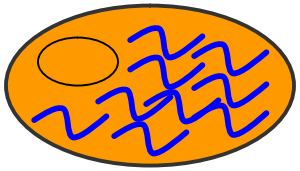
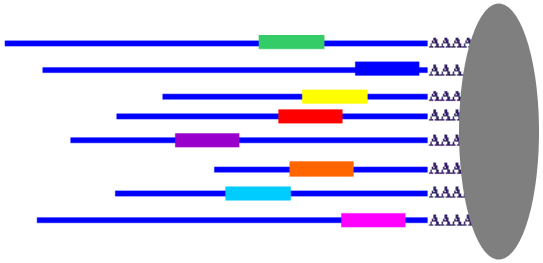
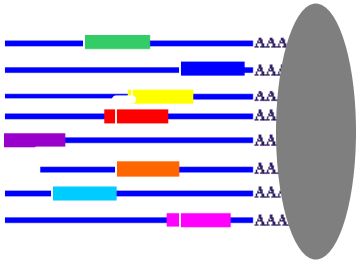
SAGE: Procedure



Isolate mRNA, RT to cDNA



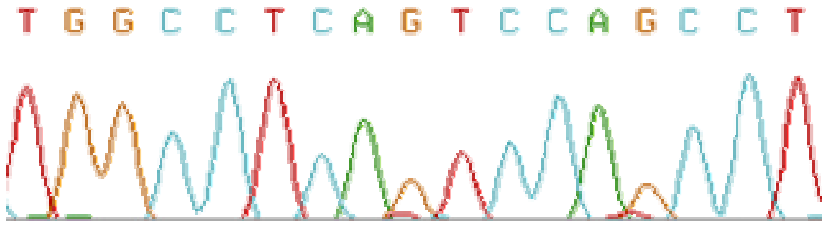
Digest with "Anchoring enzyme" NlaIII



Digest with "Tagging enzyme" BsmFI

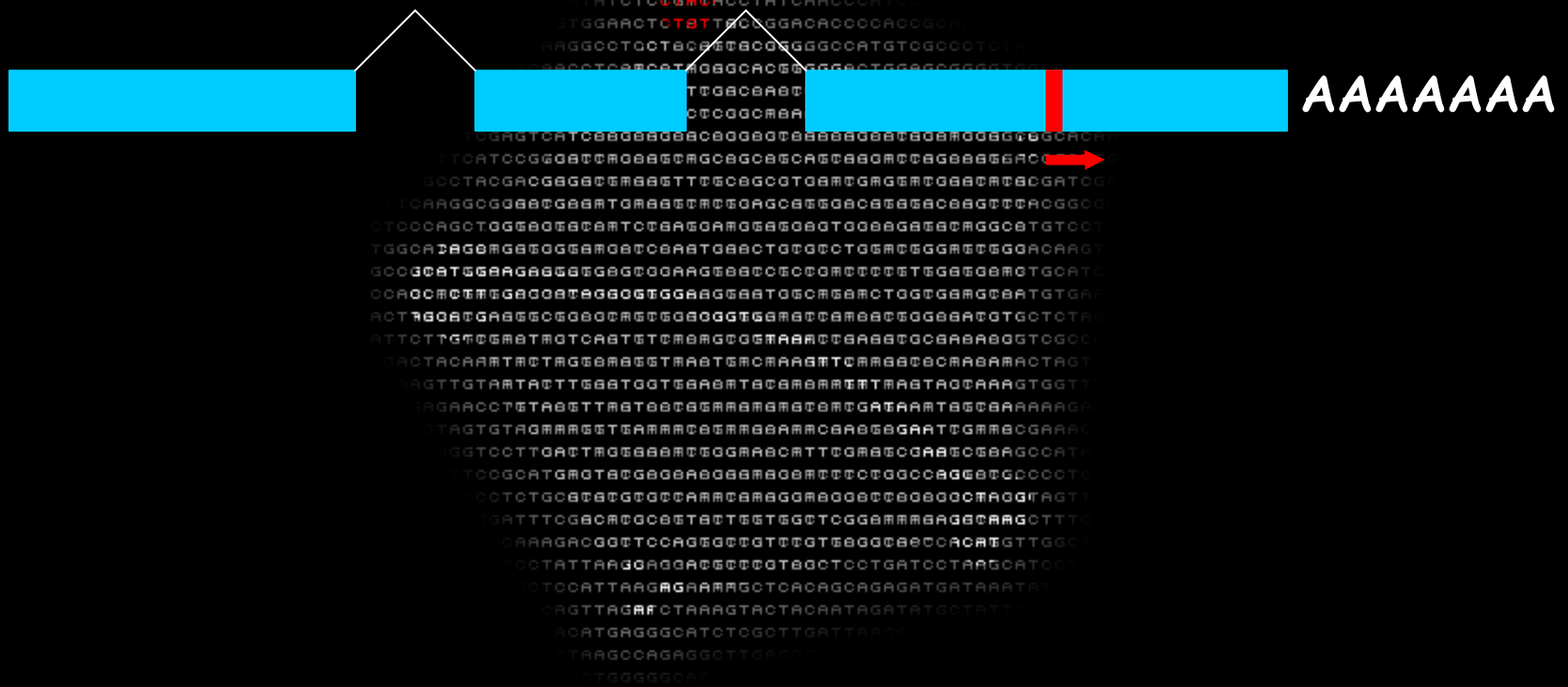
Ligate tags

Sequence

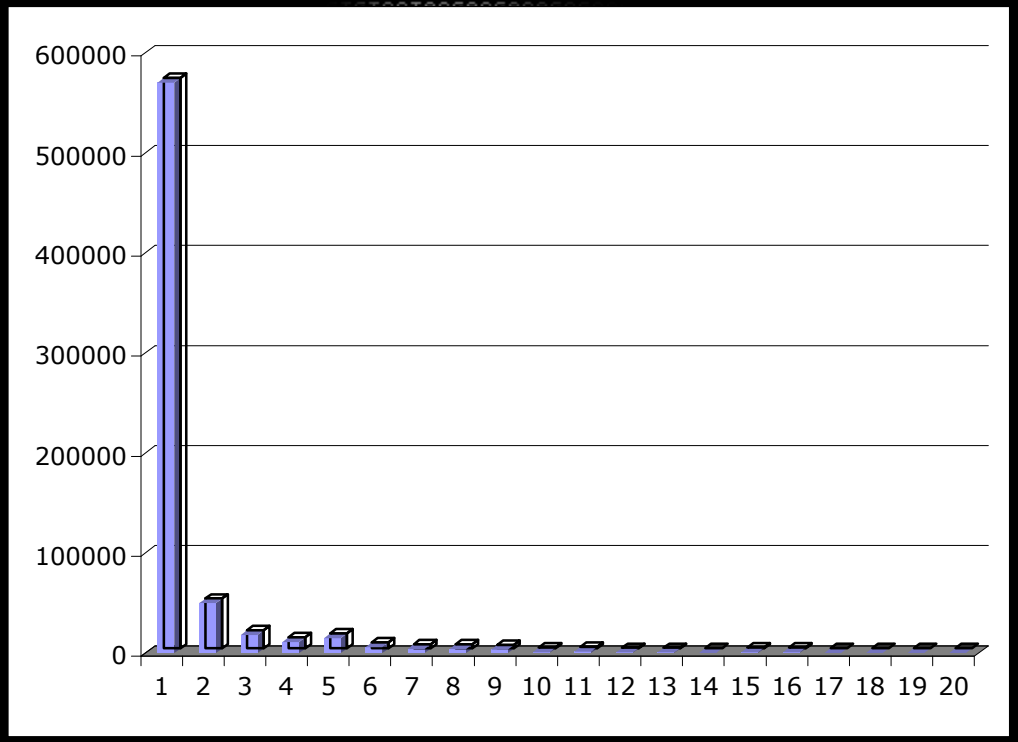


http://www.sagenet.org/home/Description.htm

