между A_1 и T добавляется новый корень A_0
- $t(A_0, T) = t(A_1, R), t(A_1, A_0) = 0$

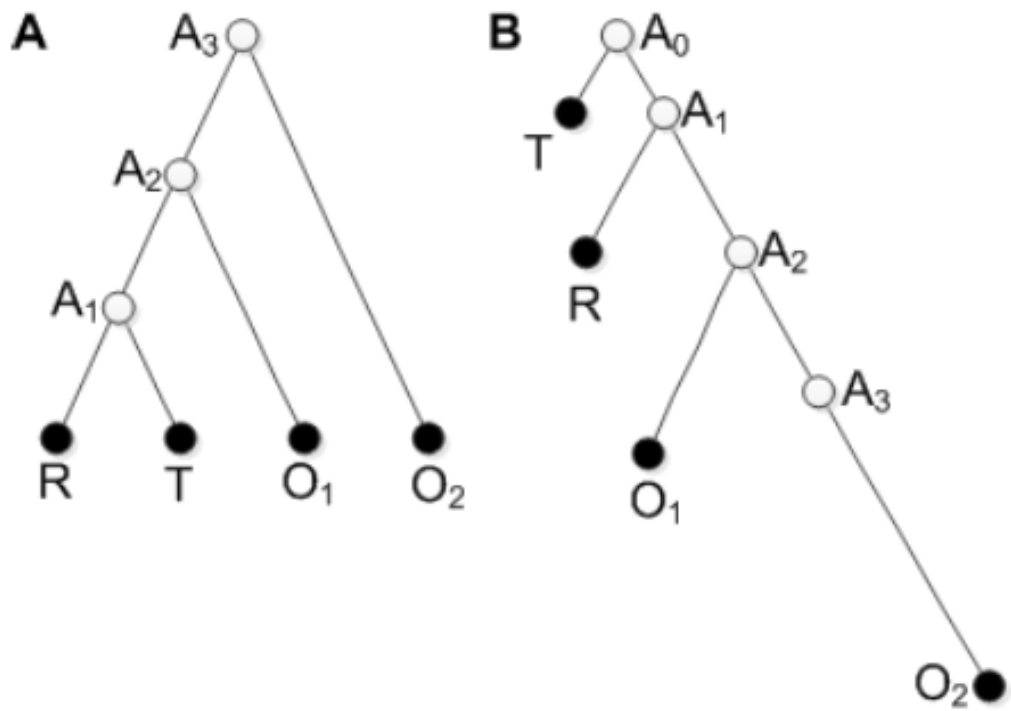


Рис. 3: Пересадка дерева

Как мы считаем $Prob(i, j)$

- Пусть в геноме T есть блок b_i, тогда $p_T(i)$ и $s_T(i)$ это следующий и предыдущий блоки
- Если $p_T(j) = i$ и $s_T(i) = j$, мы говорим, что b_i и b_j смежны в геноме T, т.е. $A_T(i, j) = 1$

$$Prob(i, j) = P(A_T(i, j) = 1 | D_T) =$$
$$= P(p_T(j) = i | D_T) P(s_T(i) = j | D_T)$$

- Посчитаем $P(p_T(j) = i | D_T)$ по формуле Байеса:

$$P(p_T(j) = i | D_T) = \frac{P(D_T | P_T(j) = i) P(P_T(j) = i)}{P(D_T)}$$

Еще одно предположение

$$P(p_T(j) = i|D_T) = \frac{P(D_T|p_T(j) = i)P(p_T(j) = i)}{\sum_k P(D_T|p_T(j) = k)P(p_T(j) = k)}$$

Предполагаем, что все априорные вероятности $P(p_T(j) = i)$ одинаковые:

$$P(p_T(j) = i|D_T) = \frac{P(D_T|p_T(j) = i)}{\sum_k P(D_T|p_T(j) = k)}$$

Если T это лист дерева, то правдобие определяется просто:

$$P(D_T|p_T(j) = i) = \begin{cases} 1 & p_T(j) = i \\ 0 & \text{иначе} \end{cases}$$

Если мы не в листе

Если T это корень поддерева с двумя дочерними узлами L и R:

$$P(p_T(j) = i|D_T) = P(D_L|p_T(j) = i)P(D_R|p_T(j) = i) =$$
$$= \sum_k P(D_L|p_L(j) = k)\, P(p_L(j) = k|p_T(j) = k)\times$$
$$\times \sum_k P(D_R|p_R(j) = k)\, P(p_R(j) = k|p_T(j) = i)$$

