Session 1 – Lecture 2

Working with Short-Read Data Files

Qi Sun

Computational Biology Service Unit
Cornell University

OUTLINE

1. What are in the data files?

2. Sequencing errors in both platforms.

3. Tools for inspect qualities and pre-processing files.

4. Computational resources available at Cornell.

Files distributed after Roche 454 run

1. SFF file (Standard flowgram format)

Binary file. Can only be opened by SFF compatible software, eg. gsMapper by Roche. It can be used for depository into NCBI SRA database.

2. FASTA file

Text file, with sequence reads.

```
>FVJHPP401DVJEI rank=0000010 x=1472.0 y=1128.0 length=245
GTTCGTGAGGAAATCTAGGCATGTCAGAAAACAATGATTACTTTTCCACCAAAGACTACT
ATTGTATACATTAAGCAAGAAAATACCTGCATATCCTGCCTGTTTAAATGAGAGCTGCAA
CAAAAAGTTAGACAGATATGGGAGATGGCAGTTGGCGCTGCGAAAAGTGTGATGCGAATC
ATCCAAGACCCGAATATCGCTATATTCTATCTCTTAACGTAAATGATCATACTGGTCAAC
TTTGG
>FVJHPP401A56GX rank=0000013 x=363.0 y=1199.0 length=46
GCTGCGACCTGCATCGGCGTCATTCTGCTGGTCATGCTGGTGGAGG
>FVJHPP401C6VFC rank=0000021 x=1191.0 y=1286.0 length=87
TATTCCACTAAACTGCGAAGATAGTAGCCTGACGAGTACTTATTAATCCCTTAACTAGAG
GCCTATTGCGCGTGCGTAACATAAGAG
```

3. QUAL file

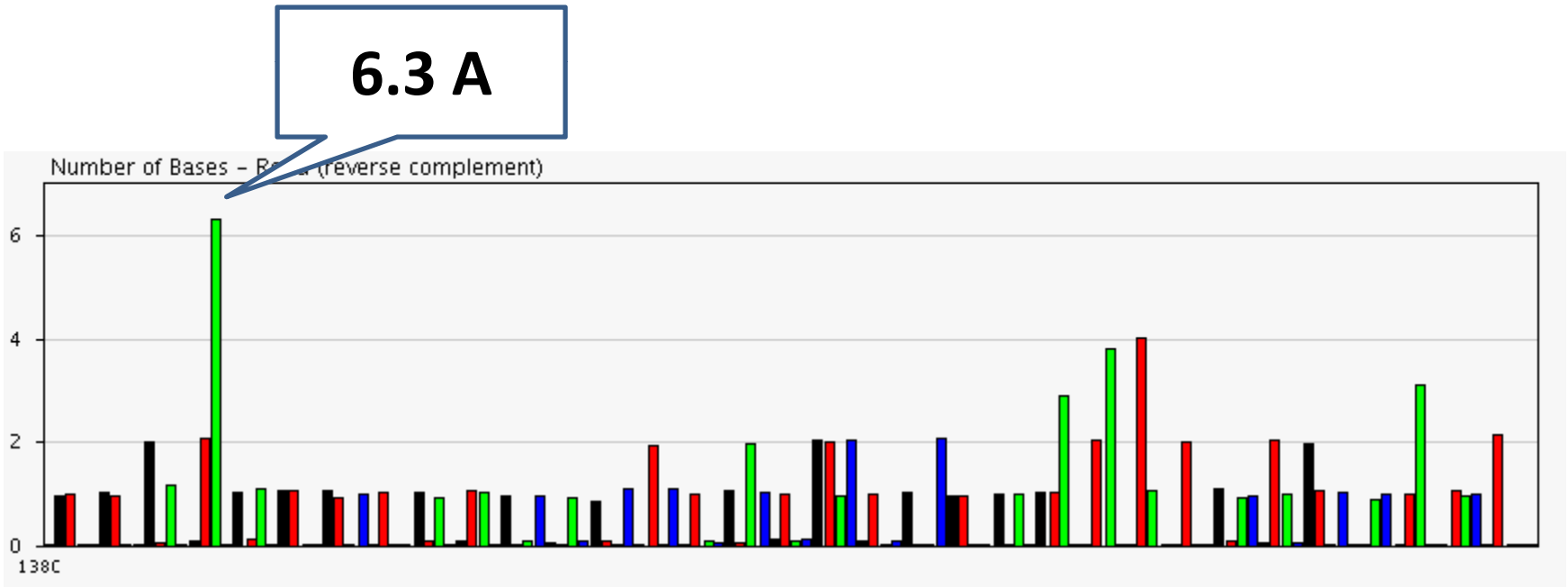
Text file, with PHRED based quality.

```
>FVJHPP401DVJEI rank=0000010 x=1472.0 y=1128.0 length=245
32 32 32 32 32 32 32 32 24 24 19 19 19 25 28 32 32 32 32 33 33 33 32 30 32 30 25
 18 18 18 18 18 14 18 27 27 27 27 27 21 23 13 13 13 13 15 15 17 15 20 15 15 15 1
7 11 23 23 28 28 28
28 28 28 24 23 24 27 28 29 30 30 30 30 27 24 24 24 28 28 30 30 30 30 30 22 31 28
 23 23 23 30 30 30 30 30 30 27 27 27 27 27 27 29 31 31 29 23 17 14 17 17 27 28 2
5 24 24 24 24 24 19
13 13 13 13 13 13 13 13 26 9  18 18 23 25 28 30 30 30 28 29 26 22 22 19 21 22 29
 29 26 26 26 25 25 28 28 28 22 24 24 24 21 21 15 15 13 13 18 19 23 23 23 23 21 2
1 25 25 26 26 26 25
27 27 21 21 21 21 21 17 13 12 12 12 12 12 21 16 23 24 21 21 21 24 26 28 21 21 21
 26 27 27 27 27 27 26 24 24 23 23 23 23 19 19 19 17 17 11 11 11 17 17 20 20 20 1
1 11 11 11 11 11 11
11 11 11 17 17
>FVJHPP401A56GX rank=0000013 x=363.0 y=1199.0 length=46
```