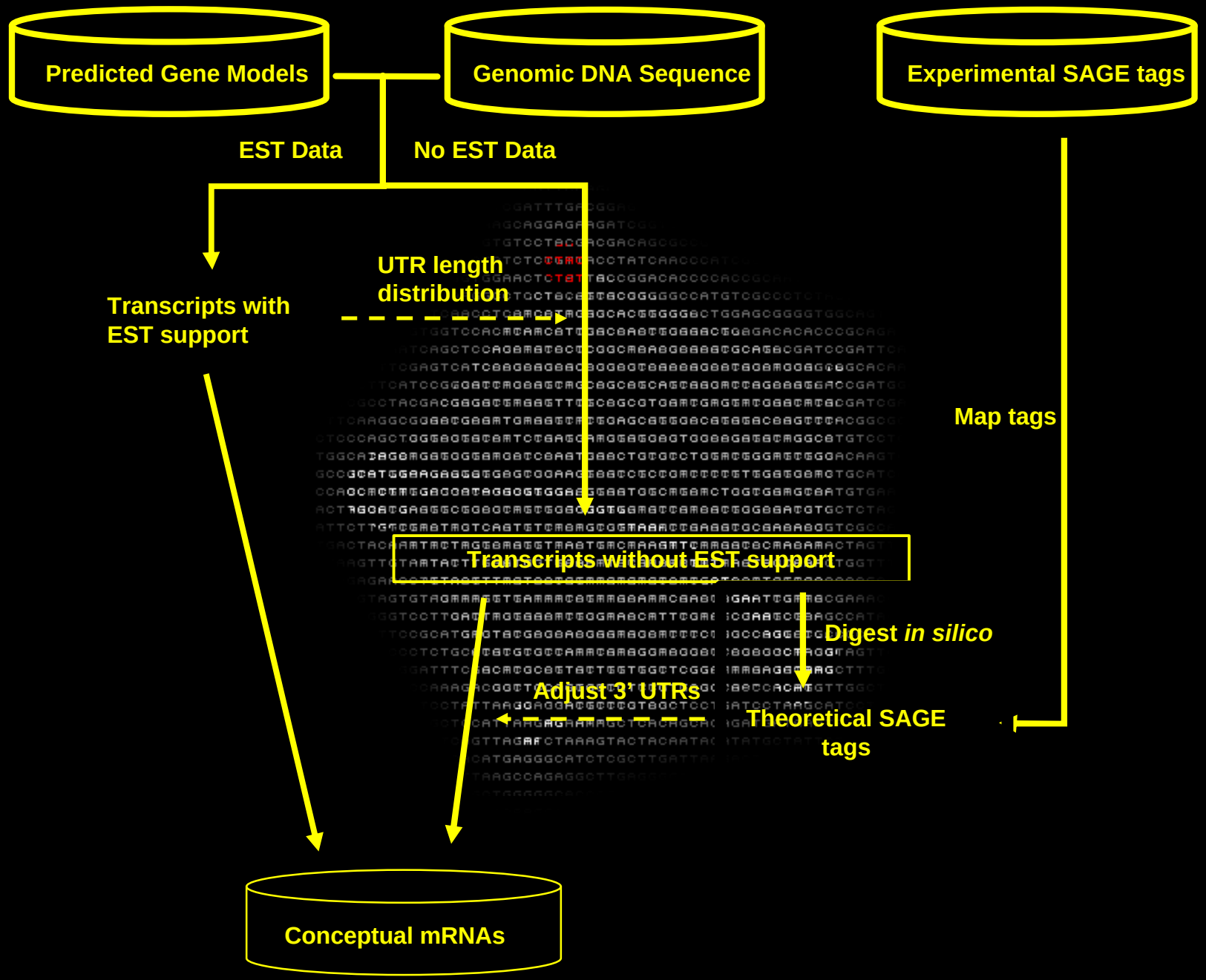
Mapping SAGE tags to genes



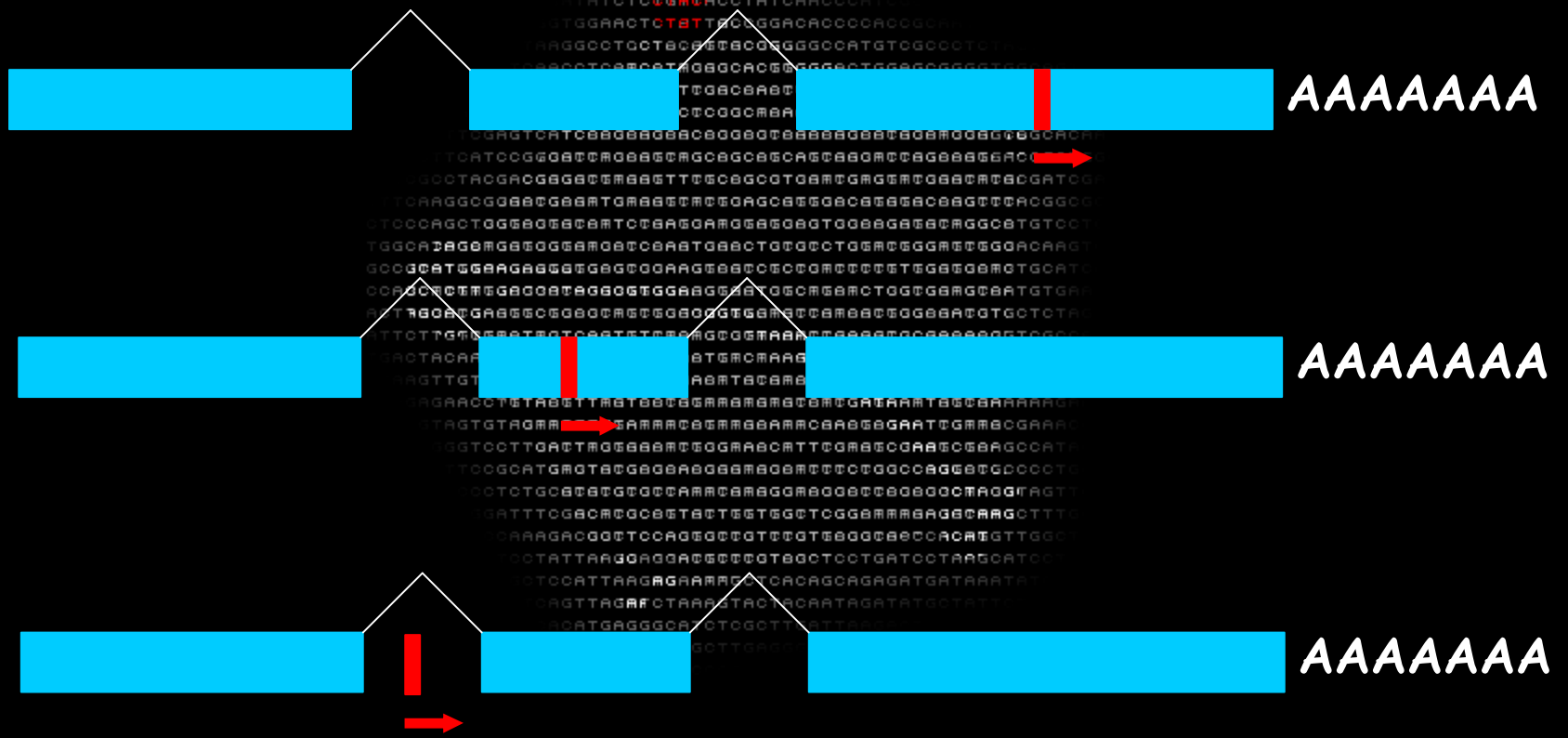
Tags



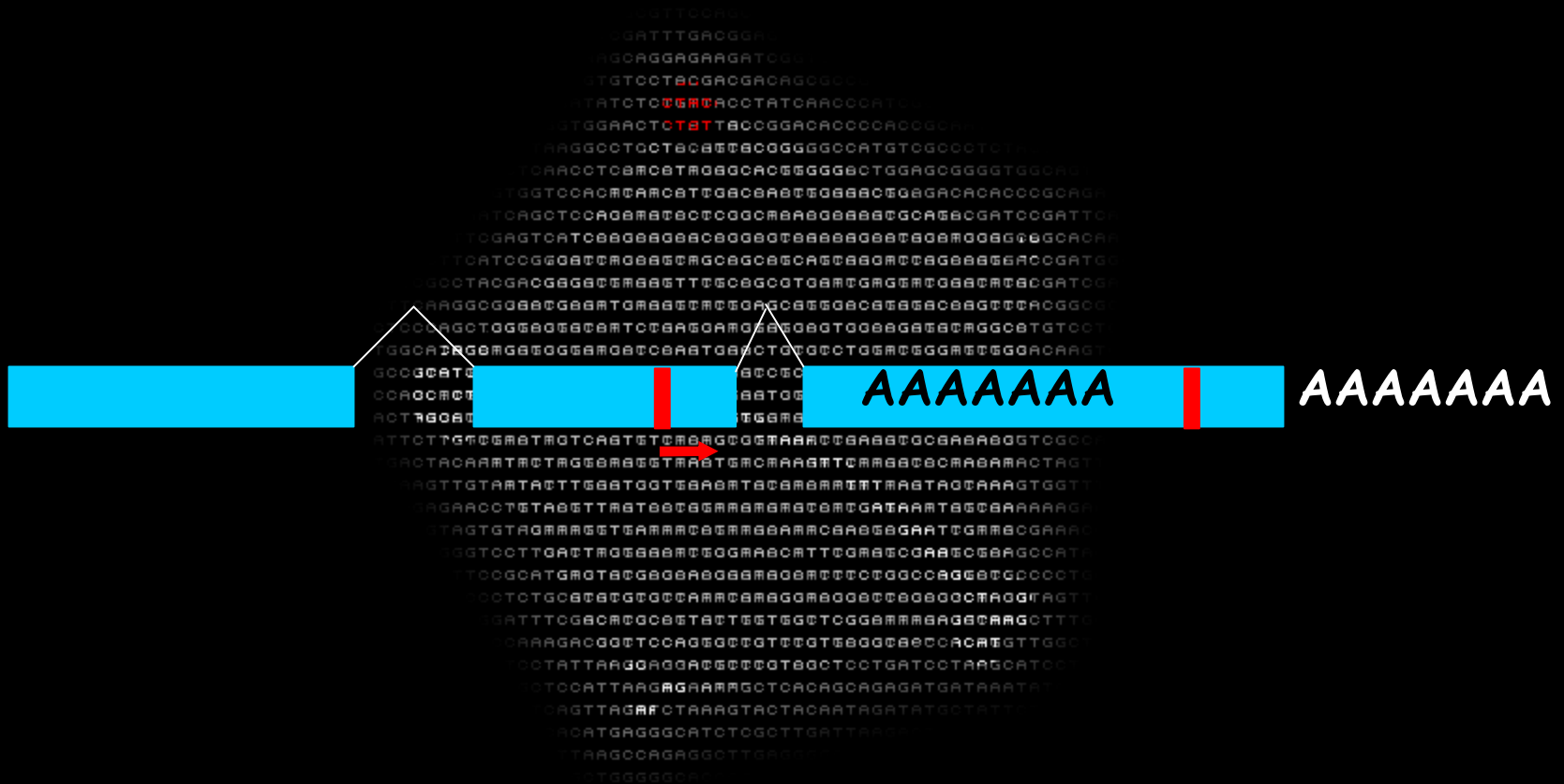
Position of NlaIII site


Predicted Gene Models
Genomic DNA Sequence
Experimental SAGE tags
EST Data
No EST Data
Transcripts with EST support
UTR length distribution
Map tags
Transcripts without EST support
Digest *in silico*
Adjust 3' UTRs
Theoretical SAGE tags
Conceptual mRNAs

