Референсные геномы. Форматы FASTA, FASTQ, и их взаимопревращение

где найти собранный геном, что обычно присутствует в собранном геноме, что такое форматы FASTA и FASTQ

Page • Tag • 1 backlink

Ссылки

1. Емкий туториал по bedtools, дополняющий приведенные в этом модуле задачи
2. Вообще страница bedtools, например, разделы advanced usage и FAQ
3. Форум seqanswers.com - разделы, обсуждающие биоинформатическую обработку данных

Необходимые программы

Ниже приведен список программ, которые необходимо установить для решения практических задач, представленных в этом модуле.

1. Samtools
2. Bedtools
3. IGV genome browser
4. FastQC

Собранный геном

NCBI : информация по виду и статистика

Модельные организмы - организмы, используемые для изучения тех или иных свойств живой природы. Геномы этих организмы, как правило, собраны и аннотированы лучше других. Для агрегации геномной информации о модельных организмах существуют специальные консорциумы и веб-порталы

Пример: Homo sapiens, Mus musculus, Rattus norvegicus, Arabidopsis thaliana, Caenorhabditis elegans,Drosophila

melanogaster, Saccharomyces cerevisiae, Escherichia coli

Геном человека

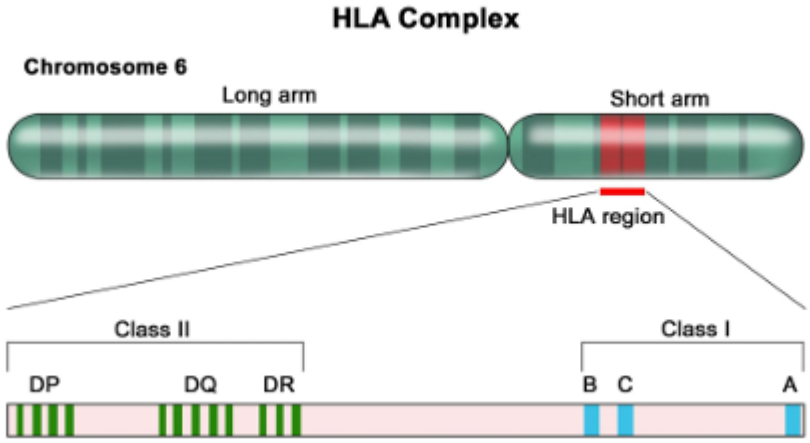
Геном человека - один из наиболее качественно собранных и аннотированных геномов. Современные версии сборки содержат, помимо основных хромосом, патчи-"заплатки" и альтернативные локусы

Геном человека: 25 апреля 2003 год, секвенирование методом
shotgun против BAC/YAC

Версии от Genome Reference Consortium - GRCh
Версии UCSC - hg
Часто используется версия от 1000 геномов - b37

Элементы

- chr1-22, chrX, chrY, chrM
- “заплатки” (patches): контиги, которые мы не знаем куда поместить (unplaced — непонятно)
- альтернативные локусы: консенсусная последовательность, содержит вариабельность (полиморфность) => разнообразие. Пример: комплекс HLA complex, участок с большой полиморфностью, единый консенсус не имеет смысла



www.ncbi.nlm.nih.gov

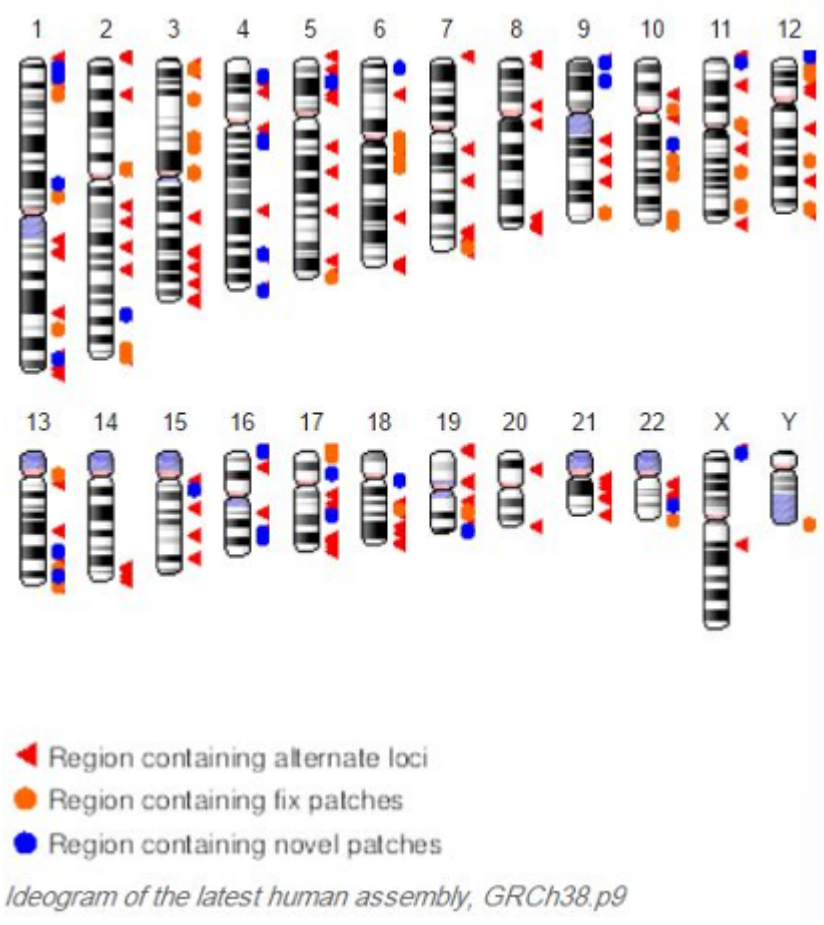
Assembly Terminology - Genome Reference Consortium

