Day Two
Qiime

Nate

the plan…

use virtual machines and terminal to run QIIME analysis.

go back and forth between the online tutorial and our terminals.

The tutorial is at (google): QIIME 454 tutorial

the bookmarks of firefox on your virtual machine

or here

http://qiime.org/tutorials/tutorial.html

our broad goals:

- identify the bacteria in our samples to the species level (OTUs at 97% sequence identity)
- count the number of each of these OTUs in each sample
- compare the communities to each other.

Organization of Qiime Analysis

A good way to keep track of your analysis pipeline is to set up a directory system that looks like this.

- 00_rawseqs
- 01_mapping
- 02_splitlibraries
- 03_otus
- 04_rep_set
- 05_taxonomy
- 06_otu_table
- 07_aligned_seqs
- 08_phylogeny
- 09_beta_diversity
- 10_otu_network
- 11_taxa_summary
- 12_alpha_rarefaction
- 13_jknifed_bdiv
- 14_3d_biplot

for right now, we are going to have you follow the tutorial only

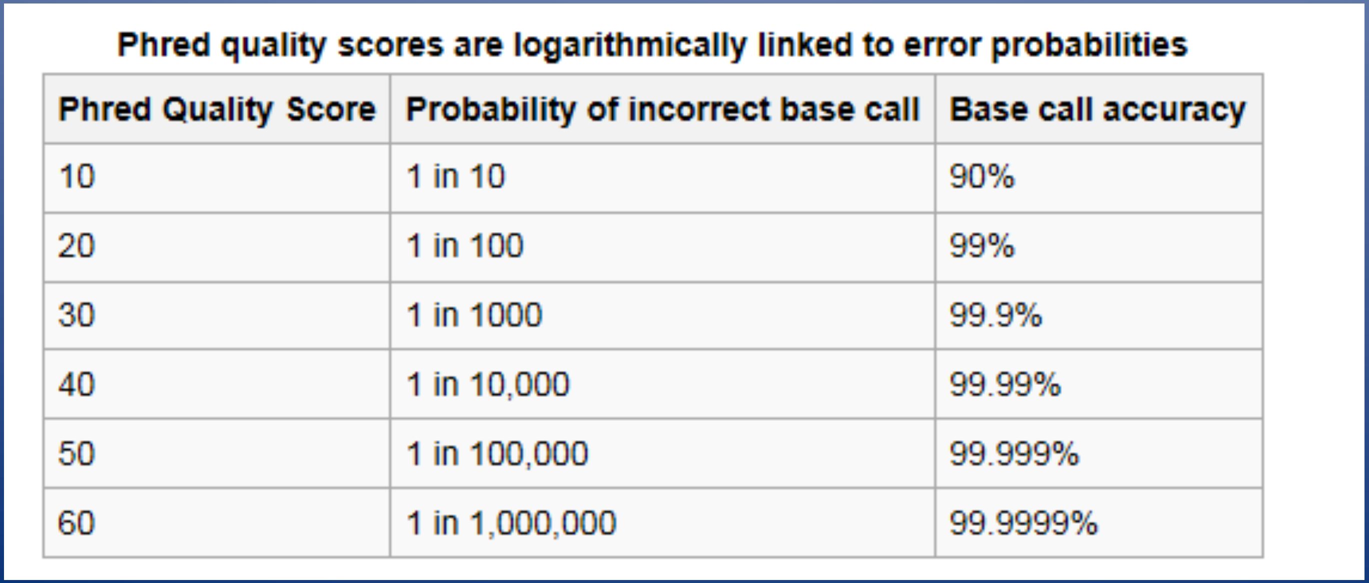
our data

```
$ cd qiime_overview_tutorial

$ head Fasting_Example.fna

$ head Fasting_Example.qual

$ head Fasting_Example.fastq
```

What does the .qual file show? how about the fastq?