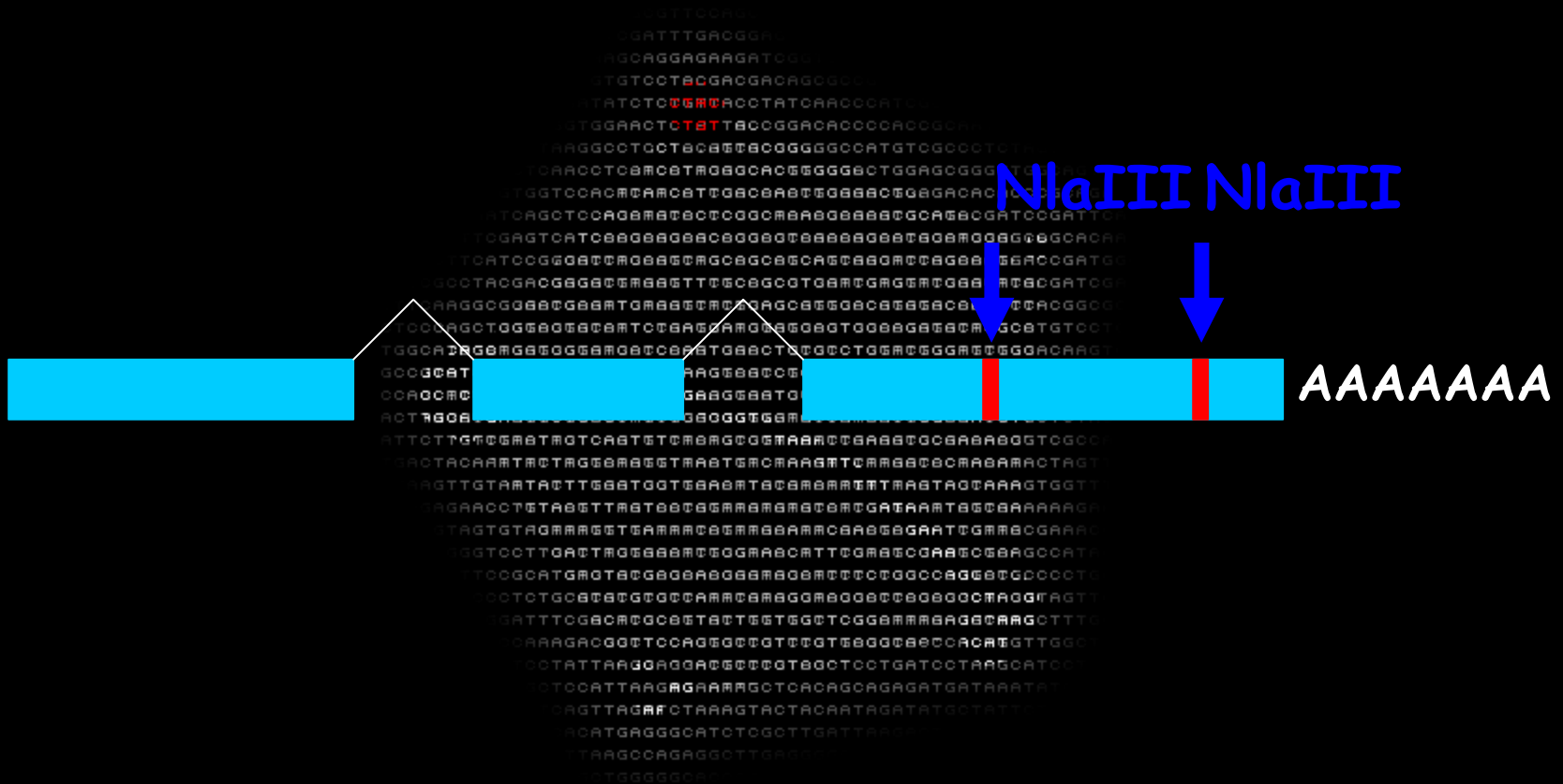
Mapping SAGE tags to genes



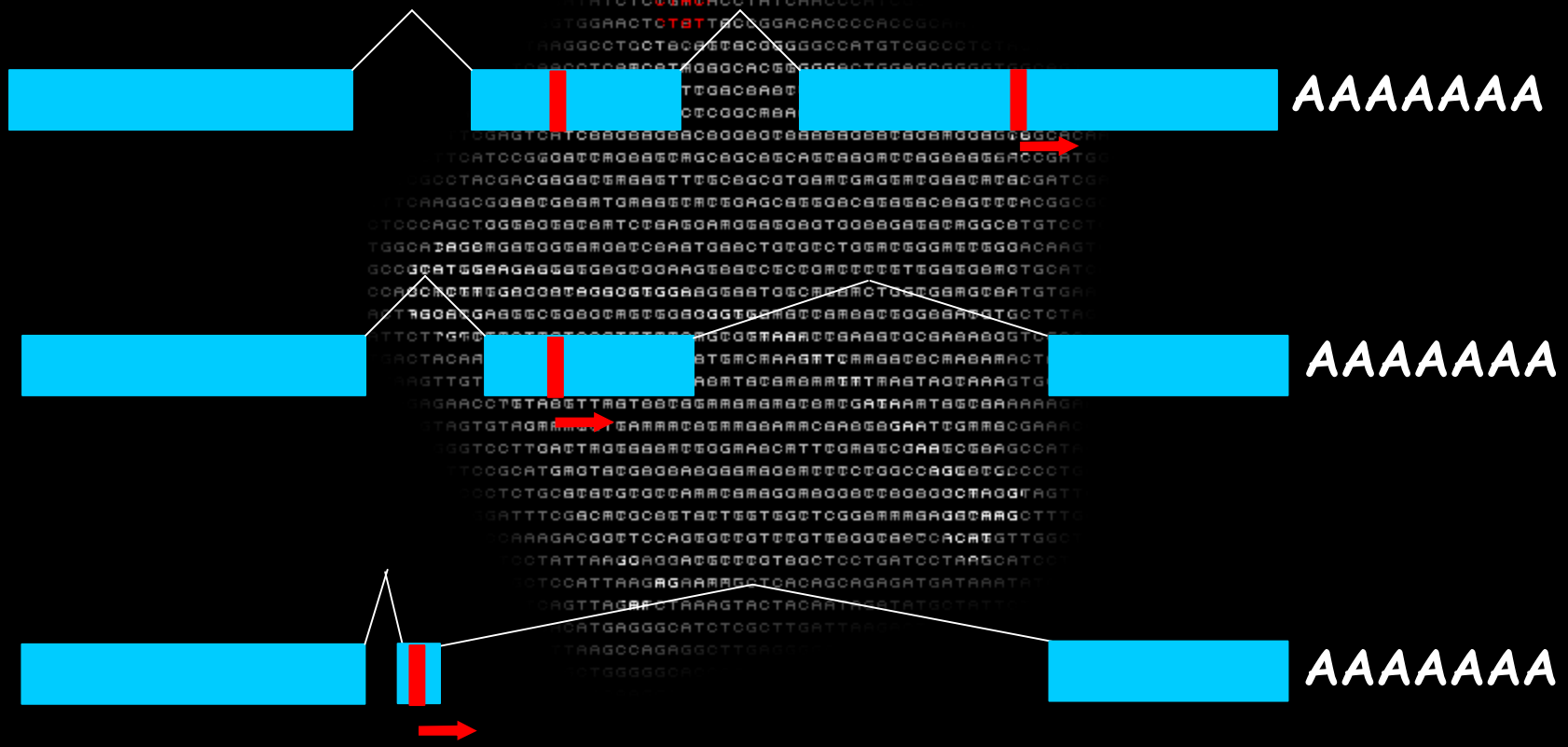
Internal polyA tracts



NlaIII Partial Digestion



Alternative Splicing



C. elegans SAGE DATA -- WS110

Summarize library quality and tag-to-gene mapping

Tag filtering criteria

SWN21 N2 Embryos — Library

Descending — Tag count sorting order

0.99 — Overall Sequence Quality

None — Small Family Sequence Quality

None — Singleton Sequence Quality

5 — Minimum tag count to show

- [x] Remove duplicate ditags
- [ ] Show unmatched tags
- [x] Resolve ambiguity to lowest position match (apply to Coding RNA only)
- [ ] Show only unmatched tags
- [x] Show only unambiguous tags (apply to Coding RNA, Other RNA and Mitochondrial only)
- [x] Hide antisense tags (apply to Coding RNA only)
- [ ] Output raw text only

