Занятие №9. Обработка данных секвенирования

Покрытие

- Покрытие (глубина секвенирования) – важный параметр методов NGS: кратность прочтения каждого нуклеотида. Для каждой задачи необходимо своё покрытие (обычно устанавливают не менее, чем 30-тикратное покрытие).
- Таким образом, “эффективный” объём данных равен выходу секвенирования, делённому на покрытие.

Оценка необходимого покрытия

- Вероятность того, что нуклеотид не будет определён (P), исходя из глубины покрытия (c) вычисляется по формуле Ландела–Ватермана:

$$P=e^{-c}$$

- Теоретически достаточное покрытие должно позволять определить все нуклеотиды в геноме длиной L ($P*L<1$).
- Например, для генома человека ($L=3*10^9$ п.о.) теоретически достаточно 23-кратного покрытия

Анализ данных секвенирования

- **1. Очистка “сырых” данных (raw data) (фильтрация ридов по качеству).**
- Результат: “примесные” риды удаляются, в остальных обрезаются неточно определённые нуклеотиды
- **2. Сборка генома (слияние ридов для коротких фрагментов) с помощью специальной программы – ассемблера.**
- Результат: набор длинных фрагментов (контиги) или их упорядоченная последовательность, образующая скэффолд.
- **3. Интерпретация данных (аннотация)**
- поиск кодирующих последовательностей и их структурное и

1. Оценка качества ридов: FASTQ – формат записи ридов

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTC
AACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Каждая последовательность занимает 4 строки:

– первая начинается с @ и содержит название и описание последовательности;

– вторая содержит последовательность (знаки A,G,C,T);

– третья начинается с + и может содержать примечания (технические комментарии секвенирования);

– четвёртая содержит столько же символов, что и вторая, каждый символ указывает вероятность ошибочного определения соответствующего нуклеотида по шкале Phred.

Определение качества ридов по шкале Phred

- Каждый символ означает какое-то число (Q) от 0 до 100. Вероятность ошибочного определения нуклеотида (P), качество которого оценивается как Q равна:
  - $P = 10^{-Q/10}$
- “Хорошее” качество при $Q \geq 30$ ($P \leq 0{,}001 = 0{,}1\%$)

Phred и ASCII:
номера присваивают начиная с 33 символа (!=0) (Phred+33) или с 64 (@=0) (Phred+64)

ASCII control characters		
00	NULL	(Null character)
01	SOH	(Start of Header)
02	STX	(Start of Text)
03	ETX	(End of Text)
04	EOT	(End of Trans.)
05	ENQ	(Enquiry)
06	ACK	(Acknowledgement)
07	BEL	(Bell)
08	BS	(Backspace)
09	HT	(Horizontal Tab)
10	LF	(Line feed)
11	VT	(Vertical Tab)
12	FF	(Form feed)
13	CR	(Carriage return)
14	SO	(Shift Out)
15	SI	(Shift In)
16	DLE	(Data link escape)
17	DC1	(Device control 1)
18	DC2	(Device control 2)
19	DC3	(Device control 3)
20	DC4	(Device control 4)
21	NAK	(Negative acknowl.)
22	SYN	(Synchronous idle)
23	ETB	(End of trans. block)
24	CAN	(Cancel)
25	EM	(End of medium)
26	SUB	(Substitute)
27	ESC	(Escape)
28	FS	(File separator)
29	GS	(Group separator)
30	RS	(Record separator)
31	US	(Unit separator)
127	DEL	(Delete)

ASCII printable characters					
32	space	64	@	96	`
33	!	65	A	97	a
34	"	66	B	98	b
35	#	67	C	99	c
36	$	68	D	100	d
37	%	69	E	101	e
38	&	70	F	102	f
39	'	71	G	103	g
40	(	72	H	104	h
41	)	73	I	105	i
42	*	74	J	106	j
43	+	75	K	107	k
44	,	76	L	108	l
45	-	77	M	109	m
46	.	78	N	110	n
47	/	79	O	111	o
48	0	80	P	112	p
49	1	81	Q	113	q
50	2	82	R	114	r
51	3	83	S	115	s
52	4	84	T	116	t
53	5	85	U	117	u
54	6	86	V	118	v
55	7	87	W	119	w
56	8	88	X	120	x
57	9	89	Y	121	y
58	:	90	Z	122	z
59	;	91	[	123	{
60	<	92	\	124	|
61	=	93	]	125	}
62	>	94	^	126	~
63	?	95	_		

