

Выравнивание геномов

Сходство и гомология

Сходство – характеристика последовательностей в соответствии с некоторыми критериями

Гомология – сходство последовательностей, вызванное их общим происхождением.

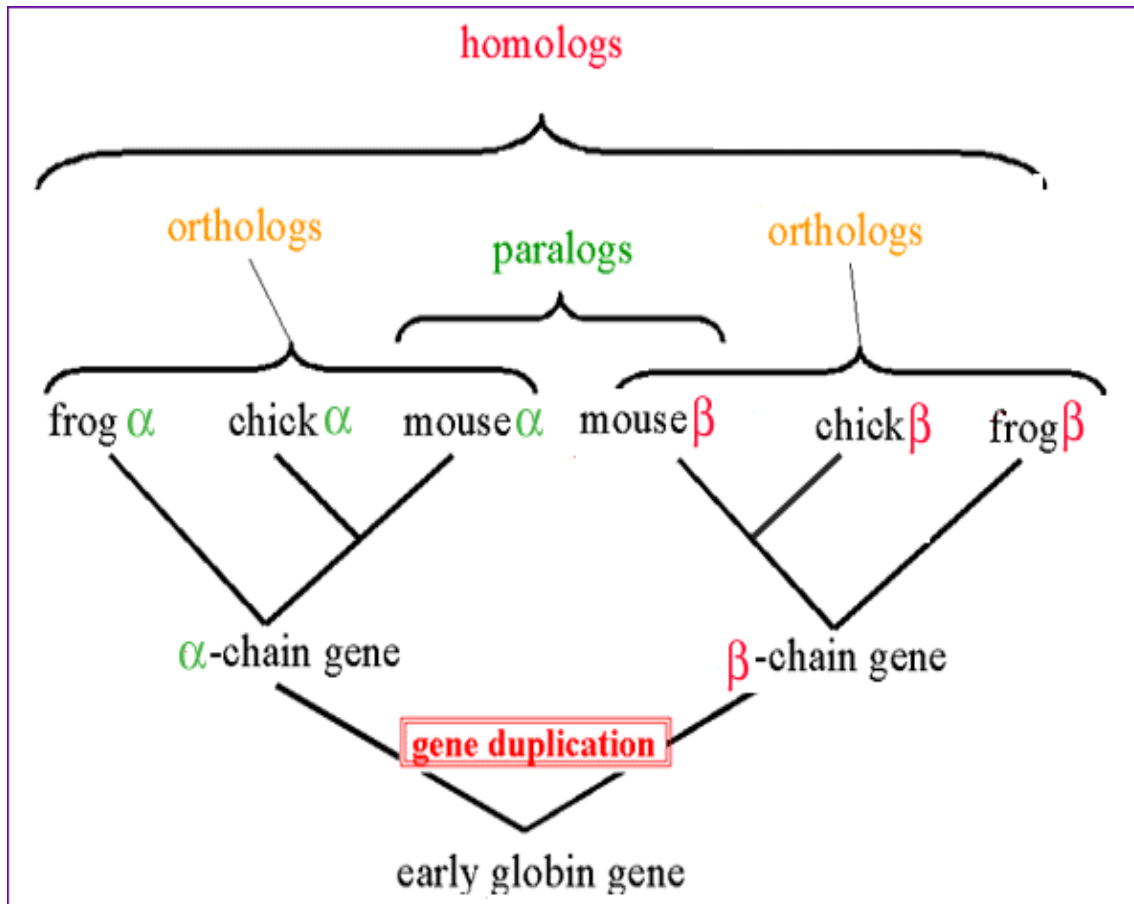
Сходство и гомология часто употребляются, как синонимы, но ими не являются.

Сходство и функция

Консервативные участки соответствуют функционально важным участкам белка

Менее функционально значимые участки легко накапливают случайные мутации

Гомологи глобина



Эволюционные события

Локальная эволюция - точечные замены, вставки, делеции

Глобальная эволюция – одномоментное изменение больших фрагментов генома (вставки, делеции, дупликации, инверсии, транслокации, слияние-разделение хромосом, горизонтальный перенос генов)

Парное выравнивание

Задача: найти ортологичные участки

Вход: две последовательности

Результат:

-набор пар гомологичных участков и их выравниваний

-визуализация

Пример: программа blast2seq и карта локального сходства

blast2seq

База – одна последовательность

Вход – другая последовательность

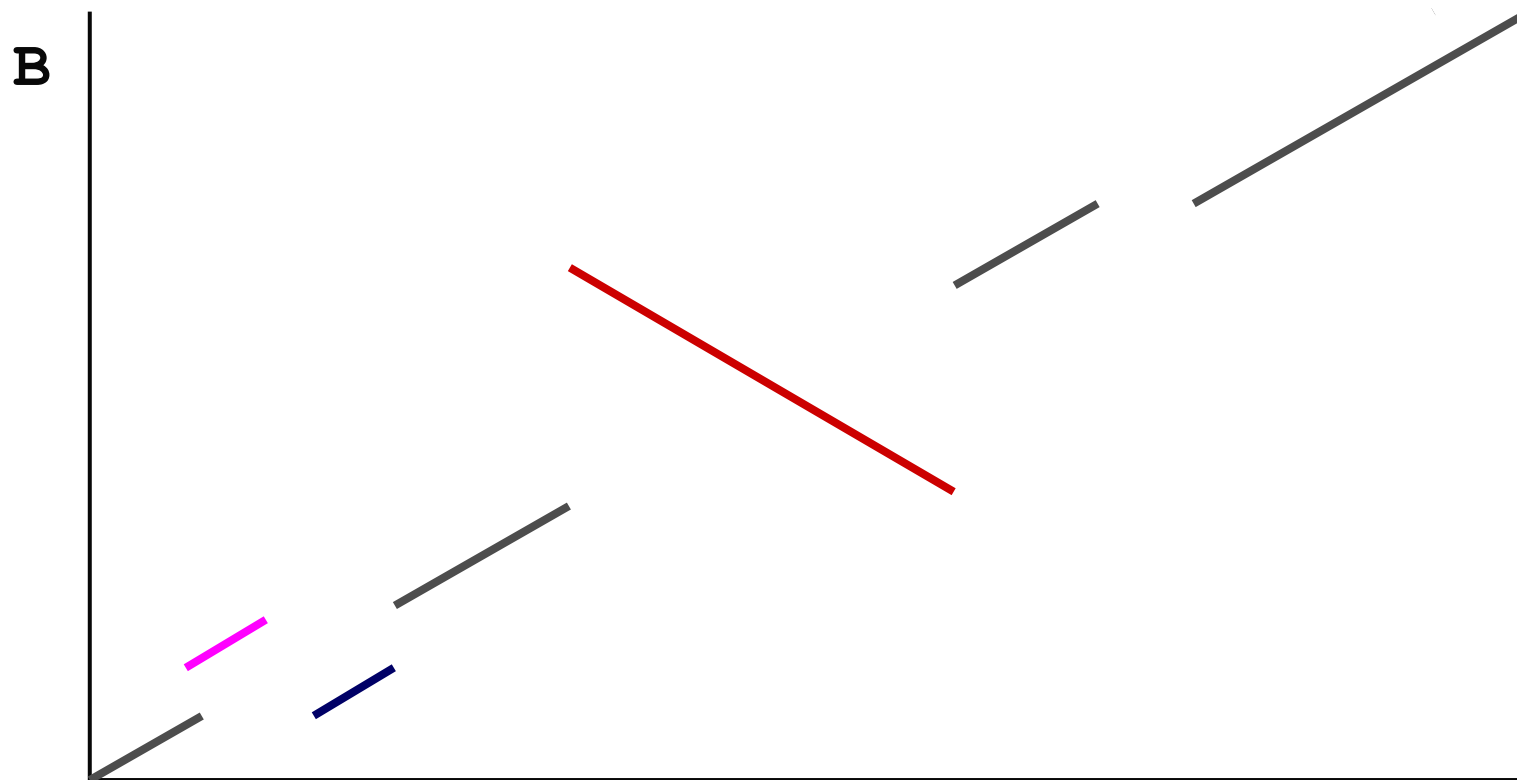
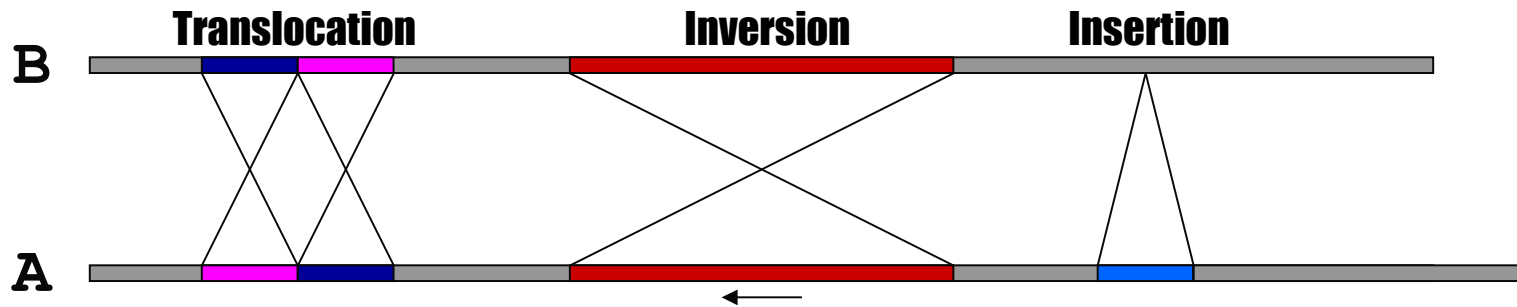
Результат – набор находок, т.е. пар фрагментов, похожих друг на друга

Алгоритм – тот же blast (обычно blastn)

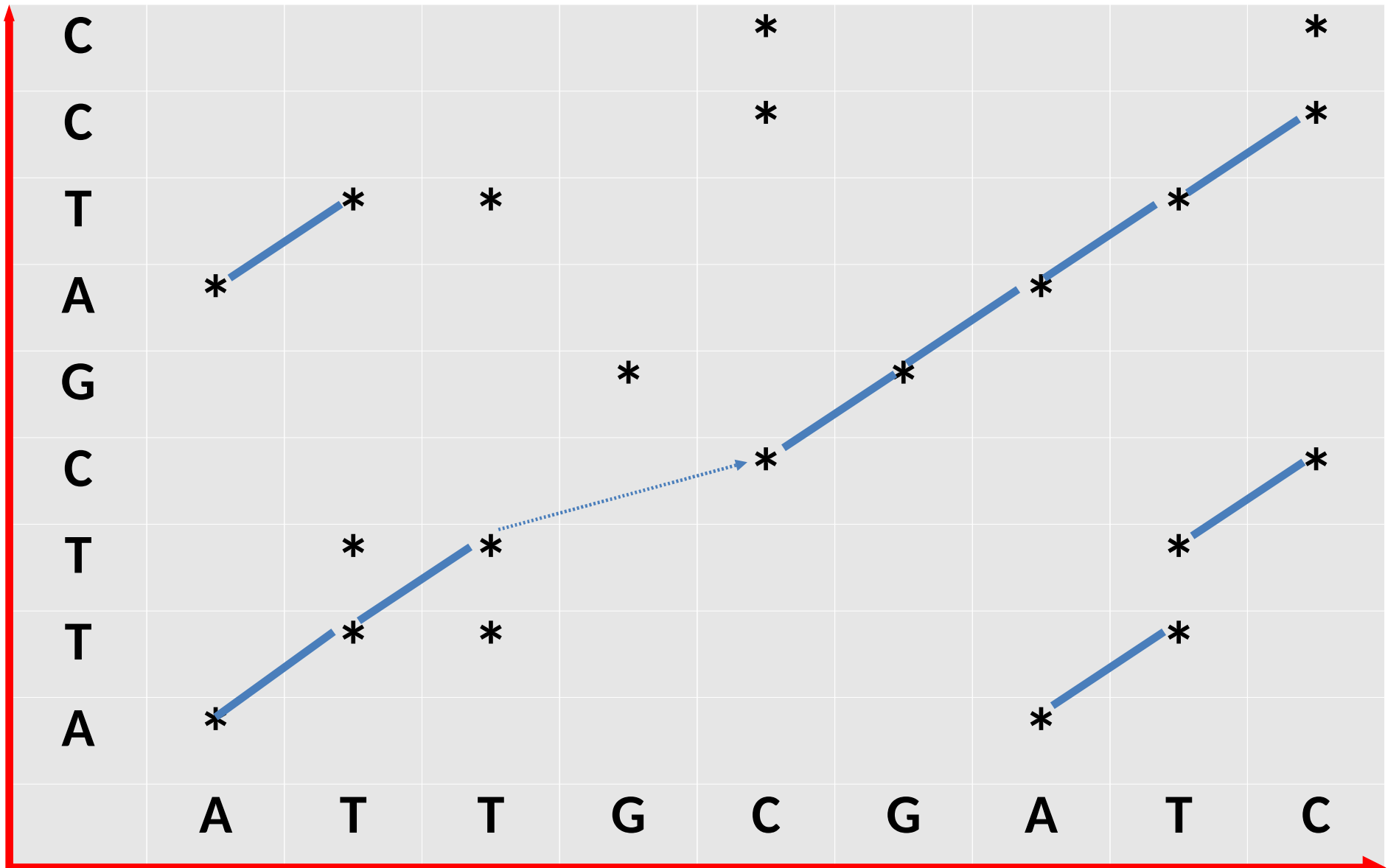
Визуализация – карта локального сходства:

Каждая находка изображается отрезком

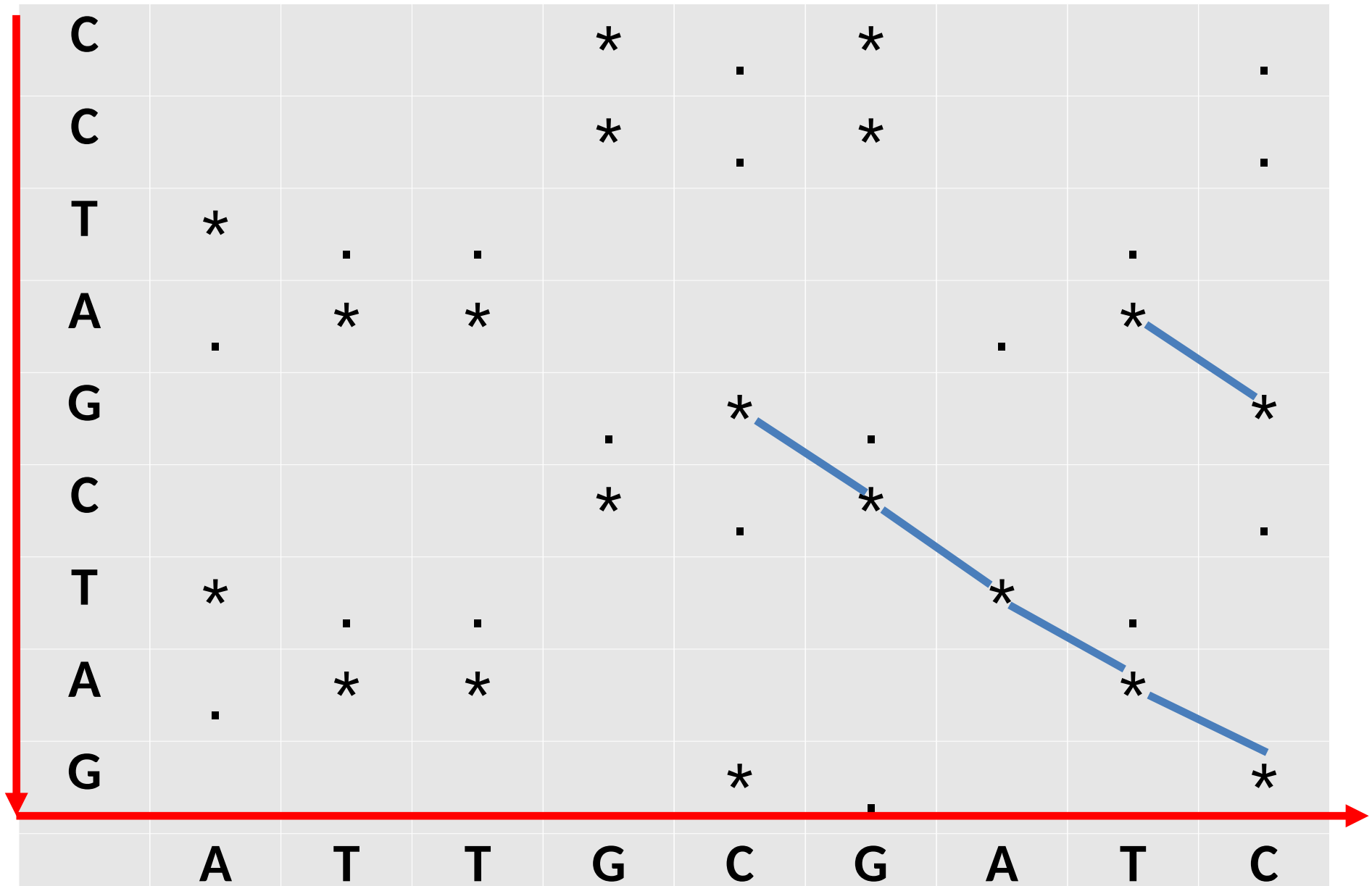
Карта локального сходства



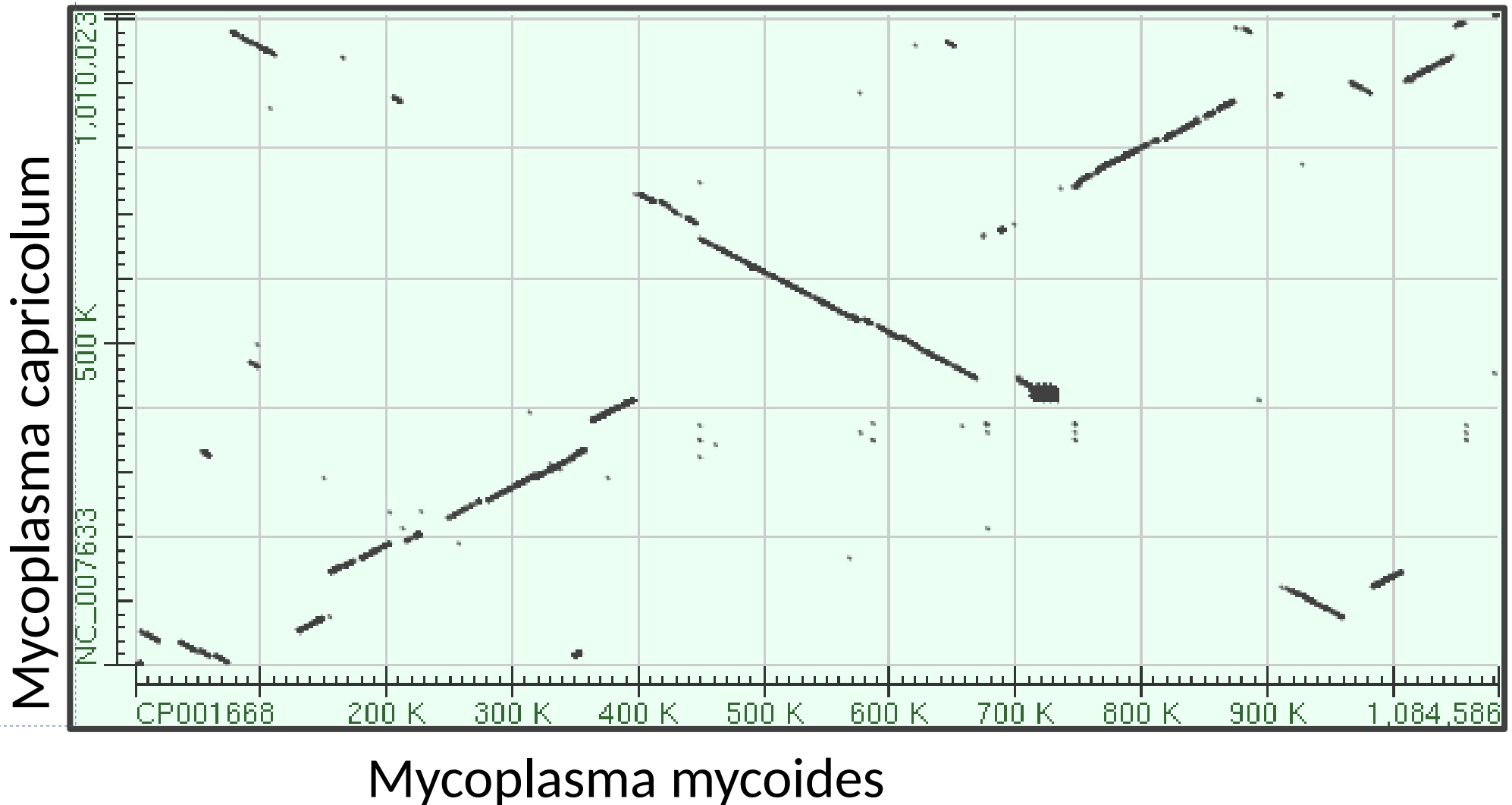