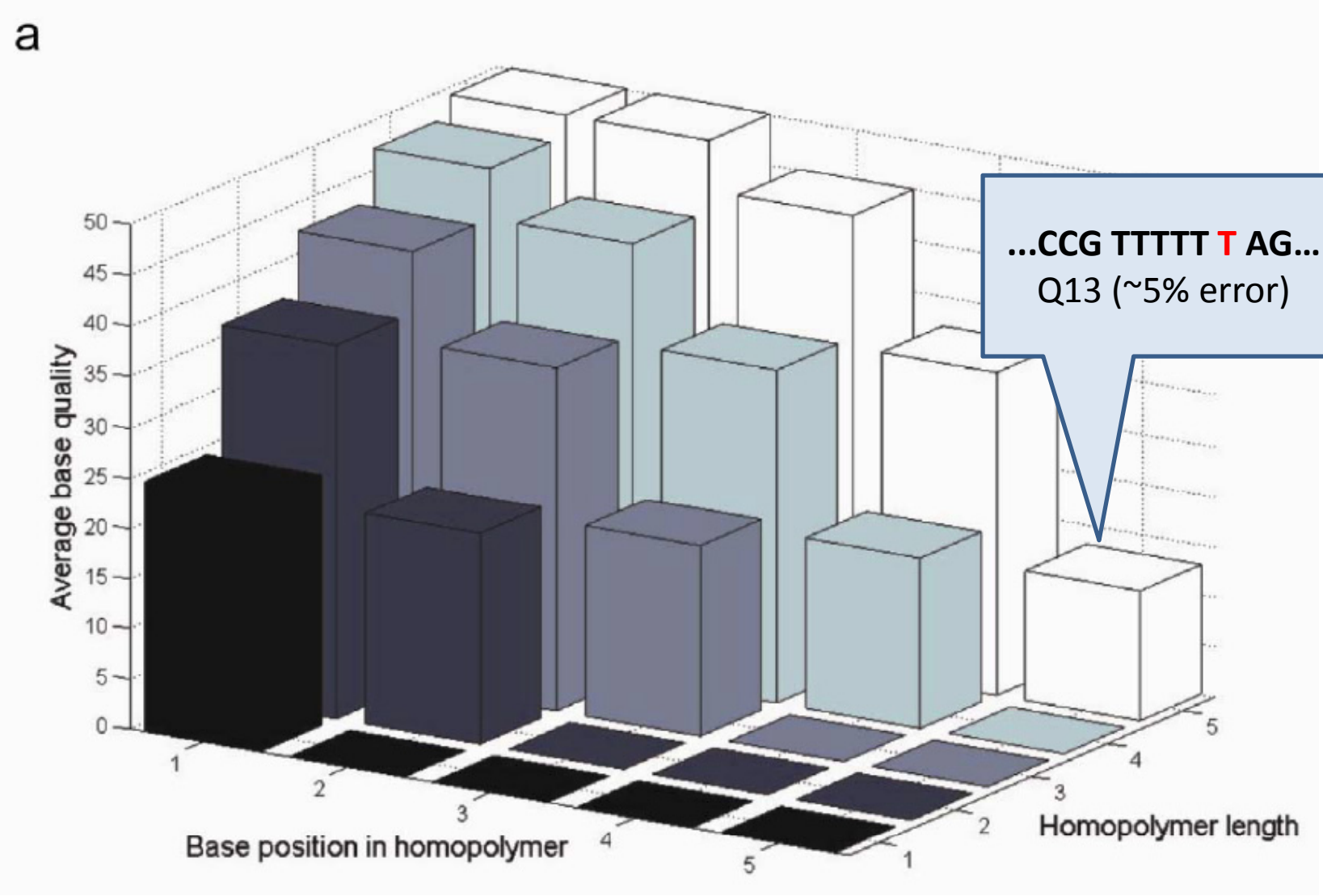
Homopolymer errors in 454 base calling results

Sequence Read: … GTGTGGATT AAAAAA GAGT …

Flowgram:



Distribution of base quality scores within homopolymer runs



Quinlan AR, Marth GT et al. (2008) *Nature Methods* **5**:179 - 181

Alignment of 454 reads to the reference

Illumina sequencing data are distributed as FASTQ files

Each sequence read is represented by 4 lines

```
@HWUSI-EAS690_0001:8:3:274:1957#0/1
CAGTAGCAATCCAAACTTTGTTACTCGTCAGAAAATCGAAATCATCTTCGGTTAAATCCAAAA
+HWUSI-EAS690_0001:8:3:274:1957#0/1
aabaaaaaaaaaa_Y_aaabaaaaaaaaa`a`[a`a```[``[``^`\\^\\_aZOY^^\_BB
```

Line1. Read Identifier 1

Line 2. Sequence

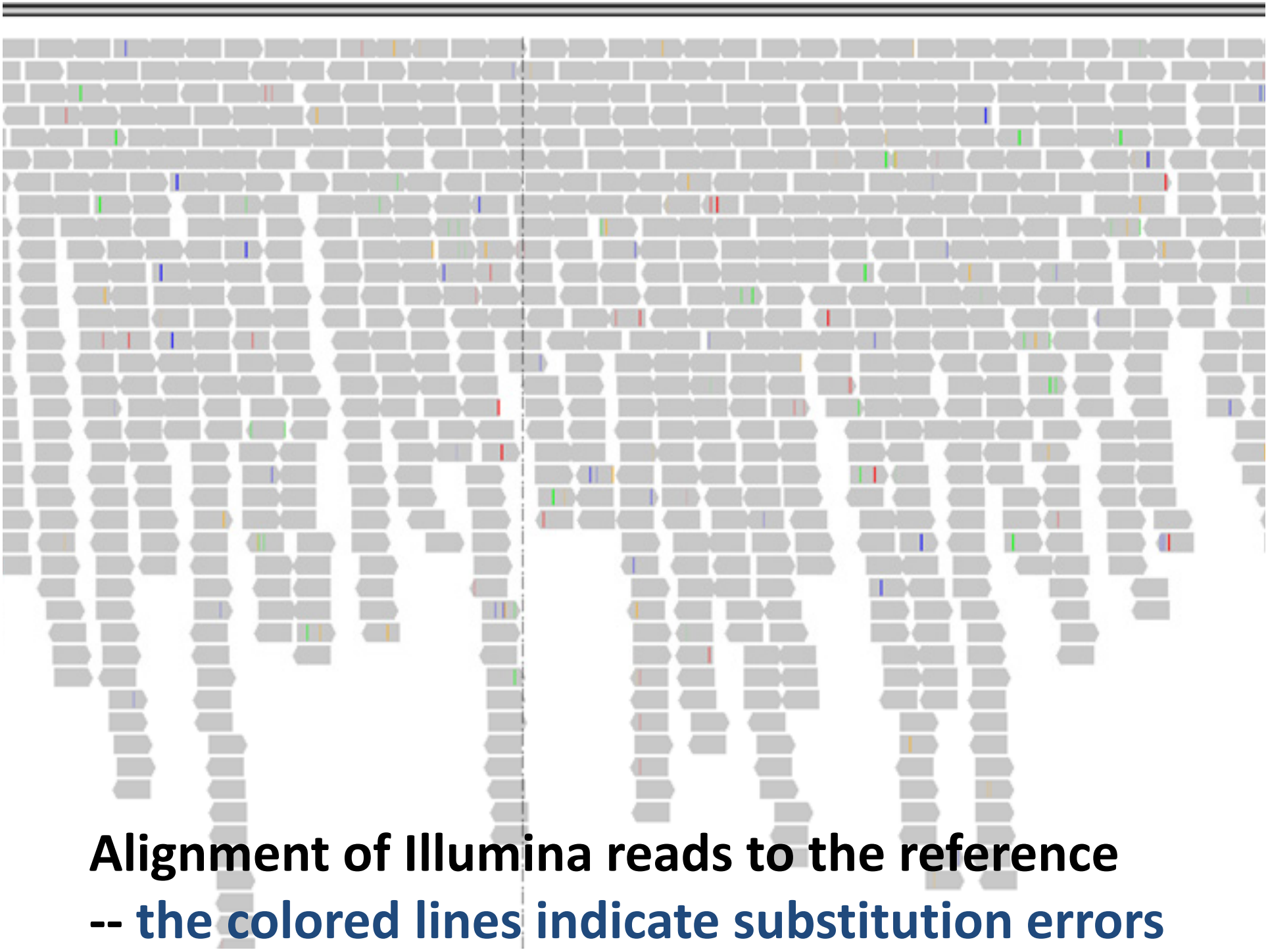
Line 3. Read Identifier 2

Line 4. Quality code

Illumian Quality Code (post v 1.3)