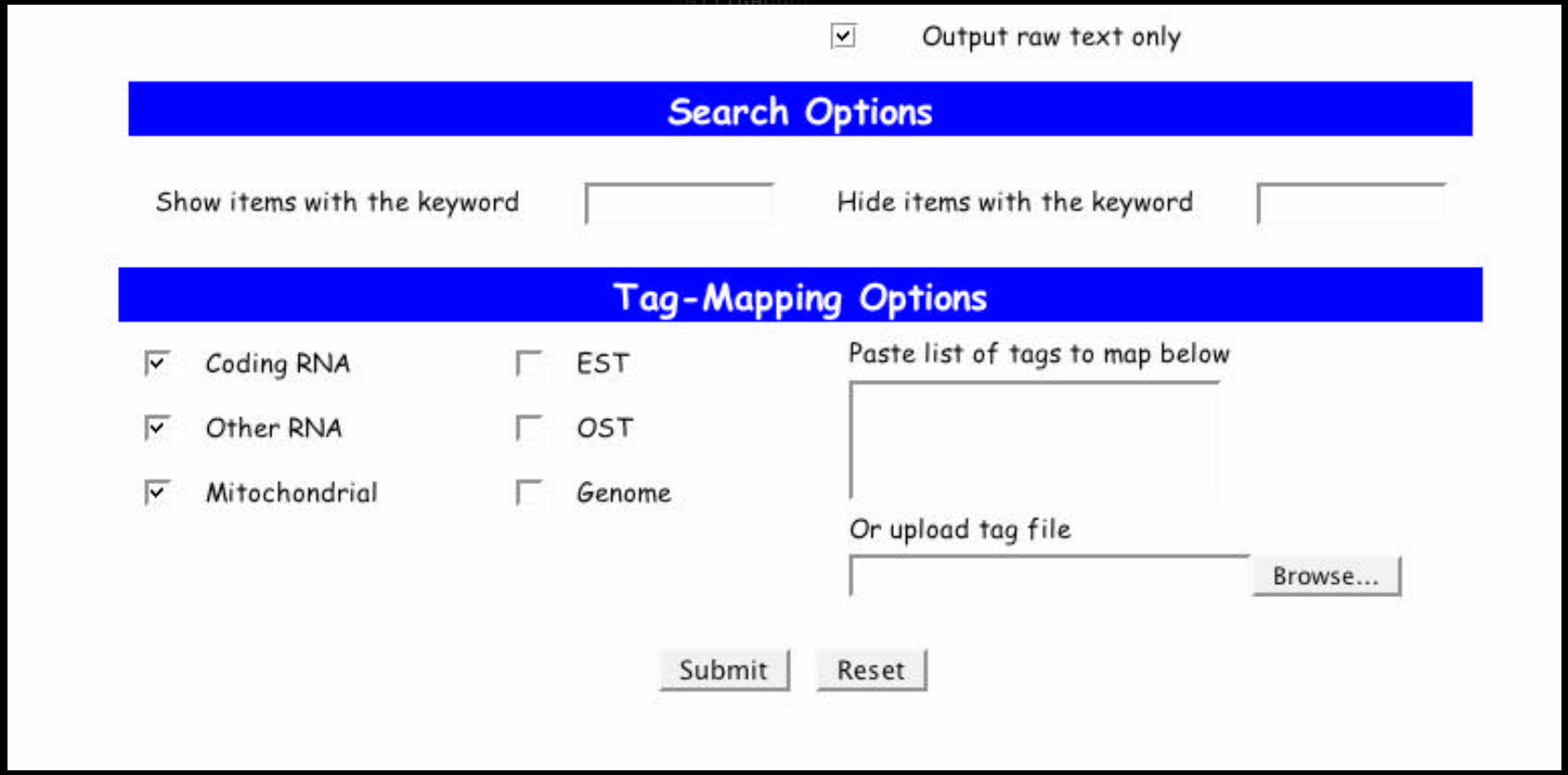
Search Options

Show items with the keyword

Hide items with the keyword

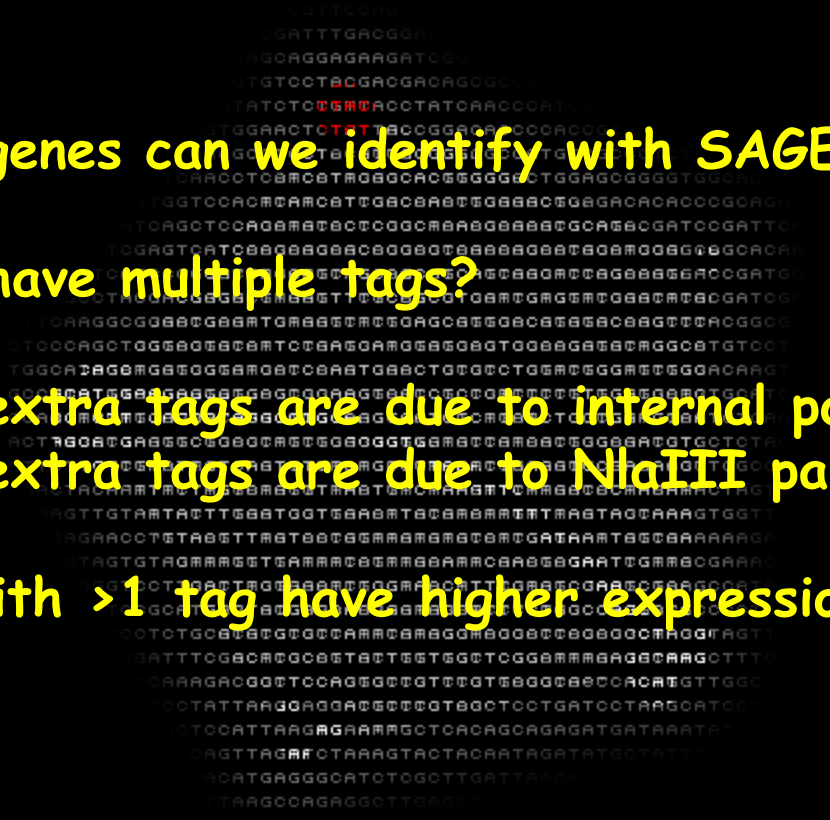
SWN21 Summary -- WS110							
Library	Total tags	Dup ditags	Overall Qual > 0.99		Singleton Qual >		Clean
SWN21	133825	10559	90696		60797		57684

Freq	Tag	Source	Position	Strand	Gene	*Locus*	"Description"(Antisense)
1335	TCTTGTGTGG	Mitochondrial	1	+	cytochrome c oxidase subunit_III	.	.
941	TGCGTTGTCT	Other_RNA	3	+	F31C3.9	.	"26s rRNA"
907	TTGTTACCTT	Coding_RNA	1	+	Y37E3.8b	.	"similar to Ribosomal protein L15"
		Coding_RNA	1	+	Y37E3.8a	.	"similar to Ribosomal protein L15"
696	CCCAACGAGA	Coding_RNA	1	+	Y77E11A.15	*col-106*	"C. elegans COL-106 protein, similar to Collagen triple helix repeat (20 copies), PF01484 (Nematode cuticle collagen N-terminal domain)"
674	AATAAACGAA	Coding_RNA	1	+	C32D5.9	*lgg-1*	"C. elegans LGG-1 protein, similar to Microtubule associated protein 1A/1B, light chain 3"
558	CCGTTAAATT	Coding_RNA	1	+	K06A4.7	.	"hypothetical gene model"
463	ATACTTATTA	Mitochondrial	1	+	cytochrome c oxidase subunit_I	.	.
462	CTTGGGCATT	Mitochondrial	1	+	cytochrome c oxidase subunit_II	.	.
445	GTCTATTCTG	Coding_RNA	1	+	F23A7.4	.	"glutamate receptor epsilon subunit


Output raw text only
Search Options
Show items with the keyword
Hide items with the keyword
Tag-Mapping Options
Coding RNA
EST
Other RNA
OST
Mitochondrial
Genome
Paste list of tags to map below
Or upload tag file
Browse...
Submit
Reset

```
1335    TCTTGTGTGG      Mitochondrial   1   +   cytochrome_c_oxidase_subunit_III  .   .
941     TGCGTTGTCT      other_RNA   3   +   F31C3.9  .   "26s rRNA"
907     TTGTTACCTT      coding_RNA   1   +   Y37E3.8b  .   "similar to Ribosomal protein L15" ; coding_RNA   1
696     CCCAACGAGA      coding_RNA   1   +   Y77E11A.15  col-106   "C. elegans COL-106 protein, similar to Coll
674     AATAAACGAA      coding_RNA   1   +   C32D5.9  lgg-1   "C. elegans LGG-1 protein, similar to Microtubule
558     CCGTTAAATT      coding_RNA   1   +   K06A4.7  .   "hypothetical gene model"
463     ATACTTATTA      Mitochondrial   1   +   cytochrome_c_oxidase_subunit_I  .   .
462     CTTGGGCATT      Mitochondrial   1   +   cytochrome_c_oxidase_subunit_II  .   .
445     GTCTATTCTG      coding_RNA   1   +   F23A7.4  .   "glutamate receptor epsilon subunit like"
401     TACAATAGTG      coding_RNA   1   +   Y119D3B.21  .   "similar to Plasmodium falciparum DNA-directed RNA
385     TAACCATTGA      coding_RNA   1   +   F23D12.1  .   "hypothetical gene model"
380     AAATCGTTAT      coding_RNA   1   +   R09B3.3  .   "RNA recognition motif. (aka RRM, RBD, or RNP domain)
334     GACCACTCAC      coding_RNA   1   +   F10B5.1  rpl-10   "ribosomal protein L10 (QM protein)"
319     TGTTGGCAAA      coding_RNA   1   +   ZK1010.1  ubq-2   "UBQ-2 ubiquitin, 60S Ribosomal protein L40"
307     CAACTCAGAA      coding_RNA   1   +   C08F11.11  .   "hypothetical gene model"
291     GGATTCGGTC      coding_RNA   1   +   F25H2.10  rpa-0   "deoxyribonuclease"
269     ACCTGTAGAA      coding_RNA   5   +   ZK380.1  tbx-32   "DNA-binding protein"
241     AAGTACAATG      coding_RNA   1   +   C26F1.9  rpl-39   "ribosomal protein L39"
235     CACAAATCTG      coding_RNA   1   +   M01F1.2  rpl-16   "L13P family ribosomal protein"
234     CGGAGAGGGA      coding_RNA   1   +   Y105E8A.16  rps-20   "rps-20 encodes a small ribosomal subunit S20
220     AAAAAAAAAA      coding_RNA   5   +   C46H11.1  .   "similar to Homo sapiens Splice isoform 3 of Q96JB8
218     CGGATGCGGA      coding_RNA   1   +   F41F3.3  .   "cuticlin"
217     TTAATGAAAA      Mitochondrial   1   +   rRNA  .   .
216     TGACTTCTGA      other_RNA   1   +   F09E10.11a  .   "RNA gene of unknown function locus:tts-1" ; other_
205     AAGCGAAAGG      coding_RNA   2   +   Y48B6A.2  rpl-43   "similar to Ribosomal L37ae protein family, Sco
194     GGAAAGAAGA      coding_RNA   1   +   C53H9.1  rpl-27   "60S ribosomal protein L27"
186     TACAAACTCA      coding_RNA   1   +   C01B10.5a  hil-7   "C. elegans HIL-7 protein, similar to Saccharom
186     GACGAAGTTT      coding_RNA   1   +   Y22D7AR.10  .   "similar to Homo sapiens DC33, TR:Q9H2L7"
```

Serial Analysis of Gene Expression



- How many genes can we identify with SAGE?
- How many have multiple tags?
- How many extra tags are due to internal polyAs?
- How many extra tags are due to NlaIII partial digestion
- Do genes with >1 tag have higher expression levels?

1. Extract theoretical SAGE tags from the virtual transcriptome -- make a note of which ones are upstream of a polyA tract

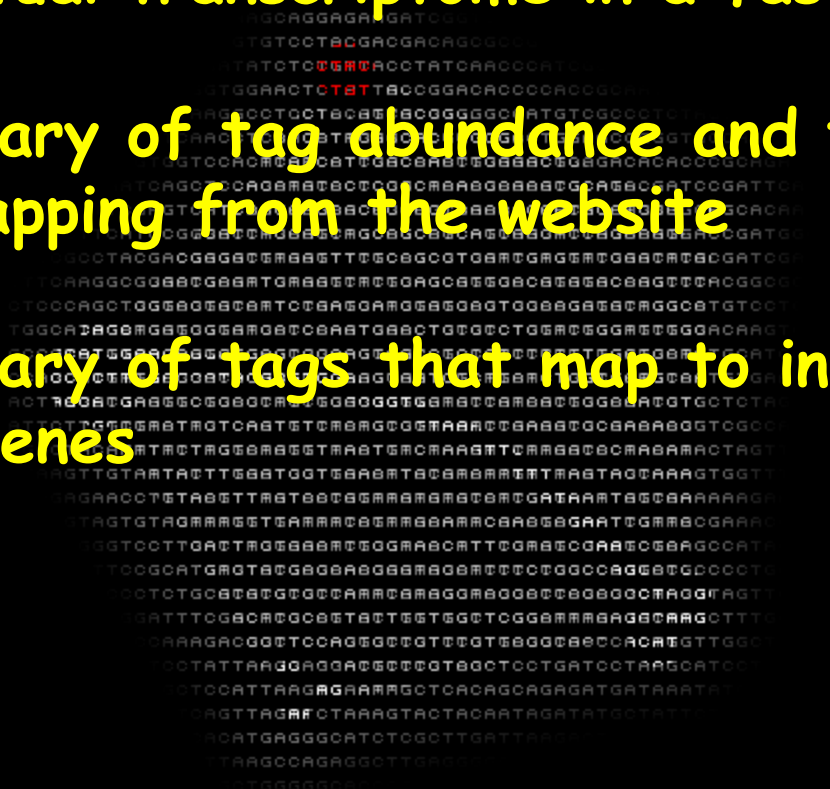
2. Identify which experimental SAGE tags could be due to partial digestion with NlaIII