Extended ASCII characters							
128	Ç	160	á	192	└	224	Ó
129	ü	161	í	193	┴	225	ß
130	é	162	ó	194	┬	226	Ô
131	â	163	ú	195	├	227	Ò
132	ä	164	ñ	196	─	228	õ
133	à	165	Ñ	197	┼	229	Õ
134	å	166	ª	198	ã	230	µ
135	ç	167	º	199	Ã	231	þ
136	ê	168	¿	200	╚	232	Þ
137	ë	169	®	201	╔	233	Ú
138	è	170	¬	202	╩	234	Û
139	ï	171	½	203	╦	235	Ù
140	î	172	¼	204	╠	236	ý
141	ì	173	¡	205	═	237	Ý
142	Ä	174	«	206	╬	238	¯
143	Å	175	»	207	¤	239	´
144	É	176	░	208	ð	240	≡
145	æ	177	▒	209	Ð	241	±
146	Æ	178	▓	210	Ê	242	‗
147	ô	179	│	211	Ë	243	¾
148	ö	180	┤	212	È	244	¶
149	ò	181	Á	213	ı	245	§
150	û	182	Â	214	Í	246	÷
151	ù	183	À	215	Î	247	¸
152	ÿ	184	©	216	Ï	248	°
153	Ö	185	╣	217	┘	249	¨
154	Ü	186	║	218	┌	250	·
155	ø	187	╗	219	█	251	¹
156	£	188	╝	220	▄	252	³
157	Ø	189	¢	221	¦	253	²
158	×	190	¥	222	Ì	254	■
159	ƒ	191	┐	223	▀	255	nbsp

Кодировка качества Phred+33

- Качество в формате FASTQ закодировано в ASCII – символах
- Существовало несколько стандартов записи качества
- Многим программам важно, чтения во Phred+33 или Phred+64

```
SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................
.....................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX...
.....................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII...
.....................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ...
..LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijk
 |                         |    |        |                              |
33                        59   64       73                            104
 0........................26...31.......40
                          -5....0........9.............................40
                               0........9.............................40
                                 3.....9.............................40
 0.2......................26...31........41

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