$P(p_L(j) = k|p_T(j) = k)$ это вероятность того, что в геноме L блок, стоящий перед j вдруг заменился на k

Как оценить вероятность замены блока

Эта вероятность оценивается при помощи модели эволюции ДНК Jukes-Cantor расширенной для точек разлома:

$$P(p_L(j) = k | p_T(j) = k) = \frac{1}{2n-1} - \frac{2n-2}{2n-1} e^{-(2n-1)\mu t_{TL}}$$

Где:

- n — число блоков
- μ — параметр модели (для всех узлов один)
- t_{TL} — длина ветви

Как узнать значение μ

$$
\begin{aligned}
P(p_L(j) = p_R(j) = i) &= \sum_{k \neq j,-j} P(p_L(j) = i | p_T(j) = k)\, P(p_R(j) = i | p_T(j) = k) \\
&= P(p_L(j) = i | p_T(j) = i) P(p_R(j) = i | p_T(j) = i) \\
&\quad + (2n-2) P(p_L(j) = i | p_T(j) = k) P(p_R(j) = i | p_T(j) = k) \\
&= \left(\frac{1}{2n-1} + \frac{2n-2}{2n-1} e^{-(2n-1)\mu t_{TL}}\right)\left(\frac{1}{2n-1} + \frac{2n-2}{2n-1} e^{-(2n-1)\mu t_{TR}}\right) \\
&\quad + (2n-2)\left(\frac{1}{2n-1} - \frac{1}{2n-1} e^{-(2n-1)\mu t_{TL}}\right)\left(\frac{1}{2n-1} - \frac{1}{2n-1} e^{-(2n-1)\mu t_{TR}}\right) \\
&= \frac{1}{2n-1} + \frac{2n-2}{2n-1} e^{-(2n-1)\mu (t_{TL} + t_{TR})}
\end{aligned}
\tag{S8}
$$

Since,

$$
P(p_L(j) \neq p_R(j)) = 1 - P(p_L(j) = p_R(j)) = \frac{2n-2}{2n-1}\left(1 - e^{-(2n-1)\mu (t_{TL} + t_{TR})}\right) \tag{S9}
$$

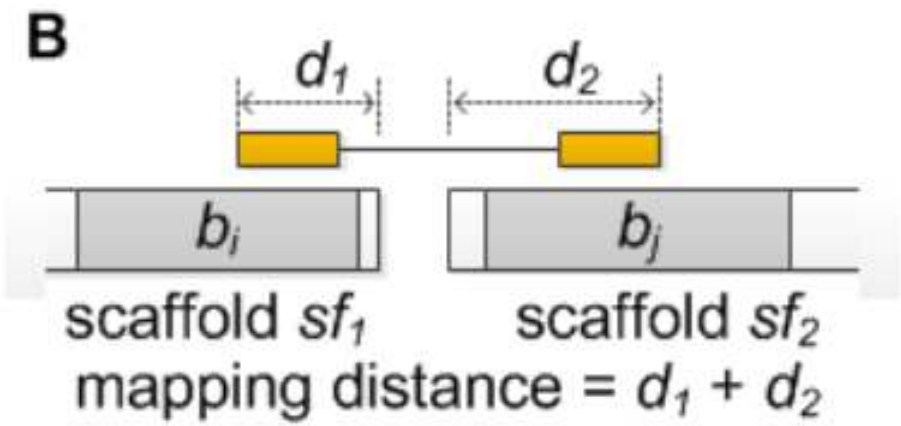
Therefore,

$$
\begin{aligned}
\mu &= -\frac{1}{(2n-1)(t_{TL} + t_{TR})} \ln\left(1 - \frac{2n-1}{2n-2} P(p_L(j) \neq p_R(j))\right) \\
&\approx -\frac{1}{(2n-1)(t_{TL} + t_{TR})} \ln\left(1 - \frac{2n-1}{2n-2} \frac{d(L,R)}{n}\right)
\end{aligned}
\tag{S10}
$$

where $d(L,R)$ is a breakpoint distance between L and R. Here, we assumed that the probability $P(p_L(j) \neq p_R(j))$ can be approximated by $d(L,R)/n$, which is the number of breakpoints per synteny blocks.

