Resolving Isoform Expression using Digital Gene Expression Data

Naomi Altman

Joint work with:

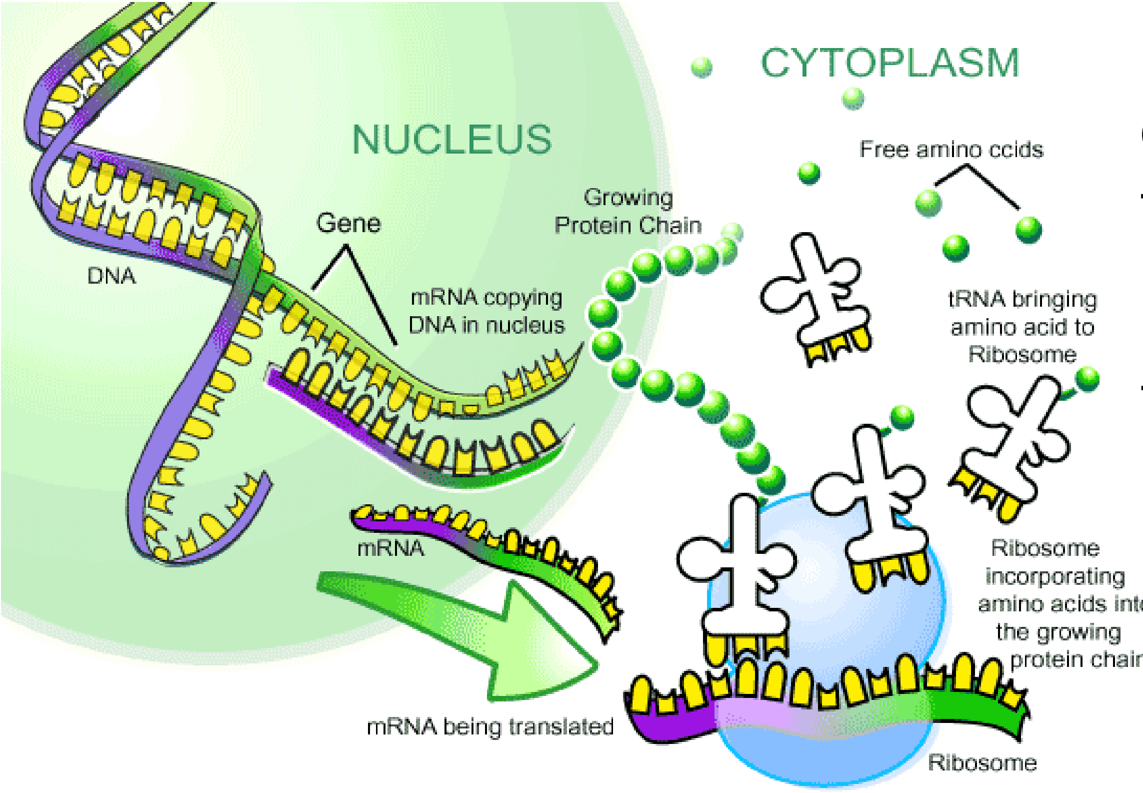
Q. Wang, V. Karwa, A. Slavkovic

Penn State University

presented: April 30, 2010

DASF III

Why Study Gene Expression?

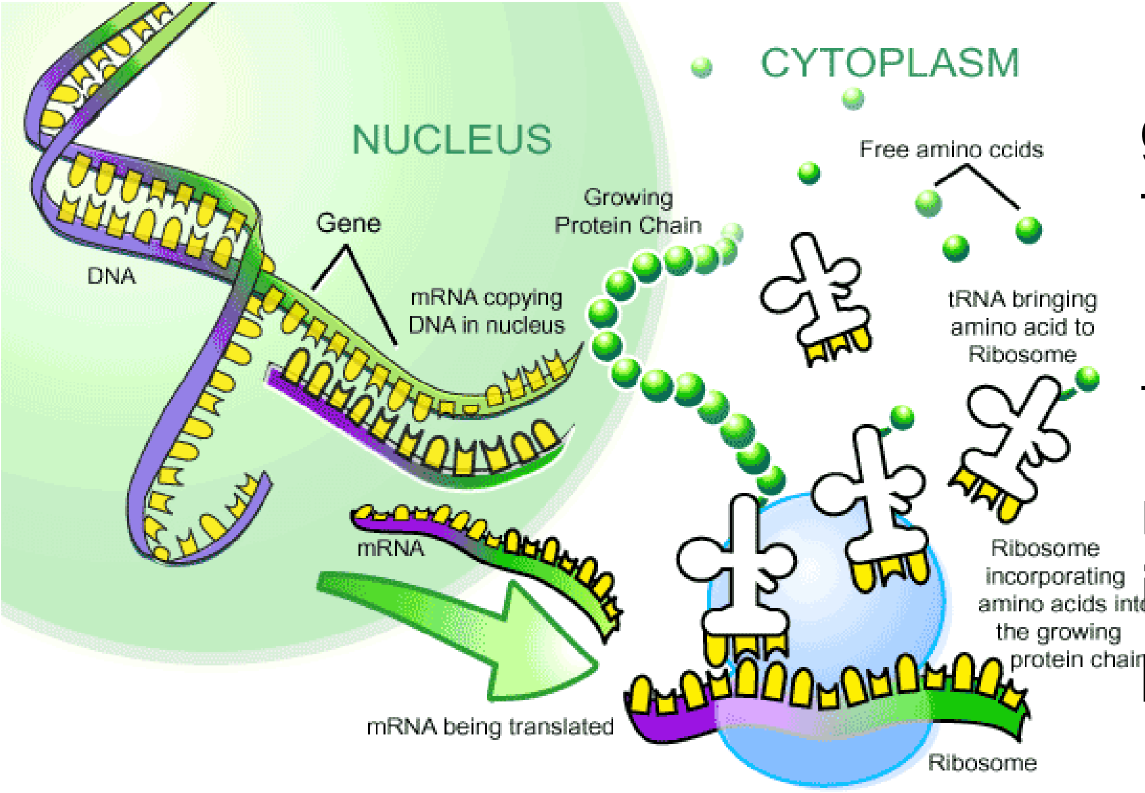

CYTOPLASM
NUCLEUS
Free amino ccids
Growing Protein Chain
Gene
DNA
mRNA copying DNA in nucleus
tRNA bringing amino acid to Ribosome
mRNA
Ribosome incorporating amino acids int the growing protein chai
mRNA being translated
Ribosome

Proteins are the product of gene expression through the intermediary of mRNA.

Each protein type comes from a unique mRNA.

mRNA is much easier to identify and quantify than proteins.

Why Study Gene Expression?


CYTOPLASM
NUCLEUS
Free amino ccids
Growing Protein Chain
Gene
DNA
mRNA copying DNA in nucleus
tRNA bringing amino acid to Ribosome
mRNA
Ribosome incorporating amino acids into the growing protein chain
mRNA being translated
Ribosome

Proteins are the product of gene expression through the intermediary of mRNA.

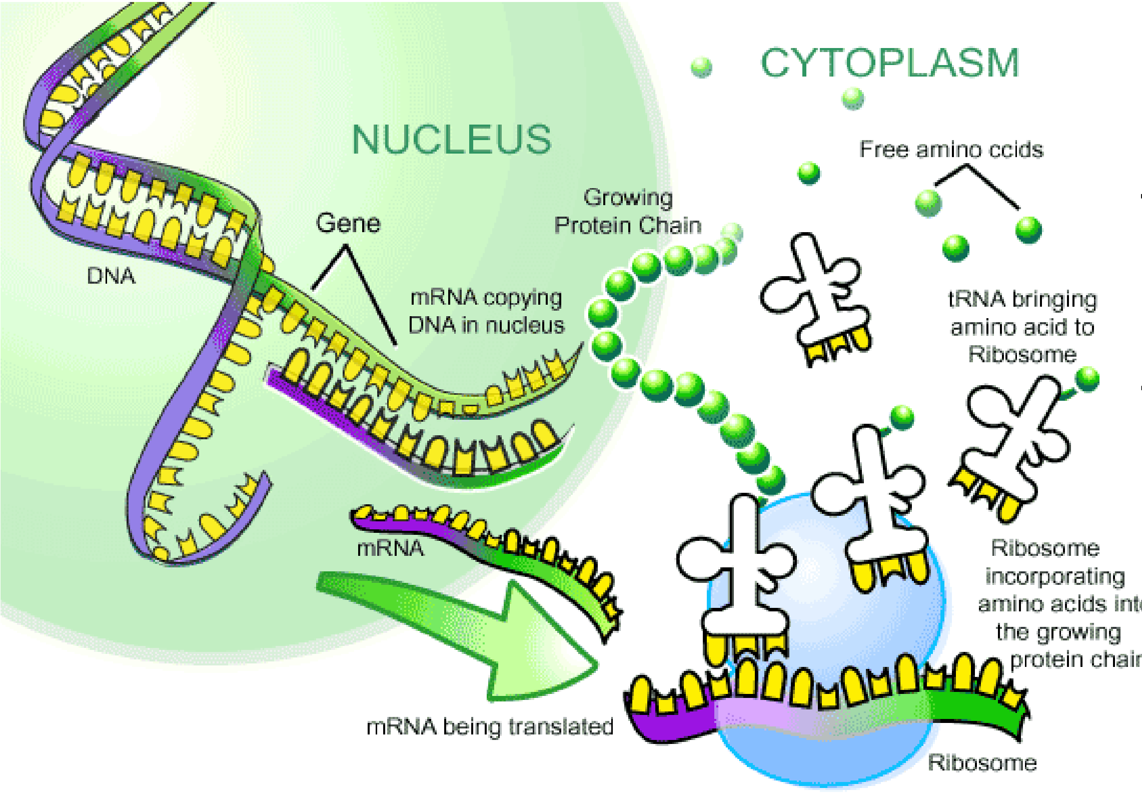
Each protein type comes from a unique mRNA.

mRNA is much easier to identify and quantify than proteins.

BUT ...

The correspondence among genes, mRNA and proteins is more complex than we imagined only a few years ago.

Why Study Gene Expression?


CYTOPLASM
NUCLEUS
Free amino ccids
Growing Protein Chain
Gene
DNA
mRNA copying DNA in nucleus
tRNA bringing amino acid to Ribosome
mRNA
Ribosome incorporating amino acids into the growing protein chain
mRNA being translated
Ribosome

Proteins are the product of gene expression through the intermediary of mRNA.

Each protein type comes from a unique mRNA.

mRNA is much easier to identify and quantify than proteins.