Below is a list of commonly used terms and definitions in the field of genomics and used by the NCBI Assembly Model. Describing Assemblies Alternate locus: A sequence that provides an alternate repres...

www.ncbi.nlm.nih.gov

Introduction to Patches - Genome Reference Consortium

What are patches? What types of patches are there? Do patches result in changes to chromosome coordinates? What is a patch release? How often does the GRC release patches? Why does the GRC release pat...



Ideogram of the latest human assembly, GRCh38.p9

Формат .fa (fasta)

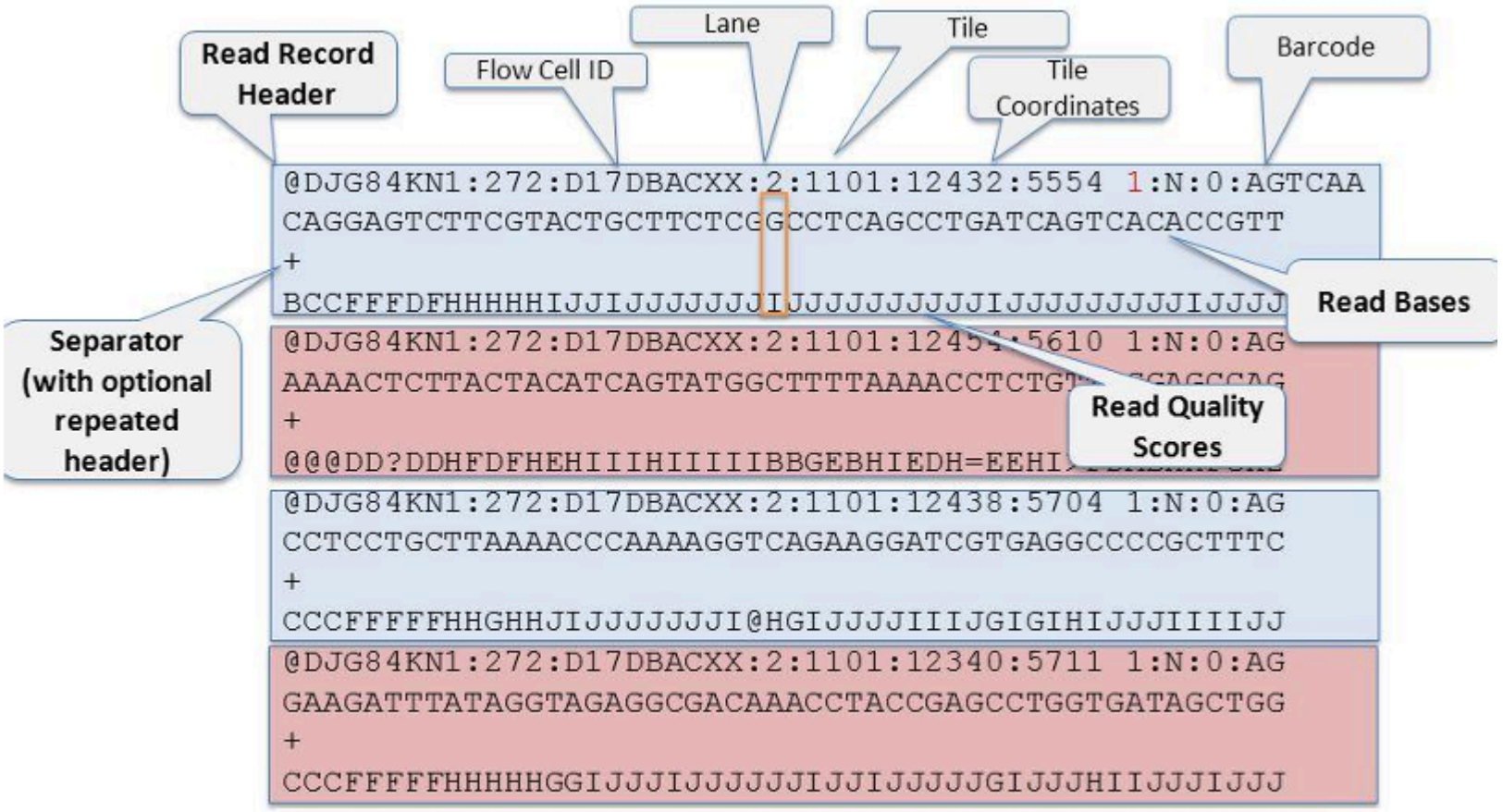
- Нуклеотидные и протеиновые последовательности
- Используется для консенсусного генома
- Содержит две строки: название (начинается с “>”) и сама последовательность

```
>read123456
GGGCCAAAGGAGCTTTCAAGGAGAGA
>read123457
GGGCAGTAGAGGCTTTCAAGGAGAGATTT
```