3. Identify genes with multiple SAGE tags, subtract tags encountered in steps 1 and 2

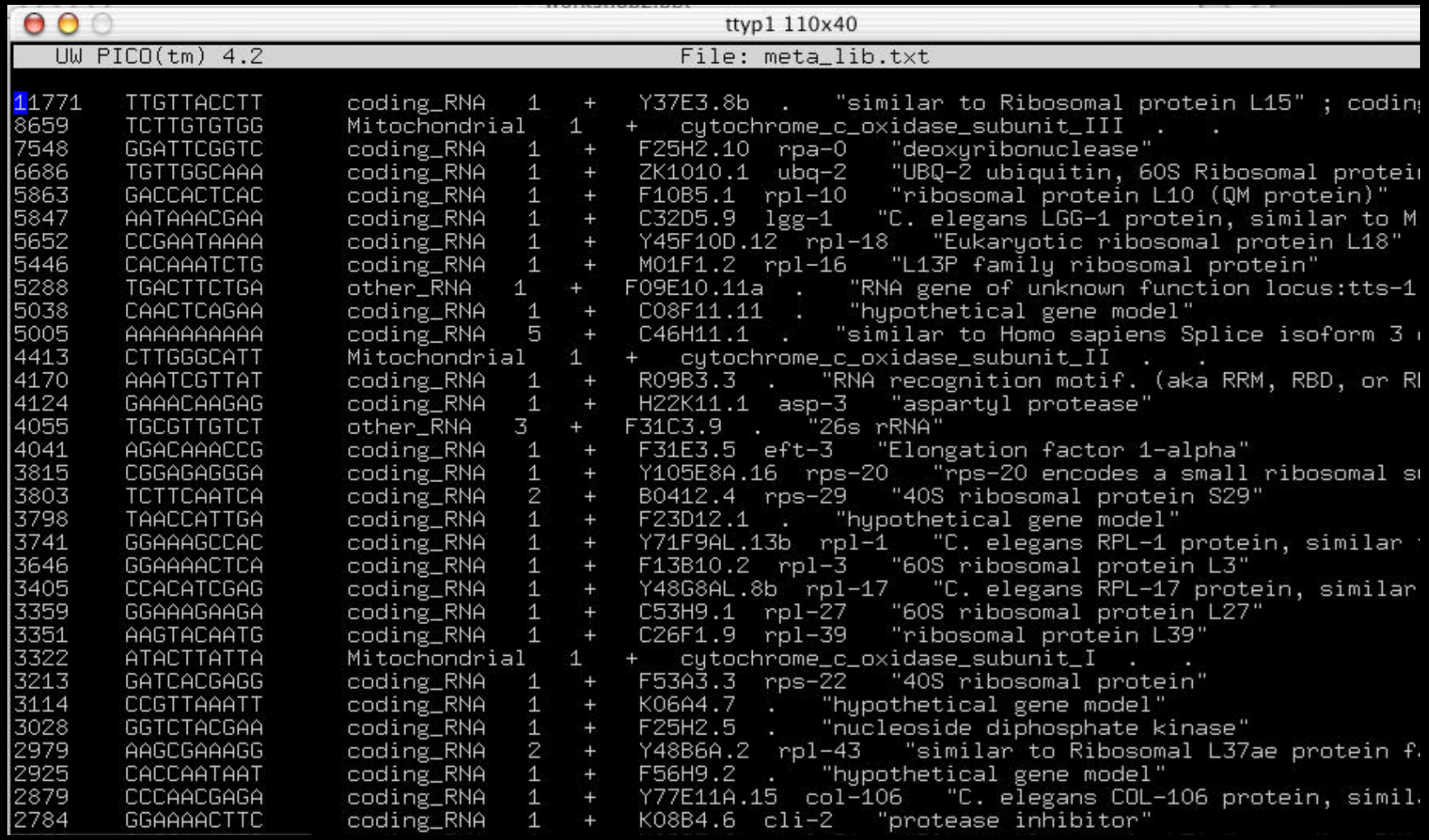
4. Calculate the tag abundance for genes with just one tag and genes with multiple tags

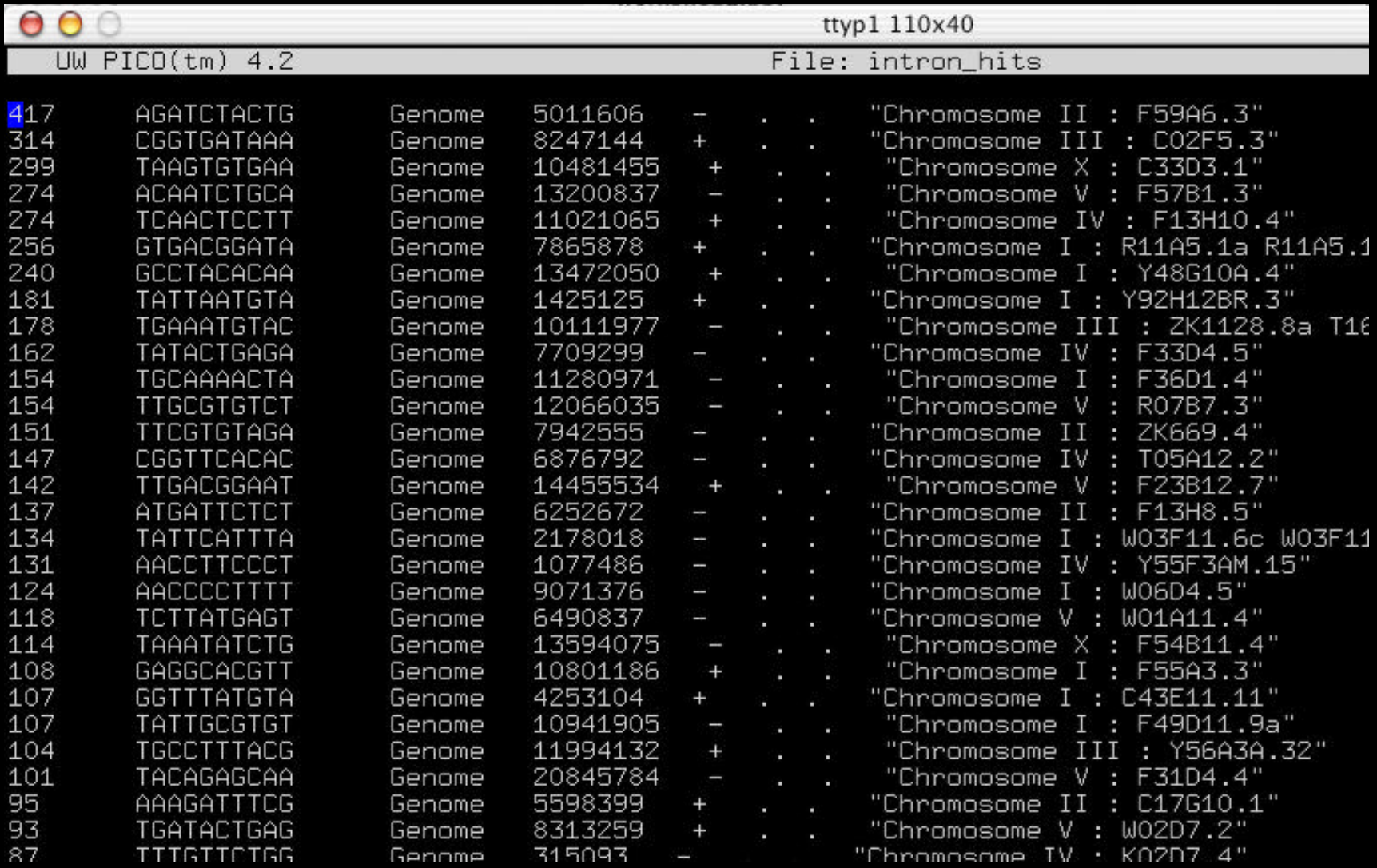
```
{xhost01}~/libstats> ./altsplice.pl meta_lib.txt

Library meta_lib.txt: 14680 genes detected

Transcripts      Genes   Abundance (average)
1                8446                 32
2                3540                 67
3                1205                114
4                385                 106
5                135                 234
6                46                  641
7                16                  210
8                8                   286
9                5                   399
>10              3                   799
{xhost01}~/libstats>
```

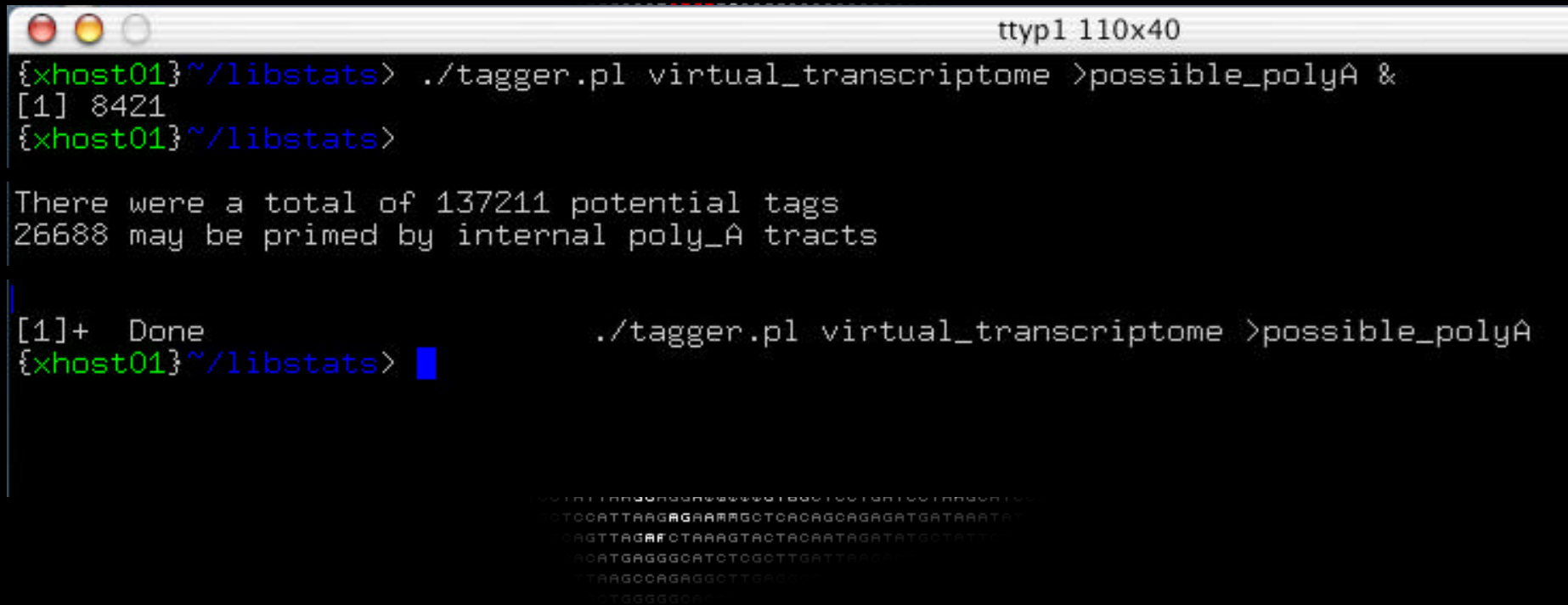
1. The virtual transcriptome in a fasta file
2. A summary of tag abundance and tag-to-gene mapping from the website
3. A summary of tags that map to intron of known genes