```
@M03106:3:000000000-AAWHP:1:1101:17129:1029 1:N:0:84
NCTACGGGAGGCAGCAGTAGGGAATCTTCCGCAATGGACGAAAGTCTGACGGAGCAACGCCGCGTGAGTGATGAAGG
TTTTCGGATCGTAAAACTCTGTTGTTAGGGAAGAATAAGTATCGGAGTAACTGCCGGTACCTTGACGGTACCTAACC
AGAAAGCCACGGCTAACTACGTGCCAGCAGCCGCGGTAATACGTAGGTGGCAAGCGTTGTCCGGAATTATTGGGCGT
AAAGCGCGCGCAGGTGGTTCCTTAAGGCTGATGTGAAAGCCCACGGCTCAACCAGGGGGGGTCATTGGGA
+
#3BCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGFGGGGGGFGGGGGGGGGGGGGGGGGGGGGGGF=BEGGGGGGGGGGGGGGGGGGGGFGGGGEGGGFGC
FFGGGFDAF7>DD>BGGEFFFFFAFEF2:B@=@@2DFFBEEEF)3==@FFD@EF0<DD;?AE?F603F?AA2CFB9B
F:?<296924(21>3>8(4>0))69<*425(5<BB>B4).((4*:E096F<6:**1;(,<;-8*++23**
```

Базовая таблица кодировки ASCII и ее соответствие качеству прочтения

Q	Шкала	символ	Q	Шкала	символ	Q	Шкала	символ
0	33	!	20	53	5	40	73	I
1	34	“	21	54	6	41	74	J
2	35	#	22	55	7	42	75	K
3	36	$	23	56	8	43	76	L
4	37	%	24	57	9	44	77	M
5	38	&	25	58	:	45	78	N
6	39	‘	26	59	;	46	79	O
7	40	(	27	60	<	47	80	P
8	41	)	28	61	=	48	81	Q
9	42	*	29	62	>	49	82	R
10	43	+	30	63	?	50	23	S
11	44	,	31	64	@	51	84	T
12	45	-	32	65	A	52	85	U
13	46	.	33	66	B	53	86	V
14	47	/	34	67	C	54	87	W
15	48	0	35	68	D	55	88	X
16	49	1	36	69	E	56	89	Y
17	50	2	37	70	F	57	90	Z
18	51	3	38	71	G	58	91	[
