Statistical analysis of RNA-seq

Expression estimation

Ernest Turro

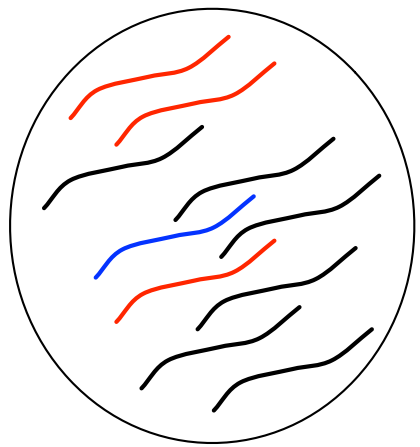
University of Cambridge

28 Oct 2016

Gene expression

An important aim in genomics is the characterisation of RNA samples. Specifically:

1. What is the sequence of each distinct RNA in a sample?
2. What is the concentration of each RNA in a sample?



6/10

3/10

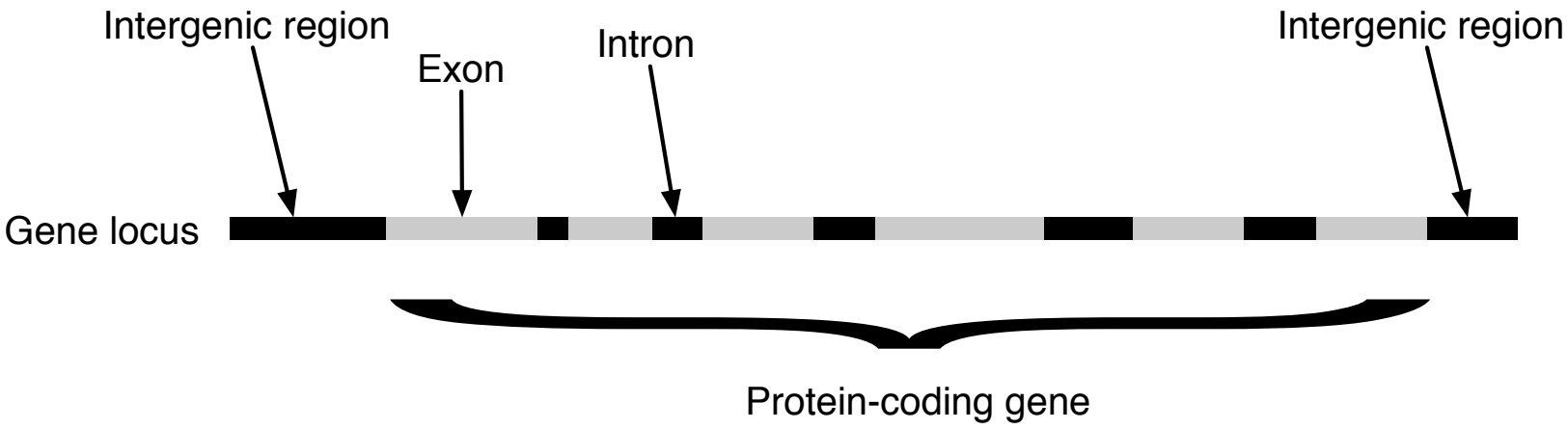
1/10

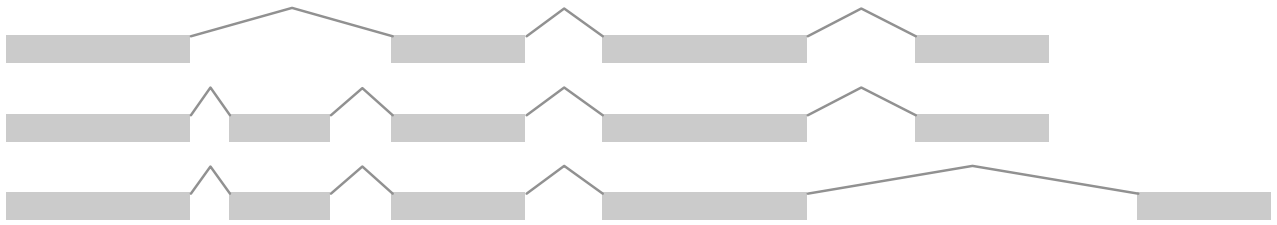
NB: in general only relative proportions available

Gene expression

Different kinds of RNAs (tRNAs, rRNAs, mRNAs, other ncRNAs...).