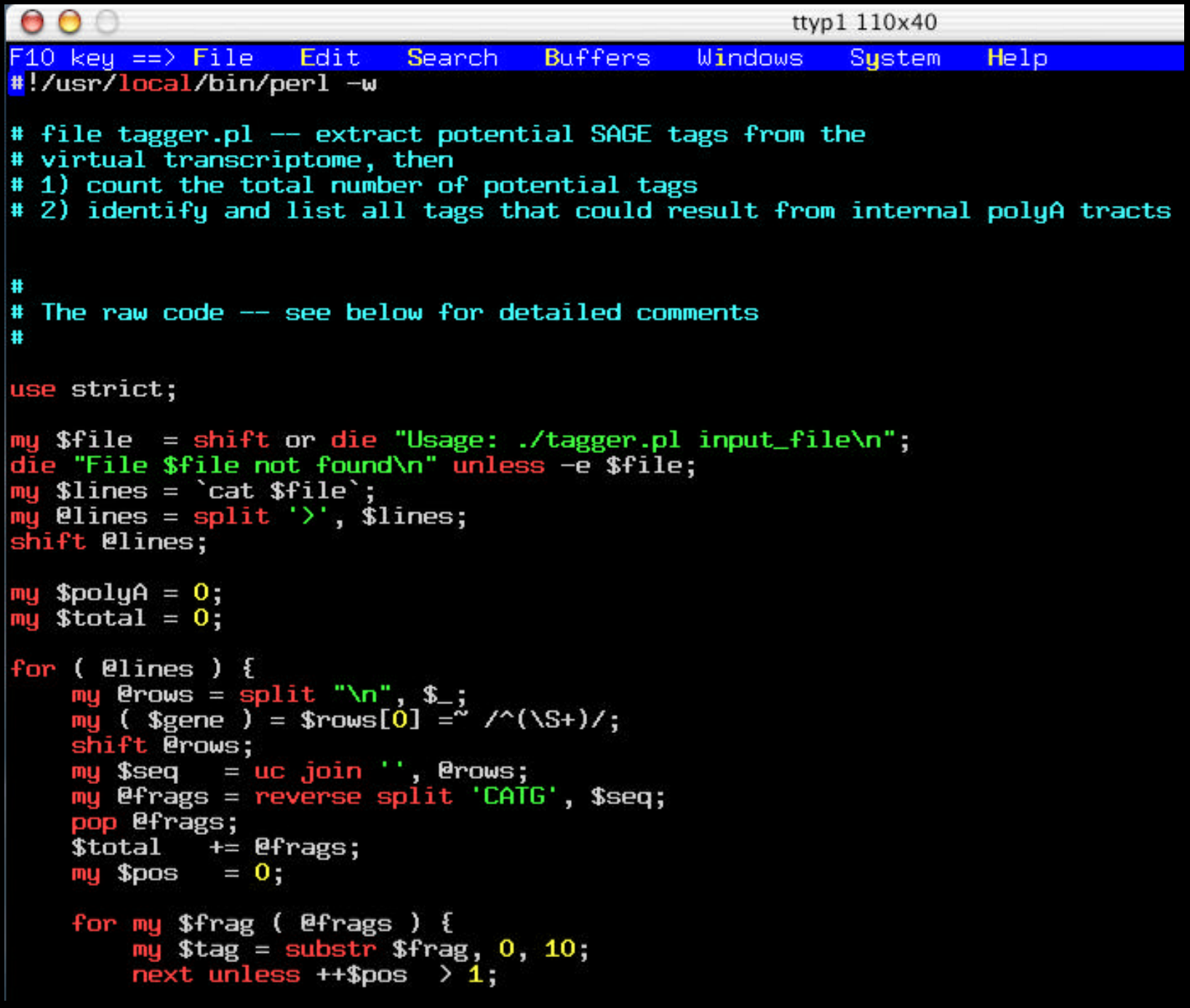



ttyp2 110x40
>Y74C9A.3 Real3=40 Real5=84 - 4181 10232 No polyA test needed
agtatcatcatcatatccgtgacctctcgtcttttcgtcacaaaagtctccagattgctcggagcagctgtcaca
TCACAATGGTGAAGATGTTTATGAAAAGGCGGAGGAATACTGGAGCCGCGCGAGCCAGGACGTCAACGGAATGCT
CGGCGTCGAAACGATTTATTGAAGGACTGAAGAAAAAGAATCTATTCGGCTACTTTGACTATGCACTGGACTGCC
ATGCCATTCTTCTCGAAAGTTGATATGGAAGACGTCGTCGAGGAGTTGATCACGAAAAGTGATCAATATATTGGA
ACTGCAGACGTTTGCACCGCCCGAACGACGTTATGATTTGATATGGATTCAATGGGTTTCAGGGCATTTGGTTGA
AAGGACTGAAACCTGGTGGATGTATTGTGCTCAAGGATAATGTGACAAATCACGAGAAACGGTTATTCGACGATC
CTTAAAGCGTTCGCCGATTCTCAACTGGACATGGTCTCGAAAGCACTGCAAACCGGATTCCCAAAGGAGATTTAT
CGGATTCACCAATAATTGAtttaaatatcgatttttattcgtttaattgcaattttccc
>Y74C9A.2 Real3=243 Real5=25 + 11616 16828 No polyA test needed
agtaagccaaacatacacaatcaacATGAAACTCGTAATTCTGCTATCTTTTGTTGCGACAGTTGCGGTTTTTGC
TGCGTGCTCTTCAGGAGCAACTGTACAGTCTGGAGAAAGAGAACGGAGTTGATGTGAAGCAAAAGGAGCAACCAC
AAGAGAATGGTCGCGTGGCAGCCGATGAAGCGGTCGATGATCAATGAGGATTCTAGAGCTCCATTGCTCCACGCA
AGAACGCCTCGGAGTCAACCCGGAGGAAGTTTTGGCGGATCTTCGTGCTCGTAATCAATTCCAATAAatattct
aaaatttcggttcttcttggcttcttctatttgtgaaatggtttatttttcccccgaactctcaaaaggtttaaat
tcaatttcttatttatcattatttttctaaacgaagacggatgtgattttaaattatgttaatggactattttac
>Y74C9A.4b Real3=361 ESTIMATE5=749 - 17550 27527 No polyA test needed
tctttgaaactacagtaatccgagagattttaaaggcgcatataggattctggaaaagaatcattttgcgcatt
tttaatgggtgttttatagaaaaactatagtatttacgcaaaagtatagaaagtagcctaaaaaatacaaaaat
tacctaaatatttgaagcagttttcctctaaatttgaatattgaacgcaaaaacaccaatatgttcccgaaaaa
cctgtttgaaaaaccgccaattgtacatttgcgatagagtgcgcttgcagcagtatagactcgcccttccgcgg
ttttgtagtttttttaagttttaatggaagaaaaaatacattataagtttcattcaaatactaaaaacattttt
aaaaattaccggagaatctgcgtctccggggtgacgattcctcaaaatccggagagcctccaaaaacagatttt
tctcgtgtttcaacataacttctagtgtttcgaatgttttctctataaagtttgtttaaaaacattcaaaaatc
GACGAAGACGCCTCTCGAAAAGAAAATATTTCAGACGAAGGCTTGAATATGTTGAATGCATCGCCGGAGCCAAT
AGAAGAAACCAGCAGAATGGCTCGTCCTATAAGATCCATGAGAAAACGCGAAACAACGTCTGGGGAATCAATGG