BUT ...

The correspondence among genes, mRNA and proteins is more complex than we imagined only a few years ago.

New technologies for measuring mRNA can improve on microarrays in providing measurements that are closer to quantifying protein expression.

Outline

- Biology of protein expression
- Massively parallel sequencing technologies
- Digital gene expression (DGE) and RNA-seq
- Statistics
- Example
- Simulation
- Closing comments

- **Protein expression**
- Massively parallel sequencing
- DGE and RNA-seq
- Statistics
- Example
- Simulation
- Closing comments

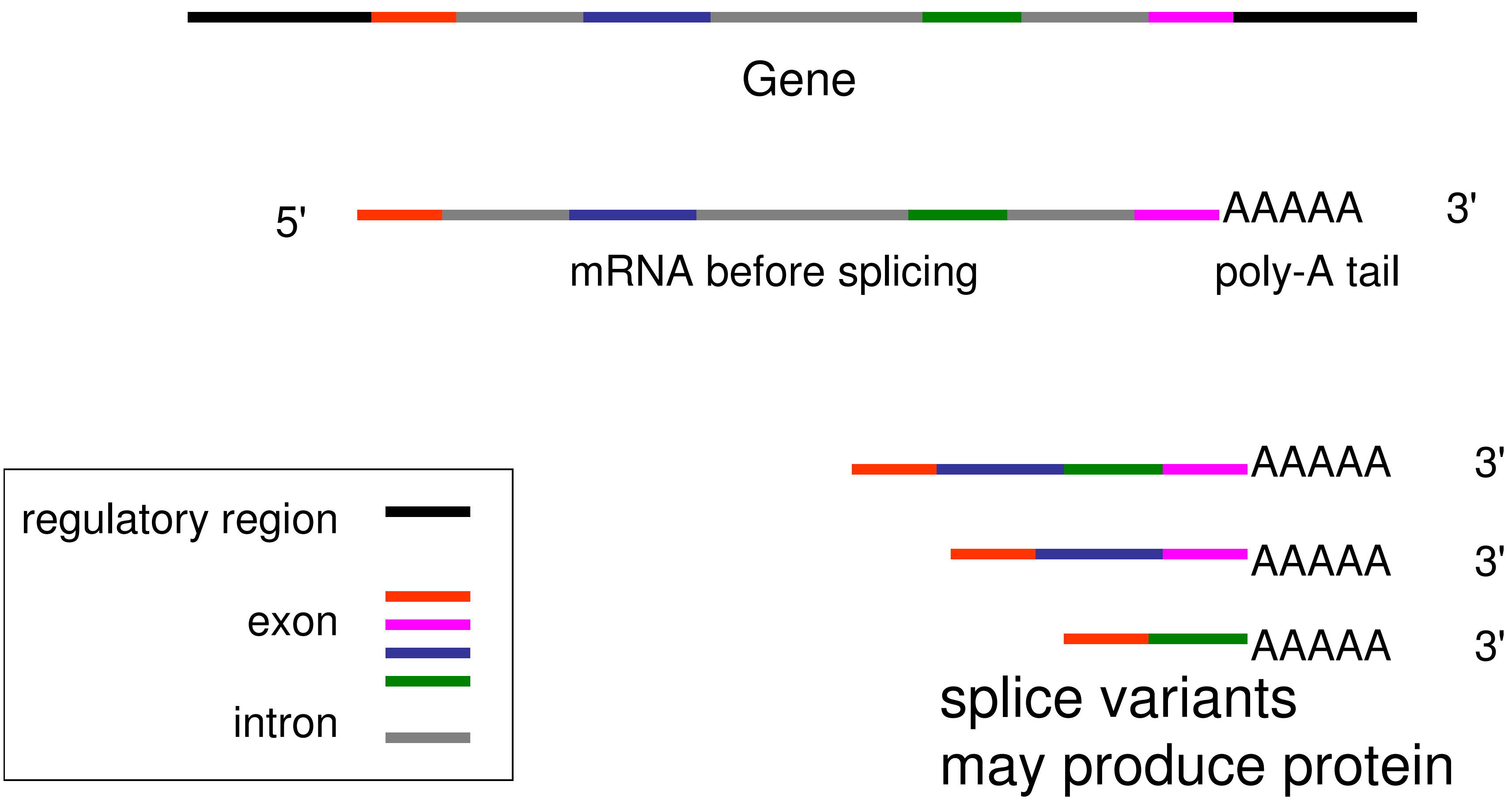
Biology of Protein Expression

Vocabulary

- transcription - gene creates mRNA
- transcript – an mRNA transcribed from a gene
- translation - mRNA creates protein
- exon - pieces of gene which may be transcribed
- intron - pieces of gene which are not transcribed
- poly-A tail - a string of "A" bases at the end of an RNA marking it as mRNA

- **Protein expression**
- Massively parallel sequencing
- DGE and RNA-seq
- Statistics
- Example
- Simulation
- Closing comments

Biology of Protein Expression



Gene

5' AAAAA 3'

mRNA before splicing

poly-A tail

AAAAA 3'

AAAAA 3'

AAAAA 3'

regulatory region

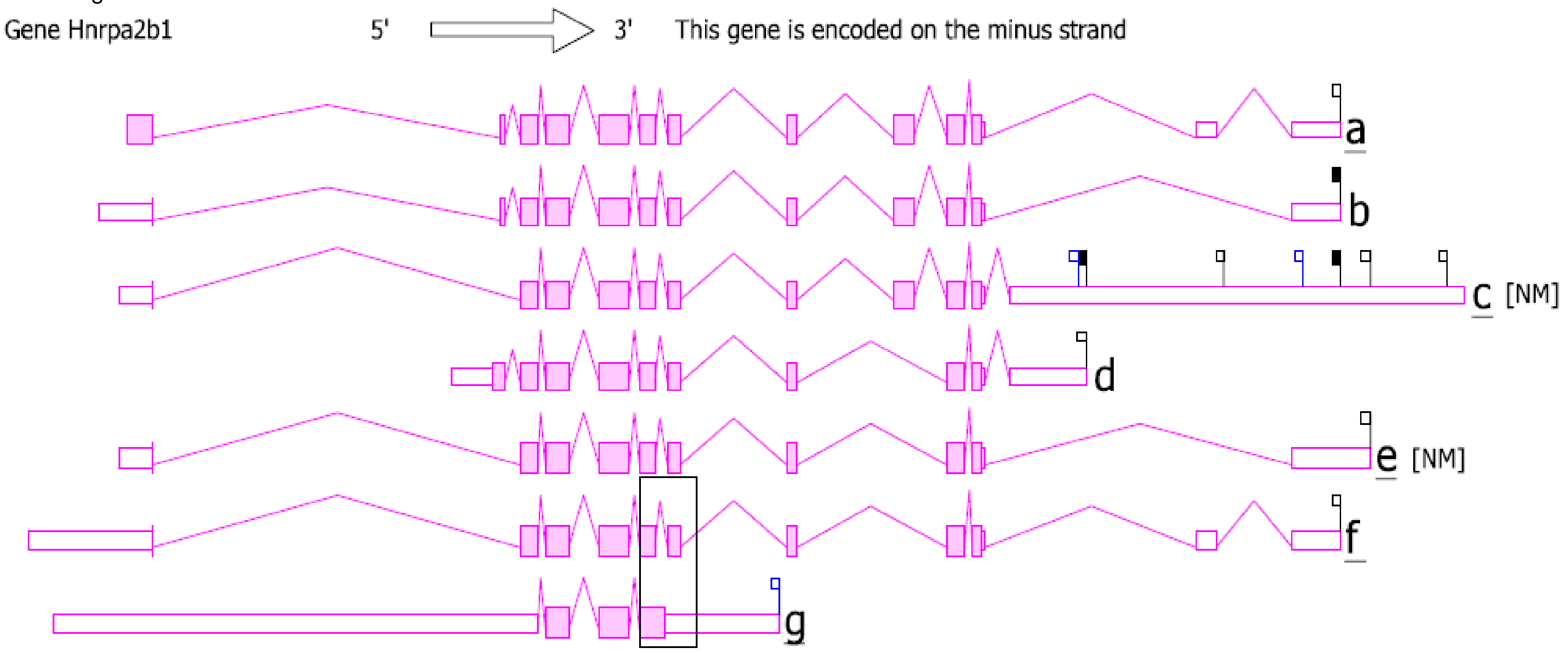
exon

intron

splice variants
may produce protein
isoforms

- **Protein expression**
- Massively parallel sequencing
- DGE and RNA-seq
- Statistics
- Example
- Simulation
- Closing comments

Biology of Protein Expression



Gene Hnrpa2b1 5' 3' This gene is encoded on the minus strand

Some splice Variants for Hnrpa2b1 from Aceview

Note the complexity: alternative poly-A sites

- •possible inclusion of intronic regions
- •alternative exon size

http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/av.cgi?exdb=AceView&db=mouse&term=mm.155896&submit=Go

Our Objective

Quantify the relative expression levels of each isoform in a sample of mRNA.

Our Objective

Quantify the relative expression levels of each isoform in a sample of mRNA.

How can we identify and quantify the expression level?

Microarrays - allow mRNA to bind to complement on substrate
- need to know what to place on the substrate

Sequencing - read the genetic sequence of the mRNA
- expensive to obtain "full-length" sequences

Massively Parallel Sequencing Technologies

New sequencing technologies can sequence 1 - 20 million short fragments of RNA per sample.