Messenger RNAs of particular interest as they code for proteins.



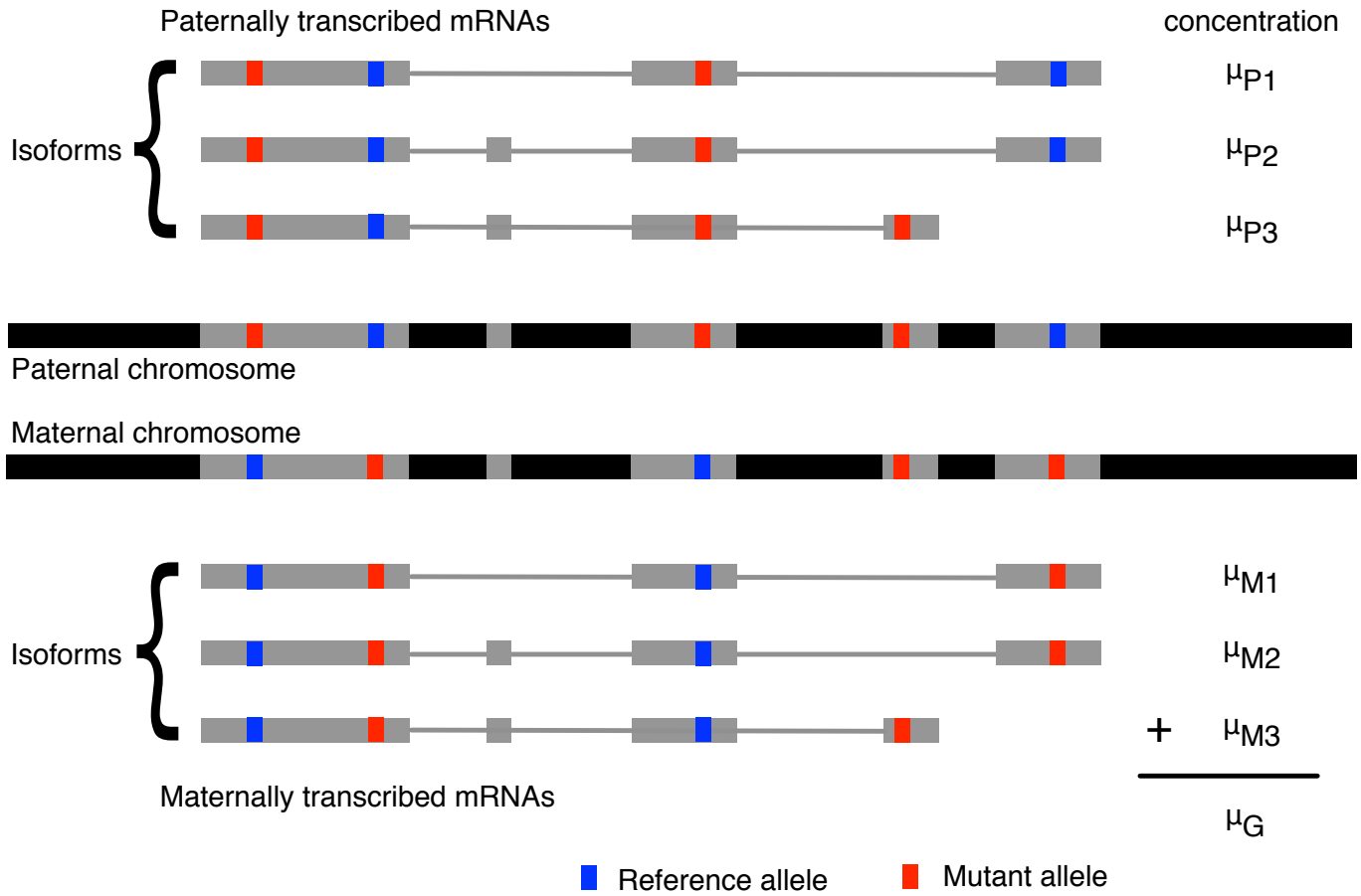


Gene expression

Different kinds of RNAs (tRNAs, rRNAs, mRNAs, other ncRNAs...).

Messenger RNAs of particular interest as they code for proteins.

1. Alternative isoforms have distinct sequences
2. Two versions of each isoform sequence in diploid organisms



RNA-seq read counts

To infer concentrations, we need to identify

- the set of transcript sequences in the sample
- the set of reads (potentially) emanating from each transcript