Формат .fastq (fastq)

Строка 1: название (@) — дополнительная информация
Строка 2: прочтение
Строка 3: комментарий
Строка 4: качества
прочтения оснований
(basecall qualities)

Заголовки



@ - название: id секвенатора : N запуска : cell ID : lane (дорожка) : tile (участок на lane): tile cood (координаты):1 (read bases — первый файл из двух фастку):::ATGCGGT (баркод)

Качества прочтения Phred

- *Логарифмическая шкала*
  10 и ниже - плохо, 20-приемлемо, 30 и выше - отлично

Quality value	Chance it is wrong	Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

- $Q = -10 \log_{10} P \iff P = 10^{-Q/10}$
  - Q = Phred quality score
  - P = probability of base call being incorrect

1. Закодировано качество определения каждого основания в прочтении
2. Качество оценивается в минус десять логарифма вероятности ошибки
3. Буквы удобнее строками и единым символом закодировать по таблице ASCII
4. К типичному Phred (0-40) добавляется 33

000	(nul)	016	►	(dle)	032	sp	048	0	064	@	080	P	096	`	112	p
001	☺	(soh)	017	◄	(dc1)	033	!	049	1	065	A	081	Q	097	a	113	q
002	☻	(stx)	018	↕	(dc2)	034	"	050	2	066	B	082	R	098	b	114	r
003	♥	(etx)	019	‼	(dc3)	035	#	051	3	067	C	083	S	099	c	115	s
004	♦	(eot)	020	¶	(dc4)	036	$	052	4	068	D	084	T	100	d	116	t
005	♣	(enq)	021	§	(nak)	037	%	053	5	069	E	085	U	101	e	117	u
006	♠	(ack)	022	▬	(syn)	038	&	054	6	070	F	086	V	102	f	118	v
007	•	(bel)	023	↨	(etb)	039	'	055	7	071	G	087	W	103	g	119	w
008	◘	(bs)	024	↑	(can)	040	(	056	8	072	H	088	X	104	h	120	x
009	○	(tab)	025	↓	(em)	041	)	057	9	073	I	089	Y	105	i	121	y
010	◙	(lf)	026	→	(eof)	042	*	058	:	074	J	090	Z	106	j	122	z
011	♂	(vt)	027	←	(esc)	043	+	059	;	075	K	091	[	107	k	123	{
012	♀	(np)	028	∟	(fs)	044	,	060	<	076	L	092	\	108	l	124	¦
013	♪	(cr)	029	↔	(gs)	045	-	061	=	077	M	093	]	109	m	125	}
014	♫	(so)	030	▲	(rs)	046	.	062	>	078	N	094	^	110	n	126	~
015	☼	(si)	031	▼	(us)	047	/	063	?	079	O	095	_	111	o	127	⌂

Другие варианты кодировок

- Sanger = Phred + 33
- Solexa = Solexa + 64
- Illumina 1.3-1.5 = Phred + 64

Архивы

- GZ сжимает fastq в 5-10 раз (bgzip)
- простой, универсальный, есть на всех Unix-like системах
- SRA - формат, разработанный в NCBI для одноименной базы данных (sequence read archive)
- декодируется командой из **NCBI SRA toolkit:** сразу выдаст парноконцевые прочтения двумя разными файлами

Bash

```
fastq-dump --split-3 file.sra
```

Индексы и размеры хромосом

- файлы типа hg19.chrom.sizes
- индекс файла fasta

```bash
samtools faidx genome.fa
```

1	chr1	249250621	8	249250621	249250622
2	chr2	243199373	249250638	243199373	243199374
3	chr3	198022430	492450020	198022430	198022431
4	chr4	191154276	690472459	191154276	191154277
5	chr5	180915260	881626744	180915260	180915261
6	chr6	171115067	1062542013	171115067	171115068
7	chr7	159138663	1233657089	159138663	159138664

первая колонка — хромосома

вторая колонка — длина хромосом

третья колонка (8) — оффсет (сдвиг)

четвертая колонка — длина строки

пятая колонка — количество битов для кодирования

💡 После парно-концевого секвенирования некоторого образца на приборе Illumina, был получен единый файл формата *fastq*, содержащий 20000 строк. Какое количество фрагментов ДНК было отсеквенировано?

Ответ: 1 фрагмент==4 строки, но секвенирование paired-end, т.е. 1 фрагмент читают дважды. 2500

При нулевом качетве секвенатор заменяет нуклеотид некоторым символом (не из множества {A,T,C,G})