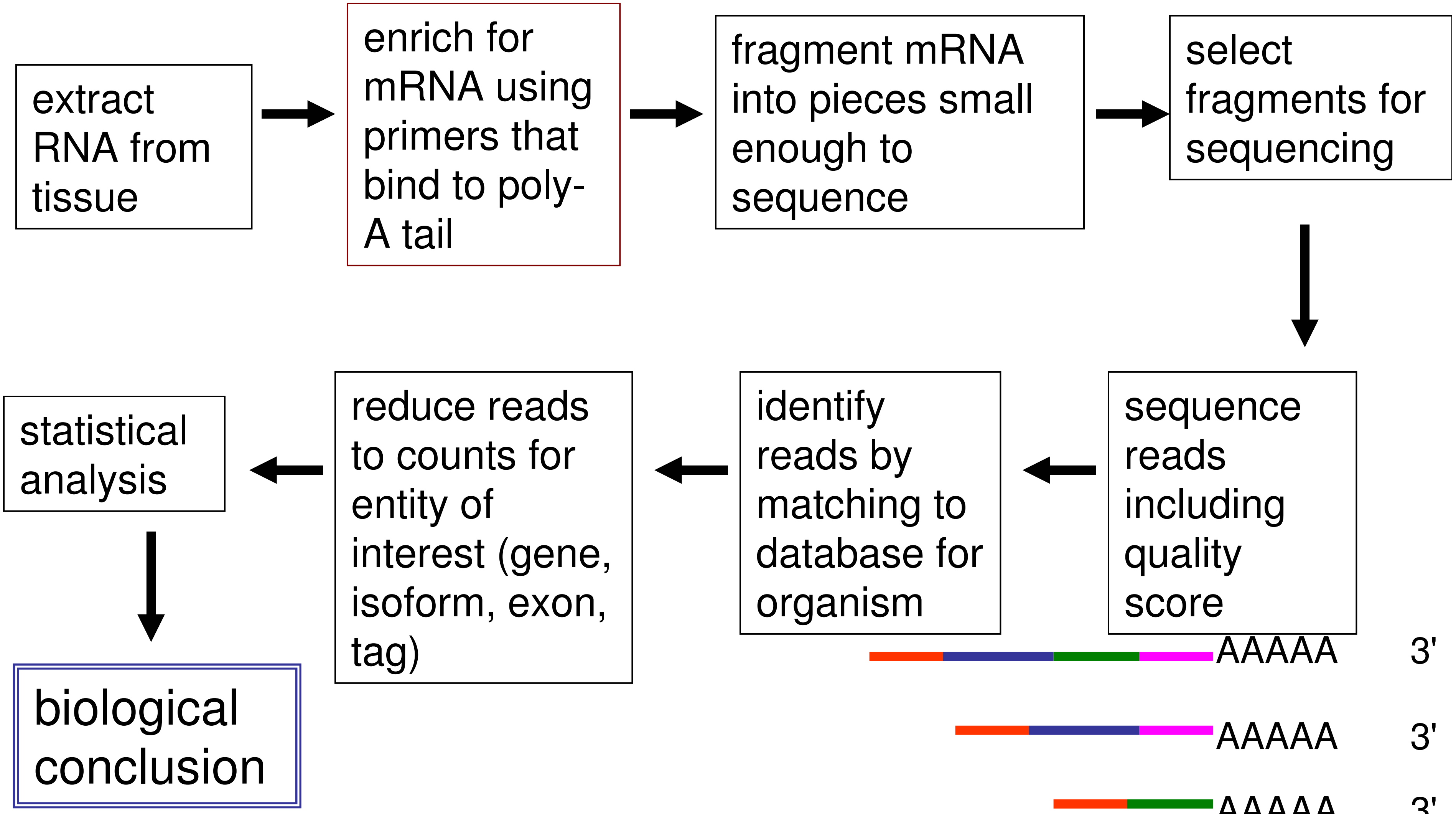
Some common brand names -

SOLiD	17 - 35 bases
Illumina (Solexa)	17 - 100 bases
454	200 - 500 bases

Between methods - short is cheaper (?)(per base) than long
Within method - short is cheaper (per mRNA) than long

- **Protein expression**
- Massively parallel sequencing
- DGE and RNA-seq
- Statistics
- Example
- Simulation
- Closing comments

Massively Parallel Sequencing Technologies



extract RNA from tissue

enrich for mRNA using primers that bind to poly-A tail

fragment mRNA into pieces small enough to sequence

select fragments for sequencing

sequence reads including quality score

identify reads by matching to database for organism

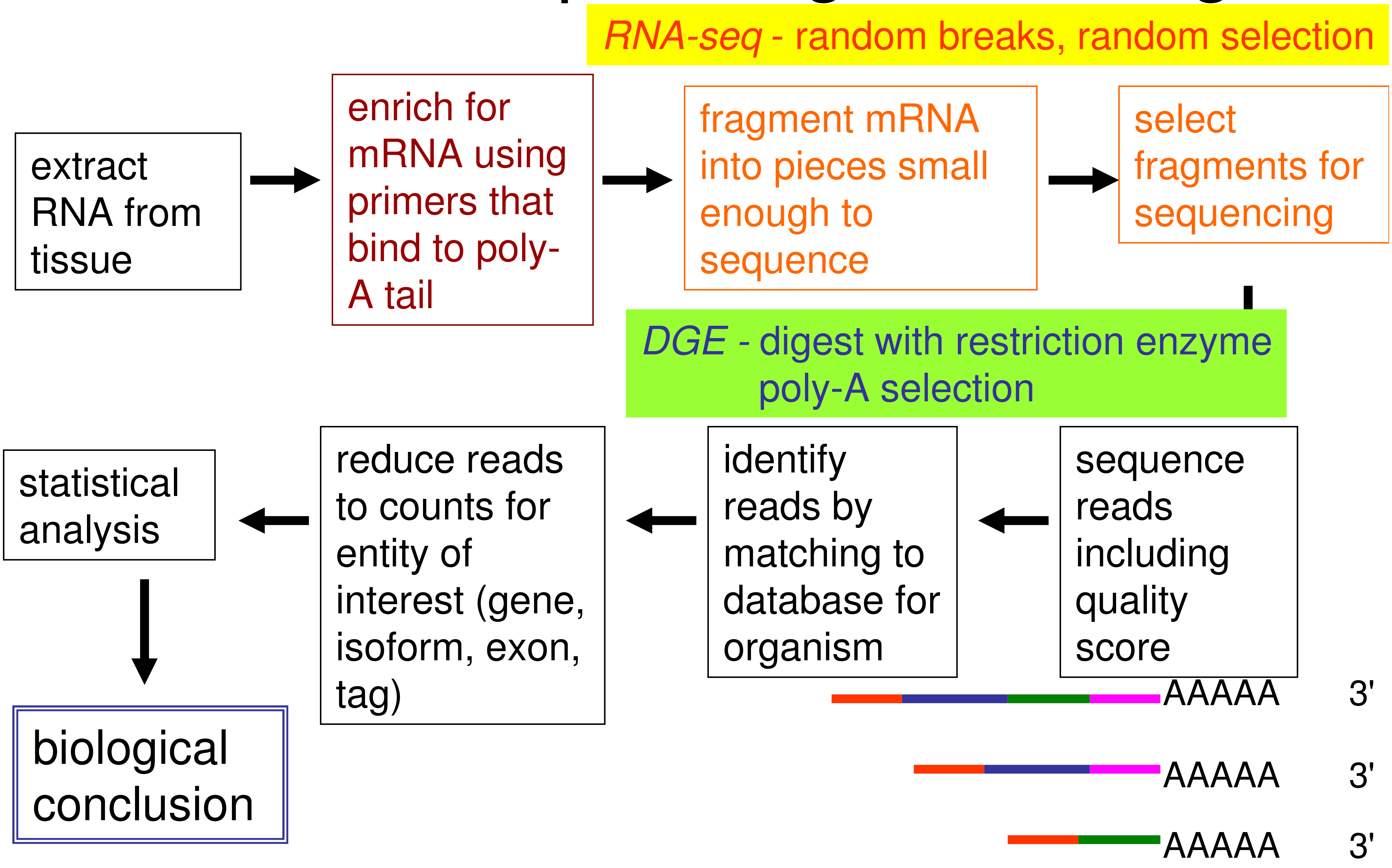
reduce reads to counts for entity of interest (gene, isoform, exon, tag)

statistical analysis

biological conclusion

- **Protein expression**
- Massively parallel sequencing
- DGE and RNA-seq
- Statistics
- Example
- Simulation
- Closing comments

Massively Parallel Sequencing Technologies



RNA-seq - random breaks, random selection

extract RNA from tissue

enrich for mRNA using primers that bind to poly-A tail

fragment mRNA into pieces small enough to sequence

select fragments for sequencing

DGE - digest with restriction enzyme poly-A selection

sequence reads including quality score

identify reads by matching to database for organism

reduce reads to counts for entity of interest (gene, isoform, exon, tag)

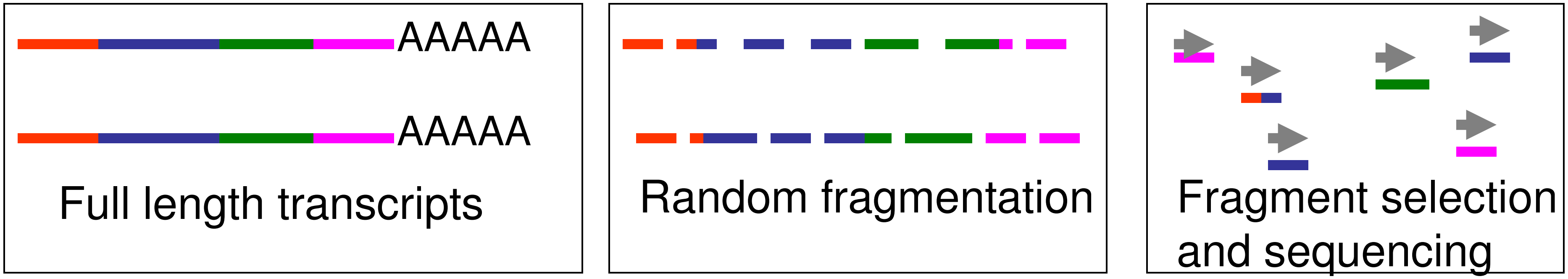
statistical analysis

biological conclusion

- **Protein expression**
- Massively parallel sequencing
- DGE and RNA-seq
- Statistics
- Example
- Simulation
- Closing comments

RNA-seq and DGE

RNA-seq - random breaks, random selection



Full length transcripts

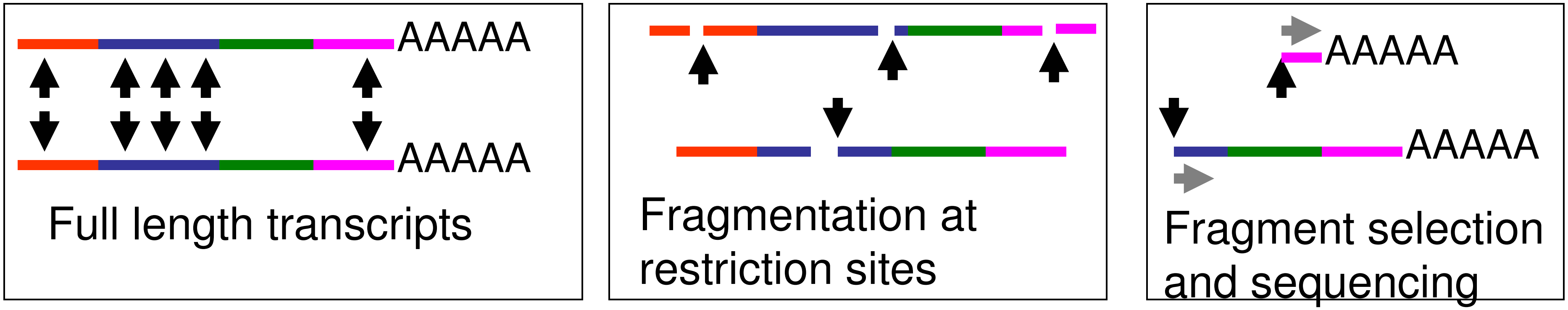
Random fragmentation

Fragment selection and sequencing

▲ restriction site

➤ sequenced read

DGE - digest with restriction enzyme, poly-A selection



Full length transcripts

Fragmentation at restriction sites

Fragment selection and sequencing

The Statistical Problem: Inferring Isoform Expression from DGE data

observe	counts/tag
genome sequence info	tag locations
exon annotation	exon boundaries
isoform annotation	exons in each isoform