Approaches:

- Select transcript sequences from a database (e.g. Ensembl) and align reads to them
- Align reads to genome and infer transcript sequences
- Assemble reads into contigs

In any case, we get a **mapping of reads to features of interest** (e.g. genes, isoforms, haplotype-specific isoforms).

How do we model the alignments?

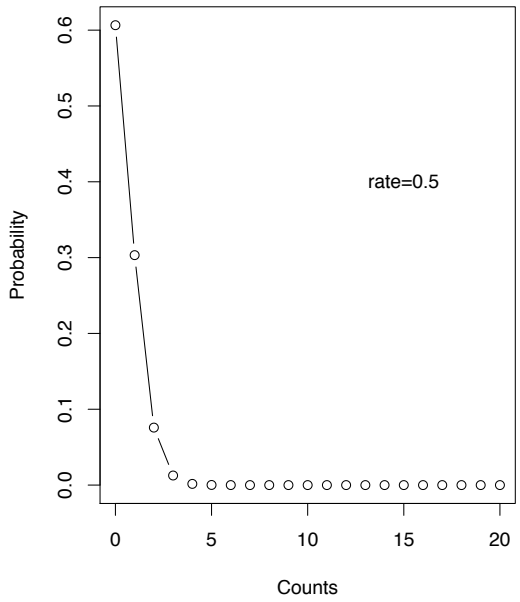
The Poisson distribution

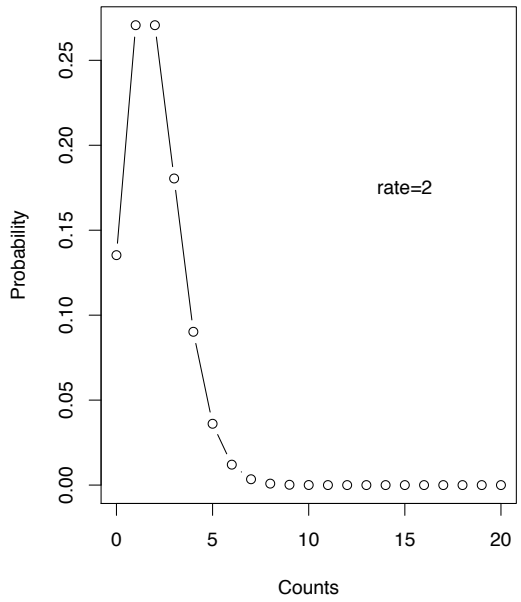
If independent events occur at a known given rate, then the number of such events follows a **Poisson distribution**.