PHRED scores in .qual files

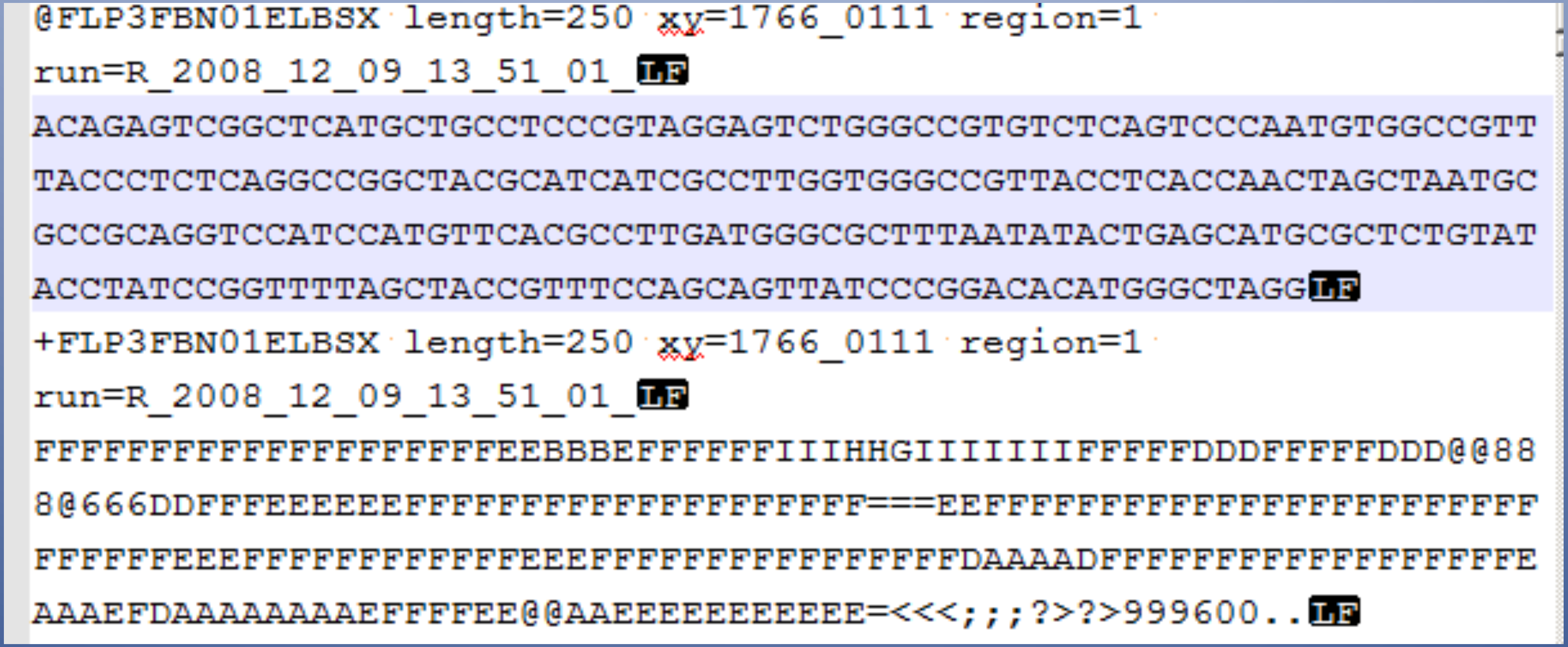
```
>FLP3FBN01ELBSX length=250 xy=1766_0111 region=1
run=R_2008_12_09_13_51_01_37 37 37 37 37 37 37 37 37 37 37 37 37 37 37 37
37 37 37 37 36 36 33 33 33 36 37 37 37 37 37 37 40 40 40 39 39 38 40 40 40 40 40
40 40 37 37 37 37 37 35 35 35 37 37 37 37 37 35 3535 31 31 23 23 23 31 21 21 21
35 35 37 37 37 36 36 36 36 36 36 37 37 37 37 37 37 37 37 37 37 37 37 37 37 37 37
37 37 37 37 28 28 28 36 36 37 37 37 37 37 37 37 37 37 37 37 37 37 3737 37 37 37
37 37 37 37 37 37 37 37 37 37 37 37 36 36 36 37 37 37 37 37 37 37 37 37 37 37 37
36 36 36 37 37 37 37 37 37 37 37 37 37 37 37 37 37 37 37 35 32 32 32 32 35 37 37
37 3737 37 37 37 37 37 37 37 37 37 37 37 37 37 36 32 32 32 36 37 35 32 32 32 32
32 32 32 32 36 37 37 37 37 36 36 31 31 32 32 36 36 36 36 36 36 36 36 36 36 36 28
27 27 27 26 26 26 30 2930 29 24 24 24 21 15 15 13 13
```