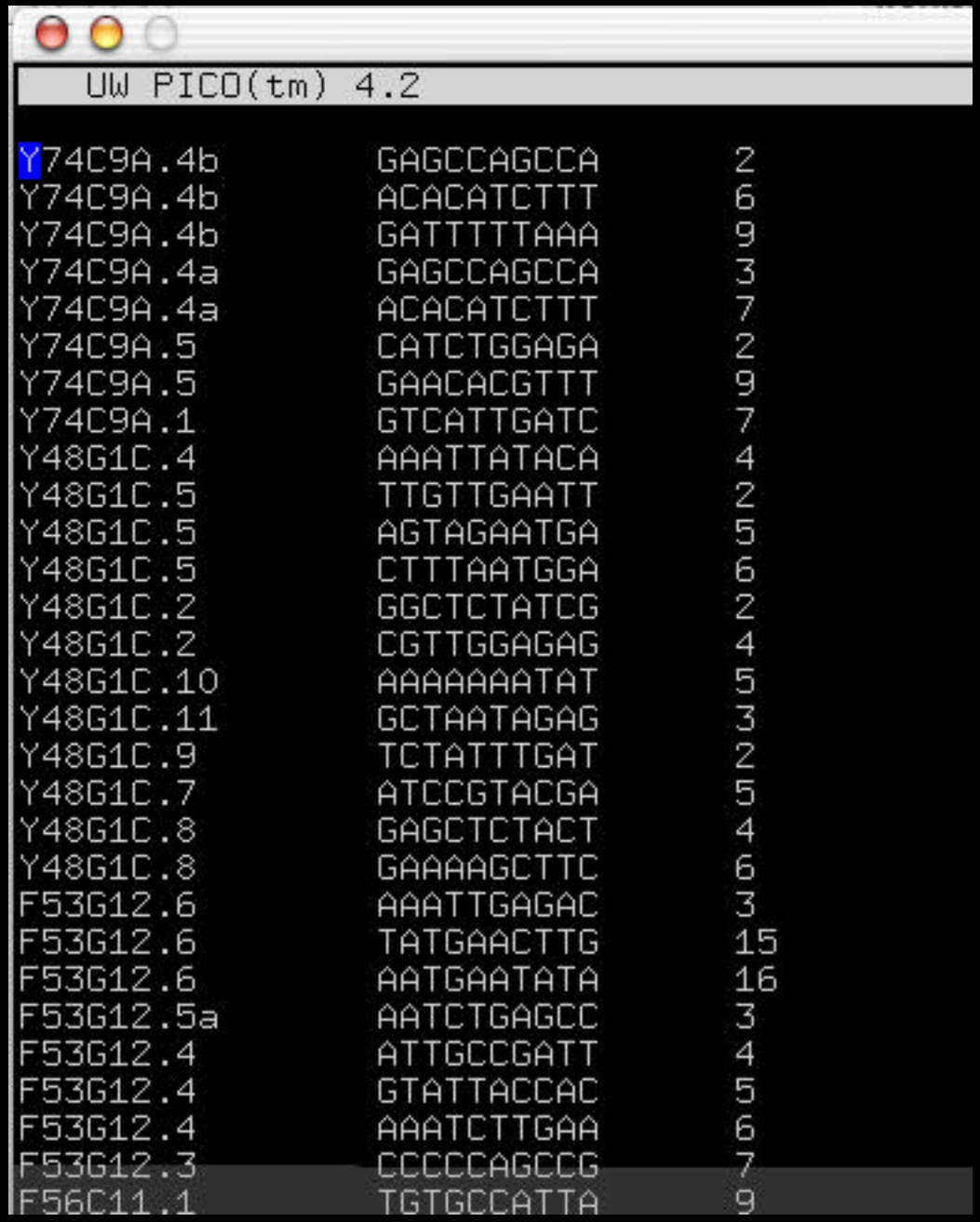

ttyp1 110x40
UW PICO(tm) 4.2
File: meta_lib.txt
11771 TTGTTACCTT coding_RNA 1 + Y37E3.8b . "similar to Ribosomal protein L15" ; codin
8659 TCTTGTGTGG Mitochondrial 1 + cytochrome_c_oxidase_subunit_III . .
7548 GGATTCGGTC coding_RNA 1 + F25H2.10 rpa-0 "deoxyribonuclease"
6686 TGTTGGCAAA coding_RNA 1 + ZK1010.1 ubq-2 "UBQ-2 ubiquitin, 60S Ribosomal protei
5863 GACCACTCAC coding_RNA 1 + F10B5.1 rpl-10 "ribosomal protein L10 (QM protein)"
5847 AATAAACGAA coding_RNA 1 + C32D5.9 lgg-1 "C. elegans LGG-1 protein, similar to M
5652 CCGAATAAAA coding_RNA 1 + Y45F10D.12 rpl-18 "Eukaryotic ribosomal protein L18"
5446 CACAAATCTG coding_RNA 1 + M01F1.2 rpl-16 "L13P family ribosomal protein"
5288 TGACTTCTGA other_RNA 1 + F09E10.11a . "RNA gene of unknown function locus:tts-1
5038 CAACTCAGAA coding_RNA 1 + C08F11.11 . "hypothetical gene model"
5005 AAAAAAAAAA coding_RNA 5 + C46H11.1 . "similar to Homo sapiens Splice isoform 3
4413 CTTGGGCATT Mitochondrial 1 + cytochrome_c_oxidase_subunit_II . .
4170 AAATCGTTAT coding_RNA 1 + R09B3.3 . "RNA recognition motif. (aka RRM, RBD, or R
4124 GAAACAAGAG coding_RNA 1 + H22K11.1 asp-3 "aspartyl protease"
4055 TGCGTTGTCT other_RNA 3 + F31C3.9 . "26s rRNA"
4041 AGACAAACCG coding_RNA 1 + F31E3.5 eft-3 "Elongation factor 1-alpha"
3815 CGGAGAGGGA coding_RNA 1 + Y105E8A.16 rps-20 "rps-20 encodes a small ribosomal s
3803 TCTTCAATCA coding_RNA 2 + B0412.4 rps-29 "40S ribosomal protein S29"
3798 TAACCATTGA coding_RNA 1 + F23D12.1 . "hypothetical gene model"
3741 GGAAAGCCAC coding_RNA 1 + Y71F9AL.13b rpl-1 "C. elegans RPL-1 protein, similar
3646 GGAAAACTCA coding_RNA 1 + F13B10.2 rpl-3 "60S ribosomal protein L3"
3405 CCACATCGAG coding_RNA 1 + Y48G8AL.8b rpl-17 "C. elegans RPL-17 protein, similar
3359 GGAAAGAAGA coding_RNA 1 + C53H9.1 rpl-27 "60S ribosomal protein L27"
3351 AAGTACAATG coding_RNA 1 + C26F1.9 rpl-39 "ribosomal protein L39"
3322 ATACTTATTA Mitochondrial 1 + cytochrome_c_oxidase_subunit_I . .
3213 GATCACGAGG coding_RNA 1 + F53A3.3 rps-22 "40S ribosomal protein"
3114 CCGTTAAATT coding_RNA 1 + K06A4.7 . "hypothetical gene model"
3028 GGTCTACGAA coding_RNA 1 + F25H2.5 . "nucleoside diphosphate kinase"
2979 AAGCGAAAGG coding_RNA 2 + Y48B6A.2 rpl-43 "similar to Ribosomal L37ae protein f
2925 CACCAATAAT coding_RNA 1 + F56H9.2 . "hypothetical gene model"
2879 CCCAACGAGA coding_RNA 1 + Y77E11A.15 col-106 "C. elegans COL-106 protein, simil
2784 GGAAAACTTC coding_RNA 1 + K08B4.6 cli-2 "protease inhibitor"


ttyp1 110x40
UW PICO(tm) 4.2
File: intron_hits
417 AGATCTACTG Genome 5011606 - . . "Chromosome II : F59A6.3"
314 CGGTGATAAA Genome 8247144 + . . "Chromosome III : C02F5.3"
299 TAAGTGTGAA Genome 10481455 + . . "Chromosome X : C33D3.1"
274 ACAATCTGCA Genome 13200837 - . . "Chromosome V : F57B1.3"
274 TCAACTCCTT Genome 11021065 + . . "Chromosome IV : F13H10.4"
256 GTGACGGATA Genome 7865878 + . . "Chromosome I : R11A5.1a R11A5.1
240 GCCTACACAA Genome 13472050 + . . "Chromosome I : Y48G10A.4"
181 TATTAATGTA Genome 1425125 + . . "Chromosome I : Y92H12BR.3"
178 TGAAATGTAC Genome 10111977 - . . "Chromosome III : ZK1128.8a T1
162 TATACTGAGA Genome 7709299 - . . "Chromosome IV : F33D4.5"
154 TGCAAAACTA Genome 11280971 - . . "Chromosome I : F36D1.4"
154 TTGCGTGTCT Genome 12066035 - . . "Chromosome V : R07B7.3"
151 TTCGTGTAGA Genome 7942555 - . . "Chromosome II : ZK669.4"
147 CGGTTCACAC Genome 6876792 - . . "Chromosome IV : T05A12.2"
142 TTGACGGAAT Genome 14455534 + . . "Chromosome V : F23B12.7"
137 ATGATTCTCT Genome 6252672 - . . "Chromosome II : F13H8.5"
134 TATTCATTTA Genome 2178018 - . . "Chromosome I : W03F11.6c W03F11
131 AACCTTCCCT Genome 1077486 - . . "Chromosome IV : Y55F3AM.15"
124 AACCCCTTTT Genome 9071376 - . . "Chromosome I : W06D4.5"
118 TCTTATGAGT Genome 6490837 - . . "Chromosome V : W01A11.4"
114 TAAATATCTG Genome 13594075 - . . "Chromosome X : F54B11.4"
108 GAGGCACGTT Genome 10801186 + . . "Chromosome I : F55A3.3"
107 GGTTTATGTA Genome 4253104 + . . "Chromosome I : C43E11.11"
107 TATTGCGTGT Genome 10941905 - . . "Chromosome I : F49D11.9a"
104 TGCCTTTACG Genome 11994132 + . . "Chromosome III : Y56A3A.32"
101 TACAGAGCAA Genome 20845784 - . . "Chromosome V : F31D4.4"
95 AAAGATTTCG Genome 5598399 + . . "Chromosome II : C17G10.1"
93 TGATACTGAG Genome 8313259 + . . "Chromosome V : W02D7.2"


ttyp1 110x40
{xhost01}~/libstats> ./tagger.pl virtual_transcriptome >possible_polyA &
[1] 8421
{xhost01}~/libstats>
There were a total of 137211 potential tags
26688 may be primed by internal poly_A tracts
[1]+ Done ./tagger.pl virtual_transcriptome >possible_polyA
{xhost01}~/libstats>