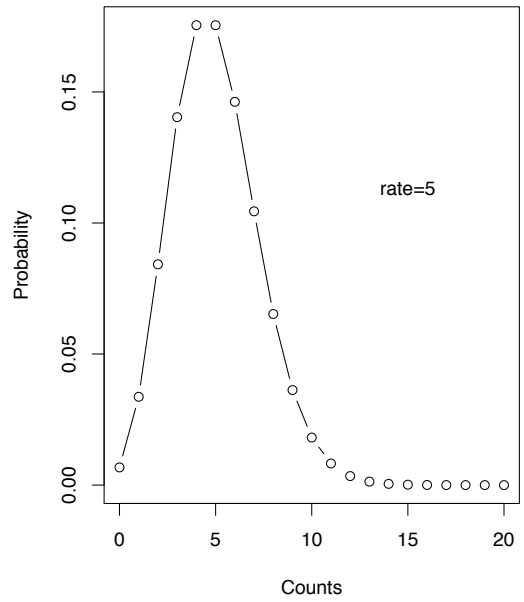
Examples:

- Number of cars crossing a milestone every hour
- Number of raindrops falling on a rooftop every minute

Single rate parameter λ (pets rate = cats rate + dogs rate).
Mean = variance = rate.







Basic Poisson model for expression quantification

Number of reads aligning to a transcript increases with

- Total number of reads
- Length of transcript
- Abundance of transcript

Number of reads from gene g captured by Poisson model (Marioni et al. 2008):

$$r_g \sim Poisson(b\mu_g l_g),$$

- μ_g: concentration of RNAs from gene g
- l_g: effective length of the gene
- b: normalisation constant (e.g. total no. of reads)

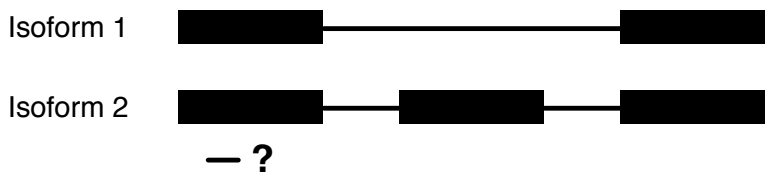
Basic Poisson model for expression quantification

Basic model is useful but:

- “gene length” ambiguous — fragments from several isoforms with different lengths are sequenced
- reads counts not always observed due to sequence sharing (e.g. paralogous families)

Can we estimate expression for each isoform?

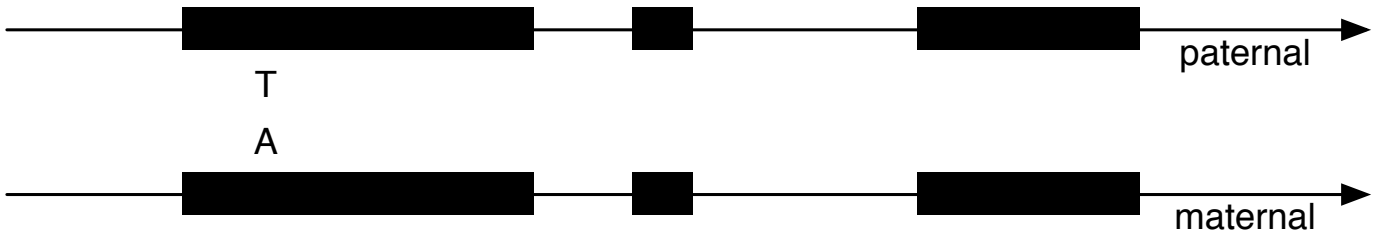
- Isoform read counts in general not observed:



- We need a **read count model for isoforms**

Basic Poisson model for expression quantification

Recall that sequencing allows us to distinguish alleles at heterozygous positions.



Can we use RNA-seq to detect allelic imbalance?