$$Q = - \log_{10} P$$

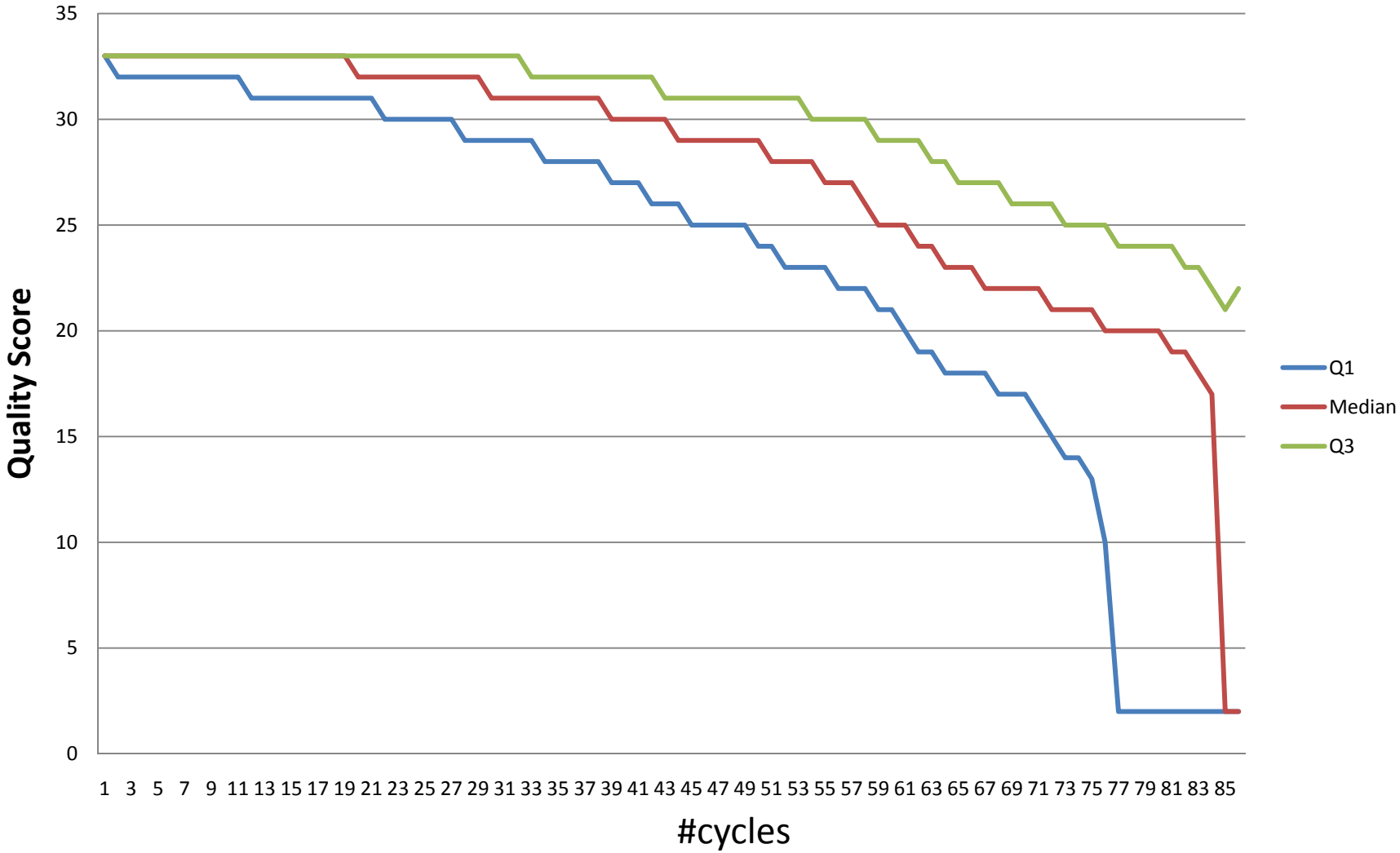
CODE	Quality	P	CODE	Quality	P
@	0	1.0000	T	20	0.0100
A	1	0.7943	U	21	0.0079
B	2	0.6310	V	22	0.0063
C	3	0.5012	W	23	0.0050
D	4	0.3981	X	24	0.0040
E	5	0.3162	Y	25	0.0032
F	6	0.2512	Z	26	0.0025
G	7	0.1995	[	27	0.0020
H	8	0.1585	\	28	0.0016
I	9	0.1259	]	29	0.0013
J	10	0.1000	^	30	0.0010
K	11	0.0794	_	31	0.0008
L	12	0.0631	`	32	0.0006
M	13	0.0501	a	33	0.0005
N	14	0.0398	b	34	0.0004
O	15	0.0316	c	35	0.0003
P	16	0.0251	d	36	0.0003
Q	17	0.0200	e	37	0.0002
R	18	0.0158	f	38	0.0002
S	19	0.0126	g	39	0.0001