The Statistical Problem: Inferring Isoform Expression from DGE data

observe	counts/tag	unambiguous
genome sequence info	tag locations	reasonably accurate
exon annotation	exon boundaries	somewhat accurate
isoform annotation	exons in each isoform	less accurate

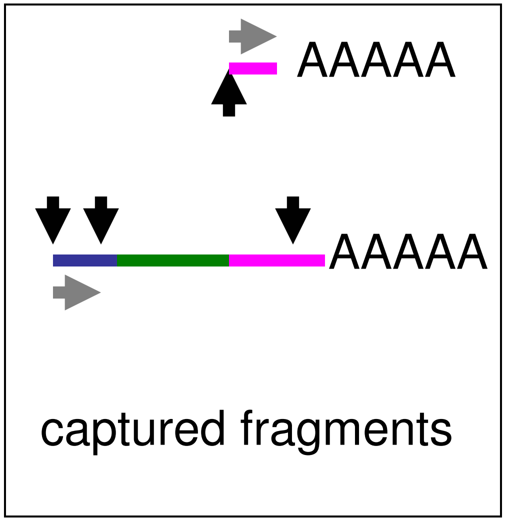
The Statistical Problem: Inferring Isoform Expression from DGE data

observe	counts/tag	unambiguous
genome sequence info	tag locations	reasonably accurate
exon annotation	exon boundaries	somewhat accurate
isoform annotation	exons in each isoform	less accurate

We want to infer counts/isoform

- **Protein expression**
- Massively parallel sequencing
- DGE and RNA-seq
- Statistics
- Example
- Simulation
- Closing comments

A model for tag retrieval



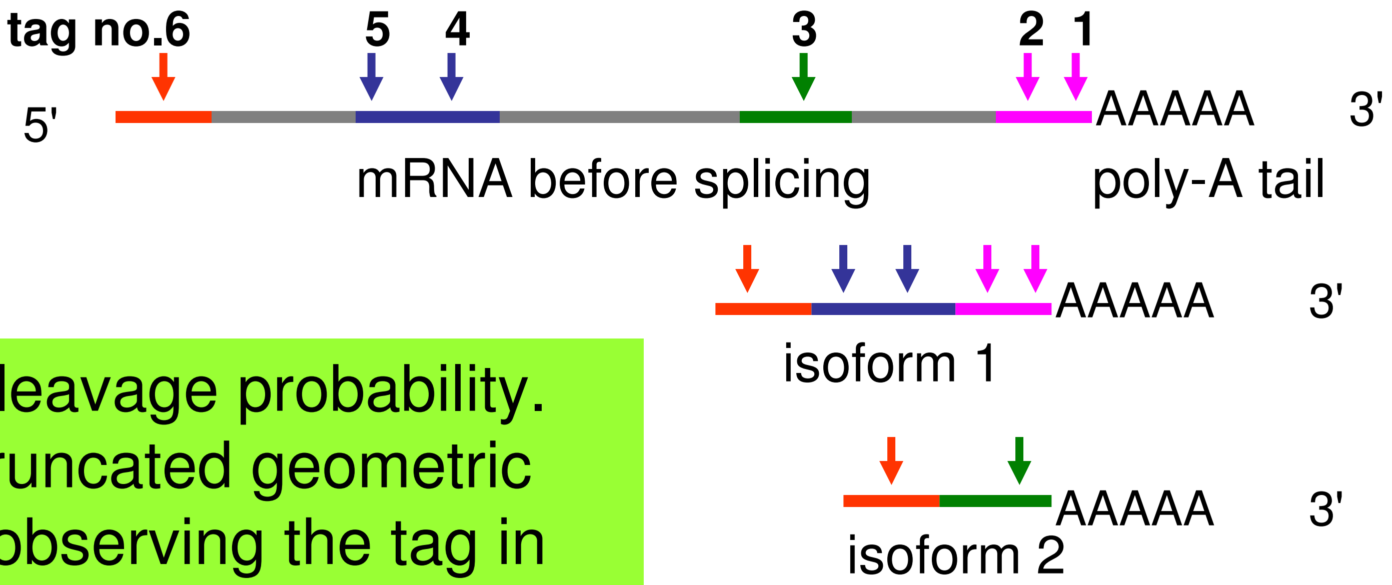
An mRNA fragment is captured if it contains the poly-A tail.

The tag is the short sequence that includes the restriction sequence (e.g. CGAT for the example) + a set number of bases (often 17 or 35) starting from the restriction site and going in the direction of the poly-A tail.

An mRNA may be fragmented at several sites, but a tag is observed only if no site closer to the poly-A tail is cut.

Gilchrist, Qin, & Zaretzki, (2007) postulate that the probability of cleavage is the same at every restriction site in the sample.

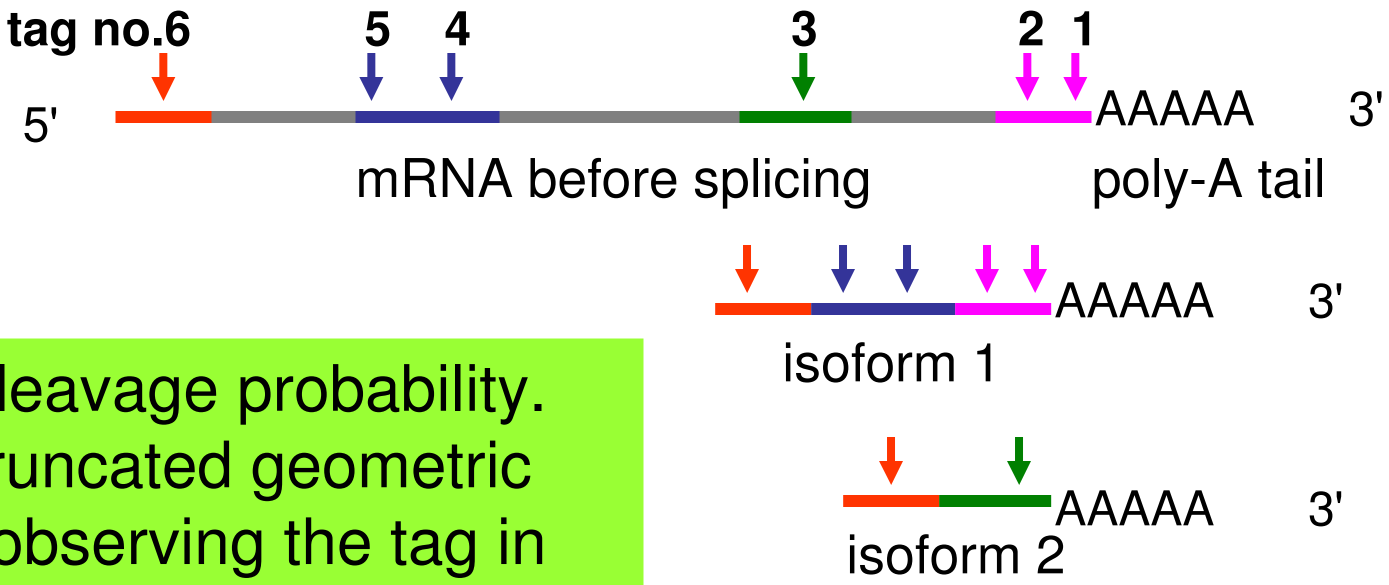
A model for tag retrieval



Let p be the cleavage probability. We obtain a truncated geometric probability of observing the tag in position s_i relative to the poly-A tail of the isoform.

- **Protein expression**
- Massively parallel sequencing
- DGE and RNA-seq
- Statistics
- Example
- Simulation
- Closing comments

A model for tag retrieval



Let p be the cleavage probability. We obtain a truncated geometric probability of observing the tag in position s_i relative to the poly-A tail of the isoform.

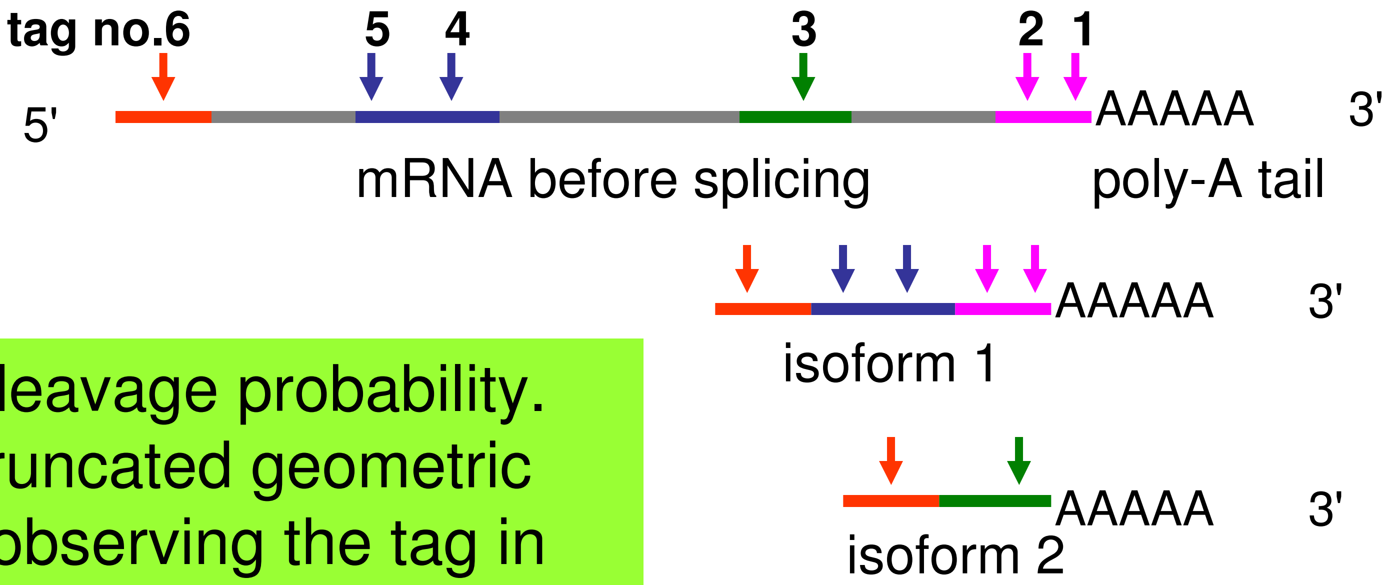
i.e. The probability of observing the red tag (6) in isoform 1 is

$\pi_{6|1}=p(1-p)^4$

but in isoform 2 it is

$\pi_{6|2}=p(1-p)$.

