DATABASES:

Refseq_genomic: HGP_set, Celera_set and CRA_chr7 and Mitochondria.

```
/Users/feng/Databases/Human.chromosomes/ref_chrMT.fa
/Users/feng/Databases/Human.chromosomes/ref_chrY.fa
/Users/feng/Databases/Human.chromosomes/alt_Celera_chrY.fa
/Users/feng/Databases/Human.chromosomes/ref_chr22.fa
/Users/feng/Databases/Human.chromosomes/alt_Celera_chr22.fa
/Users/feng/Databases/Human.chromosomes/ref_chr21.fa
/Users/feng/Databases/Human.chromosomes/alt_Celera_chr21.fa
/Users/feng/Databases/Human.chromosomes/ref_chr20.fa
/Users/feng/Databases/Human.chromosomes/alt_Celera_chr20.fa
/Users/feng/Databases/Human.chromosomes/ref_chr19.fa
/Users/feng/Databases/Human.chromosomes/alt_Celera_chr19.fa
/Users/feng/Databases/Human.chromosomes/ref_chr18.fa
/Users/feng/Databases/Human.chromosomes/alt_Celera_chr18.fa
/Users/feng/Databases/Human.chromosomes/ref_chr17.fa
/Users/feng/Databases/Human.chromosomes/alt_Celera_chr17.fa
/Users/feng/Databases/Human.chromosomes/ref_chr16.fa
/Users/feng/Databases/Human.chromosomes/alt_Celera_chr16.fa
/Users/feng/Databases/Human.chromosomes/ref_chr15.fa
/Users/feng/Databases/Human.chromosomes/alt_Celera_chr15.fa
/Users/feng/Databases/Human.chromosomes/ref_chr14.fa
/Users/feng/Databases/Human.chromosomes/alt_Celera_chr14.fa
/Users/feng/Databases/Human.chromosomes/ref_chr13.fa
/Users/feng/Databases/Human.chromosomes/alt_Celera_chr13.fa
/Users/feng/Databases/Human.chromosomes/ref_chrX.fa
/Users/feng/Databases/Human.chromosomes/alt_Celera_chrX.fa
/Users/feng/Databases/Human.chromosomes/ref_chr12.fa
/Users/feng/Databases/Human.chromosomes/alt_Celera_chr12.fa
/Users/feng/Databases/Human.chromosomes/ref_chr11.fa
/Users/feng/Databases/Human.chromosomes/alt_Celera_chr11.fa
/Users/feng/Databases/Human.chromosomes/ref_chr10.fa
/Users/feng/Databases/Human.chromosomes/alt_Celera_chr10.fa
/Users/feng/Databases/Human.chromosomes/ref_chr9.fa
/Users/feng/Databases/Human.chromosomes/alt_Celera_chr9.fa
/Users/feng/Databases/Human.chromosomes/ref_chr8.fa
/Users/feng/Databases/Human.chromosomes/alt_Celera_chr8.fa
/Users/feng/Databases/Human.chromosomes/ref_chr7.fa
/Users/feng/Databases/Human.chromosomes/alt_Celera_chr7.fa
/Users/feng/Databases/Human.chromosomes/alt_CRA_TCAGchr7.v2_chr7.fa
/Users/feng/Databases/Human.chromosomes/ref_chr6.fa
/Users/feng/Databases/Human.chromosomes/alt_Celera_chr6.fa
/Users/feng/Databases/Human.chromosomes/ref_chr5.fa
/Users/feng/Databases/Human.chromosomes/alt_Celera_chr5.fa
/Users/feng/Databases/Human.chromosomes/ref_chr4.fa
/Users/feng/Databases/Human.chromosomes/alt_Celera_chr4.fa
/Users/feng/Databases/Human.chromosomes/ref_chr3.fa
/Users/feng/Databases/Human.chromosomes/alt_Celera_chr3.fa
/Users/feng/Databases/Human.chromosomes/ref_chr2.fa
/Users/feng/Databases/Human.chromosomes/alt_Celera_chr2.fa
/Users/feng/Databases/Human.chromosomes/ref_chr1.fa
/Users/feng/Databases/Human.chromosomes/alt_Celera_chr1.fa
```

Refseq_RNA: 11.29.07 version human Refseq_RNA, limited with_GI list.

```
/Volumes/BACKUP/Databases/refseq.RNA/refseq_rna
```

Viral.genomic: 3.13.07 version of viral genomic sequences

```
/Users/feng/Databases/viral.genomic/Mar.13.07/viral1.genomic.fna
```

NOTE:
TAGE LENGTH: 20mer tags were 5'primer of 21mer tags extracted from 32-38bp ditags.
SAVING FOLDER: shyadminpcp1454:~/Documents/Lab_Data/SAGE/Manuscript/DTS
PARAMETERS: r=5 q=-4 G=25 E=10 for blastall
CONVERT TAGS: cat not_found I awk '{print $2}' > ???
Cat not_found I awk '{print ">"$2"\n"$2}' > ??? with fasta format

Part A: DTS analysis on 994 L-SAGE tags from BCBL1 cells

Step1: exact match to human Refseq_RNA sequences

```
$ ./run_parse_blast.v2.pl -p blastn -noprt -e 10000 -W 20 -F F -dbfile DBFILE.human.refseq.RNA -l
human.refseq.rna.gi.11.29.06 < BCBL1_2X_20mer.tags > BCBL1_2X_20mer.refseq.20.log &

$ wc -l found not_found
    872 found             122 not_found            994 total
```

Step 2: exact match to human Refseq_genomic sequences

```
$ ./run_parse_blast.v2.pl -p blastn  -noprt -e 10000 -W 20 -F F -dbfile DBFILE.human.genome <
BCBL1_20mer_refseq_not_found > BCBL1_refseq20_genomic20.log &

$ wc -l found not_found
     88 found             34 not_found             122 total
```