Alignment of Illumina reads to the reference

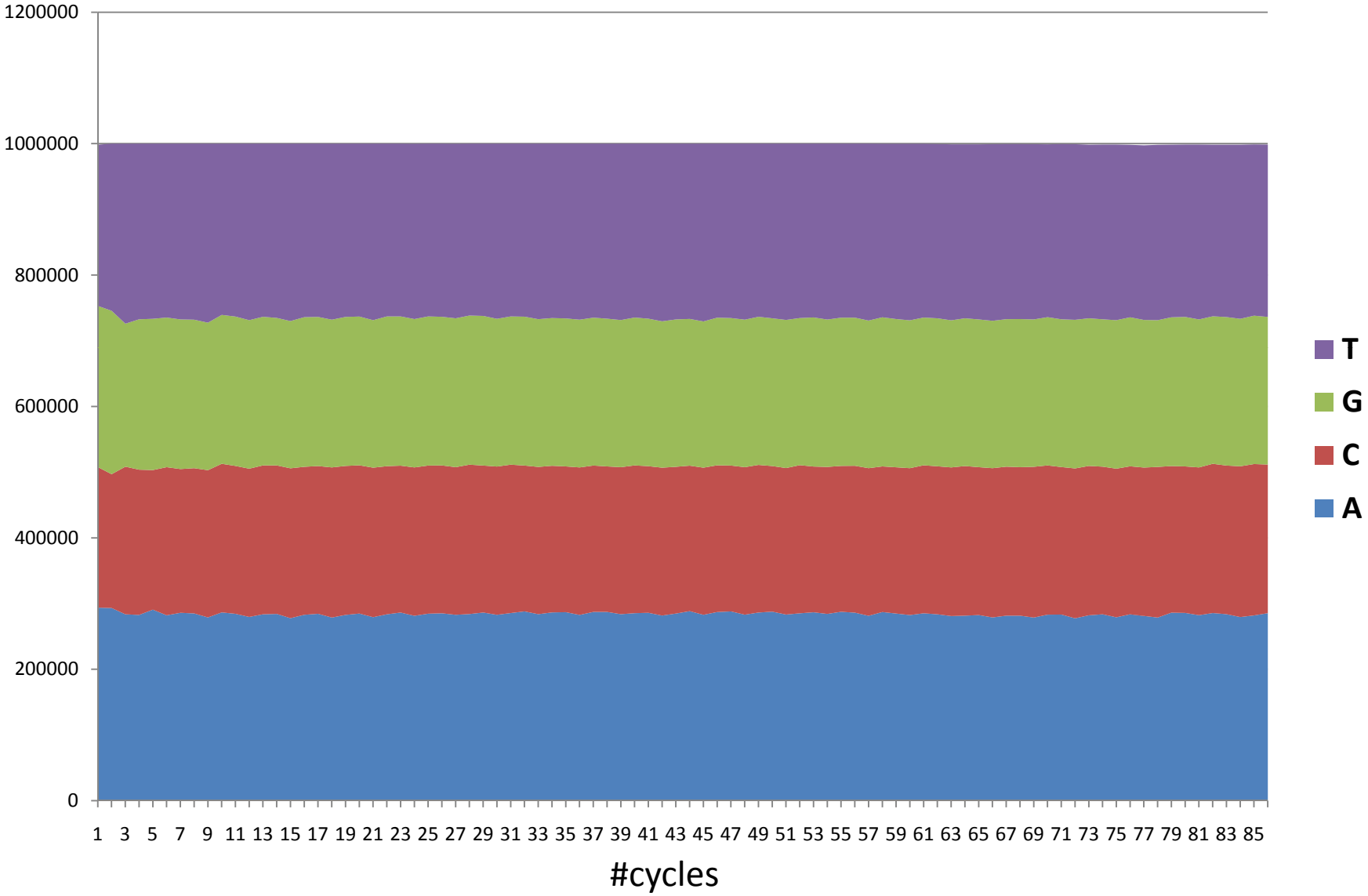
-- the colored lines indicate substitution errors

Quality Score Distribution of Illumina Sequencing Cycles



Q1, Median, Q3 refer to 25%, 50% or 75% of reads with Quality below this level

Nucleotides Distribution of Illumina Sequencing Cycles

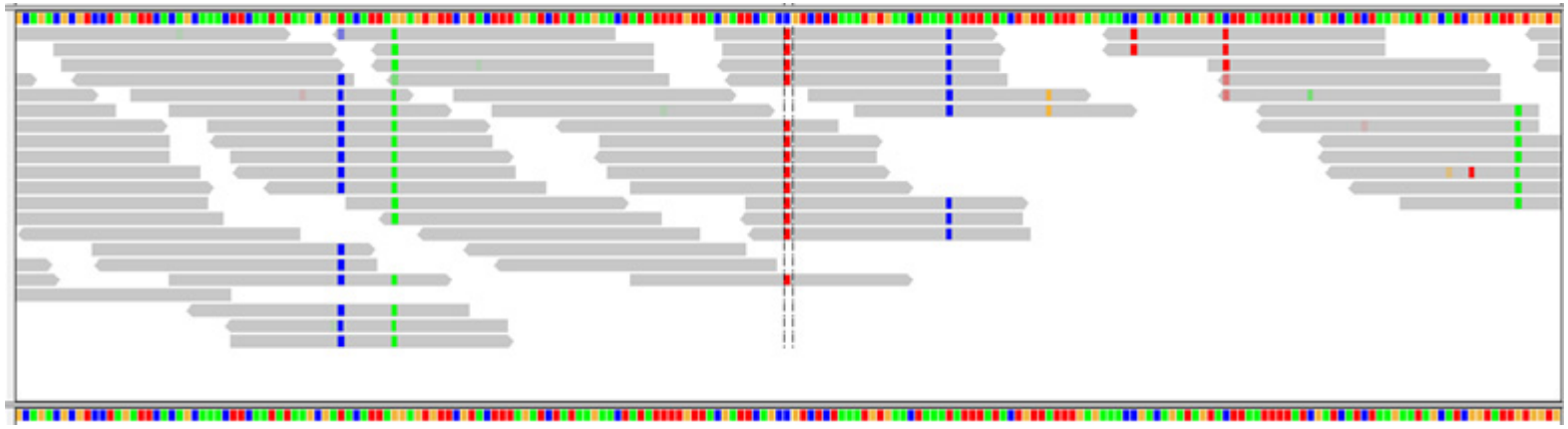


When should sequencing errors be a concern?

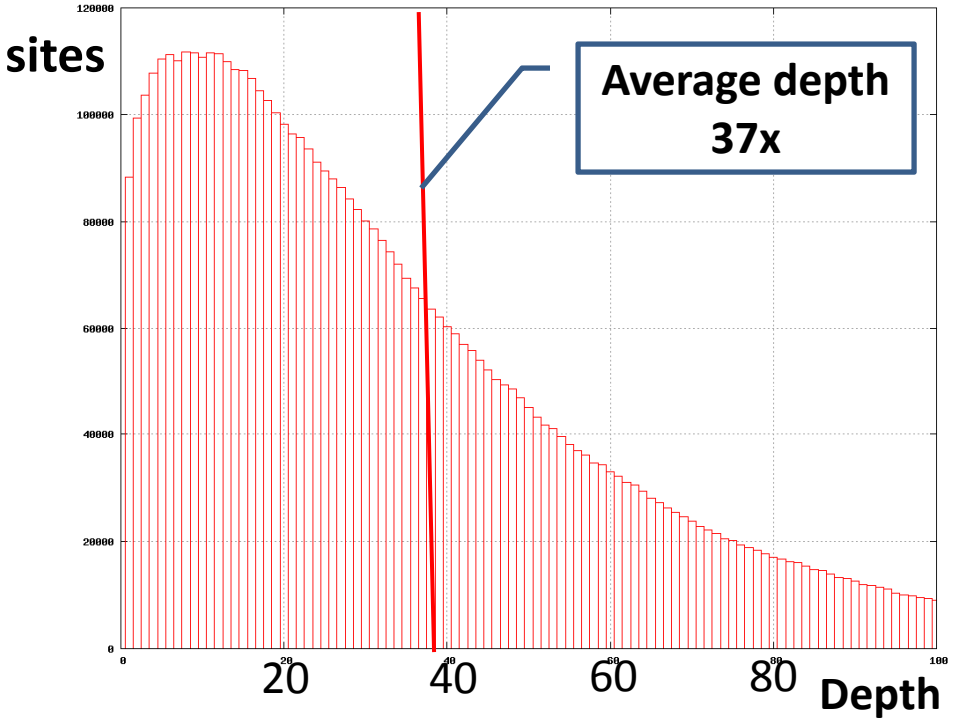
1.For RNAseq/ChIPSeq data, sequencing errors are not a major concern.

A limited number of mismatches can be tolerate. However, for some data files, trimming of low quality data are recommended.

2. For SNP detection, especially when coverage depth is low, the quality score can be used, but not always reliable.



The coverage depth are not evenly distributed across the genome.