Карта сходства 2х последовательностей



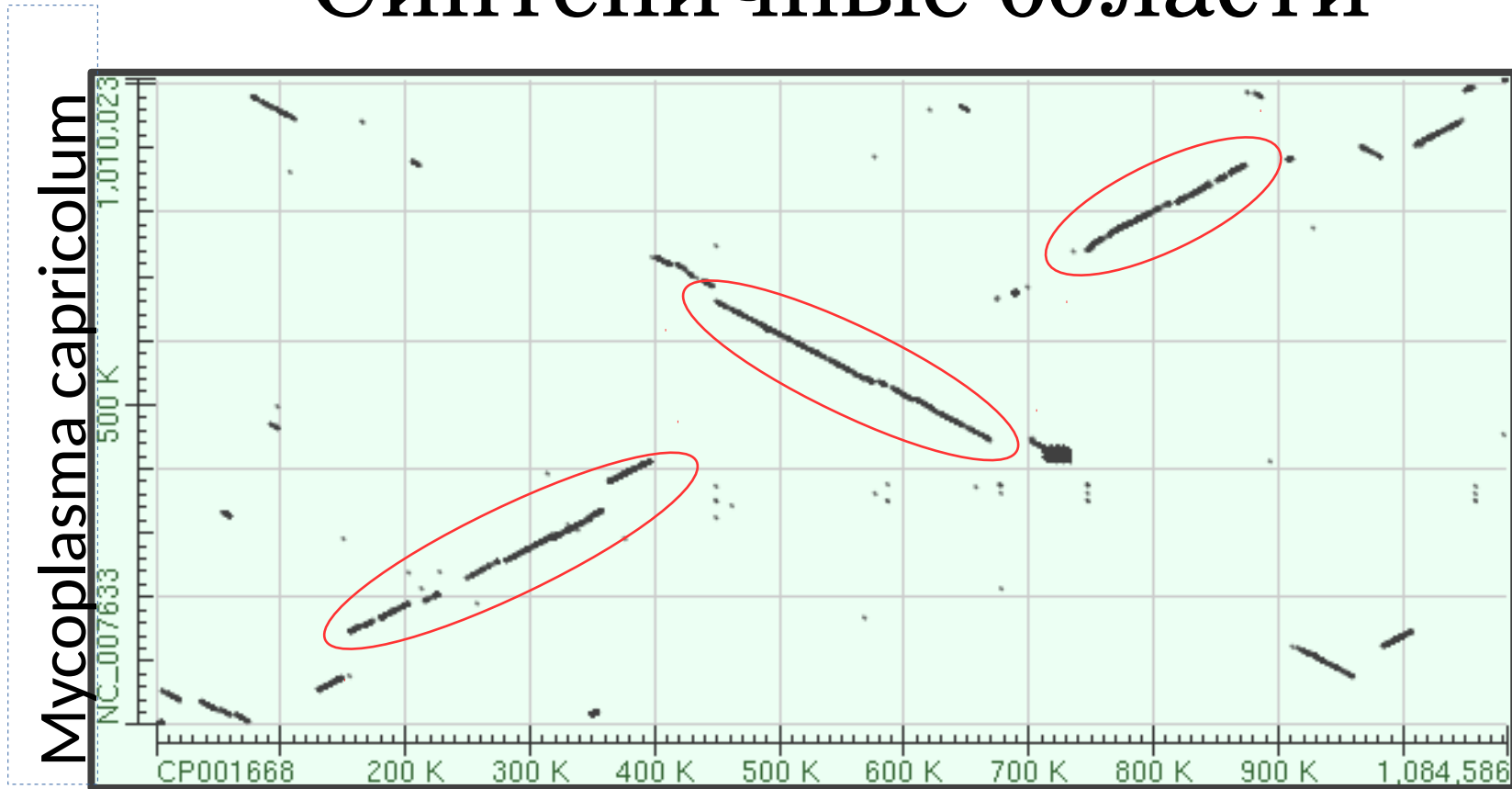
Карта сходства 2х последовательностей с учетом комплементарной цепочки