Step 3: short, nearly match (19/20) to human Refseq_RNA sequences

```
$ ./run_parse_blast.v2.lower.20.19.pl -p blastn -noprt -e 10000 -W 7 -F F -dbfile DBFILE.human.refseq.RNA -l
human.refseq.rna.gi.11.29.06 < BCBL1_20mer_refseq_genomic_not_found >
BCBL1_20mer_refseq20_genomic20_refseq19.log &

$ wc -l found_20 found_19 not_found
     0 found_20          21 found_19             13 not_found           34 total
```

Step 4: short, nearly match (19/20) to human genomic sequences

```
$ ./run_parse_blast.v2.lower.20.19.pl -p blastn -noprt -e 10000 -W 7 -F F -dbfile DBFILE.human.genome <
BCBL1_20mer_refseq2019_genomic20_not_found > BCBL1_20mer_refseq2019_genomic2019.log &

$ wc -l found_20 found_19 not_found
     0 found_20          10 found_19             3 not_found            13 total
```

Step 5: short, nearly match (20/20, 19/20, 18/20) to viral genomic sequences

```
$ ./run_parse_blast.v2.lower.20.19.18.pl -p blastn -e 10000 -W 7 -F F -dbfile DBFILE.viral1.genomic <
BCBL1_20mer_refseq2019_genomic2019_not_found > BCBL1_refseq_genomic_2019_viral.log

$ wc -l found_20 found_19 found_18 not_found
     2 found_20 (KSHV)          0 found_19       0 found_18       1 not_found       3 total
```

Step 6: Short nearly match (20/.20, 19/20, 18/20) to human nr and est (March.16.07)

```
$ ./blastcl3 -p blastn -e 10000 -W 7 -r 5 -q -4 -G 25 -E 10 -b 1 -v 1 -d nr -u human[organism] -I T -F F -i
not_found.BCBL1 -o not_found.nr
```

3 tags with 18/20 to nr

```
./blastcl3 -p blastn -e 10000 -W 7 -r 5 -q -4 -G 25 -E 10 -b 1 -v 1 -d "est" -u human[organism] -I T -F F -i
not_found.BCBL1 -o not_found.est
```

1 tag with **20/20**   1 tag with **19/20** (KSHV T0.7)   1 tag with **18/20** (KSHV T1.1)

Part B: DTS analysis on 18,204 L-SAGE tags from SCCC.ACB

Step1: exact match to human Refseq_RNA sequences

```
$ ./run_parse_blast.v2.pl -p blastn -noprt -e 10000 -W 20 -F F -dbfile DBFILE.human.refseq.RNA -l
human.refseq.rna.gi.11.29.06 < SCCC.ACB.NON.2X.20mer.tags > SCCC.ACB.20mer.refseq20.log &
```

11655 found   6549 not_found   18204 total

Step 2: exact match to human Refseq_genomic sequences

```
$ ./run_parse_blast.v2.pl -p blastn -noprt  -e 10000 -W 20 -F F -dbfile DBFILE.human.genome <
SCCC.ACB.refseq20.not_found > SCCC.ACB.refseq.genome.20.log &
```

5429 found   1120 not_found   6549 total

Step 3: short, nearly match (19/20) to human Refseq_RNA sequences

```
$ ./run_parse_blast.v2.lower.20.19.pl -p blastn -noprt -e 10000 -W 7 -F F -dbfile DBFILE.human.refseq.RNA -l
human.refseq.rna.gi.11.29.06 < SCCC.ACB.Refseq.genome.20.not_found >
SCCC.ACB.Refseq.genome.20.Refseq.19.log &
```

0 found_20   508 found_19   612 not_found   1120 total

Step 4: short, nearly match (19/20) to human genomic sequences

```
$ ./run_parse_blast.v2.lower.20.19.pl -p blastn -noprt -e 10000 -W 7 -F F -dbfile DBFILE.human.genome
<SCCC.ACB.Refseq.genome.20.Refseq.19.not_found >SCCC.ACB.Refseq.genome.20.19.log
```

0 found_20   466 found_19   146 not_found   612 total

Step 5: short, nearly match (20/20, 19/20, 18/20) to viral genome

```
$ ./run_parse_blast.v2.lower.20.19.18.pl -p blastn -noprt -e 10000 -W 7 -F F -dbfile DBFILE.viral1.genomic <
not_found.refseq.genomic.20.19.SCCC > SCCC.refseq.genomic.20.19.virus.log &
```

1 found_20 (KSHV)   1 found_19   23 found_18   121 not_found   146 total

Step 6: short, nearly match to human nr and est (Dec.23.06)

```
$ ./blastcl3 -p blastn -e 10000 -W 7 -r 5 -q -4 -G 25 -E 10 -b 1 -v 1 -d nr -u human[organism] -I T -F F -i
not_found.fasta -o not_found.fasta.nr &
```

31 (20/20)   20 (19/20)   79 (18/20)   14 (17/20)   2 (16/20)   total 2.+ mismatch (95)

```
$ ./blastcl3 -p blastn -e 10000 -W 7 -r 5 -q -4 -G 25 -E 10 -b 1 -v 1 -d "est" -u human[organism] -I T -F F -i
not_found.fasta -o not_found.fasta.est &
```

67 (20/20)   25 (19/20)   40 (18/20)   9 (17/20)   2 (16/20)   2 (15/20)   1 (not_found)
total 2.+ mismatch (54)

Combined together (nr.est):   50 tags with 2.+ mismatch

Step 7: Quality and duplicated ditags
2 with sequencing read error
24 from duplicated ditags
2 from adaptors
1 from KSHV (T1.1)

21 Candidate sequences