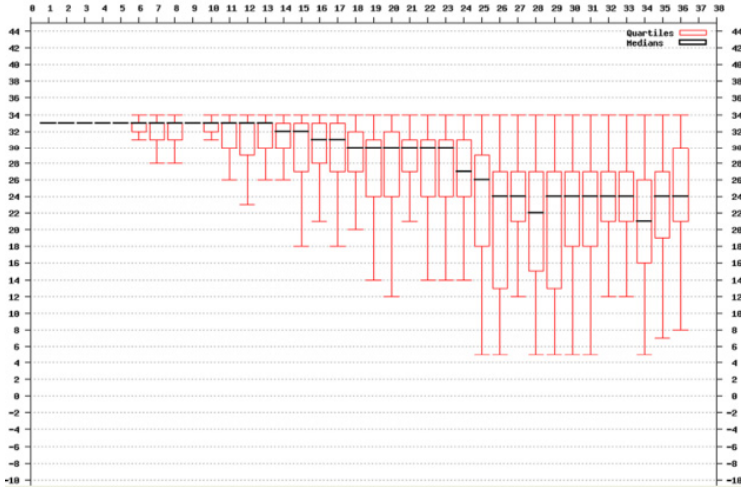


Coverage Depth	% base pairs
<=1	1.1 %
<=3	4.3 %
<=5	10%
<=10	20%
<=37	63%

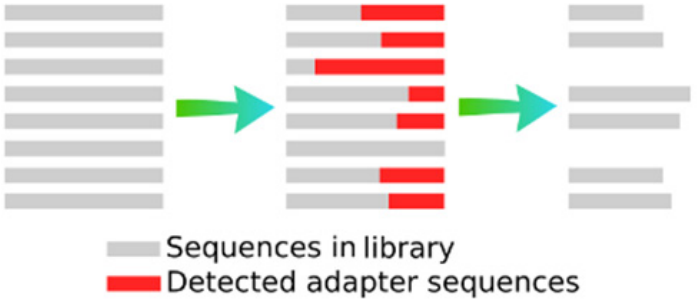
Distribution of Coverage depth in a seq-cap experiment.

FASTX ToolKit – a package for pre-processing Illumina FASTQ files and evaluating data quality (by Hannon Lab of CSHL)

Plot quality scores and trim low-quality data



Clip off the 3' adaptor



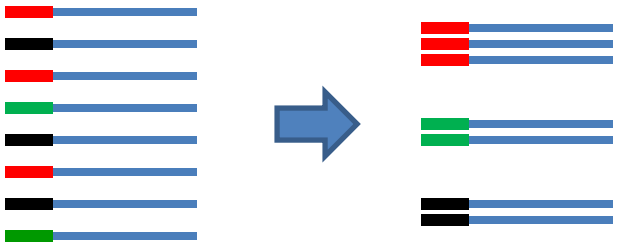
FASTX ToolKit – a package for pre-processing Illumina FASTQ files and evaluating data quality (by Hannon Lab of CSHL)

FASTQ -> FASTA

```
@HWUSI-EAS690_0001:8:3:274:1957#0/1
CAGTAGCAATCCAAACTTTGTTACTCGTCAGAAAATCGAAA
+HWUSI-EAS690_0001:8:3:274:1957#0/1
aabaaaaaaaaaa_Y_aaabaaaaaaaaa`a`[a`a```[`
```

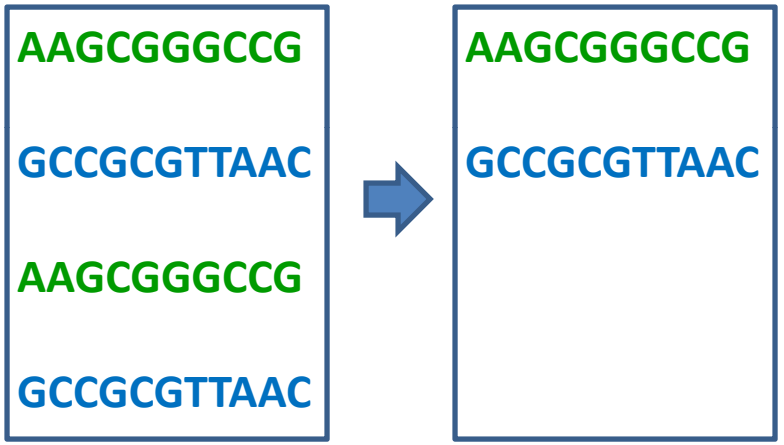
```
>1
CAGTAGCAATCCAAACTTTGTTACTCGTCAGAAAATCGAAA
>2
GATTTAACCGGCCGGAAAACCCTCGACATACGGATACCGAT
```

De-multiplexing



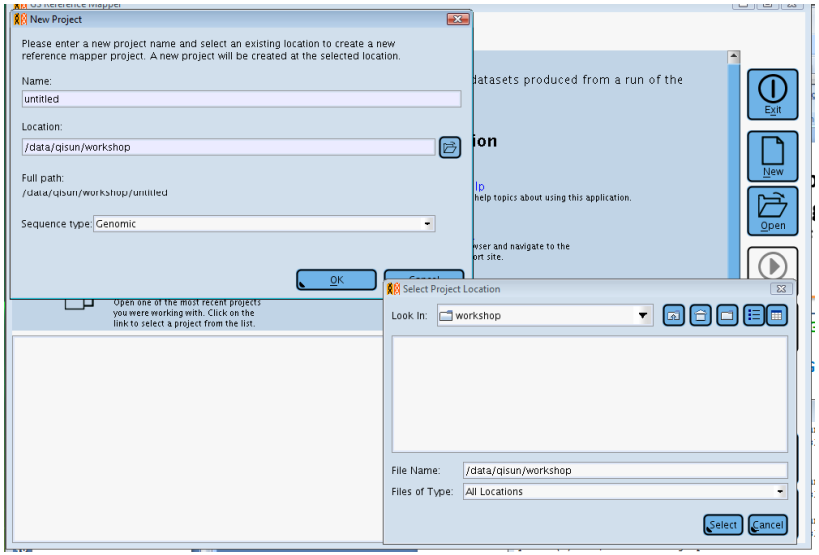
FASTX ToolKit – a package for pre-processing Illumina FASTQ files and evaluating data quality (by Hannon Lab of CSHL)