Карта локального сходства геномов *M. capricolum* и *M. mycoides*



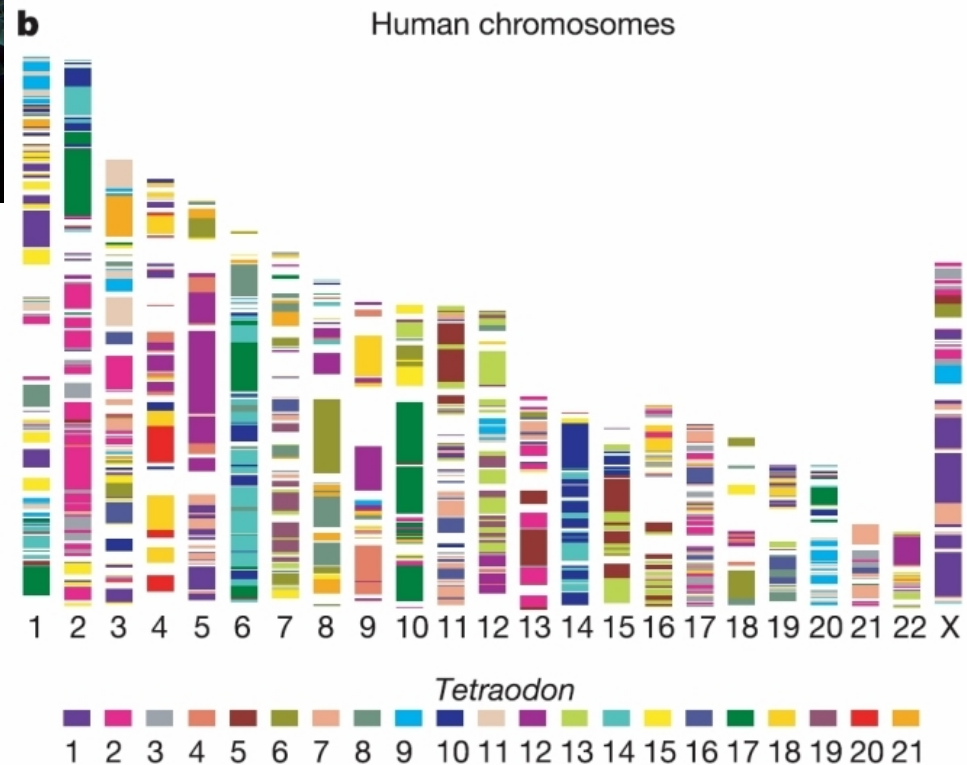
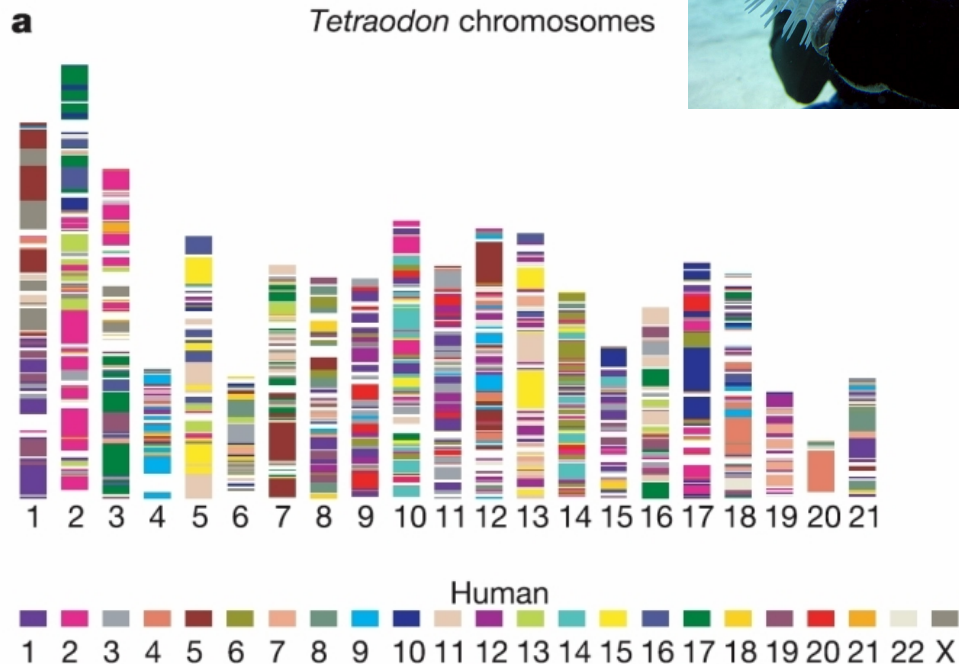
Синтеничные области



Mycoplasma mycoides

Синтеничные области - участки геномов, состоящие из ортологичных областей с сохранением их порядка на хромосоме для сравниваемых геномов

Синтеничные фрагменты на хромосомах человека и скалозуба



Множественное выравнивание

- Вход:
 - несколько геномов
 - каждый геном представлен одной или несколькими хромосомами и плазмидами
 - вариант: набор контигов или скэффолдов
- Результат:
 - Наборы ортологичных участков из всех или части геномов и их выравнивание (блок)
 - последовательность наборов в каждой ДНК

Пангеном

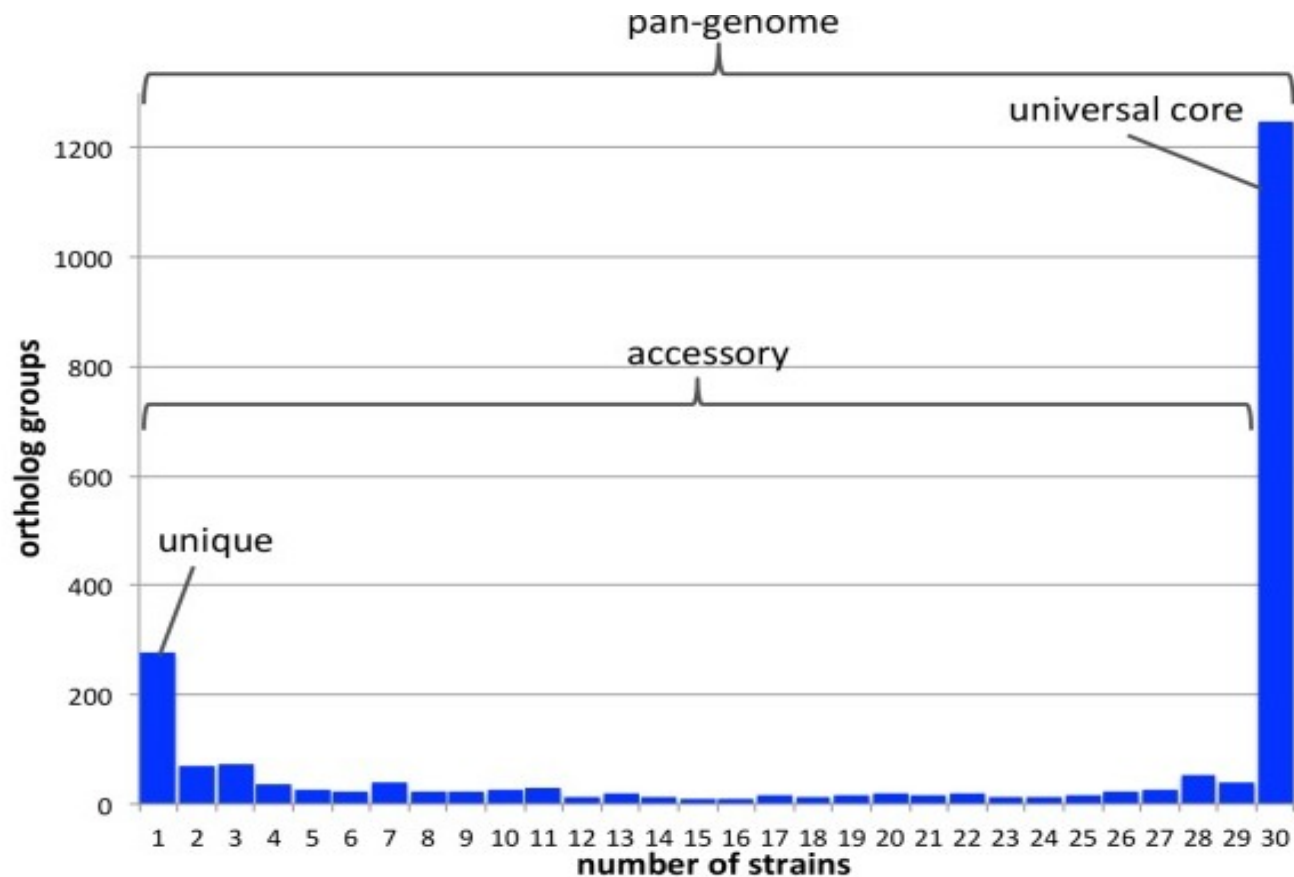
Пангеном может быть построен на основе ортологичных генов или ортологичных участков генома (нуклеотидный пангеном)

