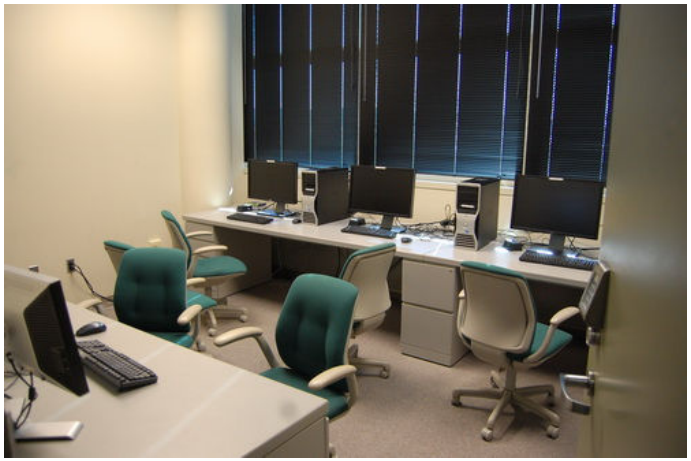
Using BioHPC Lab Software

Qi Sun
Computational Biology Service Unit
Cornell University

What is the BioHPC Lab

- **625 Rhodes Hall**
- **31 Linux remote workstations**
- **2 Large RAM workstations**

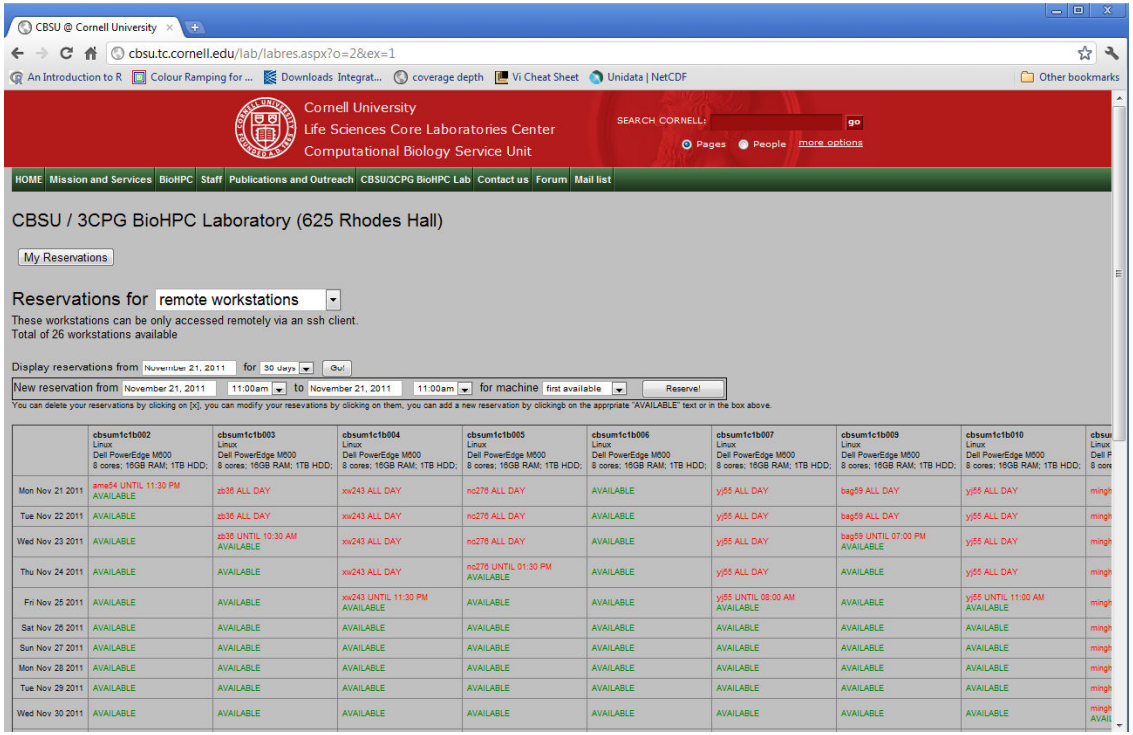


Open Hours: 24/7

Office Hours: 2-4PM, Mondays

Using BioHPC Lab:

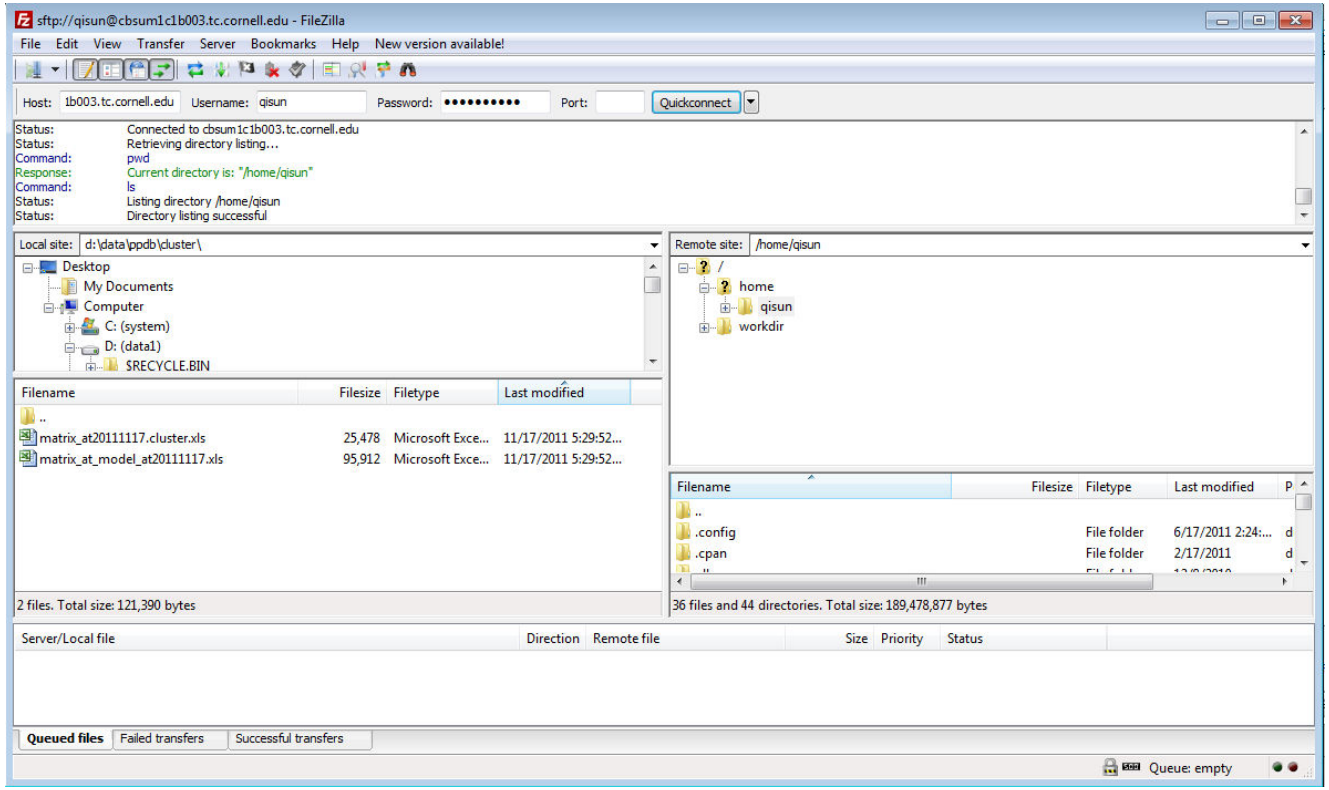
Step 1: Reserve a computer



http://cbsu.tc.cornell.edu

Using BioHPC Lab:

Step 2: Transfer files to the computer



Software: FileZilla (Win) or Fetch (Mac).

Host: machine_name.tc.cornell.edu

Port: 22 (sftp)

Using BioHPC Lab:

Step 3: run software

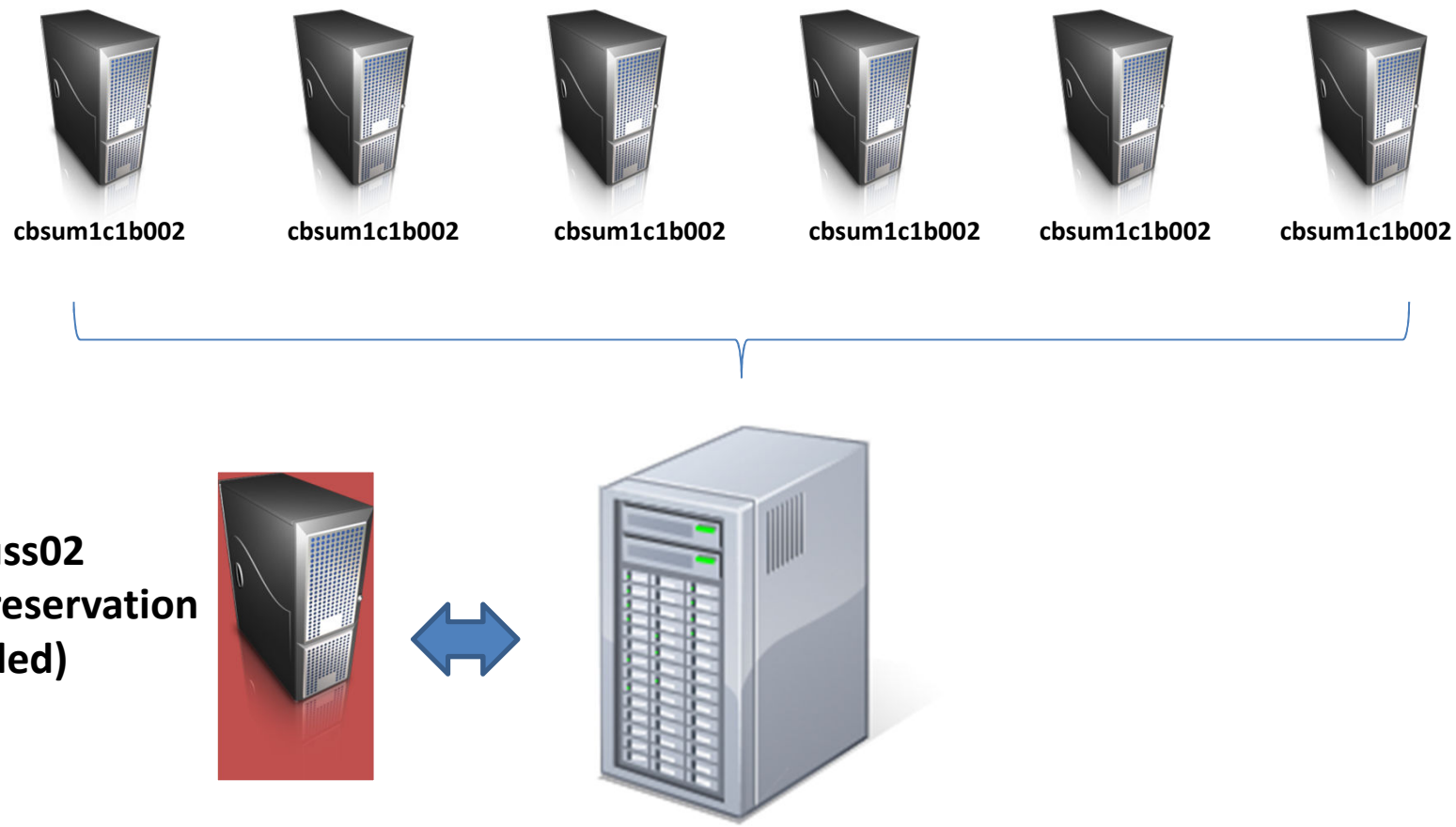
```
qisun@cbsum1c2b010:/workdir/qisun
[qisun@cbsum1c2b010 qisun]$ bwa

Program: bwa (alignment via Burrows-Wheeler transformation)
Version: 0.5.9-r16
Contact: Heng Li <lh3@sanger.ac.uk>

Usage:   bwa <command> [options]

Command: index         index sequences in the FASTA format
         aln           gapped/ungapped alignment
         samse         generate alignment (single ended)
         sampe         generate alignment (paired ended)
         bwasw         BWA-SW for long queries

         fa2pac        convert FASTA to PAC format
         pac2bwt       generate BWT from PAC
         pac2bwtgen    alternative algorithm for generating BWT
         bwtupdate     update .bwt to the new format
         pac_rev       generate reverse PAC
         bwt2sa        generate SA from BWT and Occ
         pac2cspac     convert PAC to color-space PAC
         stdsw         standard SW/NW alignment

[qisun@cbsum1c2b010 qisun]$
```

Windows: PUTTY **MAC: Terminal**

Command-line example:

bwa aln -t 7 maize s_1_sequence.txt.gz > s1.sai

Data storage in BioHPC Lab



Local Drive: /workdir/qs24 & /local_data

Network Drive: /home/qs24 & /shared_data

What software are available?