A model for tag retrieval



Let p be the cleavage probability. We obtain a truncated geometric probability of observing the tag in position s_i relative to the poly-A tail of the isoform.

i.e. The probability of observing the red tag (6) in isoform 1 is

$\pi_{6|1}=p(1-p)^4$

but in isoform 2 it is

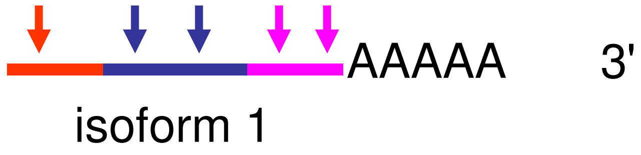
$\pi_{6|2}=p(1-p)$.

If the mRNA is not cut, no tag is observed. If isoform i has r_i sites, the probability that no tag is observed is

$1-(1-p)^{ri} =1-\Sigma_{|i}$

- **Protein expression**
- Massively parallel sequencing
- DGE and RNA-seq
- Statistics
- Example
- Simulation
- Closing comments

Estimating tag retrieval



AAAAA 3'

isoform 1

If an exon has 2 or more tags, the relative frequency of adjacent tags is 1-p. We use the median of this statistic to estimate p.
We prefer this robust estimator, because we have to rely on the exon annotation to determine which tags are in the exon. We have already seen that exon boundaries are not fully known.

- **Protein expression**
- Massively parallel sequencing
- DGE and RNA-seq
- Statistics
- Example
- Simulation
- Closing comments

Inferring Isoform Expression from DGE data

isoform	tag 1	tag 2	tag 3	tag 4	tag 5	…	tag K	No tag	isoform count
iso1	■	■		■	■		■		n_{+1}
iso2	■		■	■		■			n_{+2}
:									:
isoI		■	■	■			■		n_{+I}
tag total	T_{1+}	T_{2+}	T_{3+}	T_{4+}	T_{5+}	…	T_{K+}	**0**	**N**

■ tag k is in isoform i

We observe T and we want to infer n_{+i}.
If the transcript is not fragmented by the enzyme, it cannot be observed. We need to account for this, as isoforms with more tags are more likely to be fragmented.

- **Protein expression**
- Massively parallel sequencing
- DGE and RNA-seq
- Statistics
- Example
- Simulation
- Closing comments

Inferring Isoform Expression from DGE data

isoform	tag 1	tag 2	tag 3	tag 4	tag 5	…	tag K	No tag	isoform percent
iso1	$\pi_{1\|1}$	$\pi_{2\|1}$	0	$\pi_{4\|1}$	$\pi_{5\|1}$	$\pi_{k\|1}$	$\pi_{K\|1}$	$1-\Sigma_{\|1}$	π_{+1}
iso2	$\pi_{1\|2}$	0	$\pi_{3\|2}$	$\pi_{4\|2}$	0	$\pi_{k\|2}$	0	$1-\Sigma_{\|2}$	π_{+2}
:	$\pi_{1\|i}$	$\pi_{2\|i}$	$\pi_{3\|i}$	$\pi_{4\|i}$	$\pi_{5\|i}$	$\pi_{k\|i}$	$\pi_{K\|i}$	$1-\Sigma_{\|i}$	:
isoI	0	$\pi_{2\|I}$	$\pi_{3\|I}$	$\pi_{4\|I}$	0	$\pi_{k\|I}$	$\pi_{K\|I}$	$1-\Sigma_{\|I}$	π_{+I}
tag percent	π_{1+}	π_{2+}	π_{3+}	π_{4+}	π_{5+}	…	π_{K+}	$1-\Sigma_{\|+}$	**1**

Note that $\pi_{k+}=\Sigma\pi_{k|i}\pi_{+i}$

From Bayes' rule, the row margins can be computed from the conditional probabilities and column margins.

Slavkovic, 2004: If the matrix of conditional probabilities has full row rank, the row margins are unique.

Inferring Isoform Expression from DGE data

From Bayes' rule, we can compute the row margins from the conditional probabilities and column margins.
Slavkovic, 2004: If the matrix of conditional probabilities has full row rank, the row margins are unique.

The uniqueness condition can fail if there are more isoforms than tags, if there are isoforms that differ only in exons that have no tags, or (in practice) if there are isoforms that differ only in tags that have very low probability of being observed.

$\pi_{k|i}$ is a function of p, the cutting probability which is determined by protocols for restriction enzyme digestion and can be manipulated by the investigator. For exon detection, it is preferable to have low p, so that there is a high probability of observing tags far from the poly-A tail.

- **Protein expression**
- Massively parallel sequencing
- DGE and RNA-seq
- Statistics
- Example
- Simulation
- Closing comments

Inferring Isoform Expression from DGE data

isoform	tag 1	tag 2	tag 3	tag 4	tag 5	...	tag K	No tag	isoform count
iso1	$\pi_{1\|1}$	$\pi_{2\|1}$	0	$\pi_{4\|1}$	$\pi_{5\|1}$	$\pi_{k\|1}$	$\pi_{K\|1}$	$1-\Sigma_{\|1}$	$\mathbf{n_{+1}}$
iso2	$\pi_{1\|2}$	0	$\pi_{3\|2}$	$\pi_{4\|2}$	0	$\pi_{k\|2}$	0	$1-\Sigma_{\|2}$	$\mathbf{n_{+2}}$
:	$\pi_{1\|i}$	$\pi_{2\|i}$	$\pi_{3\|i}$	$\pi_{4\|i}$	$\pi_{5\|i}$	$\pi_{k\|i}$	$\pi_{K\|i}$	$1-\Sigma_{\|i}$	:
isoI	0	$\pi_{2\|I}$	$\pi_{3\|I}$	$\pi_{4\|I}$	0	$\pi_{k\|I}$	$\pi_{K\|I}$	$1-\Sigma_{\|I}$	$\mathbf{n_{+I}}$
tag count	$\mathbf{T_{1+}}$	$\mathbf{T_{2+}}$	$\mathbf{T_{3+}}$	$\mathbf{T_{4+}}$	$\mathbf{T_{5+}}$	**...**	$\mathbf{T_{K+}}$	**0**	**N**

We observe T with $E(T_{k+})=N\pi_{k+}$ but not the number of transcripts that did not produce a tag. We note that $E(n_{+i})=N\pi_{+i}$ so $E(\Sigma\pi_{k|i}n_{+i})=N\pi_{+i}=E(T_{k+})$.

We use the least squares estimator to estimate n_{+i} from the T's.

- **Protein expression**
- Massively parallel sequencing
- DGE and RNA-seq
- Statistics
- Example
- Simulation
- Closing comments

Inferring Isoform Expression from DGE data

P	n_{+i}
T_{k+}	N

Let P be the matrix of conditional probabilities.

We note that $E(Pn_{+i})=E(T_{+k})$.

We use the least squares estimator to estimate n_{+i} from the T's.

I.e. $\hat{n}_{+i} = (P'P)^{-1}P'T_{k+}$

This also suggests the use of the estimated sandwich estimator of variance

$$\hat{V}ar(\hat{n}_{+i}) = (P'P)^{-1}P'\hat{V}ar(T_{k+})P(P'P)^{-1}$$

Inferring Isoform Expression from DGE data

P	n_{+i}
T_{k+}	N

Let P be the matrix of conditional probabilities.

We note that $E(Pn_{+i})=E(T_{+k})$.

We use the least squares estimator to estimate n_{+i} from the T's.

I.e. $\hat{n}_{+i} = (P'P)^{-1}P'T_{k+}$

This also suggests the use of the estimated sandwich estimator of variance but this needs improvement.

$$\hat{V}ar(\hat{n}_{+i}) = (P'P)^{-1}P'\hat{V}ar(T_{k+})P(P'P)^{-1}$$

- **Protein expression**
- Massively parallel sequencing
- DGE and RNA-seq
- Statistics
- Example
- Simulation
- Closing comments

The 't Hoen Mouse Data

't Hoen et al, 2008 collected RNA from mouse brain tissue for wild-type and transgenic mice. mRNA was extracted and processed for DGE analysis.

Sample	W1	M1	W2	M2	W3	M3	W4	M4
total reads (millions)	2.7	3.5	3.2	3.5	2.4	0.3	0.6	3.1
total matched reads (millions)	0.7	1.1	0.9	1.1	0.6	0.1	0.2	0.9
max reads/tag (thousands)	10	15	12	17	12	1	2	12

We did not use M3 or W4.

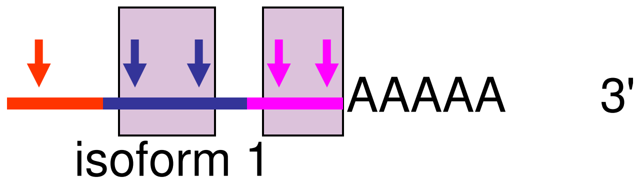
The 't Hoen Mouse Data

Step 0: The initial step in the analysis is to map the tags to the genes and exons.
Illumina@ kindly provided us with the tag database used in the original study, which greatly reduced the work by matching the tags to the genes.