Phred quality scores are logarithmically linked to error probabilities

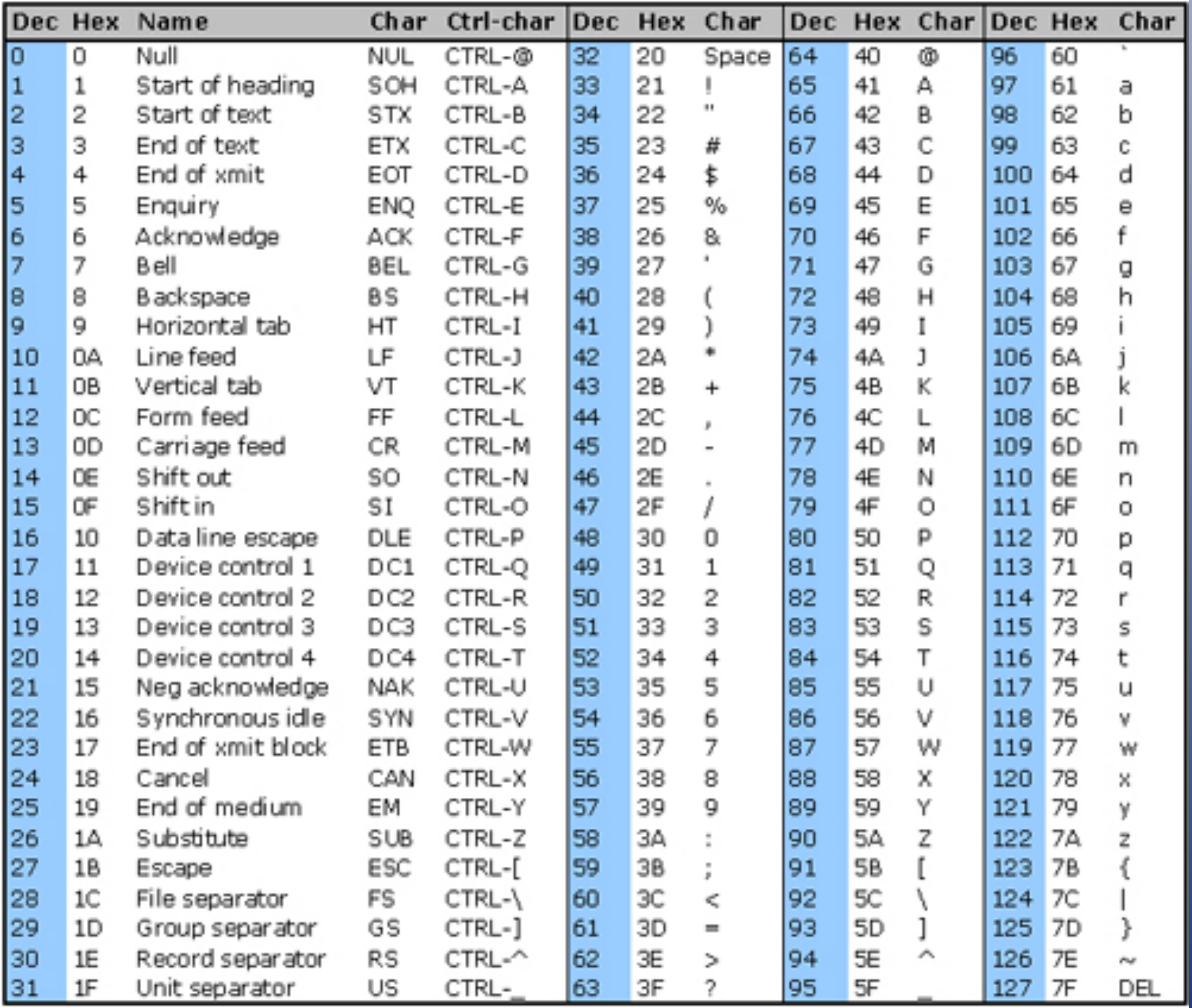
Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