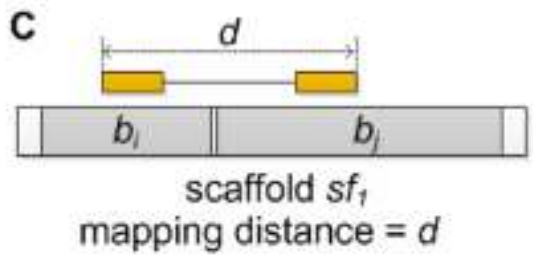
Теперь считаем score для парной информации

- Рассмотрим случай для разных scaffold'ов
- $N_{ir}(i, j)$ — количество парных ридов, приложившихся в блоки i и j
- Расстояние не должно превышать размер вставки + $2SD$



Случай одинаковых scaffold'ов

- Найдем парные риды, дистанция между которыми = расстояние вставки +/- $2SD$
- По регионам между блоками пройдемся скользящим окном
- Окна считают покрытие и "заезжают"в блоки на $L_f = 50$ Kbp
- Размер окна $L_w = 1$ Kbp, перекрытие = $L_w/2$
- Для каждого окна считаем p_a = покрытие относительно среднего по всем скаффолдам
- $P_{ia}(i, j)$ = минимальное значение p_a



Теперь считаем $Link(i,j)$

- Соберем все в кучу
- $P_{ir}(i,j)$ — значение $N_{ir}(i,j)$ относительно среднего по всем возможным ребрам

$$P(i,j) = \begin{cases} P_{ir}(i,j) & sf(i) \neq sf(j) \\ P_{ia}(i,j) & sf(i) = sf(j) \end{cases}$$

$$Link(i,j) = \frac{P(i,j) - \min\left(P(i',j'), \forall i', j'\right)}{\max(P(i',j'), \forall i', j') - \min(P(i',j'), \forall i', j')}$$

Склеиваем контиги направо и налево

1: **Begin**

2: *C: initially empty set of connected components*

3: **For each** *edge e in descending order of weight* ≥ 0.1*:*

4: **If** *e is not inconsistent with any previously used edges*

5: **AND** *does not introduce a cycle*

6: **If** *e can be added to any connected component in C*

7: *Add e to that connected component*

8: **Else**

9: *Create a new connected component with e and add it to C*

10: **End**

Синтетический тест

- Возьмем две человеческие хромосомы
- Попросим Evolver сгенерировать нам 12 синтетических геномов
- Один геном всегда будет референсом
- Какой-то другой будет собираемым
- Остальные будут внешними