19	52	4	39	72	H	59	92	\

Программа FastQC

http://www.bioinformatics.babraham.ac.uk/

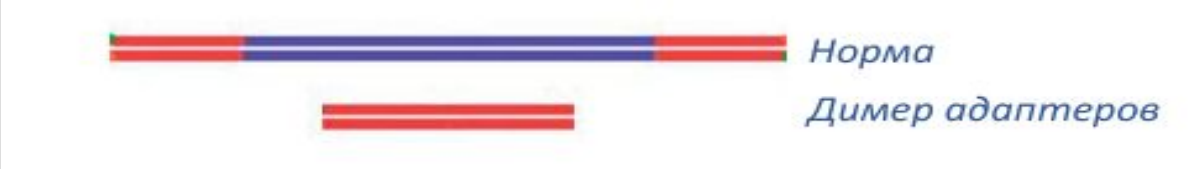
Примеры качества по шкалам Phred+33 и Phred+64

Качество по Phred, Q	Символ ASCII (Phred33)	Символ ASCII (Phred64)	Вероятность ошибки	Точность

Источники ошибок в ридах: примеси

- Примеси бывают:
- 1. Артефактные (ошибки секвенирования)

- образование димеров адаптеров

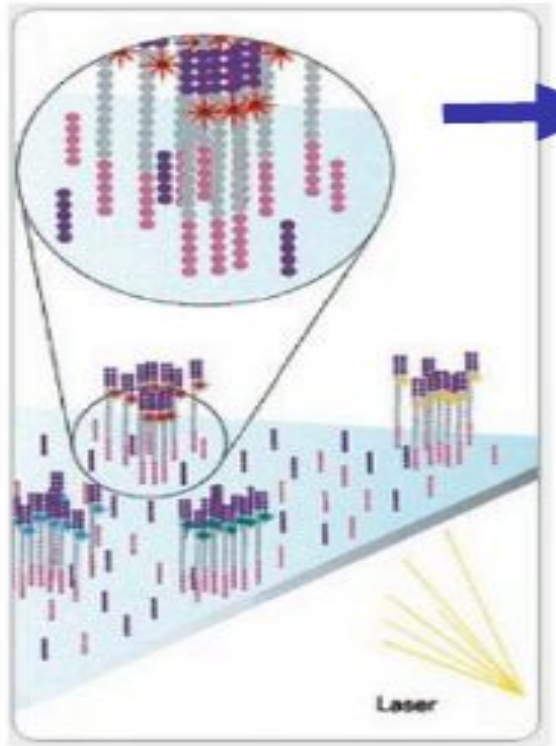


- чтение сквозь – вставки слишком короткие

- 2. Биологические – контаминация

Источники ошибок в ридах: фазировка

- Фрагменты в одном кластере строятся с разной скоростью – секвенатору сложно определить верный нуклеотид.



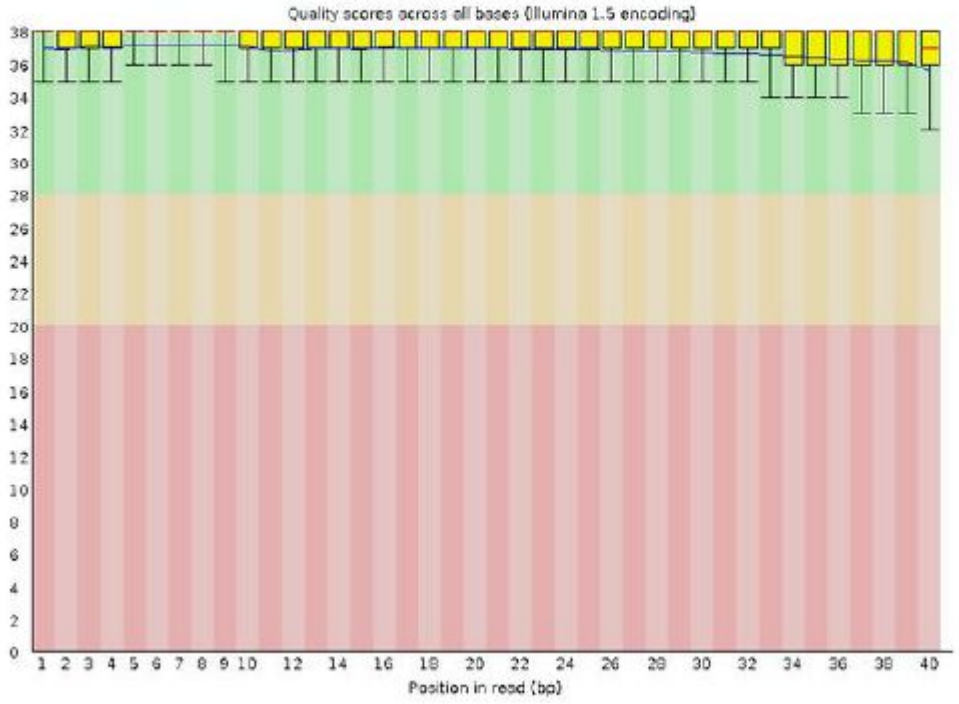
CGCAAGGAAGCTGATTG
CGCAAGGAAGCTGATTG
CGCAAGGAAGCTGATTG
CGCAAGGAAGCTGATT
CGCAAGGAAGCTGATTGC
CGCAAGGAAGCTGATTG

Чем дольше идёт прогон, тем больше будет накапливаться отстающих и опережающих олигонуклеотидов

Программа FastQC – контроль качества ридов: 1. Среднее нуклеотидное качество – хорошее (все Me>25, все Q1>10)

Нуклеотидное качество

Per base sequence quality



Качество высокое на протяжении всего чтения

- Синяя линия – среднее качество
- Красная линия – медиана