PHRED scores in .fastq



```
@FLP3FBN01ELBSX length=250 xy=1766_0111 region=1
run=R_2008_12_09_13_51_01_LF
ACAGAGTCGGCTCATGCTGCCTCCCGTAGGAGTCTGGGCCGTGTCTCAGTCCCAATGTGGCCGTT
TACCCTCTCAGGCCGGCTACGCATCATCGCCTTGGTGGGCCGTTACCTCACCAACTAGCTAATGC
GCCGCAGGTCCATCCATGTTCACGCCTTGATGGGCGCTTTAATATACTGAGCATGCGCTCTGTAT
ACCTATCCGGTTTTAGCTACCGTTTCCAGCAGTTATCCCGGACACATGGGCTAGGLF
+FLP3FBN01ELBSX length=250 xy=1766_0111 region=1
run=R_2008_12_09_13_51_01_LF
FFFFFFFFFFFFFFFFFFFFEEBBBEFFFFFFIIIHHGIIIIIIIFFFFFDDDFFFFFDDD@@88
8@666DDFFFEEEEEEFFFFFFFFFFFFFFFFFFFF===EEFFFFFFFFFFFFFFFFFFFFFFFF
FFFFFFEEEFFFFFFFFFFFFEEEFFFFFFFFFFFFFFFFDAAAADFFFFFFFFFFFFFFFFFFE
AAAEFDAAAAAAAAEFFFFEE@@AAEEEEEEEEEEE=<<<;;;?>?>999600..LF
```

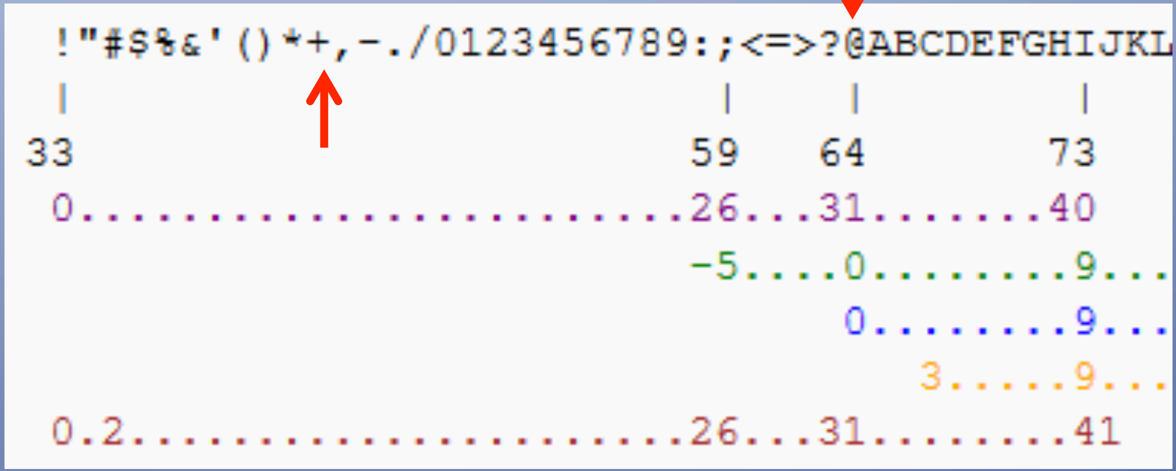
ASCII characters are used to indicate quality scores