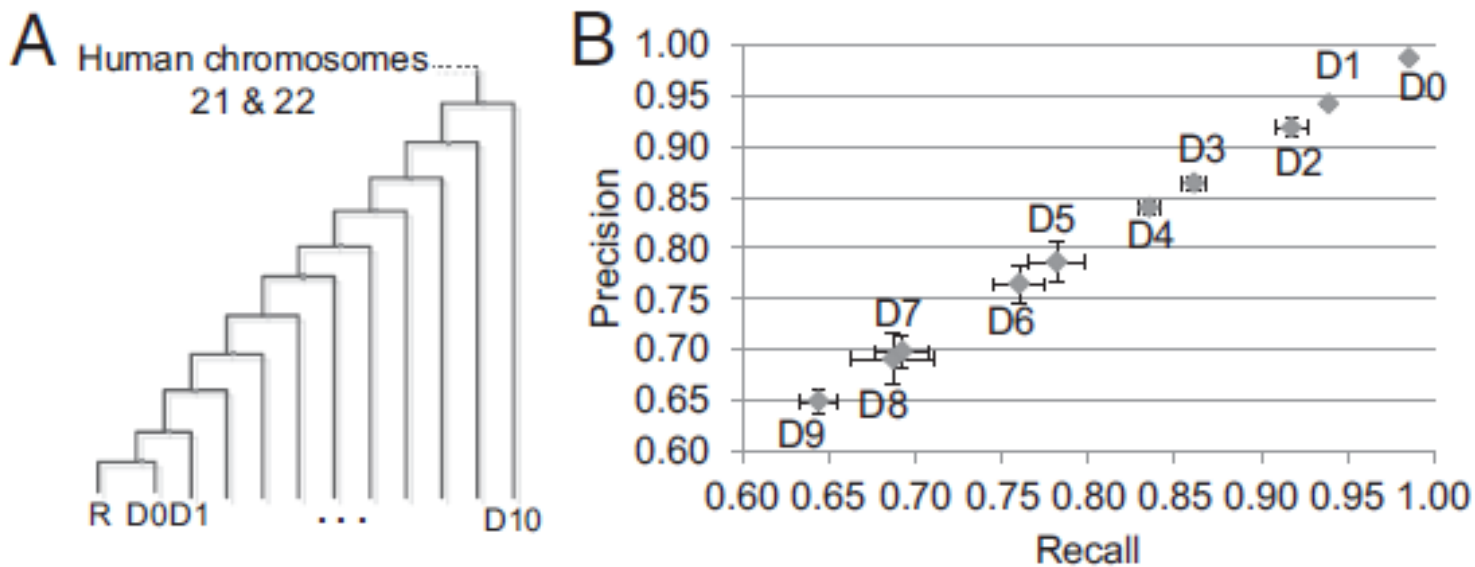


Рис. 4: Синтетический тест

Попробуем улучшить реальные сборки

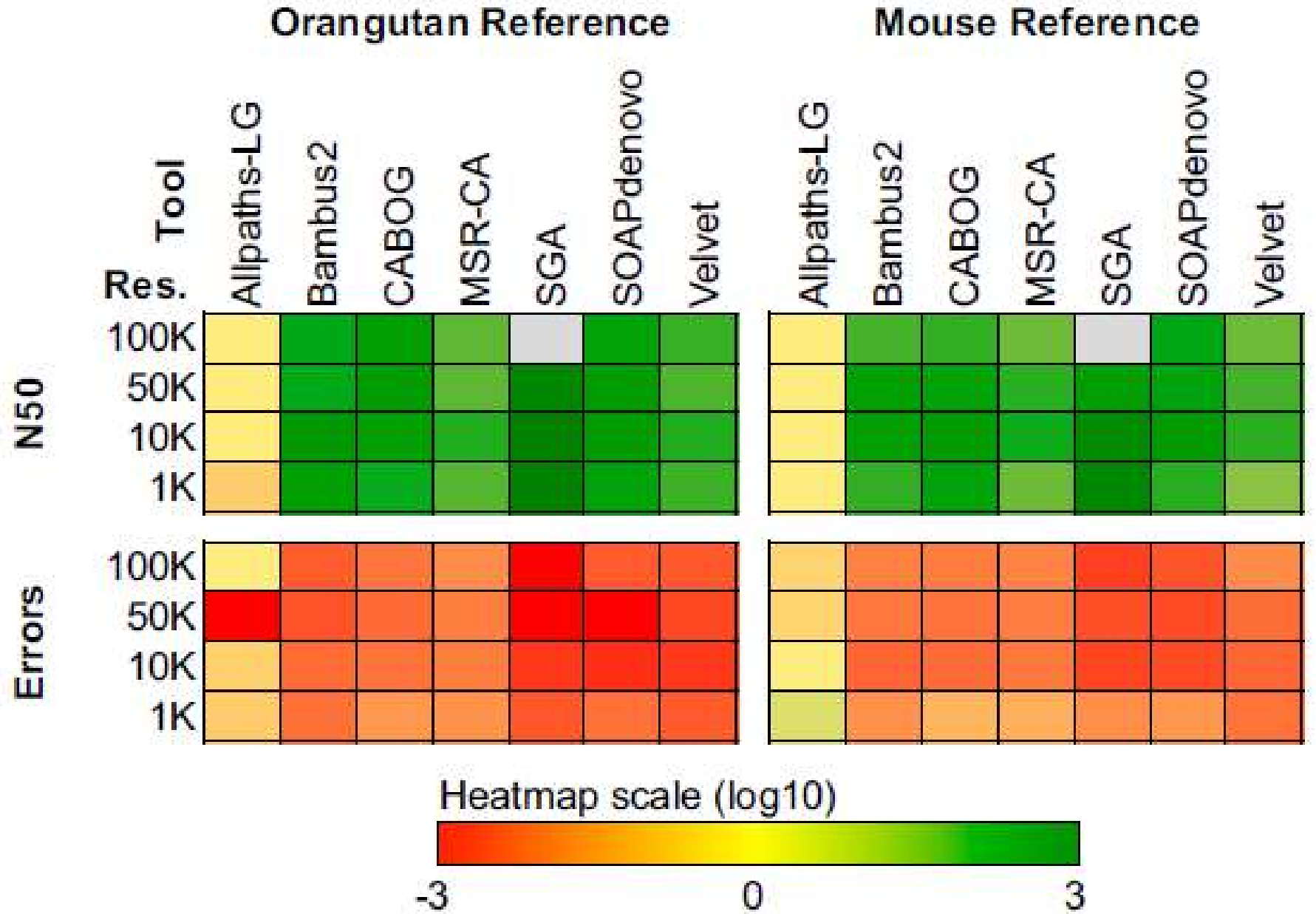
- Данные от Genome Assembly Gold-Standard Evaluations (GAGE)
- Семь сборок 14-й хромосомы человека
- Мышь и орангутанг в качестве референса
- Крупный рогатый скот в качестве внешнего генома

Улучшаем реальные сборки

Supplementary Table S3. Statistics of the original assemblies in the GAGE data sets and RACA assemblies using orangutan (ponAbe2 assembly) as a reference genome.

