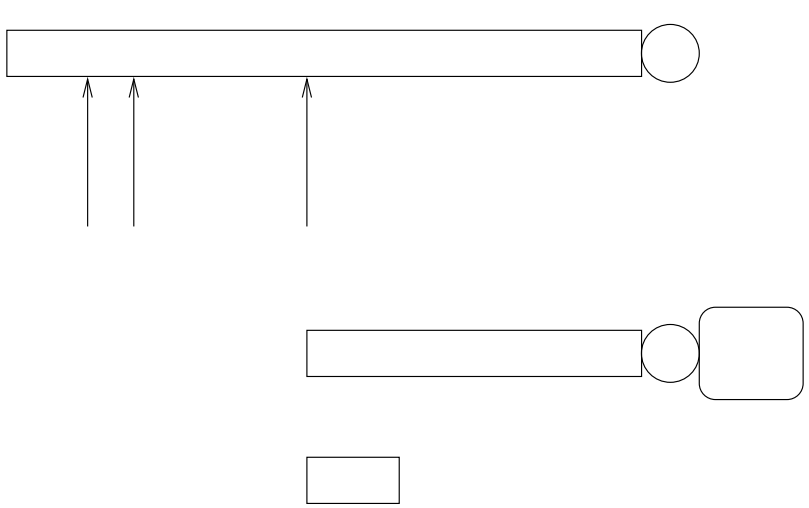
SAGE — Serial Analysis of Gene Expression

1. Create cDNA sequences from poly−A+ RNA with biotinylated ends
2. Cut at every CATG (Nla3 site)
3. Isolate 3'−most fragments using streptavidin beads
4. Chop off the CATG plus next 10 nucleotides (SAGE tags)
5. Sequence the tags



SAGE Output

For a given cDNA library, e.g., from a certain tissue or tumor, we get a list of how many times each of the possible 4^{10} 10-mer tags was found:

AAAAAAAAAA	14
AAAAAAAAAC	1074
AAAAAAAAAG	0
...	
TTTTTTTTTT	192

Frequently, we want to know how many mRNA molecules for a particular **gene** are in a certain cell type. How do we map SAGE tags onto known genes?

Mapping SAGE Tags to Genes

For each known gene sequence, *which must include all of the* $3'$ *UTR*, find the $3'$-most *Nla3* restriction site (CATG). The 10 bp immediately $3'$ of that site is the gene's SAGE tag.

What can go wrong with (a) determining tag counts and (b) mapping tags to genes?

What Can Go Wrong?

Problems with determining tag counts:

- A gene might not have a *Nla3* site (CATG).
- Sequencing errors create incorrect tags.

Problems with mapping tags to genes:

- Two different genes can have the same SAGE tag.
- A gene can have two different tags.

Also, there are the usual experimental pitfalls, e.g., internal poly-A runs and incomplete *Nla3* digestion.

What Can Go Wrong?

Problems with determining tag counts:

- **A gene might not have a CATG. How often? How will it affect the results?**
- Sequencing errors create incorrect tags.

Problems with mapping tags to genes:

- Two different genes can have the same SAGE tag.
- A gene can have two different tags.

Genes With No CATG.

Frequency:

The average length of a mammalian mRNA, including UTRs, might be around 2-3 kilobases.

On average, there will be a CATG every 256 bp.

Conclusion: The vast majority of genes will have a CATG site.

Effect on our results:

Some genes will be missed, particularly very short genes.

What Can Go Wrong?

Problems with determining tag counts:

- A gene might not have a *Nla3* site (CATG).
- **Sequencing errors create incorrect tags. How often? How will it affect our results?**

Problems with mapping tags to genes:

- Two different genes can have the same SAGE tag.
- A gene can have two different tags.

Frequency of Erroneous SAGE Tags

Typically, SAGE tags are generated by single-pass sequencing, i.e., DNA isn't re-sequenced to improve accuracy. Hence, about 1%-3% of the nucleotides are determined incorrectly. This means that about 1 of 10 (at 1% error) or 1 of 4 (at 3% error) tags is erroneous.

If an erroneous tag is the legitimate tag for a real gene, that gene's mRNA count will be artificially inflated. If it is not a tag for a real gene, then we'll be lead to think that some bogus gene is being expressed.

For an incorrectly determined SAGE tag, what is the probability that it will not represent a real gene? Will that happen more frequently than the case where a real gene's count is erroneously raised by 1?

Generally, Sequencing Errors Create Bogus Tags

$4^{10} \sim 10^6$, i.e., there are about a million possible SAGE tags.

The number of genes in a mammalian genome is about 50,000. So the odds are about 20-to-1 that an erroneous tag will not be the tag for a real gene.

Suppose that a given tag, say ACCGTATAGC, is not the tag for a real gene that is expressed in the cell type under investigation. How many bogus copies of that tag can we expect to arise from sequencing errors?

Generally, A Given Bogus Tag Will Be Rare

This depends on how many tags are sequenced. A typical mammalian cell might contain 300,000 mRNA molecules. Tags might be sampled at “0.5X”, i.e., 150,000 tags might be sequenced. (This is sampling with replacement: the tag from a given mRNA molecule can be determined several times.)

Assuming 1%/bp sequencing errors, 15,000 of these tags might be incorrectly sequenced. Since only, say, 30,0000 of the possible 1 million tags are actually expressed in that cell type, each bogus tag has probability about 15,000/1,000,000 = 0.015 of appearing to be expressed.