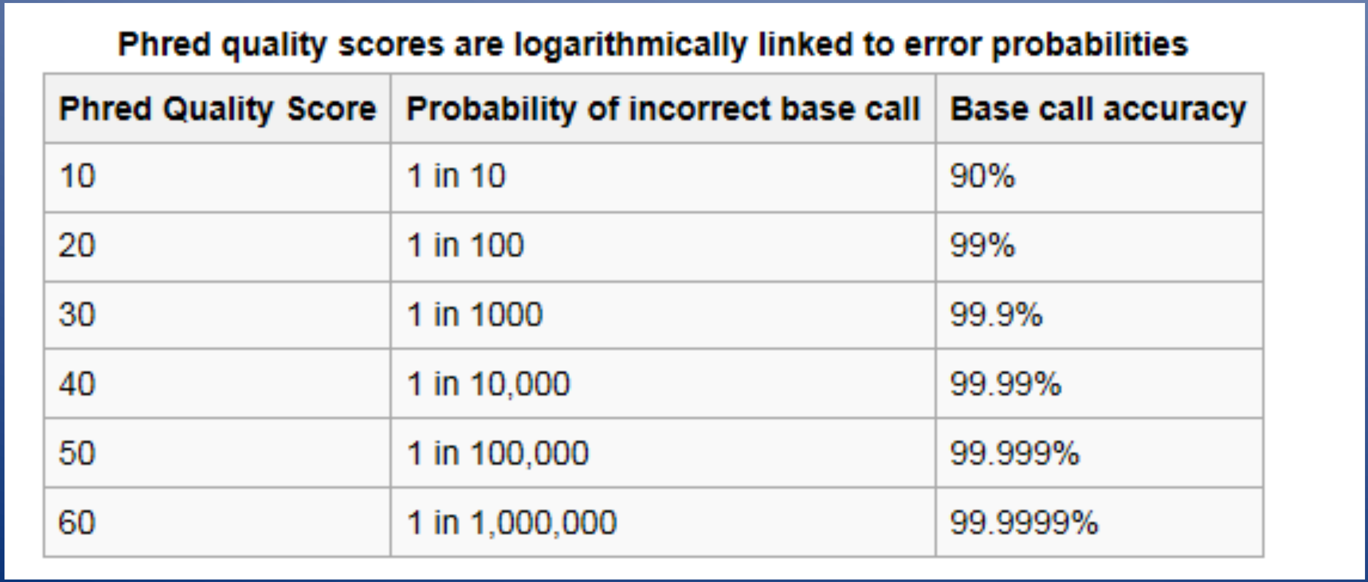
Dec	Hex	Name	Char	Ctrl-char	Dec	Hex	Char	Dec	Hex	Char	Dec	Hex	Char
0	0	Null	NUL	CTRL-@	32	20	Space	64	40	@	96	60	`
1	1	Start of heading	SOH	CTRL-A	33	21	!	65	41	A	97	61	a
2	2	Start of text	STX	CTRL-B	34	22	"	66	42	B	98	62	b
3	3	End of text	ETX	CTRL-C	35	23	#	67	43	C	99	63	c
4	4	End of xmit	EOT	CTRL-D	36	24	$	68	44	D	100	64	d
5	5	Enquiry	ENQ	CTRL-E	37	25	%	69	45	E	101	65	e
6	6	Acknowledge	ACK	CTRL-F	38	26	&	70	46	F	102	66	f
7	7	Bell	BEL	CTRL-G	39	27	'	71	47	G	103	67	g
8	8	Backspace	BS	CTRL-H	40	28	(	72	48	H	104	68	h
9	9	Horizontal tab	HT	CTRL-I	41	29	)	73	49	I	105	69	i
10	0A	Line feed	LF	CTRL-J	42	2A	*	74	4A	J	106	6A	j
11	0B	Vertical tab	VT	CTRL-K	43	2B	+	75	4B	K	107	6B	k
12	0C	Form feed	FF	CTRL-L	44	2C	,	76	4C	L	108	6C	l
13	0D	Carriage feed	CR	CTRL-M	45	2D	-	77	4D	M	109	6D	m
14	0E	Shift out	SO	CTRL-N	46	2E	.	78	4E	N	110	6E	n
15	0F	Shift in	SI	CTRL-O	47	2F	/	79	4F	O	111	6F	o