- Жёлтая рамка – интерквартиль (50% чтений попадает в эти границы)
- **Чёрные** засечки – 80% чтений попадает в эти границы

Программа FastQC – контроль качества ридов: 1. Среднее нуклеотидное качество – неудовлетворительное (есть Me<20 или Q1<5)

Нуклеотидное качество

Per base sequence quality

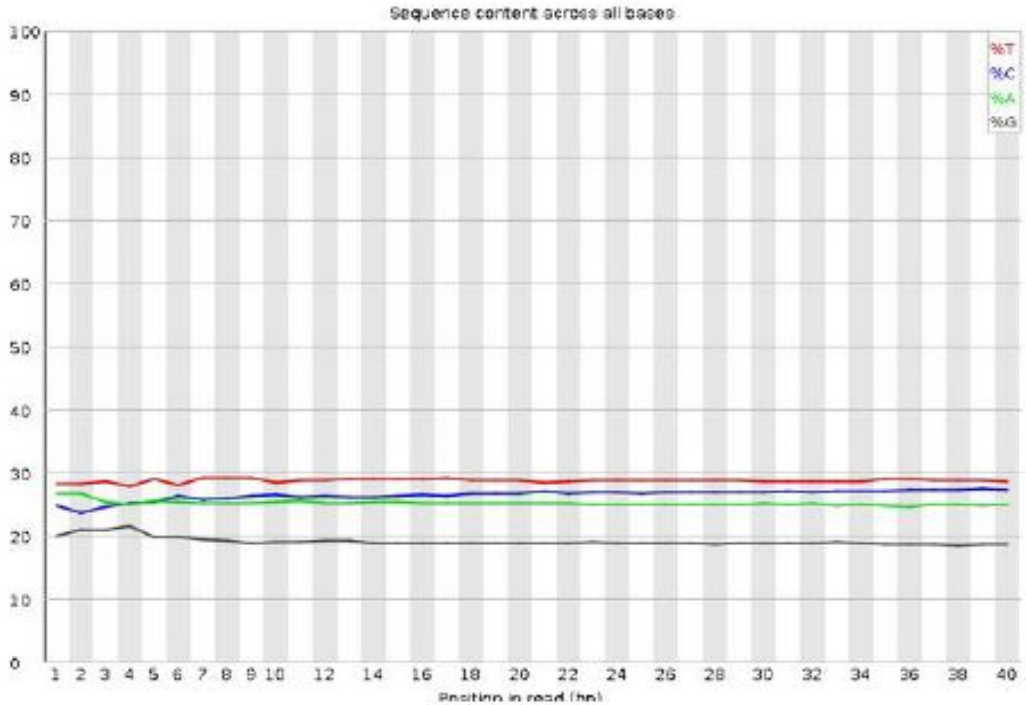
Качество сильно падает к концу

- Синяя линия – среднее качество
- Красная линия – медиана
- Жёлтая рамка – интерквартиль (50% чтений попадает в эти границы)
- **Чёрные** засечки – 80% чтений попадает в эти границы

Программа FastQC – контроль качества ридов: 2. Средний нуклеотидный состав ридов

Нуклеотидный состав

Per base sequence content

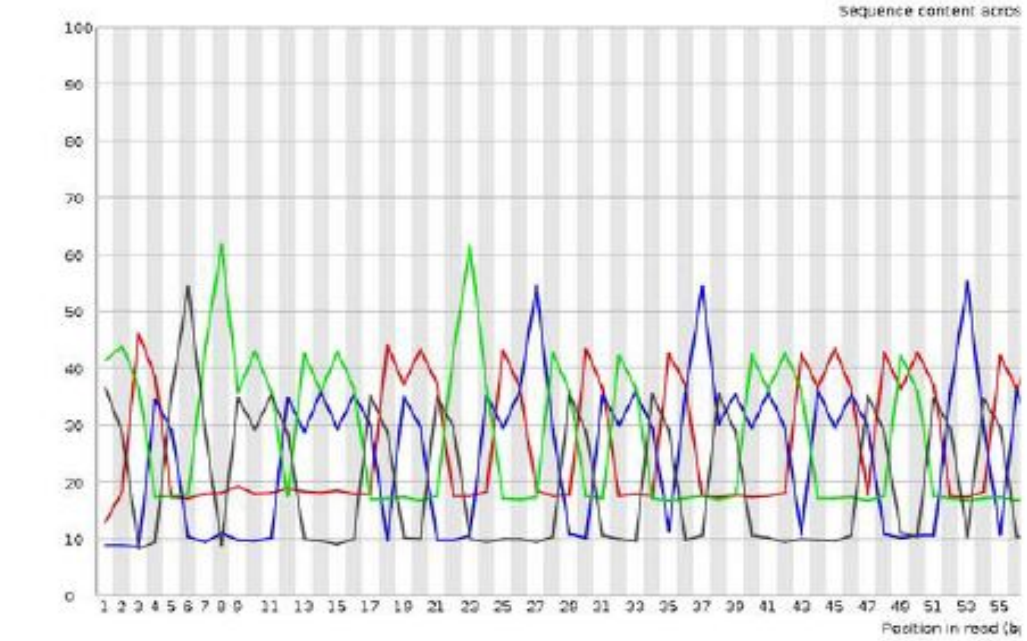


- Нуклеотидный состав примерно одинаковый на протяжении всего рида
- **В начале чтения небольшие колебания** из-за специфичности фрагментации – программа обозначает их восклицательным знаком, но это не проблема

Программа FastQC – контроль качества ридов: 2. Средний нуклеотидный состав ридов

Нуклеотидный состав

Per base sequence content



Сильные колебания нуклеотидного состава – скорее всего отсеквенированы димеры адаптеров

Программа FastQC – контроль качества ридов: 3. Чрезмерно представленные последовательности