Alignment:

- BWA
- Tophat/Bowtie
- gsMapper
- BLAST
- BLAT
- ClustalW

Data analysis software available on BioHPC lab

- **RNA-seq**
- **ChIP-seq**
- **SNP genotyping**
- **Genotyping-by-sequencing**
- ***De novo* assembly (transcriptome and genome)**

What software are available?

Assembly:

- Velvet
- AllPaths
- gsAssembler
- iAssembler

*** Some require large memory workstations**

What software are available?

Other utilities:

- SAMTOOLS
- GATK / PICARD
- CUFFLINKS
- MACS
- ANNOVAR
- MYSQL
- R

GBS /RAD tools:

- TASSEL
- STACKS

Exercise 1: RNA-seq:

1. Alignment tool: TOPHAT

- **Reads from exons;**
- **Reads across splicing junctions;**
- **Reads larger than exons;**

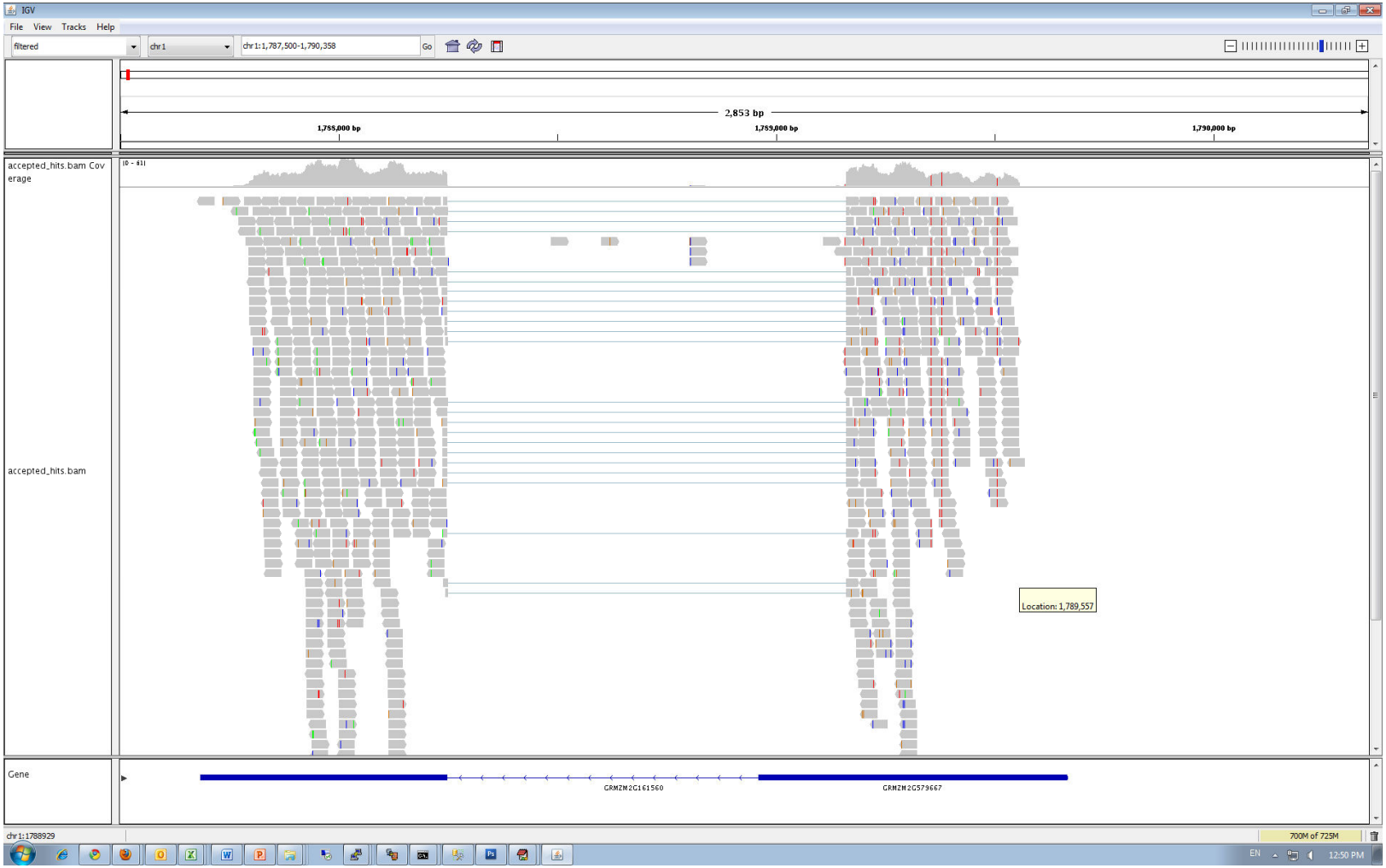
```
tophat -p 4 -o s1 /local_data/tair10/tair10 ./001_s2_sequence.txt.gz
```