- **Protein expression**
- Massively parallel sequencing
- DGE and RNA-seq
- Statistics
- Example
- Simulation
- Closing comments

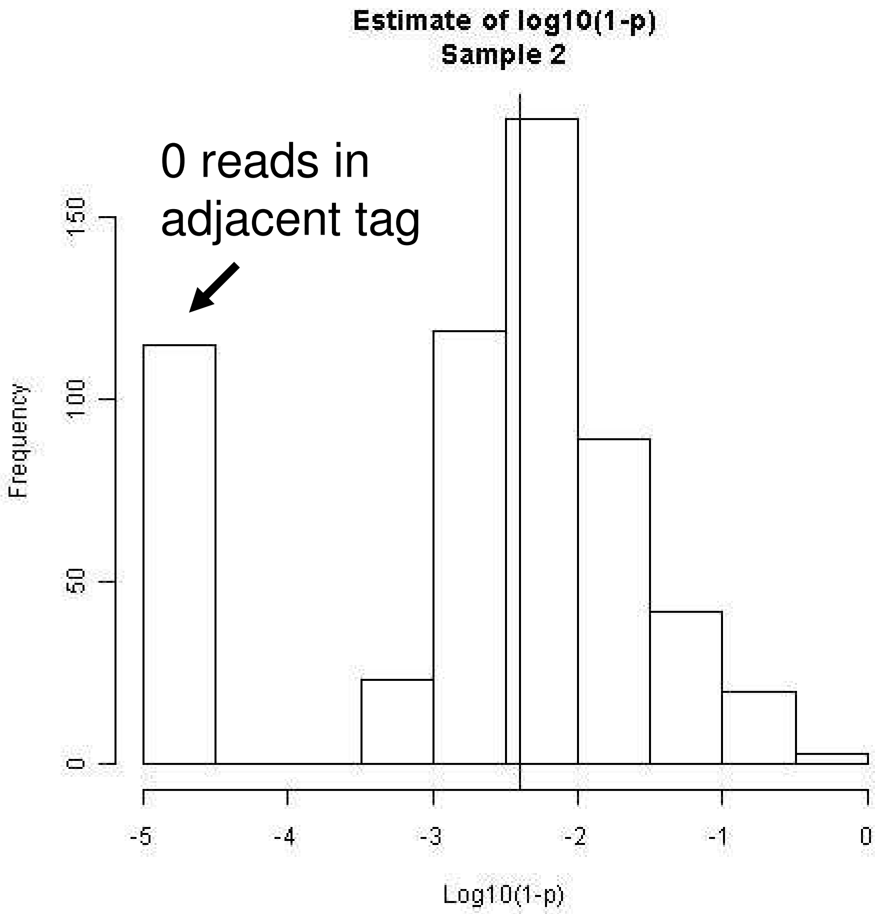
The 't Hoen Mouse Data

Step 1: Estimate p for each sample.



For each exon with a 3' tag with more than 500 reads and at least 2 tags, we took the ratio of the 3' tag count to the adjacent tag count.
The median of the ratios is an estimate of 1-p which is robust with respect to exon and tag annotation.

p=.996 for all 6 samples



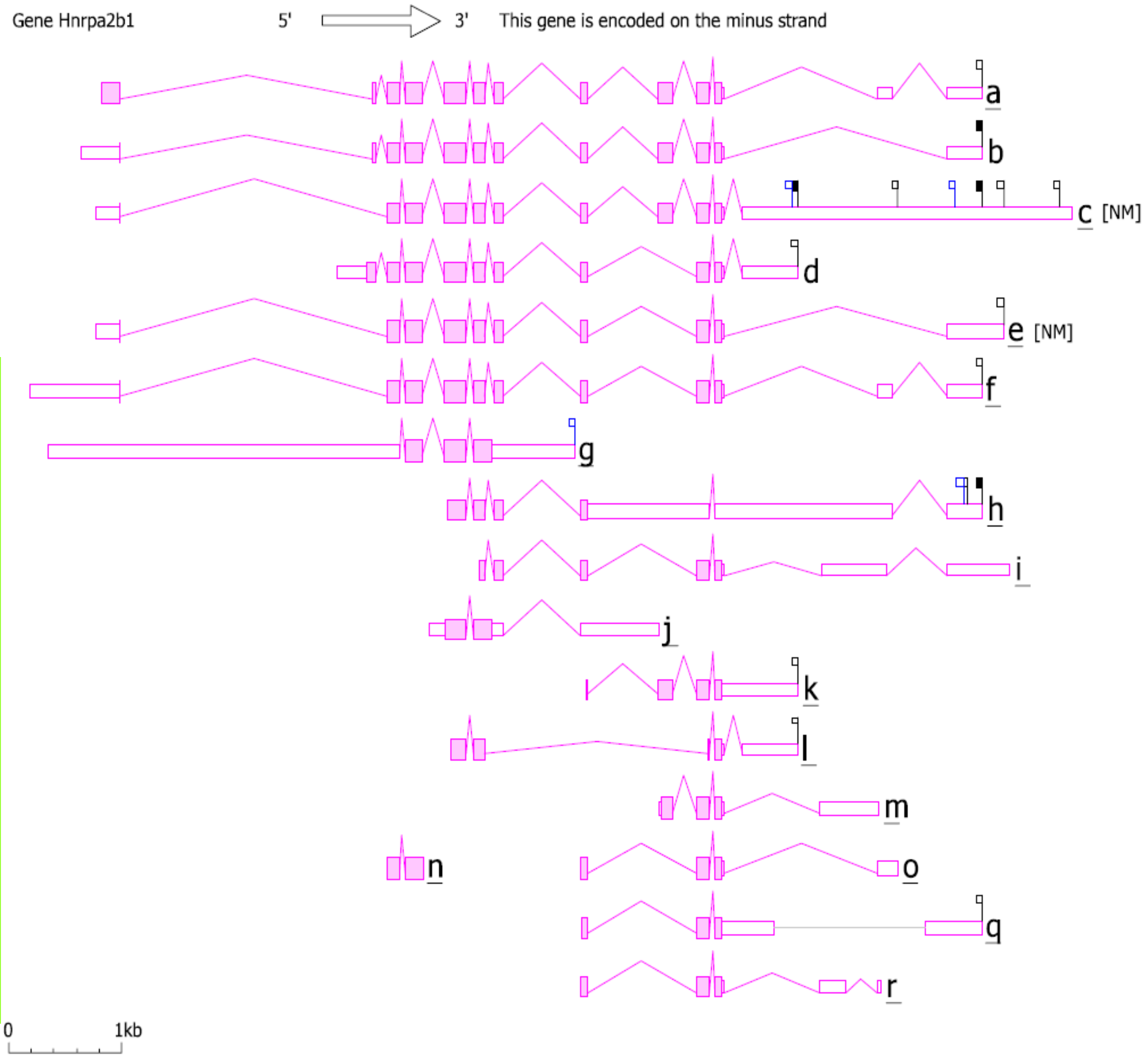
- **Protein expression**
- Massively parallel sequencing
- DGE and RNA-seq
- Statistics
- Example
- Simulation
- Closing comments

The 't Hoen Mouse Data

p=.996 for all 6 samples

This is bad news for our estimation procedure.

Only the tag at closest to the ¶ has substantial probability of being observed.

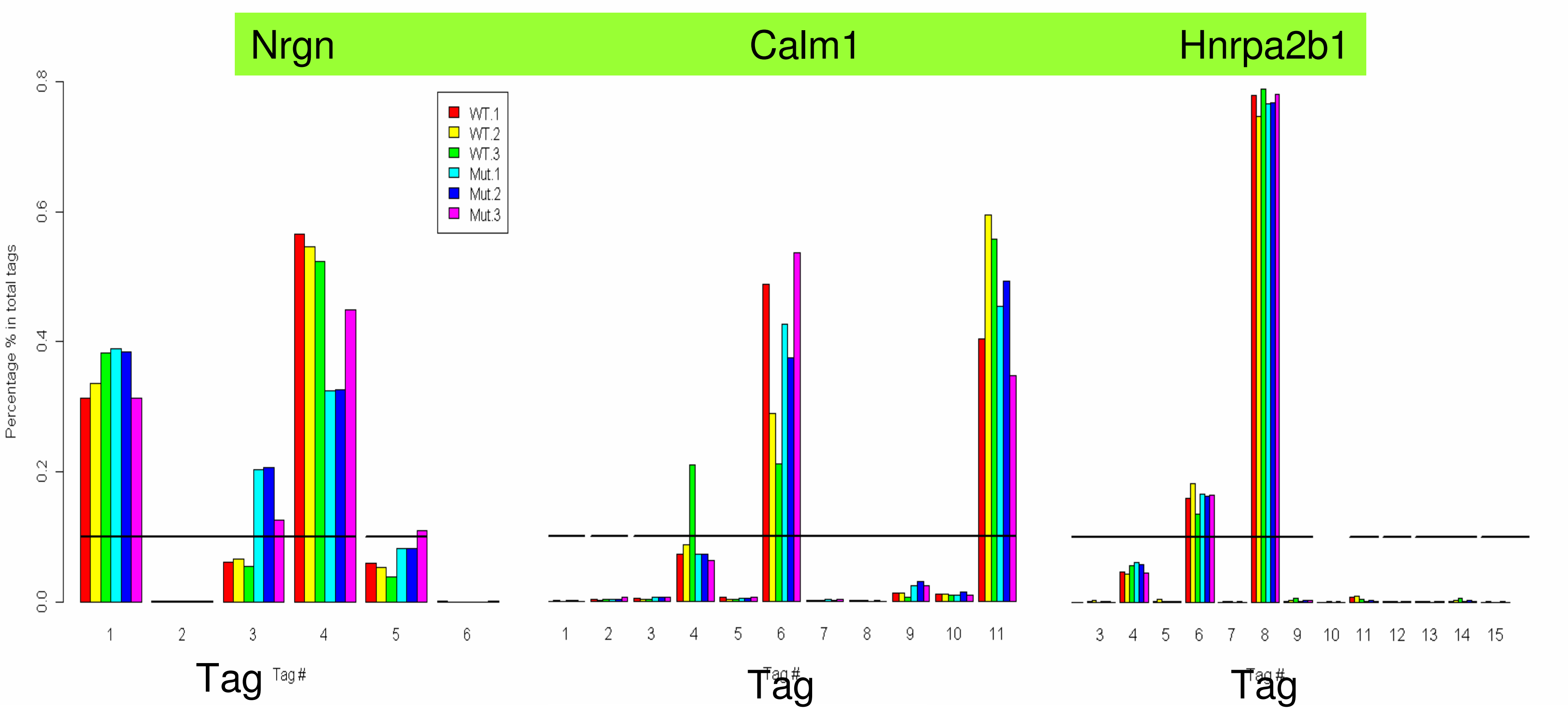


http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/av.cgi?exdb=AceView&db=mouse&term=mm.155896&submit=Go

- **Protein expression**
- Massively parallel sequencing
- DGE and RNA-seq
- Statistics
- Example
- Simulation
- Closing comments