1) **Overrepresented sequences**

No overrepresented sequences

2) **Overrepresented sequences**

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACGTCCGCACATCTCGTAT	938427	17.993033077161126	TruSeq Adapter, Index 1 (97% over 36bp)

Последовательность адаптера, из-за которого было колебание нуклеотидного состава на прошлом слайде

Адаптер найден в ~18 % чтений

Название найденного адаптера – пригодится, когда настанет пора его удалять

Очистка “сырых” ридов: тримминг

- 1. Удаление адаптерных последовательностей из ридов
- 2. Отсечение с конца ридов нуклеотидов, качество которых ниже определённого уровня (Q<20 или Q<30)

- Инструмент для тримминга: программа Trimmomatic

Особый этап для метагеномики – Сортировка данных (биннинг)

- 1. Методы, основанные на нуклеотидном составе
- GC-состав
- динуклеотидный состав
- тринуклеотидный состав
- тетрануклеотидный состав

- 2. Методы, основанные на гомологии
- сравнение с базой данных

2. Сборка генома (assemby)

- de novo (сборка не секвенированного ранее генома)
- – метод OLC (overlap layout concensus) (перекрытие фрагментов) – для малого количества длинных фрагментов (Sanger)
- – графы де Брёйна – для большого количества коротких фрагментов (NGS)
- сборка генома, аналогичного ранее собранному (ресеквенирование) референсному геному (выравнивание на геном, alignment)
- – хэш-таблицы
- – суффиксные деревья

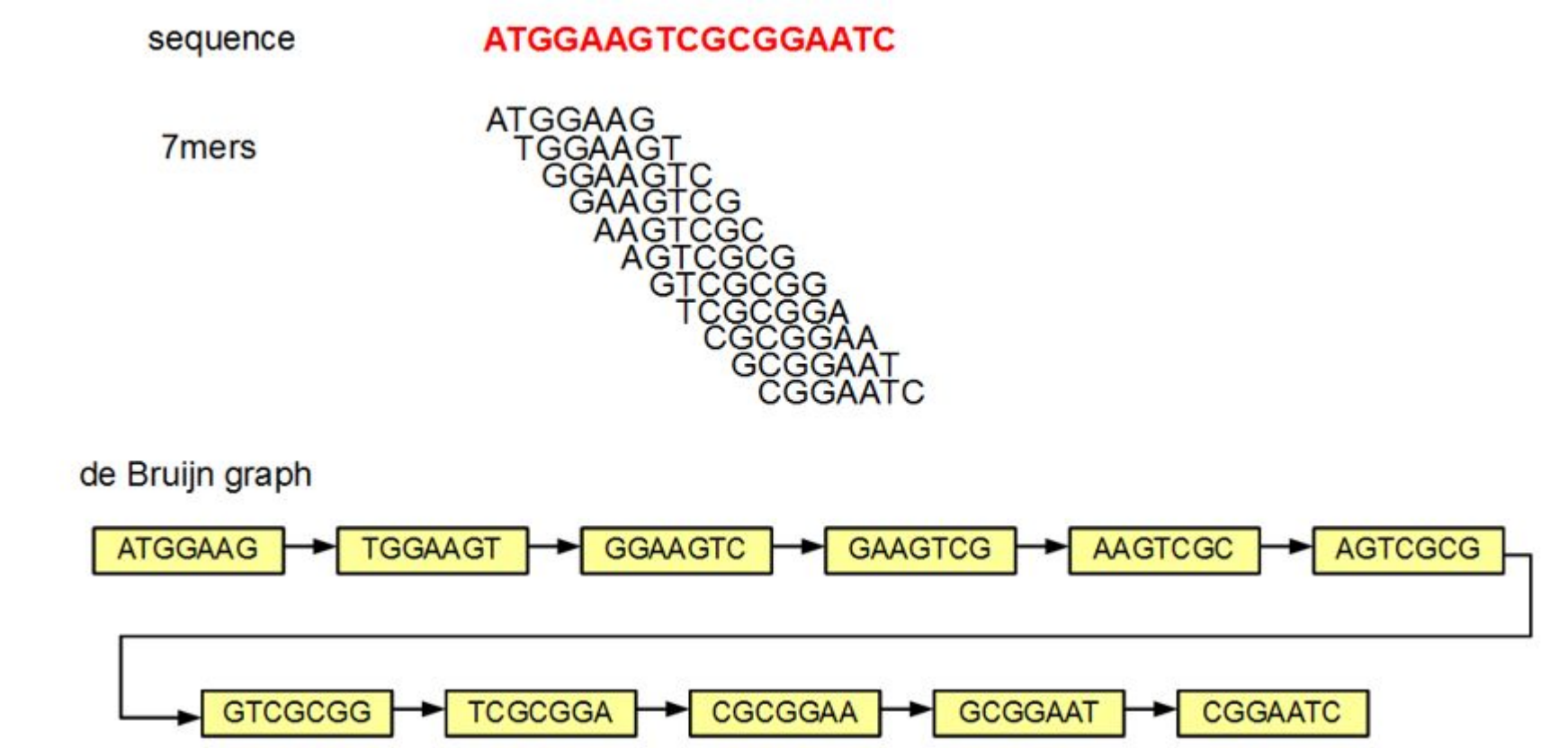
Сборка de novo: Overlap layout consensus: 1

- Поиск пар ридов, имеющих общие k-меры (последовательности длиной k, k=24), смещение двух строк относительно друг друга (выравнивание) до максимального совмещения (≥95% сходства)

Сборка de novo: Overlap layout consensus: 2

- На базе попарного выравнивания строят множественное выравнивание, корректируют ошибки

Сборка de novo: Графы де Брёйна



Результат сборки: контиги и скэффолды

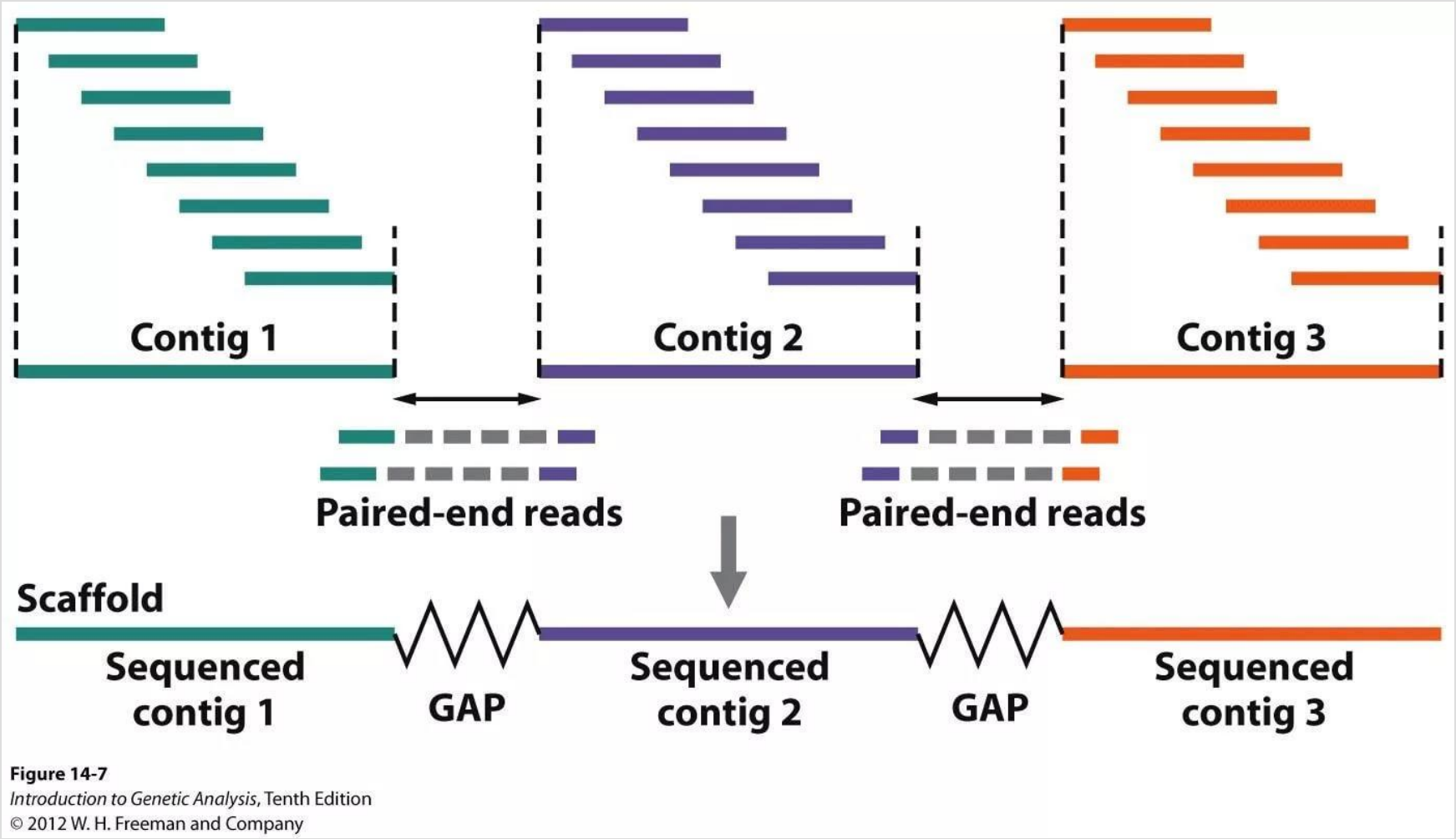
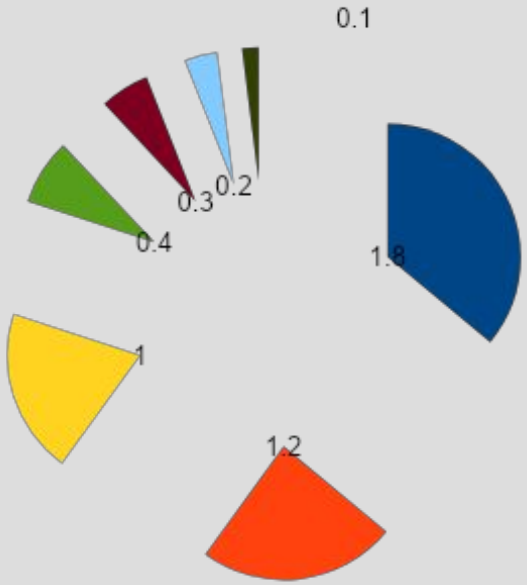
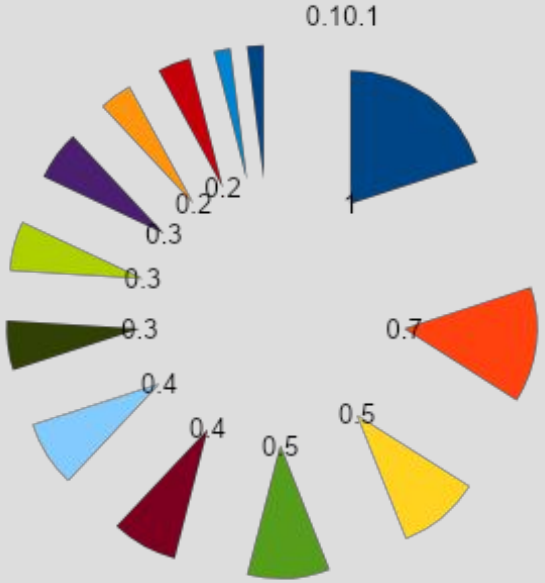


Figure 14-7
Introduction to Genetic Analysis, Tenth Edition

Качество сборки генома

- N50 – длина контига, который вместе с остальными контигами большей длины покрывает не менее 50% генома (обычно под геномом понимают суммарную длину всех контигов).
- L50 – число контигов не меньших чем N50.
- Пример: две сборки генома длиной 5 Mb





Формат представления нуклеотидных последовательностей – FASTA