2. Quantification: CUFFLINKS

- **Normalization: FPKM or Upper Quantile;**
- **CUFFDIFF: identify differentially expressed genes;**

```
cuffdiff -p 4 -o results /local_data/tair10/TAIR10_GFF3_genes.gff
s1.bam,s3.bam s2.bam,s4.bam
```

3. Visualization tool: IGV



Cufflinks output

trans_id	bundle_id	chr	left	right	FPKM	FMI	frac	FPKM_conf_lo	FPKM_conf_hi	coverage	length	effective_length	status
GRMZM2G060082_T01	99289	1	2	3807	2.05938	0.507199	0.576667	0	5.04574	0.25115	2804	2769	OK
GRMZM2G060082_T02	99289	1	2606	3754	4.0603	1	0.423333	0	8.65942	0.495171	1066	1031	OK
GRMZM2G059865_T01	99290	1	4853	9652	15.6517	1	0.471931	7.73925	23.5641	1.90879	1966	1931	OK
GRMZM2G059865_T03	99290	1	4856	6355	4.18E-09	2.67E-10	7.70E-11	0	0.000129	5.10E-10	1214	1179	OK
GRMZM2G059865_T02	99290	1	4856	9652	14.2274	0.909003	0.528069	6.68358	21.7713	1.7351	2412	2377	OK
GRMZM2G059856_T01	99291	1	9855	10388	0	0	0	0	0	0	533	498	OK
GRMZM5G888250_T01	99291	1	9881	10387	0	0	0	0	0	0	506	471	OK
GRMZM2G059843_T01	99292	1	11454	14988	0	0	0	0	0	0	1788	1788	OK
GRMZM5G866996_T01	99293	1	46227	47746	0	0	0	0	0	0	472	437	OK
GRMZM2G059818_T02	99294	1	50452	54182	0	0	0	0	0	0	3099	3099	OK
GRMZM2G059818_T01	99294	1	50452	56348	0	0	0	0	0	0	4379	4379	OK
GRMZM2G059818_T03	99294	1	52003	52543	0	0	0	0	0	0	540	540	OK
GRMZM2G360269_T01	99295	1	57418	61452	0	0	0	0	0	0	2556	2556	OK
GRMZM2G518629_T01	99296	1	62320	62588	0	0	0	0	0	0	98	98	OK
GRMZM5G811273_T02	99296	1	62501	64014	3.65473	0.943998	0.41328	0	8.63598	0.44571	279	244	OK
GRMZM5G811273_T01	99296	1	62733	64014	3.87154	1	0.58672	0	8.47176	0.472151	362	327	OK
AC177838.2_FGT002	99297	1	70594	71919	0	0	0	0	0	0	633	633	OK
GRMZM2G518627_T01	99298	1	73839	74024	0	0	0	0	0	0	185	185	OK
GRMZM2G059778_T01	99299	1	76119	76752	0	0	0	0	0	0	411	376	OK
GRMZM2G518609_T01	99300	1	90684	90815	0	0	0	0	0	0	131	96	OK
GRMZM2G059745_T01	99301	1	92353	93541	0	0	0	0	0	0	425	390	OK
GRMZM2G093344_T01	99302	1	109518	111769	8.56066	1	1	2.70894	14.4124	1.04401	1012	977	OK
GRMZM2G394757_T01	99302	1	110764	111506	0	0	0	0	0	0	419	419	OK