We need a **read count model for alleles**

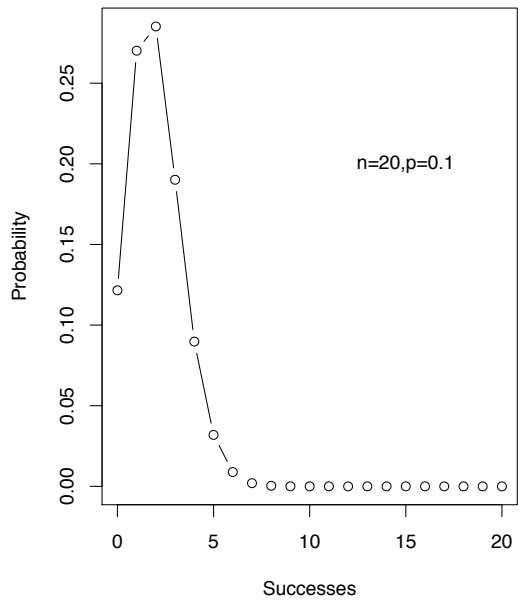
The binomial distribution

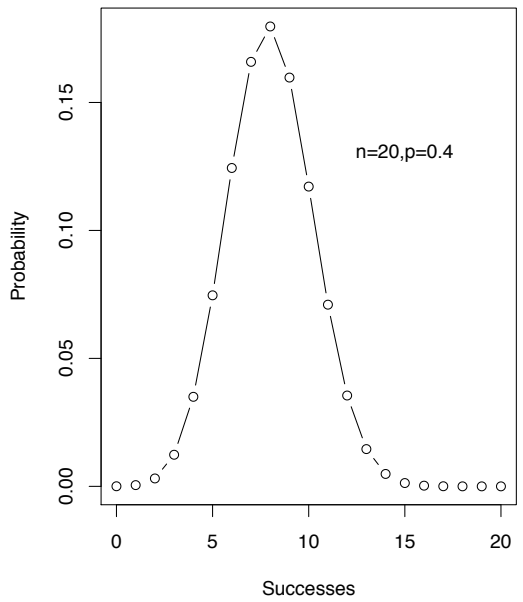
A Bernoulli trial is an experiment in which "success" occurs with probability p and "failure" occurs with probability $1 - p$.

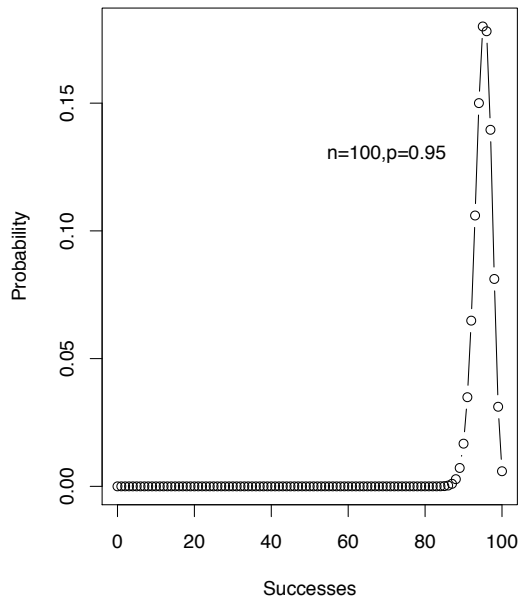
The number of successes given n Bernoulli trials follows a **binomial distribution** with parameters n and p. $\mathbb{E}(X) = np$.

Examples:

- Number of heads after n coin tosses. $p \neq 0.5$ if not fair
- Number of times you win the lottery (tiny p (but £££))







The multinomial distribution

A Bernoulli trial is an experiment in which "success" occurs with probability p and "failure" occurs with probability $1 - p$.

The number of successes given n Bernoulli trials follows a **binomial distribution** with parameters n and p.

Examples:

- Number of heads after n coin tosses. $p \neq 0.5$ if it is unfair
- Number of times you hit the bullseye out of n shots

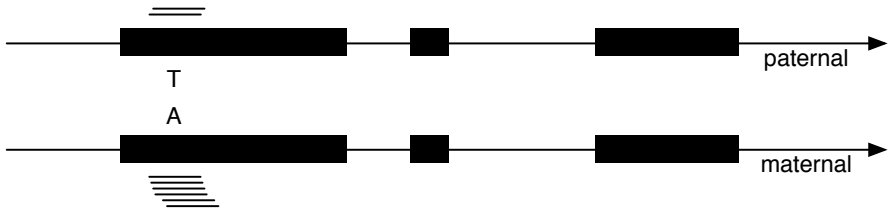
If there are > 2 categories, the per-category counts follow a **multinomial distribution** with parameters n and $(p_1, p_2, \ldots)$.