Кор — блоки, которые включают участки, встречающиеся во всех геномах

Дополнительные участки — встречаются не во всех геномах

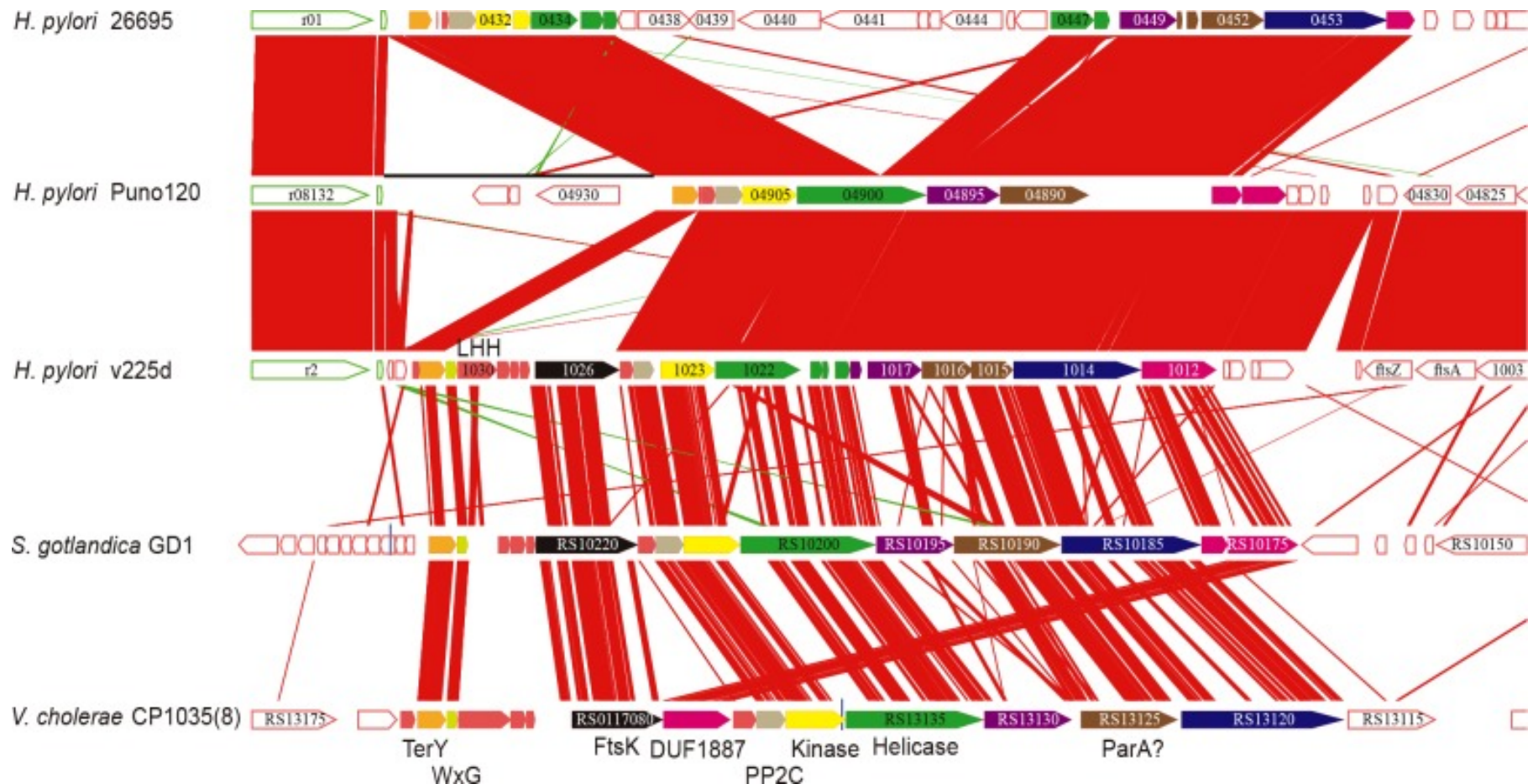
Уникальные участки — встречаются только в одном геноме

Пангеном 30 штаммов *Helicobacter pylori*



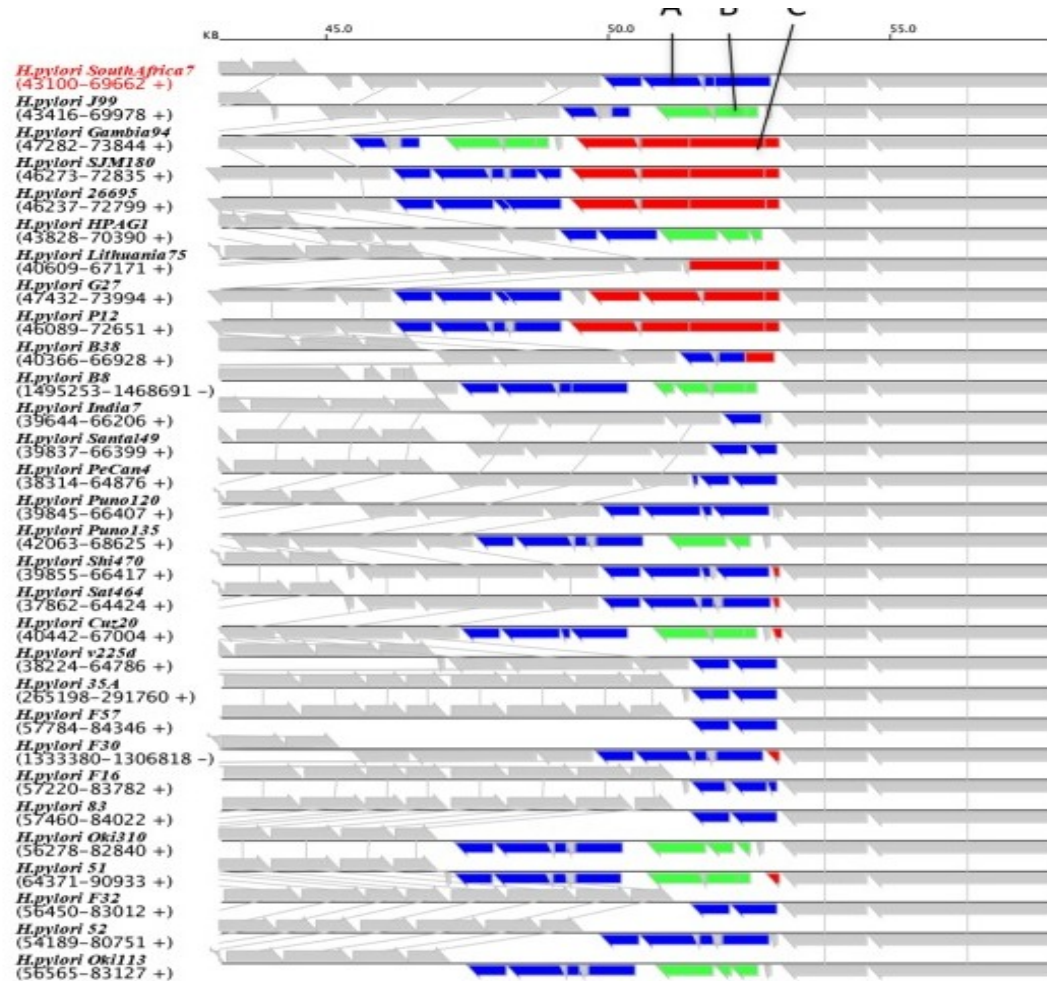
Uchiyama I, Albritton J, Fukuyo M, Kojima KK, Yahara K, Kobayashi I. A Novel Approach to *Helicobacter pylori* Pan-Genome Analysis for Identification of Genomic Islands. PLoS One. 2016 Aug 9;11(8):e0159419.