Using script to automate the batch processing

1. Make a text file with all the commands.

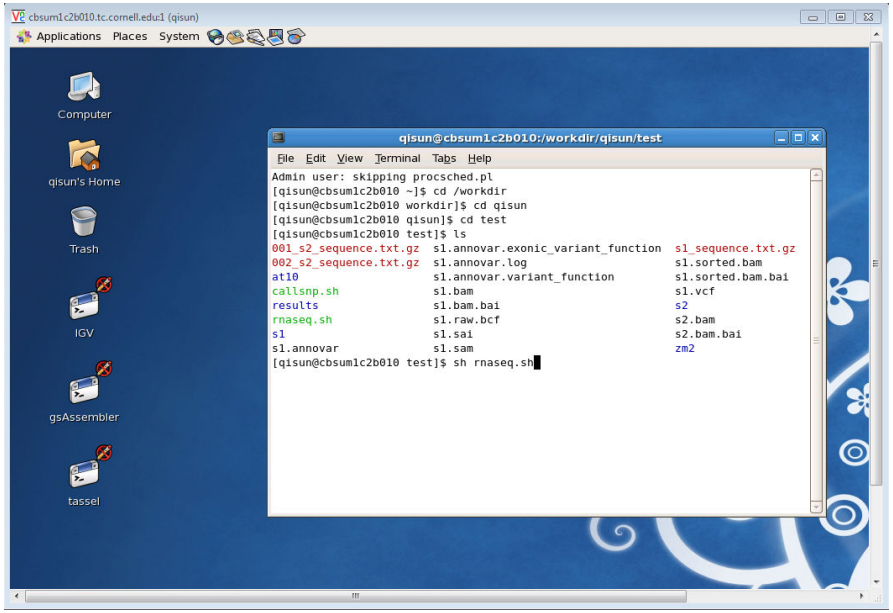
```
tophat -p 4 -o s1 /local_data/tair10/tair10 ./001_s2_sequence.txt.gz
tophat -p 4 -o s2 /local_data/tair10/tair10 ./002_s2_sequence.txt.gz
tophat -p 4 -o s3 /local_data/tair10/tair10 ./003_s2_sequence.txt.gz
tophat -p 4 -o s4 /local_data/tair10/tair10 ./004_s2_sequence.txt.gz


mv s1/accepted_hits.bam s1.bam
mv s2/accepted_hits.bam s2.bam
mv s3/accepted_hits.bam s3.bam
mv s4/accepted_hits.bam s4.bam

samtools index s1.bam
samtools index s2.bam
samtools index s3.bam
samtools index s4.bam

cuffdiff -p 4 -o results /local_data/tair10/TAIR10_GFF3_genes.gff s1.bam,s3.bam s2.bam,s4.bam
```

2. Run the script: sh *script_file_name*

Some tips for running scripts

1. You can create a script on a Windows computer and transfer to the Linux workstation. Before using the script, make sure you run “dos2unix <script name>” or “mac2unix <script name>”.

2. If the script takes long time to finish, start it through a VNC window. Then you can safely turn off your own computer without terminating the job.



* Instruction for using VNC is in the exercise instruction sheet.

Exercise 2: SNP/INDEL detection:

1. Alignment tool: BWA

```
bwa aln -t 4   /local_data/tair10/tair10   s1_sequence.txt.gz   >   s1.sai
bwa samse -n 10 /local_data/tair10/tair10   s1.sai   s1_sequence.txt.gz  > s1.sam
```

2. Call SNPs using SAMTOOLS

```
samtools view -bS -o s1.bam s1.sam
samtools sort s1.bam s1.sorted
samtools index s1.sorted.bam
samtools mpileup -uf /local_data/tair10/tair10 s1.sorted.bam
|bcftools view -bvc
g - > s1.raw.bcf
bcftools view s1.raw.bcf | vcfutils.pl varFilter -D100 > s1.vcf
```

Two pipelines available for SNP/INDEL calling

- **GATK**
  - **Optimized for 1k Human Genome project**
  - **Many filtering utilities**

- **SAMTOOLS**
  - **Not many filtering tools available**
  - **Easy to customize**

Commonly Used File Formats

Category	File Extension	Reference
Sequence	fasta	http://en.wikipedia.org/wiki/FASTA_format
Sequence	fastq	http://en.wikipedia.org/wiki/FASTQ_format
Alignment	SAM/BAM	http://samtools.sourceforge.net/SAM-1.3.pdf
Sequence variation	VCF/BCF	http://www.1000genomes.org/node/101
Genome Annotation	gff/gff3	http://gmod.org/wiki/GFF3
Genome Annotation	gtf	http://genome.ucsc.edu/FAQ/FAQformat#format4

Most files that you downloaded from a web site are compressed .gz files.
Use the gunzip command to de-compress the file. E.g.

```
gunzip s_1_sequence.txt.gz
```

A few other topics

Where to get the reference genome and annotation files?