Effect of Erroneous SAGE Tags

It all depends on what we're trying to do. Suppose we want to accurately count the number of genes being expressed in a genome that is largely unknown (e.g., a starfish). Suppose we sequence 100,000 tags, 10% of which are erroneous. It might not be easy to distinguish very low expressing genes from bogus tags.

Now suppose we're measuring expression of a well characterized set of genes. For highly or moderately expressed tags, the errors won't be much of a problem, though our counts may be low by 10-25%.

What Can Go Wrong?

Problems with determining tag counts:

- A gene might not have a *Nla3* site (CATG).
- Sequencing errors create incorrect tags.

Problems with mapping tags to genes:

- **Two different genes can have the same SAGE tag. How frequent is it? How will it affect our results?**
- A gene can have two different tags.

Two Genes With The Same SAGE Tag

Given a tag for a gene, what is the probability that it is also the tag for another gene? The reasoning is the same as before: there are about 1 million possible tags and about 50,000 genes, so the probability is roughly 0.05.

Frequently, we'll want to simply discard tags that map to more than one gene. In theory, this means losing less than 5% of the genes. But see the real-life example, below.

What Can Go Wrong?

Problems with determining tag counts:

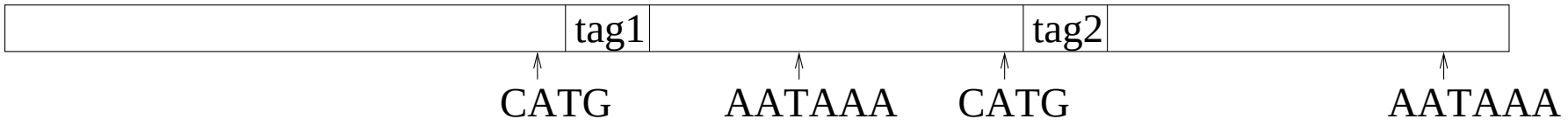
- A gene might not have a *Nla3* site (CATG).
- Sequencing errors create incorrect tags.

Problems with mapping tags to genes:

- Two different genes can have the same SAGE tag.
- **A gene can have two different tags. How can this happen? How frequent is it? How will it affect our results?**

One Gene With Several SAGE Tags

Alternative polyadenylation sites can lead to multiple SAGE tags. Consider the following $3'$ UTR:



One paper estimates that this affects 3% of all genes. We need to be aware that tag-to-gene mappings can produce duplicates.

A Real-Life Example

I am analyzing 1183 $\langle SAGE\ tag, mRNA\ count \rangle$ pairs for my BIO/CSE/STAT 598F final project. These tags were observed by Velculescu *et al.* (1999) in all of the 19 sampled tissues (both normal tissues and tumors). That is, this is the putative "housekeeping transcriptome".

AAAAAAAAAA	469	Enolase 1, (alpha)
AAAAAAAAAA	469	GFRA1
AAAACATTCT	306	Tag matches mitochondrial sequence
AAAAGAAACT	62	Polyadenylate-binding protein
AAACCAGGGC	18	ESTs
AAACTCTGTG	27	CTNND1
...		
AACCCAGGAG	110	No match
...		

Processing the SAGE Data

The first step in my project was to convert this to a list of $\langle gene\ name, mRNA\ count \rangle$ pairs.

How many of these 1183 tags don't belong to a real human gene, because of sequencing error? Basically, none. Each tag was observed at least 5 times in each of the 19 tissues. (Recall our estimation that a given bogus tag has probability 0.015 of being observed.)

Thus we can assume that each of these tags belongs to a real gene, However, how many will belong to *two* genes or to an unknown gene?

Unique Tags for Known Genes

Earlier, we argued on theoretical grounds that just a few percent of all tags belonged to several genes, *assuming each gene has a unique tag*.

Many of the 1183 $\langle SAGE\ tag, mRNA\ count \rangle$ pairs will be discarded because they don't map to named genes. There are roughly 15,000 named human genes, and at least that many unnamed genes. On the other hand, housekeeping genes are probably better known than average (because they're expressed everywhere), so a random selection of 1000 (both already known and currently uncharacterized) housekeeping genes might have, say, 600-700 known members.

Mapping SAGE Tags To UniGene IDs

The National Center for Biotechnology Information (NCBI) maintains *SAGEmap*, which includes a mapping of SAGE tags to UniGene clusters. I downloaded this information, *which contains 272131 tag-ID pairs!*

Because of the way NCBI predicts tags for UniGene clusters, a cluster can have quite a few putative tags. (One cluster that I looked at had 70 putative tags.) Hence, the frequency with which a tag is assigned to more then one sequence is much higher than our previous estimation that 5%. Indeed, when I discarded tags that weren't mapped to one and only one cluster, I was left with only 722 UniGene-mRNAcount pairs.

Mapping UniGene IDs to Known Genes

NCBI maintains information about each UniGene cluster, including gene name whenever it is known. I downloaded this huge file (49 Mb compressed), and extracted 15,311 ID-gene pairs. Merging with the 722 UniGene-mRNAcount pairs gave 697 gene-mRNAcount pairs. In some cases, the same gene name appeared in several pairs. (Recall that a gene can have several tags and several UniGene clusters.) I removed duplicates, retaining the larger mRNA count if there was a difference. This left me with 617 gene-mRNAcount pairs.