16	10	Data line escape	DLE	CTRL-P	48	30	0	80	50	P	112	70	p
17	11	Device control 1	DC1	CTRL-Q	49	31	1	81	51	Q	113	71	q
18	12	Device control 2	DC2	CTRL-R	50	32	2	82	52	R	114	72	r
19	13	Device control 3	DC3	CTRL-S	51	33	3	83	53	S	115	73	s
20	14	Device control 4	DC4	CTRL-T	52	34	4	84	54	T	116	74	t
21	15	Neg acknowledge	NAK	CTRL-U	53	35	5	85	55	U	117	75	u
22	16	Synchronous idle	SYN	CTRL-V	54	36	6	86	56	V	118	76	v
23	17	End of xmit block	ETB	CTRL-W	55	37	7	87	57	W	119	77	w
24	18	Cancel	CAN	CTRL-X	56	38	8	88	58	X	120	78	x
25	19	End of medium	EM	CTRL-Y	57	39	9	89	59	Y	121	79	y
26	1A	Substitute	SUB	CTRL-Z	58	3A	:	90	5A	Z	122	7A	z
27	1B	Escape	ESC	CTRL-[	59	3B	;	91	5B	[	123	7B	{
28	1C	File separator	FS	CTRL-\	60	3C	<	92	5C	\	124	7C	|
29	1D	Group separator	GS	CTRL-]	61	3D	=	93	5D	]	125	7D	}
30	1E	Record separator	RS	CTRL-^	62	3E	>	94	5E	^	126	7E	~
31	1F	Unit separator	US	CTRL-_	63	3F	?	95	5F	_	127	7F	DEL