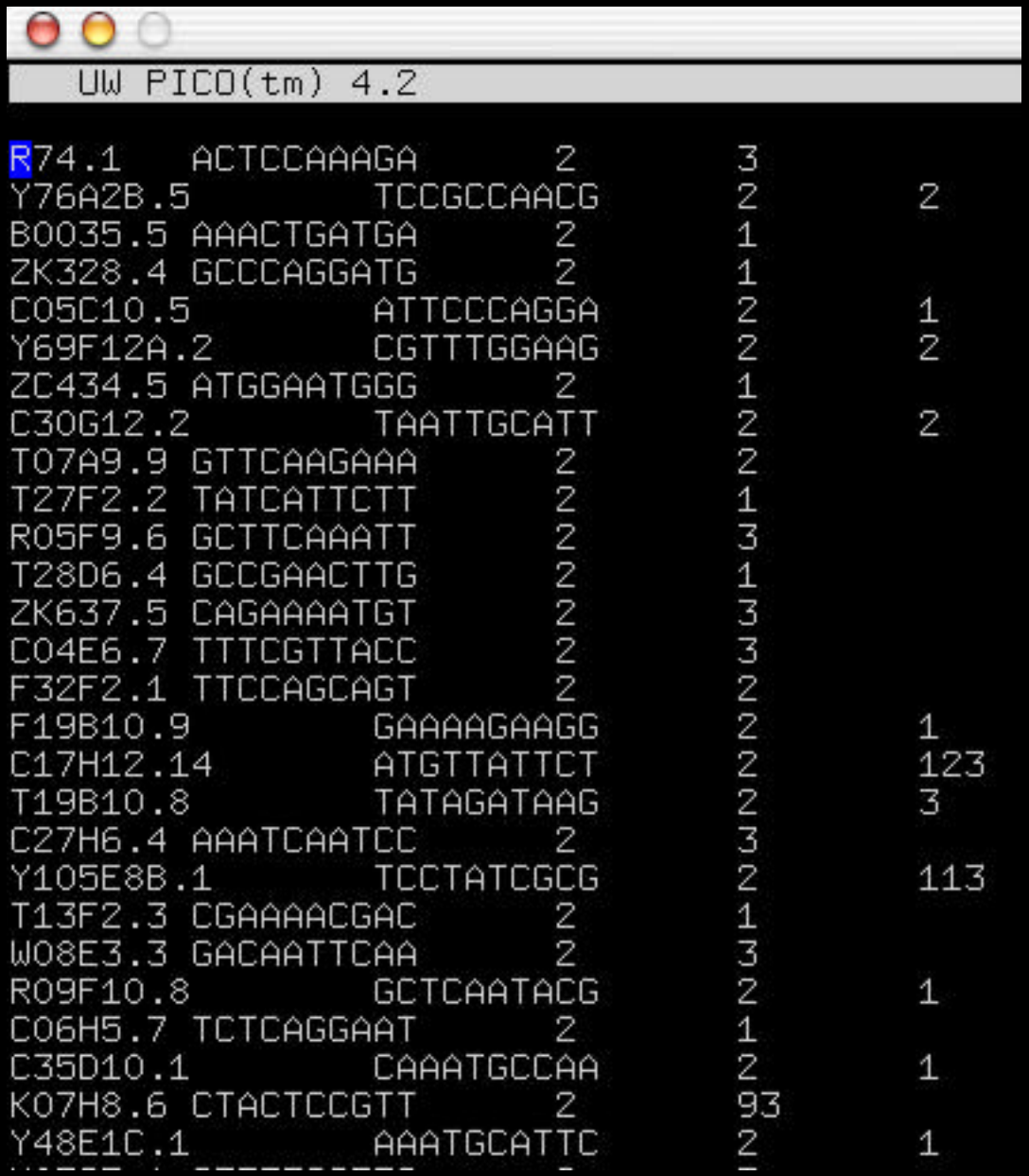

ttyp1 110x40
F10 key ==> File Edit Search Buffers Windows System Help
#!/usr/local/bin/perl -w
file tagger.pl -- extract potential SAGE tags from the
virtual transcriptome, then
1) count the total number of potential tags
2) identify and list all tags that could result from internal polyA tracts
#
The raw code -- see below for detailed comments
#
use strict;
my $file = shift or die "Usage: ./tagger.pl input_file\n";
die "File $file not found\n" unless -e $file;
my $lines = `cat $file`;
my @lines = split '>', $lines;
shift @lines;
my $polyA = 0;
my $total = 0;
for (@lines) {
my @rows = split "\n", $_;
my ($gene) = $rows[0] =~ /^(\S+)/;
shift @rows;
my $seq = uc join '', @rows;
my @frags = reverse split 'CATG', $seq;
pop @frags;
$total += @frags;
my $pos = 0;
for my $frag (@frags) {
my $tag = substr $frag, 0, 10;
next unless ++$pos > 1;

```
    for my $frag ( @frags ) {
        my $tag = substr $frag, 0, 10;
        next unless ++$pos  > 1;

        while ( $frag =~ /(\w{10})/g ) {
            pos $frag -= 9;

            if ( polyA($1) ) {
                $polyA++;
                print "$gene\t$tag\t$pos\n";
                last;
            }
        }
    }
}

warn "\n\nThere were a total of $total potential tags\n",
     "$polyA may be primed by internal poly_A tracts\n\n";

sub polyA {
    $_ = shift;
    return tr/A/A/ > 7 ? 1 : 0
}
```


UW PICO(tm) 4.2
Y74C9A.4b GAGCCAGCCA 2
Y74C9A.4b ACACATCTTT 6
Y74C9A.4b GATTTTTAAA 9
Y74C9A.4a GAGCCAGCCA 3
Y74C9A.4a ACACATCTTT 7
Y74C9A.5 CATCTGGAGA 2
Y74C9A.5 GAACACGTTT 9
Y74C9A.1 GTCATTGATC 7
Y48G1C.4 AAATTATACA 4
Y48G1C.5 TTGTTGAATT 2
Y48G1C.5 AGTAGAATGA 5
Y48G1C.5 CTTTAATGGA 6
Y48G1C.2 GGCTCTATCG 2
Y48G1C.2 CGTTGGAGAG 4
Y48G1C.10 AAAAAAATAT 5
Y48G1C.11 GCTAATAGAG 3
Y48G1C.9 TCTATTTGAT 2
Y48G1C.7 ATCCGTACGA 5
Y48G1C.8 GAGCTCTACT 4
Y48G1C.8 GAAAAGCTTC 6
F53G12.6 AAATTGAGAC 3
F53G12.6 TATGAACTTG 15
F53G12.6 AATGAATATA 16
F53G12.5a AATCTGAGCC 3
F53G12.4 ATTGCCGATT 4
F53G12.4 GTATTACCAC 5
F53G12.4 AAATCTTGAA 6
F53G12.3 CCCCCAGCCG 7
F56C11.1 TGTGCCATTA 9

```perl
#!/usr/local/bin/perl -w


# file partial.pl -- Estimate how many observed SAGE tags could be
# due to partial digestion with the NlaIII anchoring enzyme


#
# The raw code -- see below for detailed comments
#

use strict;

my $file = shift or die "Usage: ./partial.pl input_file\n";
die "File $file not found\n" unless -e $file;
my @file = sort_by_position( $file );

my ( $count, %seen, %seenpos1, $maybepartial, %pos1_freq );

for (@file) {
    my @field  = split;
    my $gene   = $field[5];
    my $pos    = $field[3];
    my $tcount = $field[0];
    my $tag    = $field[1];

    $gene =~ s/[a-z]$//;
    $count++ unless $seen{$gene};
    $seen{$gene} = 1;
    $seenpos1{$gene}  ||= $pos;
    $pos1_freq{$gene} ||= $tcount;
```

```
    if ( $pos == ( $seenpos1{$gene} + 1 ) &&
         $tcount <= ( $pos1_freq{$gene}/10 ) ) {
        $maybepartial++;
        print "$gene\t$tag\t$pos\t$tcount\n";
    }
}


$file =~ s/\.txt//;

warn "\n\n$file has $count unambiguously mapped genes\n",
     "There were a total of ", scalar( @file ), " tags mapped\n",
     "There were $maybepartial possible NlaIII partials\n\n";


sub sort_by_position {
    my $lib = shift;
    my @file = `grep coding_RNA $lib`;

    return map  { $_->[1] }
           sort { $a->[0] <=> $b->[0] }
           map  { [ (split)[3], $_ ] }
           @file;
}
```


UW PICO(tm) 4.2
R74.1 ACTCCAAAGA 2 3
Y76A2B.5 TCCGCCAACG 2 2
B0035.5 AAACTGATGA 2 1
ZK328.4 GCCCAGGATG 2 1
C05C10.5 ATTCCCAGGA 2 1
Y69F12A.2 CGTTTGGAAG 2 2
ZC434.5 ATGGAATGGG 2 1
C30G12.2 TAATTGCATT 2 2
T07A9.9 GTTCAAGAAA 2 2
T27F2.2 TATCATTCTT 2 1
R05F9.6 GCTTCAAATT 2 3
T28D6.4 GCCGAACTTG 2 1
ZK637.5 CAGAAAATGT 2 3
C04E6.7 TTTCGTTACC 2 3
F32F2.1 TTCCAGCAGT 2 2
F19B10.9 GAAAAGAAGG 2 1
C17H12.14 ATGTTATTCT 2 123
T19B10.8 TATAGATAAG 2 3
C27H6.4 AAATCAATCC 2 3
Y105E8B.1 TCCTATCGCG 2 113
T13F2.3 CGAAAACGAC 2 1
W08E3.3 GACAATTCAA 2 3
R09F10.8 GCTCAATACG 2 1
C06H5.7 TCTCAGGAAT 2 1
C35D10.1 CAAATGCCAA 2 1
K07H8.6 CTACTCCGTT 2 93
Y48E1C.1 AAATGCATTC 2 1

```
#!/usr/local/bin/perl -w


#
# file altsplice.pl -- count the number of tags/gene and relate
# to the tag abundance
#

use strict;
my $file = shift or die "Usage: ./altsplice.pl input_file\n";
die "File $file not found\n" unless -e $file;

my $count   = 0;

# initialize a hash reference
my $tags    = {};

# load data from external files
my @file    = `cat $file`;
my $partial = `cat NlaIII_partial`;
my $polyA   = `cat possible_polyA`;
my @introns = `cat intron_hits`;

# add introns to the other tags
push @file, @introns;


# iterate through the list
for (@file) {
    # split the line into 'words'
    my @line  = split;

    # this is what a line would typically look like:
    # count tag source pos'n strand gene locus descrption
```

```perl
# iterate through the list
for (@file) {
    # split the line into 'words'
    my @line  = split;

    # this is what a line would typically look like:
    # count tag source pos'n strand gene locus descrption

    # assign some variables
    my $freq  = $line[0];
    my $tag   = $line[1];
    my $pos   = $line[3];
    my $gene  = $line[5];

    # if no gene, it's an intron hit -- find the gene name
    # at the end of the line
    if ( !$gene ) {
        ($gene) = /(\S+)"$/;
    }

    # remove alternative splice suffix
    $gene =~ s/[a-z]$//;

    # intialize an array reference for the gene if we do not have one
    $tags->{$gene} ||= [];

    # skip tags that may be due to internal polyA's or
    # NlaIII partial digestion
    next if reject( $gene, $tag );

    # add the tag's count to the array
    push @{$tags->{$gene}}, $freq;
}
```

```
# this subroutine evaluates each gene/tag pair to see if
# they were previosuly identified as potential artifacts
sub reject {
    my ($gene, $tag) = @_;

    # are tag and gene in the partial digest file?
    return 1 if $partial =~ /^${gene}[a-z]?\s+$tag/m;

    # are tag and gene in the intternal polyA file?
    return 1 if $polyA   =~ /^${gene}[a-z]?\s+$tag/m;

    return 0;
}
```

```perl
# initialize some hashes we will need
my (%count, %total);


for ( sort keys %{$tags} ) {

    # get the list of tag counts for this gene
    my @tags = @{$tags->{$_}};

    # set the counter to zero
    my $sum  = 0;

    # add up all the tag counts
    for my $f (@tags) {
       $sum += $f;
    }

    # count the number of genes in each (number of tags) category
    for my $num ( 1..10 ) {
       $count{$num}++ and $total{$num} += $sum if @tags == $num;
    }
    $count{11}++ and $total{11} += $sum if @tags   >  10;
}

print "\nLibrary $file: ", scalar( keys %{$tags} ), " genes detected\n";
print "\nTranscripts\tGenes\tAbundance (average)\n";

for ( 1..11 ) {
    next unless $total{$_} ;
    my $average = int( $total{$_}/$count{$_} + 0.5 );
    my $transcripts = $_ == 11 ? '>10' : $_;
    print "$transcripts\t\t$count{$_}\t\t$average\n";
}
```

```
{xhost01}~/libstats> ./altsplice.pl meta_lib.txt

Library meta_lib.txt: 14680 genes detected

Transcripts      Genes   Abundance (average)
1                8446                 32
2                3540                 67
3                1205                114
4                385                 106
5                135                 234
6                46                  641
7                16                  210
8                8                   286
9                5                   399
>10              3                   799
{xhost01}~/libstats>
```