- >OTU-160-1 Acinetobacter baumannii
- CCTACGGGGGGCTGCAGTGGGGAATATTGGACAATGGGGGGA ACCCTGATCCAGCCATGCCGCGTGTGTGAAGAAGGCCTTATG GTTGTAAAGCACTTTAAGCGAGGAGGAGGCTACTCTAGTTAAT ACCTAGGGATAGTGGACGTTACTCGCAGAATAA

- Каждая последовательность занимает две строки:
- 1). первая строка начинается со знака > и содержит идентификатор (за которым эта последовательность закреплена в некоторой базе данных), через пробел следует опциональное словесное описание;
- 2). вторая строка – сама последовательность нуклеотидов.

3. Аннотация

- 1. Поиск белок-кодирующих последовательностей
- на основе гомологии – сравнение с уже известными генами
- аннотация ab initio – статистический поиск по характерным для белок-кодирующих участков последовательностям (ATG.....)
- 2. Поиск других кодирующих последовательностей (гены РНК)

Результаты аннотации

- 1. Структурное описание:
- – открытые рамки считывания (ORF) и из расположение
- – структура гена
- – кодирующие области
- – локализация регуляторных последовательностей
- 2. Функциональное описание
- – биохимическая функция белкового продукта
- – биологическая функция белка
- – экспрессия белка
- – участие белка в регуляторных и межбелковых