- **Using UCSC site to download genome fasta file.**

  http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/

  use “cat chr* > allchr.fa” command to concatenate the individual chromsomes into one file)

- **Using the UCSC Table Browser to create the GTF file.**

  http://genome.ucsc.edu/cgi-bin/hgTables?command=start

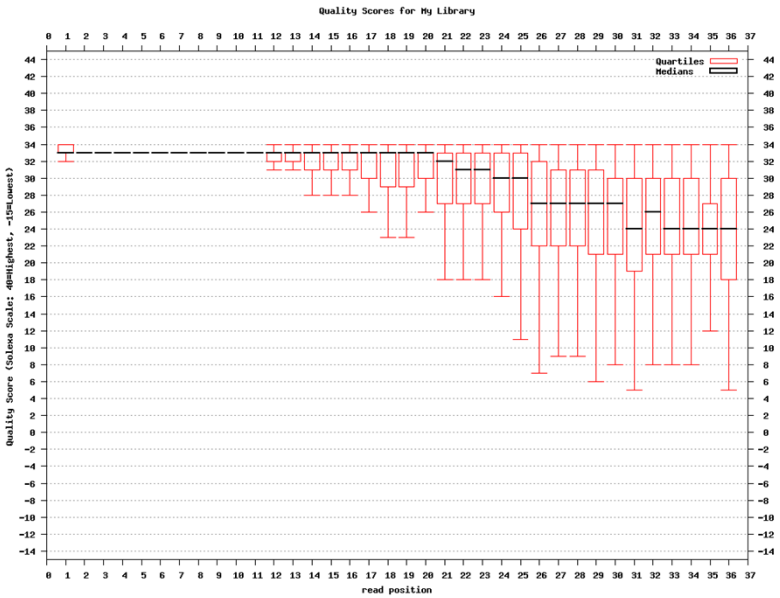
Troubleshooting 1

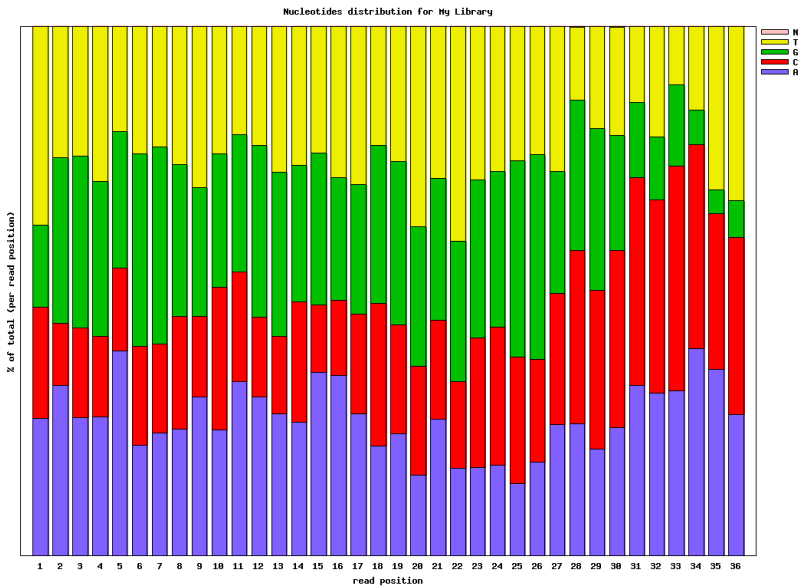
Check sequencing quality using fastx toolkit

```
fastx_quality_stats –Q33  –i s_3_sequence.txt  -o stat_report.xls &
```