Сравнение нескольких геномов: визуализация



Uchiyama I, Albritton J, Fukuyo M, Kojima KK, Yahara K, Kobayashi I. A Novel Approach to *Helicobacter pylori* Pan-Genome Analysis for Identification of Genomic Islands. PLoS One. 2016 Aug 9;11(8):e0159419.

Горизонтальный перенос генов



Uchiyama I, Albritton J, Fukuyo M, Kojima KK, Yahara K, Kobayashi I. A Novel Approach to *Helicobacter pylori* Pan-Genome Analysis for Identification of Genomic Islands. PLoS One. 2016 Aug 9;11(8):e0159419.

NPG Explorer

Программа для поиска и визуализации нуклеотидного пангенома <http://mouse.belozersky.msu.ru/tools/npge.html>

Для построения пангенома NPGE использует данные о гомологичных участках, а не генах

Один блок может включать ортологи и паралоги

Обозначения блоков:

s – такие участки присутствуют по одному у всех изучаемых геномов

h – такие участки есть в части геномов

u – уникальные участки

r – повторы

m - минорные участки, длиной меньше 100 н.п., разделяющие остальные блоки

NPG-explorer

only blocks of >= 2 fragments Search: Clear

Global blocks Normal blocks

blockset alignment:

fr	^	colu	% id	% GC	gene	split	low s	% lo
s29x3446	29	3446	96...	42...	0	0	2	5.45
s29x3443	29	3443	96...	38...	0	0	2	9.61
s29x3419	29	3419	90...	42...	0	0	4	23...
s29x3415	29	3415	91...	45...	0	0	9	39...
s29x3350	29	3350	90...	44...	0	0	5	54...
s29x3344	29	3344	98...	40...	0	0	1	0.53
s29x3337	29	3337	96...	32...	0	0	2	5.27
s29x3244	29	3244	93...	40...	0	0	1	9.86
s29x3176	29	3176	92...	43.5	0	0	4	45...
s29x3166	29	3166	93...	44...	0	0	5	47...
s29x3139	29	3139	92...	38...	0	0	2	29...
s29x3077	29	3077	97...	40...	0	0	1	2.59
s29x3066	29	3066	98...	34...	0	0	1	0.91
s29x3035	29	3035	93...	35...	0	0	2	48...
s29x3008	29	3008	96...	38...	0	0	3	5.48

	1	2	3	4	5	6	7	8
+SpnCGSP14&chr1&c	s29x3446 >	r60x101 >	m10x21 >	r16x101 >	m6x63 >	-	-	-
+SpnD39&chr1&c	s29x3446 >	-	-	-	-	-	-	-
+SpnG54&chr1&c	s29x3446 >	-	-	-	-	-	-	-
+SpnHungary19A6&chr1&c	s29x3446 >	r60x101 >	m10x21 >	r16x101 >	m6x63 >	-	-	-
+SpnINV104&chr1&c	s29x3446 >	-	m10x21 >	r16x101 >	-	m3x3 >	h10x105 >	-
+SpnINV200&chr1&c	s29x3446 >	r60x101 >	m10x21 >	r16x101 >	m6x63 >	-	-	-
+SpnJJA&chr1&c	s29x3446 >	-	m10x21 >	r16x101 >	m6x63 >	-	-	-
-SpnNCTC7465&chr1&c	s29x3446 >	-	-	-	-	-	-	-
+SpnNT11058&chr1&c	s29x3446 >	-	-	-	-	-	-	-
+SpnOXC141&chr1&c	s29x3446 >	-	m10x21 >	r16x101 >	m6x63 >	-	-	-
+SpnP1031&chr1&c	s29x3446 >	m6x95 >	-	-	-	-	-	-
+SpnR6&chr1&c	s29x3446 >	-	-	-	-	-	-	-
+SpnSNP034156&chr1&c	s29x3446 >	-	m10x21 >	r16x101 >	m1x28 >	r15x111 >	m1x15 >	-
+SpnSNP034183&chr1&c	s29x3446 >	-	m10x21 >	r16x101 >	m6x63 >	-	-	-
+SpnSNP994038&chr1&c	s29x3446 >	r60x101 >	m10x21 >	r16x101 >	-	m3x3 >	h10x105 >	-
+SpnSNP994039&chr1&c	s29x3446 >	-	m10x21 >	r16x101 >	-	m3x3 >	h10x105 >	-
+SpnSPN032672&chr1&c	s29x3446 >	-	-	-	-	-	-	-
+SpnSPN033038&chr1&c	s29x3446 >	-	-	-	-	-	-	-