illumina (and everyone else) uses +33 to indicate PHRED scores

PHRED (+33) scores in .fastq files



```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKL
|                          |    |         |
33                         59   64        73
 0.........................26...31........40
                          -5....0.........9...
                                0.........9...
                                   3......9...
 0.2.......................26...31........41
```



Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of incorrect base call	Base call 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%
60	1 in 1,000,000	99.9999%

does anyone see a problem with this?

Examining PHRED scores

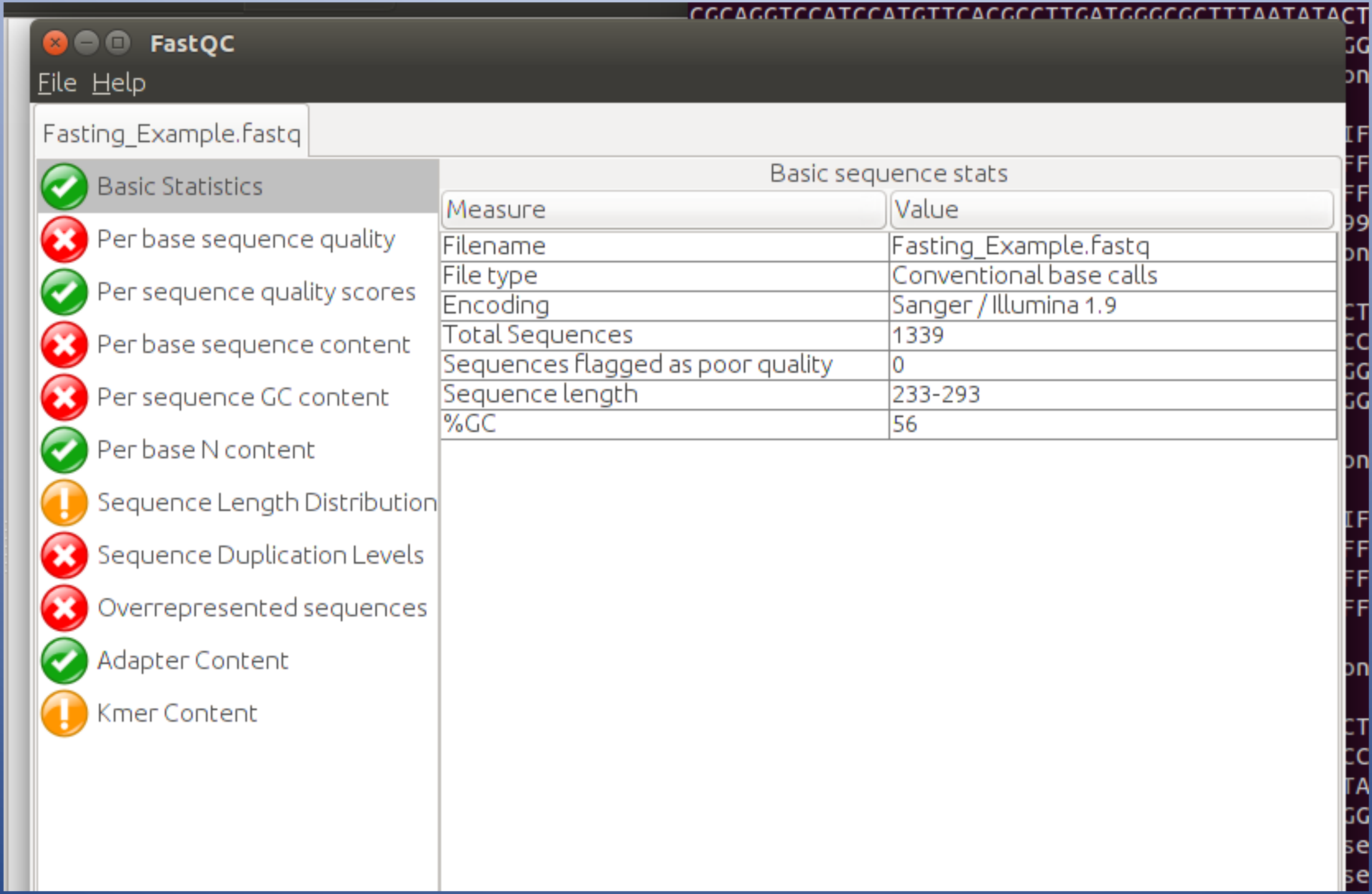
in your terminals:

```
$ fastqc
```

wait for graphics sofware to open and then navigate to the file called

Fasting_Example.fastq

select.



- ✓ Basic Statistics
- ✗ Per base sequence quality
- ✓ Per sequence quality scores
- ✗ Per base sequence content
- ✗ Per sequence GC content
- ✓ Per base N content
- ! Sequence Length Distribution
- ✗ Sequence Duplication Levels
- ✗ Overrepresented sequences
- ✓ Adapter Content
- ! Kmer Content

Basic sequence stats

Measure	Value
Filename	Fasting_Example.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1339
Sequences flagged as poor quality	0
Sequence length	233-293
%GC	56

take a few minutes to look at them and ask me questions about these options!

The tutorial is at (google): QIIME 454 tutorial

the bookmarks of firefox on your virtual machine

or here

http://qiime.org/tutorials/tutorial.html

Steps to do before tutorial:

```
$ pip install numpy
$ pip install biom-format
```

Tutorial:

- examine each step
- note how many individual scripts are being run when we enter the simple command pick_de_novo_otus.py
- run (almost) each step
- examine the output using terminal or the graphic interface
- think about how you would do this if we were using the organizational structure from the first slide.