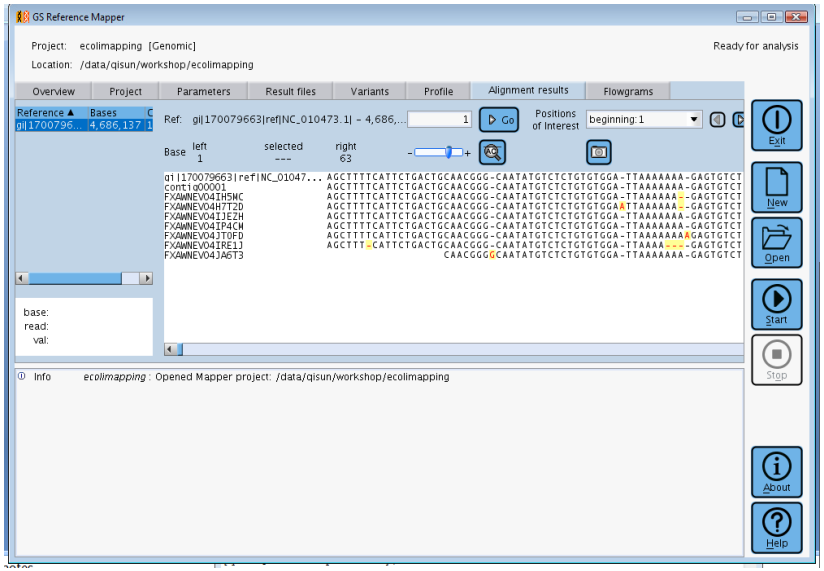
Remove redundancy



Web Interface

Roche 454 Software

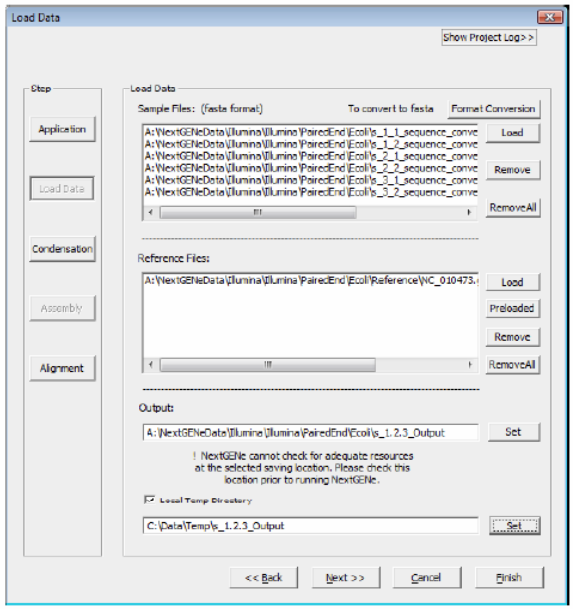


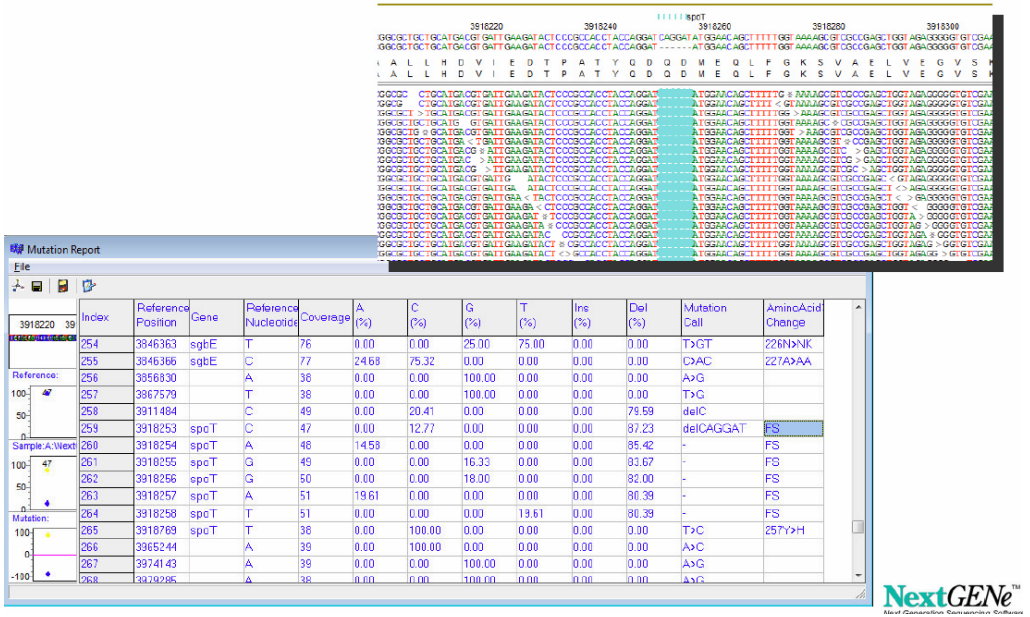


- Align reads to genome
- De novo assembly
- Amplicon diversity

Computational Resources at Cornell

1.Commercial Software workstation (624 Rhodes Hall)





SoftGenetics NextGENe
DNAStar NGen
Illumina Genome Studio

Computational Resources at Cornell

2.CAC Linux Cluster

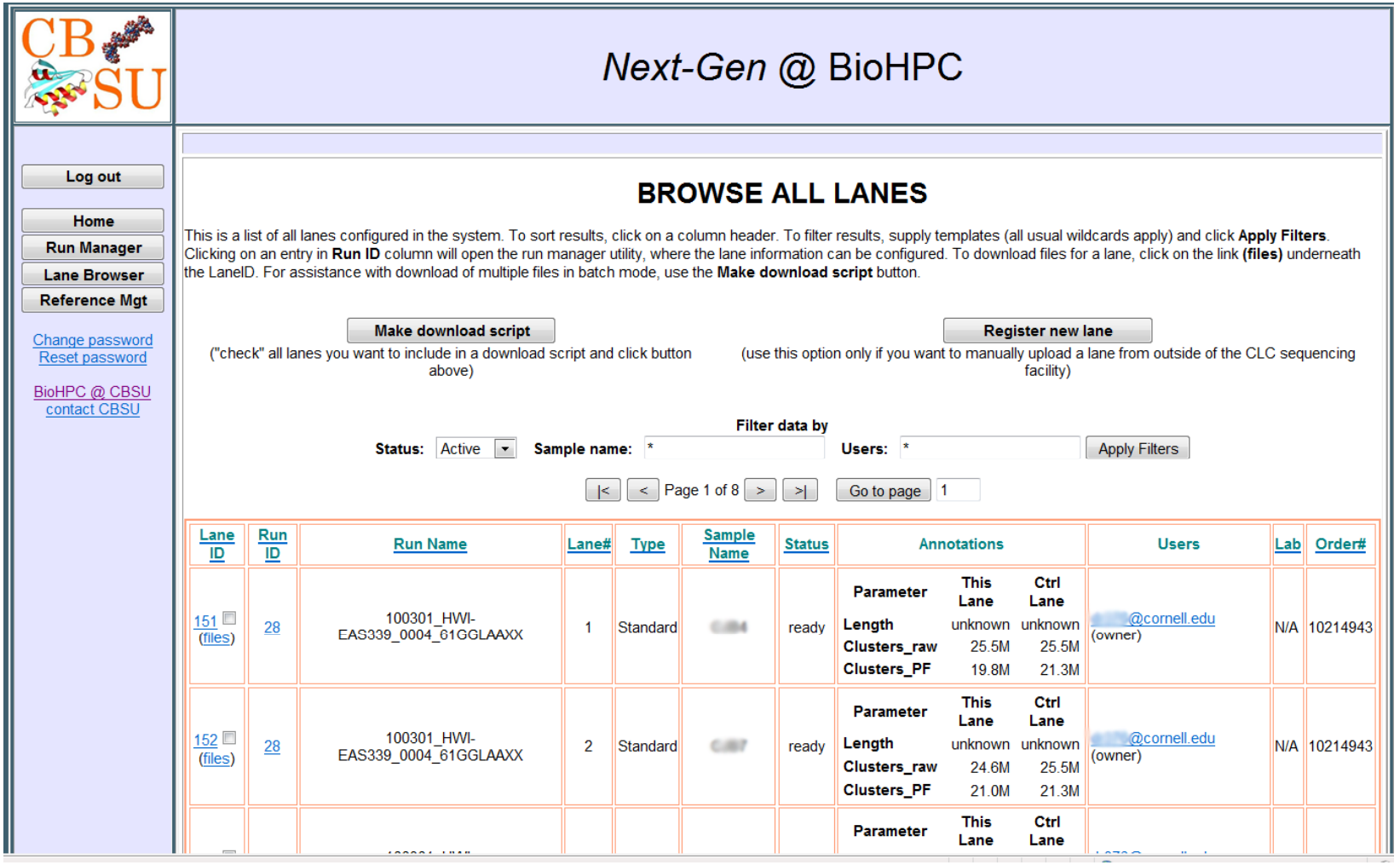
- **Linux**
- **Common open source software preinstalled**