10 20 30 40 50 60 70 80

AACTCTATCAGGAAAGTCAAATTAATTTATAGAAATATTTTAGCAGTCAAAGGTGTACTATTATAGATTCAATATACATATAT

Son6706B&chr1&c 84929 88366 AACTCTATCAGGAAAGTCAAATTAATTTATAGAAATATTTTAGCAGTCAAAGGTGTACTATTATAGATTCAATATACATATAT

Son70585&chr1&c 68971 72416 AACTCTATCAGGAAAGTCAAATTAATTTATAGAAATATTTTAGCAGTCAAAGGTGTACTATTATAGATTCAATATACATATAT

SonA45&chr1&c 2026819 2023382 AACTCTATCAGGAAAGTCAAATTAATTTATAGAAATATTTTAGCAGTCAAAGGTGTACTATTATAGATTCAATATACATATAT

SonA66&chr1&c 30299 33744 AACTCTATCAGGAAAGTCAAATTAATTTATAGAAATATTTTAGCAGTCAAAGGTGTACTATTATAGATTCAATATACATATAT

SonAP200&chr1&c 66418 69855 AACTCTATCAGGAAAGTCAAATTAATTTATAGAAATATTTTAGCAGTCAAAGGTGTACTATTATAGATTCAATATACATATAT

SonATCC700669&chr1&c 38907 42344 AACTCTATCAGGAAAGTCAAATTAATTTATAGAAATATTTTAGCAGTCAAAGGTGTACTATTATAGATTCAATATACATATAT

SonCGSP14&chr1&c 32689 36126 AACTCTATCAGGAAAGTCAAATTAATTTATAGAAATATTTTAGCAGTCAAAGGTGTACTATTATAGATTCAATATACATATAT

SonD39&chr1&c 31094 34539 AACTCTATCAGGAAAGTCAAATTAATTTATAGAAATATTTTAGCAGTCAAAGGTGTACTATTATAGATTCAATATACATATAT

SonG54&chr1&c 30266 33703 AACTCTATCAGGAAAGTCAAATTAATTTATAGAAATATTTTAGCAGTCAAAGGTGTACTATTATAGATTCAATATACATATAT

SonHunaarv19A6&chr1&c 105643 109080 AACTCTATCAGGAAAGTCAAATTAATTTATAGAAATATTTTAGCAGTCAAAGGTGTACTATTATAGATTCAATATACATATAT

SonINV104&chr1&c 31313 34750 AACTCTATCAGGAAAGTCAAATTAATTTATAGAAATATTTTAGCAGTCAAAGGTGTACTATTATAGATTCAATATACATATAT

SonINV200&chr1&c 32686 36123 AACTCTATCAGGAAAGTCAAATTAATTTATAGAAATATTTTAGCAGTCAAAGGTGTACTATTATAGATTCAATATACATATAT

SonJJA&chr1&c 37546 40991 AACTCTATCAGGAAAGTCAAATTAATTTATAGAAATATTTTAGCAGTCAAAGGTGTACTATTATAGATTCAATATACATATAT

SonNCTC7465&chr1&c 430580 427143 AACTCTATCAGGAAAGTCAAATTAATTTATAGAAATATTTTAGCAGTCAAAGGTGTACTATTATAGATTCAATATACATATAT

SonNT11058&chr1&c 41162 44599 AACTCTATCAGGAAAGTCAAATTAATTTATAGAAATATTTTAGCAGTCAAAGGTGTACTATTATAGATTCAATATACATATAT

SonOXC141&chr1&c 64784 68221 AACTCTATCAGGAAAGTCAAATTAATTTATAGAAATATTTTAGCAGTCAAAGGTGTACTATTATAGATTCAATATACATATAT

SonP1031&chr1&c 66799 70236 AACTCTATCAGGAAAGTCAAATTAATTTATAGAAATATTTTAGCAGTCAAAGGTGTACTATTATAGATTCAATATACATATAT

SonR6&chr1&c 31094 34539 AACTCTATCAGGAAAGTCAAATTAATTTATAGAAATATTTTAGCAGTCAAAGGTGTACTATTATAGATTCAATATACATATAT

SonSNP034156&chr1&c 1171657 1175094 AACTCTATCAGGAAAGTCAAATTAATTTATAGAAATATTTTAGCAGTCAAAGGTGTACTATTATAGATTCAATATACATATAT

SonSNP034183&chr1&c 64784 68221 AACTCTATCAGGAAAGTCAAATTAATTTATAGAAATATTTTAGCAGTCAAAGGTGTACTATTATAGATTCAATATACATATAT

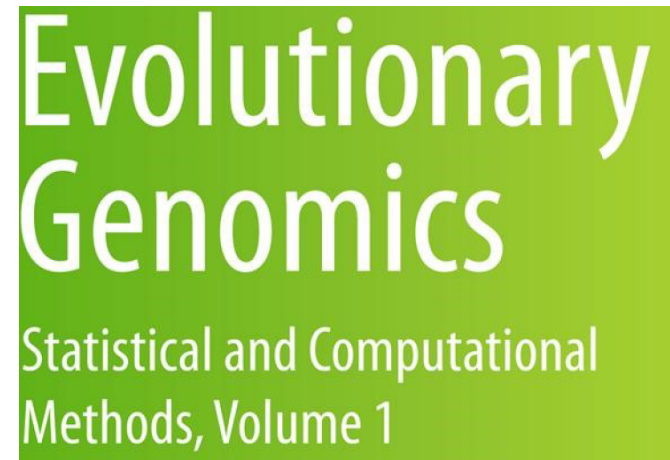
SonSNP994038&chr1&c 64784 68221 AACTCTATCAGGAAAGTCAAATTAATTTATAGAAATATTTTAGCAGTCAAAGGTGTACTATTATAGATTCAATATACATATAT

SonSNP994039&chr1&c 64784 68221 AACTCTATCAGGAAAGTCAAATTAATTTATAGAAATATTTTAGCAGTCAAAGGTGTACTATTATAGATTCAATATACATATAT

SonSPN032672&chr1&c 1107113 1200550 AACTCTATCAGGAAAGTCAAATTAATTTATAGAAATATTTTAGCAGTCAAAGGTGTACTATTATAGATTCAATATACATATAT

Литература (книга открыта)

Method	Category	Relationships predicted	Pairwise or multiple
BLAST	Local alignment	Homology	Pairwise
LASTZ	Local genomic alignment	Homology	Pairwise
MUMmer	Local genomic alignment	Orthology	Pairwise
CHAOS	Local genomic alignment	Homology	Pairwise
GRIMM-synteny	Orthology mapping	Toporthology	Multiple
DRIMM-synteny	Orthology mapping	Orthology, paralogy	Multiple
Mercator	Orthology mapping	Toporthology	Multiple
Enredo	Orthology mapping	Orthology, paralogy	Multiple
OSfinder	Orthology mapping	Orthology	Multiple
SuperMap	Orthology mapping	Orthology, paralogy	Multiple
progressiveMauve	Hierarchical WGA	Toporthology	Multiple
MUGSY	Hierarchical WGA	Toporthology	Multiple
MAVID	Global genomic alignment	Colinear homology	Multiple
LAGAN/Multi-LAGAN	Global genomic alignment	Colinear homology	Pairwise/ multiple
DIALIGN	Global genomic alignment	Colinear homology	Multiple
SeqAn::T-Coffee	Global genomic alignment	Colinear homology	Multiple
FSA	Global genomic alignment	Colinear homology	Multiple
Pecan	Global genomic alignment	Colinear homology	Multiple
NUCmer/PROmer	Local WGA	Orthology	Pairwise
MULTIZ/TBA	Local WGA	Orthology, paralogy	Multiple
AXTCHAIN/ CHAINNET	Alignment chaining and filtering	Orthology	Pairwise



Chapter 8

Whole-Genome Alignment

Colin N. Dewey