Resolution (Kbp)	Tool	Assembly[1]	Total Scaffolds	N50[2] (bp)	Misjoin Errors	Unjoin Errors	Total Errors[3]	Coverage[4]
100	ALLPATHS-LG	org	3	81,646,936	0	2	2	0.98
		raca	2	85,369,765	0	1	1	0.98
	Bambus2	org	186	324,289	26	174	200	0.87
		raca	1	67,596,306	0	0	0	0.87
	CABOG	org	206	392,605	25	195	220	0.87
		raca	1	75,284,819	3	0	3	0.87
	MSR-CA	org	109	893,428	158	59	217	0.92
		raca	2	52,689,077	8	0	8	0.92
	SGA	org	216	-	0	215	215	0.40
		raca	1	-	0	0	0	0.40
	SOAPdenovo	org	211	453,540	3	203	206	0.75
		raca	1	78,803,340	0	0	0	0.75
	Velvet	org	143	1,190,421	208	83	291	0.81
		raca	1	114,571,702	0	0	0	0.81
50	ALLPATHS-LG	org	4	81,646,936	0	3	3	0.99
		raca	1	86,776,395	0	0	0	0.99
	Bambus2	org	238	324,289	31	227	258	0.92
		raca	1	72,104,657	0	0	0	0.92
	CABOG	org	285	392,605	38	268	306	0.94
		raca	1	81,393,373	3	0	3	0.94
	MSR-CA	org	135	893,428	206	71	277	0.94
		raca	3	28,099,128	9	0	9	0.94
	SGA	org	499	81,968	0	498	498	0.61
		raca	1	57,458,484	0	0	0	0.61
	SOAPdenovo	org	281	453,540	3	273	276	0.80
		raca	1	84,423,938	0	0	0	0.80
	Velvet	org	177	1,190,421	396	64	460	0.86
		raca	1	123,014,014	0	0	0	0.86

Не все референсы одинаково полезны



Сборка генома тибетских антилоп

- Pantholops hodgsonii; 2N = 60
- Коровы в качестве референса
- Человек в роли внешнего генома
- Минимальный размер synteny-блока — 150 KBP
- Выбрали 1 434 scaffold'ов из всего 15 996 штук (покрытие 96%)
- Нашли 1 597 synteny-блоков
- Эти блоки покрывают 95% генома антилоп, 29 коровьих автосом и X хромосомы
- Нашлось 1,537 соединений между блоками, из которых 73 были найдены только RACA

Результаты

Table 1. Statistics of Tibetan antelope predicted chromosome fragments

Category	Value
No. PCFs	60
No. PCFs that are homologous to complete cattle chromosomes*	16
No. PCFs without outgroup (human) matches†	1
Total length of PCFs	2.601 gbp
Maximum length of PCFs	193 mbp
Minimum length of PCFs	251 kbp
PCF N50	87 mbp
Maximum no. Tibetan antelope scaffolds in PCFs	111
Minimim no. Tibetan antelope scaffolds in PCFs	1
No. cattle EBRs	64
No. other EBRs	411
No. Tibetan antelope scaffolds that have more than one SF	130 (9%)‡
No. Tibetan antelope scaffolds predicted as chimeric	84 (6%)§

EBRs, evolutionary breakpoint regions; PCFs, predicted chromosome fragments; SFs, syntenic fragments.

*These correspond to entire cattle chromosomes 2, 3, 6, 8, 9, 11, 12, 15, 18, 19, 20, 23, 24, 25, 28, and 29.

†There was no mapped human genome fragment to these PCFs.

‡Percentage of the total number of aligned Tibetan antelope scaffolds.

§Among 84 scaffolds, 6 were mapped to three different PCFs, 69 were mapped to two different PCFs, and the remaining 9 were mapped to the same PCF at different and nonadjacent locations.

Валидация

- 14 сочленений были валидированы с помощью PCR
- Из них 11 сгенерировали единственный фрагмент
- Четыре PCR продукта были очень похожи на примерный размер пробела
- RACA нашла также два ошибочных соединения
- PCR подтвердила ошибку

Заключение

- Теперь у нас есть способ склеивать контиги без картирования
- Сама идея — использовать несколько внешних геномов очень привлекательна
- Используется информация изо всех геномов сразу
- Явно используется филогенетическое дерево
- Как показывают результаты, RACA работает

Спасибо за внимание!