The 't Hoen Mouse Data

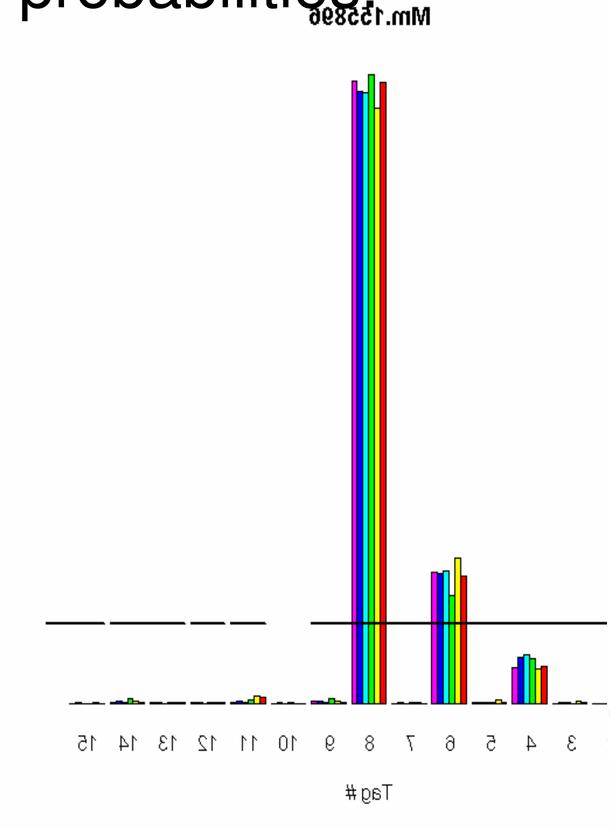
Step 2: For each gene of interest create the matrix of conditional probabilities

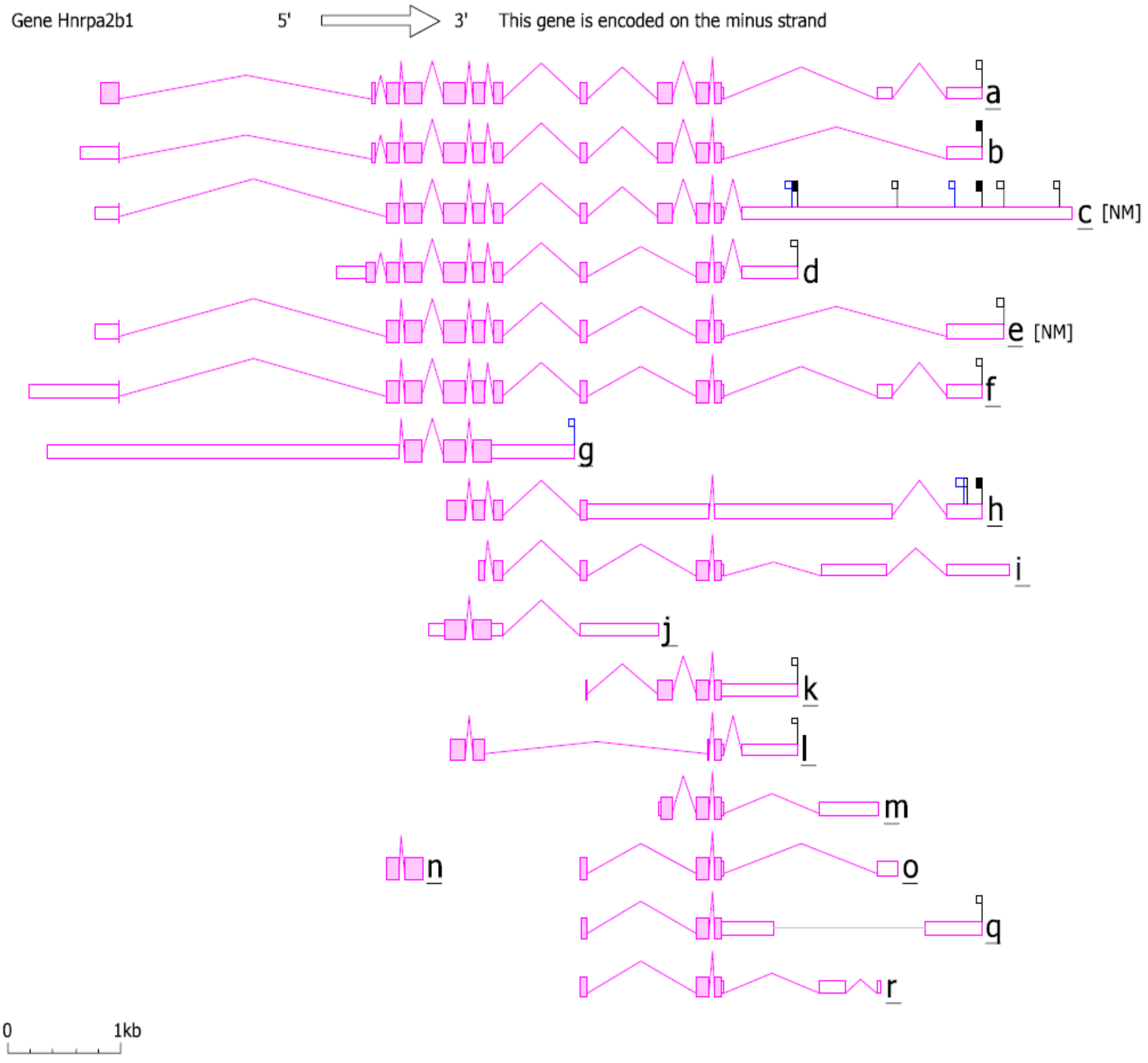


- **Protein expression**
- Massively parallel sequencing
- DGE and RNA-seq
- Statistics
- Example
- Simulation
- Closing comments

The 't Hoen Mouse Data

Step 2: For each gene of interest create the matrix of conditional probabilities.





The 't Hoen Mouse Data

Step 2: For each gene of interest create the matrix of conditional probabilities.

The high value of p makes it very difficult to distinguish among isoforms with the same first tag. We assumed that every tag with more than 5 reads in a sample was a 3' tag, and that the 3 nearest tags were in the same isoform.

The 't Hoen Mouse Data

Step 2: For each gene of interest create the matrix of conditional probabilities.

The high value of p makes it very difficult to distinguish among isoforms with the same first tag. We assumed that every tag with more than 5 reads in a sample was a 3' tag, and that the 3 nearest tags were in the same isoform.

At this point, the data become a demo rather than a real data analysis - more on this later!

☹

The 't Hoen Mouse Data

Step 3: Estimate the isoform counts.
e.g. Hnrpa2b1 (counts per 10 thousand reads, rounded)

	W1	W2	W3	M1	M2	M3
total	22	19	16	14	14	14
most abundant	13	11	9	5	5	6
discordant	.13	.13	.09	.29	.28	.17

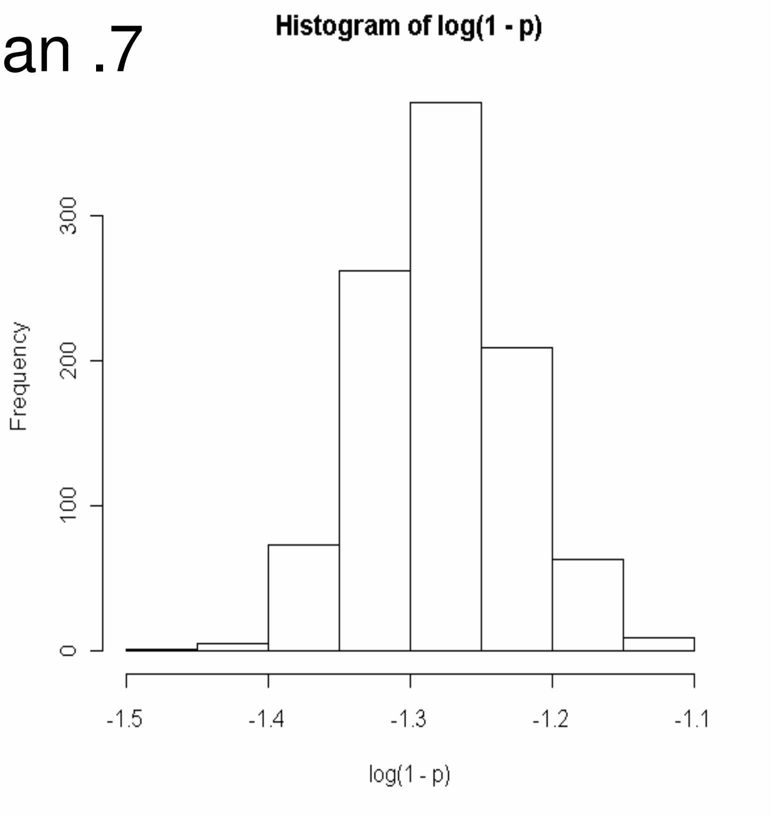
Note that by any rank based test, the gene and isoform expression is a significantly different as possible. But at least one isoform is discordant.

When in doubt, simulate

Since the p in the 't Hoen study was too high for isoform resolution, we simulated data from a gene with 4 exons with 2 tags/exon.
For each tag, we simulated a cutting probability

p_i~ Beta(700,700*28/72) which has mean .7

tag	1	2	3	4	5	6	7	8
isoe1	🟪	🟪						
isoe1e2	🟪	🟪	🟪	🟪				
isoe2e3			🟪	🟪	🟪	🟪		



- **Protein expression**
- Massively parallel sequencing
- DGE and RNA-seq
- Statistics
- Example
- Simulation
- Closing comments

When in doubt, simulate

true n	1000	500	500	50
🟪🟪⬜⬜⬜⬜ 🟪🟪🟪🟪⬜⬜	1000.79 126.07 84.14	499.47 135.16 90.18	500.73 28.36 43.31	49.13 32.02 40.74
🟪🟪🟪🟪⬜⬜ ⬜⬜🟪🟪🟪🟪	999.99 14.84 23.39	499.87 14.26 23.12	499.92 3.27 7.49	50.07 10.42 12.48

XXX mean estimated count

XXX SD of simulated estimates

XXX Sandwich estimator SD

- **Protein expression**
- Massively parallel sequencing
- DGE and RNA-seq
- Statistics
- Example
- Simulation
- Closing comments

DGE and Isoform Expression

Gilchrist et al suggested that investigators should use lower p in gene expression studies.

DGE and Isoform Expression

Gilchrist et al suggested that investigators should use lower p in gene expression studies.

Their reason was non-uniqueness of tags. If a tag occurs in multiple locations in the genome, it cannot be attributed to the gene.
So, for p close to 1.0, the expression of genes with non-unique 3’ tags cannot be estimated.