- **V4 16GB RAM 19 servers**
- **V4-64G 64GB RAM 3 servers**

3. Research collaboration with CBSU

Next-Gen@BioHPC

QC **RNA-seq** **ChIP-seq** **SNP/Indel** **Assembly**



Next-Gen @ BioHPC

Log out

Home
Run Manager
Lane Browser
Reference Mgt

Change password
Reset password

BioHPC @ CBSU
contact CBSU

BROWSE ALL LANES

This is a list of all lanes configured in the system. To sort results, click on a column header. To filter results, supply templates (all usual wildcards apply) and click **Apply Filters**. Clicking on an entry in **Run ID** column will open the run manager utility, where the lane information can be configured. To download files for a lane, click on the link **(files)** underneath the LaneID. For assistance with download of multiple files in batch mode, use the **Make download script** button.

Make download script
("check" all lanes you want to include in a download script and click button above)

Register new lane
(use this option only if you want to manually upload a lane from outside of the CLC sequencing facility)

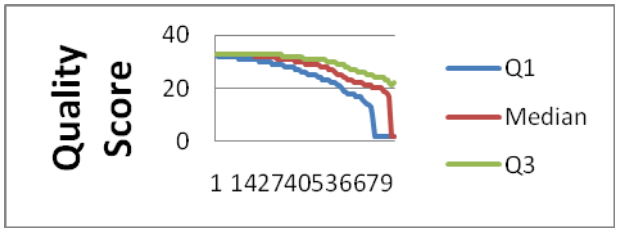
Filter data by
Status: Active Sample name: * Users: * Apply Filters

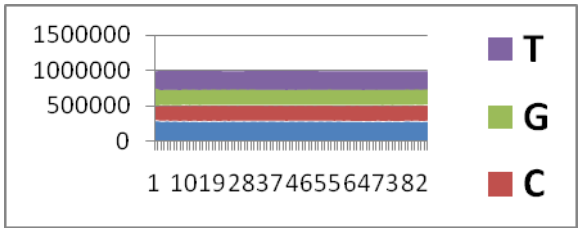
|< < Page 1 of 8 > >| Go to page 1

Lane ID	Run ID	Run Name	Lane#	Type	Sample Name	Status	Annotations	Users	Lab	Order#
151 (files)	28	100301_HWI-EAS339_0004_61GGLAAXX	1	Standard	[illegible]	ready	Parameter / This Lane / Ctrl Lane Length unknown unknown Clusters_raw 25.5M 25.5M Clusters_PF 19.8M 21.3M	[illegible]@cornell.edu (owner)	N/A	10214943
152 (files)	28	100301_HWI-EAS339_0004_61GGLAAXX	2	Standard	[illegible]	ready	Parameter / This Lane / Ctrl Lane Length unknown unknown Clusters_raw 24.6M 25.5M Clusters_PF 21.0M 21.3M	[illegible]@cornell.edu (owner)	N/A	10214943
							Parameter / This Lane / Ctrl Lane			

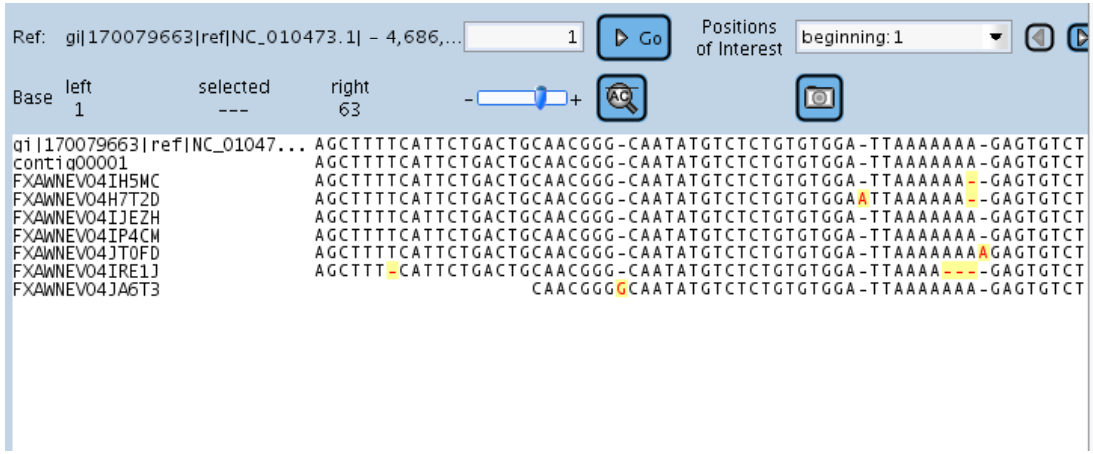
Exercises:

1.Using FASTX to get statistic report of Illimina sequencing data





2.Using Roch GSMapper to align 454 reads to E.coli genome.



1. **Download workshop slides and exercise instructions from the workshop web site. http://cbsu.tc.cornell.edu**