column	count	min	max	sum	mean	Q1	med	Q3	IQR	lW	rW	A_Count	C_Count	G_Count	T_Count	N_Count
1	6362991	-4	40	250734117	39.41	40	40	40	0	40	40	1396976	1329101	678730	2958184	0
2	6362991	-5	40	250531036	39.37	40	40	40	0	40	40	1786786	1055766	1738025	1782414	0
3	6362991	-5	40	248722469	39.09	40	40	40	0	40	40	2296384	984875	1443989	1637743	0
4	6362991	-5	40	247654797	38.92	40	40	40	0	40	40	1683197	1410855	1722633	1546306	0
5	6362991	-4	40	248214827	39.01	40	40	40	0	40	40	2536861	1167423	1248968	1409739	0
6	6362991	-5	40	248499903	39.05	40	40	40	0	40	40	1598956	1236081	1568608	1959346	0
7	6362991	-4	40	247719760	38.93	40	40	40	0	40	40	1692667	1822140	1496741	1351443	0
8	6362991	-5	40	245745205	38.62	40	40	40	0	40	40	2230936	1343260	1529928	1258867	0
9	6362991	-5	40	245766735	38.62	40	40	40	0	40	40	1702064	1306257	1336511	2018159	0
10	6362991	-5	40	245089706	38.52	40	40	40	0	40	40	1519917	1446370	1450995	1945709	0
11	6362991	-5	40	242641359	38.13	40	40	40	0	40	40	1717434	1282975	1387804	1974778	0
12	6362991	-5	40	242026113	38.04	40	40	40	0	40	40	1662872	1202041	1519721	1978357	0
13	6362991	-5	40	238704245	37.51	40	40	40	0	40	40	1549965	1271411	1973291	1566681	1643
14	6362991	-5	40	235622401	37.03	40	40	40	0	40	40	2101301	1141451	1603990	1515774	475
15	6362991	-5	40	230766669	36.27	40	40	40	0	40	40	2344003	1058571	1440466	1519865	86
16	6362991	-5	40	224466237	35.28	38	40	40	2	35	40	2203515	1026017	1474060	1651582	7817
17	6362991	-5	40	219990002	34.57	34	40	40	6	25	40	1522515	1125455	2159183	1555765	73
18	6362991	-5	40	214104778	33.65	30	40	40	10	15	40	1479795	2068113	1558400	1249337	7346
19	6362991	-5	40	212934712	33.46	30	40	40	10	15	40	1432749	1231352	1769799	1920093	8998
20	6362991	-5	40	212787944	33.44	29	40	40	11	13	40	1311657	1411663	2126316	1513282	73
21	6362991	-5	40	211369187	33.22	28	40	40	12	10	40	1887985	1846300	1300326	1318380	10000
22	6362991	-5	40	213371720	33.53	30	40	40	10	15	40	542299	3446249	516615	1848190	9638
23	6362991	-5	40	221975899	34.89	36	40	40	4	30	40	347679	1233267	926621	3855355	69
24	6362991	-5	40	194378421	30.55	21	40	40	19	-5	40	433560	674358	3262764	1992242	67
25	6362991	-5	40	199773985	31.40	23	40	40	17	-2	40	944760	325595	1322800	3769641	195
26	6362991	-5	40	179404759	28.20	17	34	40	23	-5	40	3457922	156013	1494664	1254293	99
27	6362991	-5	40	163386668	25.68	13	28	40	27	-5	40	1392177	281250	3867895	821491	178
28	6362991	-5	40	156230534	24.55	12	25	40	28	-5	40	907189	981249	4174945	299437	171
29	6362991	-5	40	163236046	25.65	13	28	40	27	-5	40	1097171	3418678	1567013	280008	121
30	6362991	-5	40	151309826	23.78	12	23	40	28	-5	40	3514775	2036194	566277	245613	132
31	6362991	-5	40	141392520	22.22	10	21	40	30	-5	40	1569000	4571357	124732	97721	181
32	6362991	-5	40	143436943	22.54	10	21	40	30	-5	40	1453607	4519441	38176	351107	660
33	6362991	-5	40	114269843	17.96	6	14	30	24	-5	40	3311001	2161254	155505	734297	934
34	6362991	-5	40	140638447	22.10	10	20	40	30	-5	40	1501615	1637357	18113	3205237	669
35	6362991	-5	40	138910532	21.83	10	20	40	30	-5	40	1532519	3495057	23229	1311834	352
36	6362991	-5	40	117158566	18.41	7	15	30	23	-5	40	4074444	1402980	63287	822035	245

FASTX output





Troubleshooting 2

1. **Total number of reads.**

2. **% Reads that can be aligned to the genome.**

```
samtools flagstat myBAMfile.bam
```

- The flagstat tool does not give the accurate count with BAM files created by Tophat. The reason is that Tophat would report ambiguous alignments in many rows. The following command would give the number of reads that are aligned:

```
samtools view myBAMfile.bam   | awk -F"\t" '{print $1}' |sort|uniq|wc
```

Downstream analysis

- **RNA-seq**
  - **DAVID**
  - **Mapman**
  - **GeneSpring/Ingenuity**

- **SNP/INDEL**
  - **Annotate SNP/INDEL with Annovar**
  - **QTL, GWAS**
  - **CBSU tool for analyzing pooled segregated F2 population**

CBSU Office Hours

Every Monday 2 to 4 PM

Office hour schedule:

http://cbsu.tc.cornell.edu/lab/office.aspx