DGE and Isoform Expression

Gilchrist et al suggested that investigators should use lower p in gene expression studies.

Their reason was non-uniqueness of tags. If a tag occurs in multiple locations in the genome, it cannot be attributed to the gene.
So, for p close to 1.0, the expression of genes with non-unique 3’ tags cannot be estimated.

However, they considered all locations in the genome – really they only need to consider uniqueness among 3’ tags.

DGE and Isoform Expression

There is little incentive for investigators to induce lower values of p if they are interested in overall gene expression.

Even for p=.7 the probability of observing any but the first few 3’ tags is vanishingly small.

If p is too small, there is a high probability that transcripts with few tags will not be cut.

DGE and Isoform Expression

There is little incentive for investigators to induce lower values of p.

Even for p=.7 the probability of observing any but the first few 3’ tags is vanishingly small.

Our study started in an attempt to verify the Gilchrist et al model.

If the model is correct, DGE is not as powerful as RNA-seq for estimating isoform expression.

DGE and Isoform Expression

There is little incentive for investigators to induce lower values of p.

Even for p=.7 the probability of observing any but the first few 3’ tags is vanishingly small.

Our study started in an attempt to verify the Gilchrist et al model. ☺

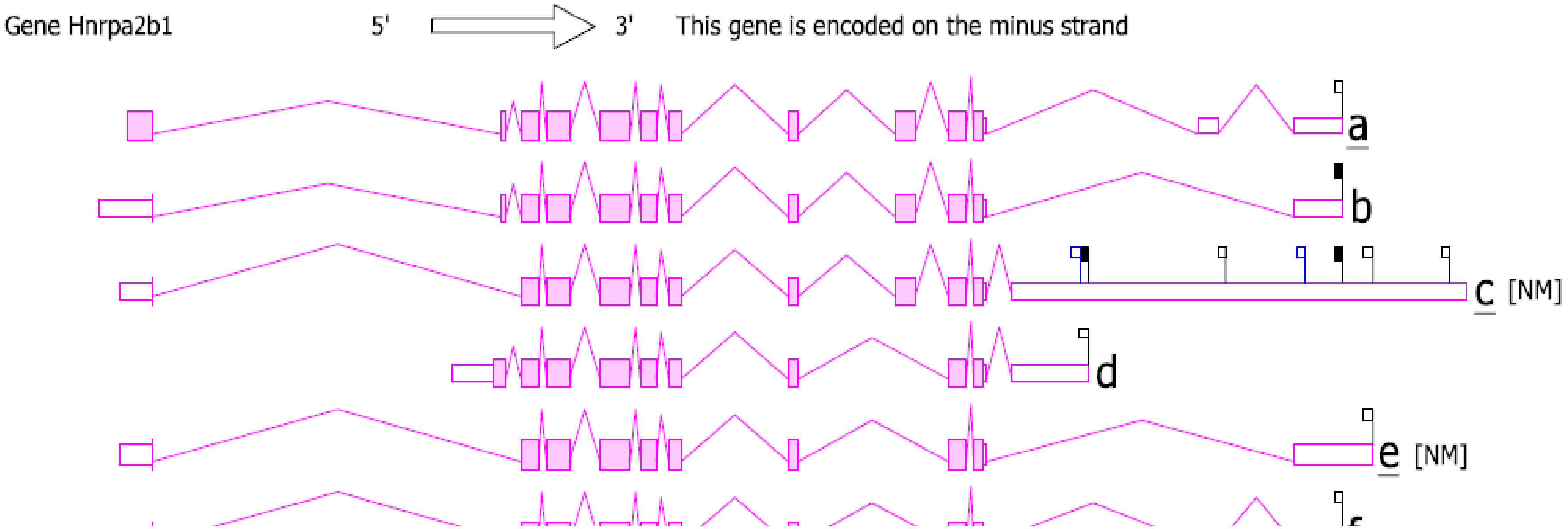
If the model is correct, DGE is not are powerful as RNA-seq for estimating isoform expression. ☹

Searching for a Solution to the Isoform Expression Problem

DGE data are much easier to work with than RNA-seq data.

But RNA-seq data have more relevant information about isoform expression.

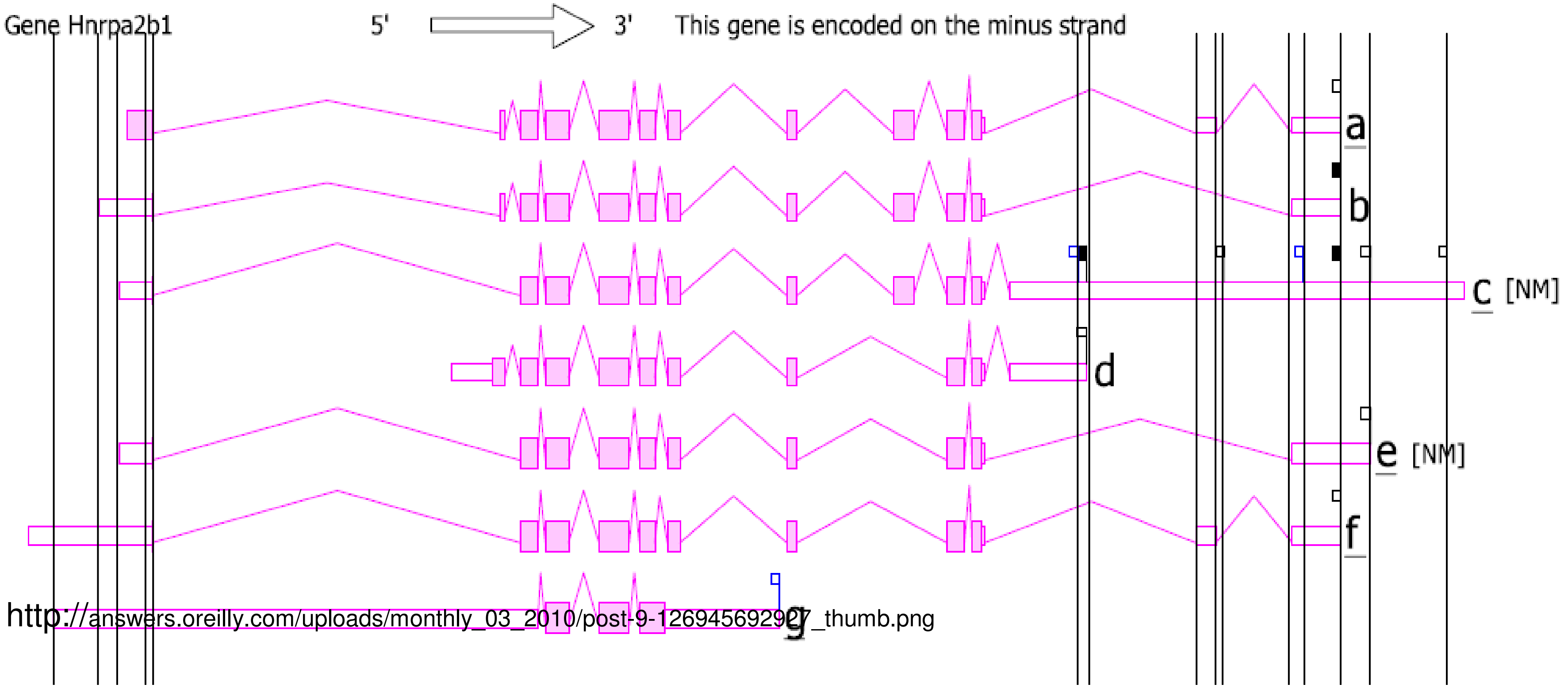
We no longer have tags. Each read maps to the genome. We can replace tags by exon segments.



- **Protein expression**
- Massively parallel sequencing
- DGE and RNA-seq
- Statistics
- Example
- Simulation
- Closing comments

Searching for a Solution to the Isoform Expression Problem

We no longer have tags. Each read maps to the genome.
We can replace tags by exon segments.
Lets call these extags.
We need a model for detecting extags.



http://answers.oreilly.com/uploads/monthly_03_2010/post-9-126945692927_thumb.png

Searching for a Solution to the Isoform Expression Problem

We no longer have tags. Each read maps to the genome.
We can replace tags by exon segments.
Lets call these extags.
We need a model for detecting extags.

e.g. We may assume that the probability of detecting a read in extag j is proportional to some feature of the extag. (e.g. length of uniquely mappable section; CG content ...)

Searching for a Solution to the Isoform Expression Problem

Let S_i be the set of extags in isoform i.

Then if extag j is in isoform i,the conditional detection probability is

$$\pi_{j|i} = \frac{x_j}{\sum_{k \in S_i} x_k}$$

Searching for a Solution to the Isoform Expression Problem

We are back to the previous situation (except every extag has non-zero detection probability).

$$\pi_{j|i} = \frac{x_j}{\sum_{k \in S_i} x_k}$$

The covariate x will depend on the sample preparation protocol.

isoform	tag 1	tag 2	tag 3	tag 4	tag 5	...	tag K	isoform count
iso1	$\pi_{1\|1}$	$\pi_{2\|1}$	0	$\pi_{4\|1}$	$\pi_{5\|1}$	$\pi_{k\|1}$	$\pi_{K\|1}$	$\mathbf{n}_{+1}$
iso2	$\pi_{1\|2}$	0	$\pi_{3\|2}$	$\pi_{4\|2}$	0	$\pi_{k\|2}$	0	$\mathbf{n}_{+2}$
:	$\pi_{1\|i}$	$\pi_{2\|i}$	$\pi_{3\|i}$	$\pi_{4\|i}$	$\pi_{5\|i}$	$\pi_{k\|i}$	$\pi_{K\|i}$	:
isoI	0	$\pi_{2\|I}$	$\pi_{3\|I}$	$\pi_{4\|I}$	0	$\pi_{k\|I}$	$\pi_{K\|I}$	$\mathbf{n}_{+I}$
tag count	$\mathbf{T}_{1+}$	$\mathbf{T}_{2+}$	$\mathbf{T}_{3+}$	$\mathbf{T}_{4+}$	$\mathbf{T}_{5+}$	**...**	$\mathbf{T}_{K+}$	**N**

Thanks to: NSF for partial funding via a variety of grants.
R. Schilder for help with the biology.
Loren Honaas for help with RNA-seq data.
Illumina@ for providing the tag database.

References: 't Hoen et al, 2008, *Nucleic Acids Research*
: Gilchrist et al, 2007, *BMC Bioinformatics*
: Altman et al, 2010, *J. Indian Society of Agricultural Statistics*