2. **Setup your desktop computer to work with Linux server.**

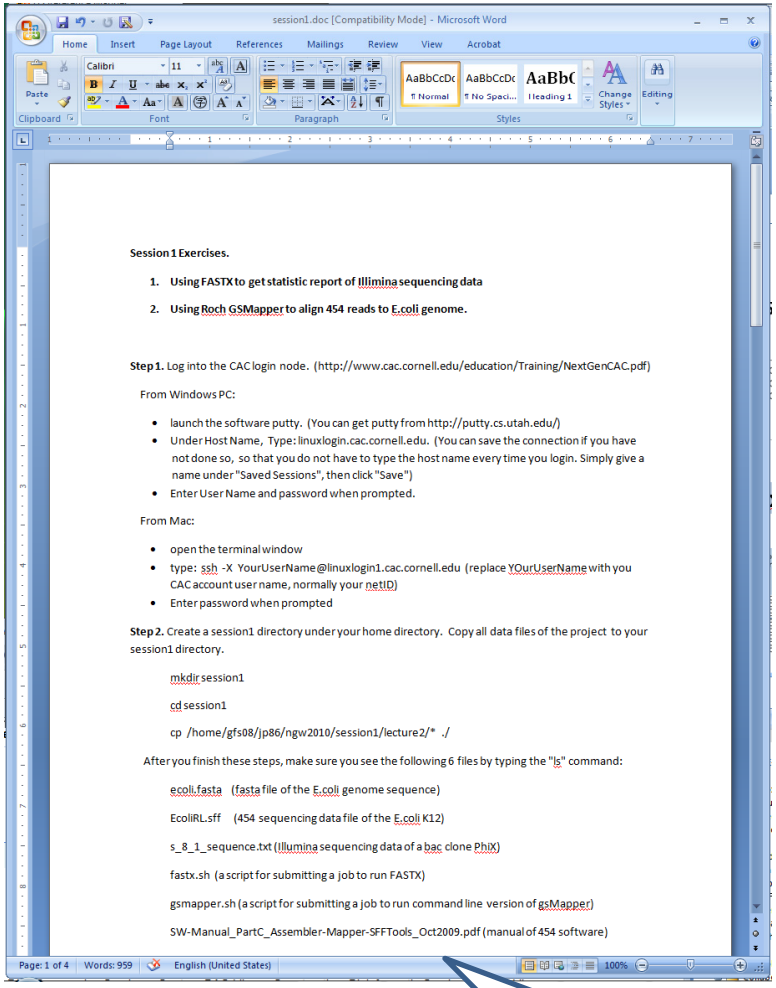
   <u>Window Users:</u> Install Putty, WinSCP, Xming, EditPlus

   <u>Mac users:</u> Install TextWrangler, FETCH

3. **Log into the linuxlogin.cac.cornell.edu**

4. **Modify the fastx.sh script as described.**

5. **Submit the job.**



session1.doc [Compatibility Mode] - Microsoft Word

Session 1 Exercises.

1. **Using FASTX to get statistic report of Illimina sequencing data**
2. **Using Roch GSMapper to align 454 reads to E.coli genome.**

Step 1. Log into the CAC login node. (http://www.cac.cornell.edu/education/Training/NextGenCAC.pdf)

From Windows PC:

- launch the software putty. (You can get putty from http://putty.cs.utah.edu/)
- Under Host Name, Type: linuxlogin.cac.cornell.edu. (You can save the connection if you have not done so, so that you do not have to type the host name every time you login. Simply give a name under "Saved Sessions", then click "Save")
- Enter User Name and password when prompted.

From Mac:

- open the terminal window
- type: ssh -X YourUserName@linuxlogin1.cac.cornell.edu (replace YOurUserName with you CAC account user name, normally your netID)
- Enter password when prompted

Step 2. Create a session1 directory under your home directory. Copy all data files of the project to your session1 directory.

mkdir session1

cd session1

cp /home/gfs08/jp86/ngw2010/session1/lecture2/* ./

After you finish these steps, make sure you see the following 6 files by typing the "ls" command:

ecoli.fasta (fasta file of the E.coli genome sequence)

EcoliRL.sff (454 sequencing data file of the E.coli K12)

s_8_1_sequence.txt (Illumina sequencing data of a bac clone PhiX)

fastx.sh (a script for submitting a job to run FASTX)

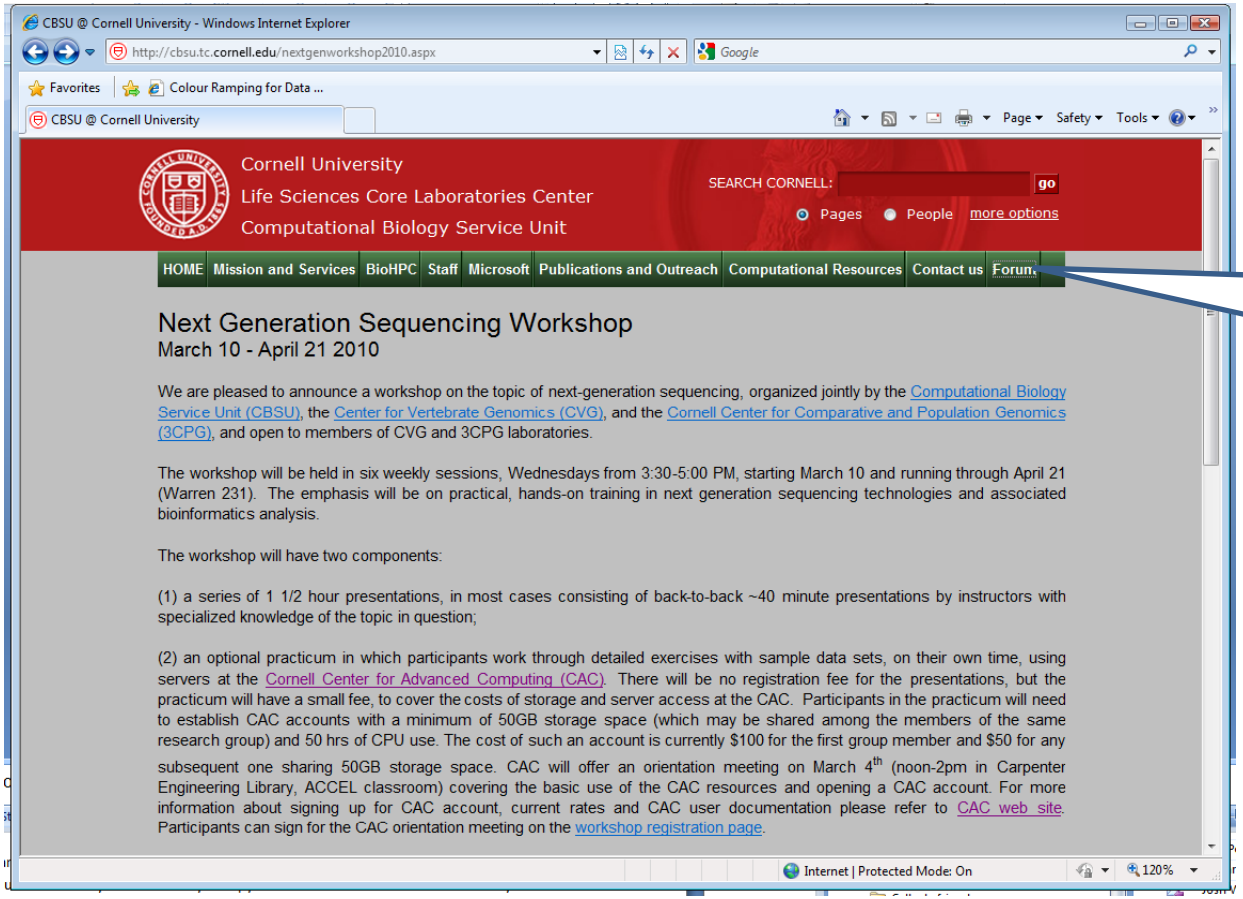
gsmapper.sh (a script for submitting a job to run command line version of gsMapper)

SW-Manual_PartC_Assembler-Mapper-SFFTools_Oct2009.pdf (manual of 454 software)

A few tips of using Linux system is available at the end of the document.

Forum

Please post your questions to the online forum. You need register to post questions.



Click to enter forum

Office hours:

Friday 3:00PM. 102 Weill (small conference room)