Example:

- Number of 1s, 2s, 3s, 4s, 5s, 6s if you roll a die n times. If $\{p_i\} \neq \frac{1}{6}$ then the die is not fair.

Basic Binomial model for allelic imbalance

- Reads permit discrimination between two copies of an isoform

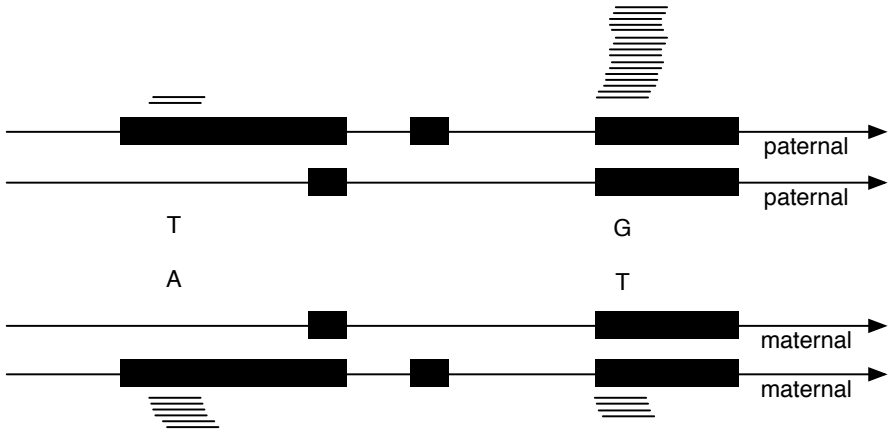


- Binomial test: $\sum_{r=0}^{r_0} P(r|p = 0.5, n = r_0 + r_1) < \alpha$? (Degner et al. 2009). E.g. suppose $r_0 = 2; r_1 = 6$:

$$P(r = 0|p = 0.5, n = 8) = 0.00390625$$
$$P(r = 1|p = 0.5, n = 8) = 0.03125$$
$$P(r = 2|p = 0.5, n = 8) = 0.109375$$
$$\sum_{r=0}^{2} P(r|p = 0.5, n = 8) = 0.1445312 \text{ (not significant)}$$

Basic Binomial model for allelic imbalance

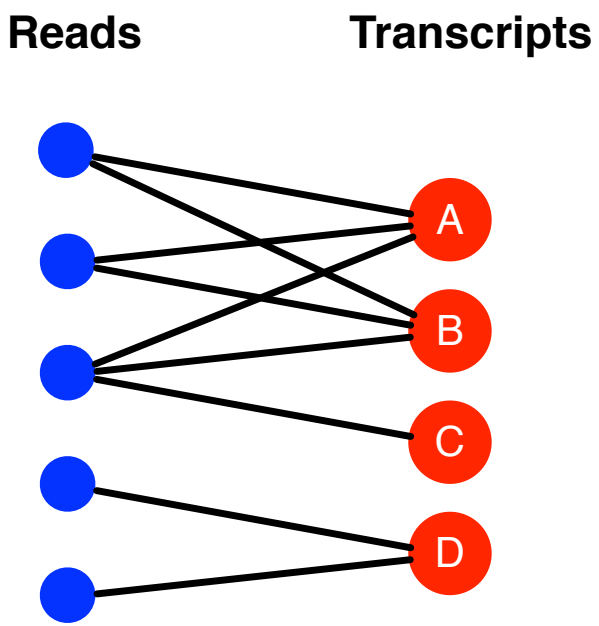
- What if there are multiple SNPs and isoforms?



- Binomial test not appropriate
- We need a **read count model for haplotype-specific isoforms**

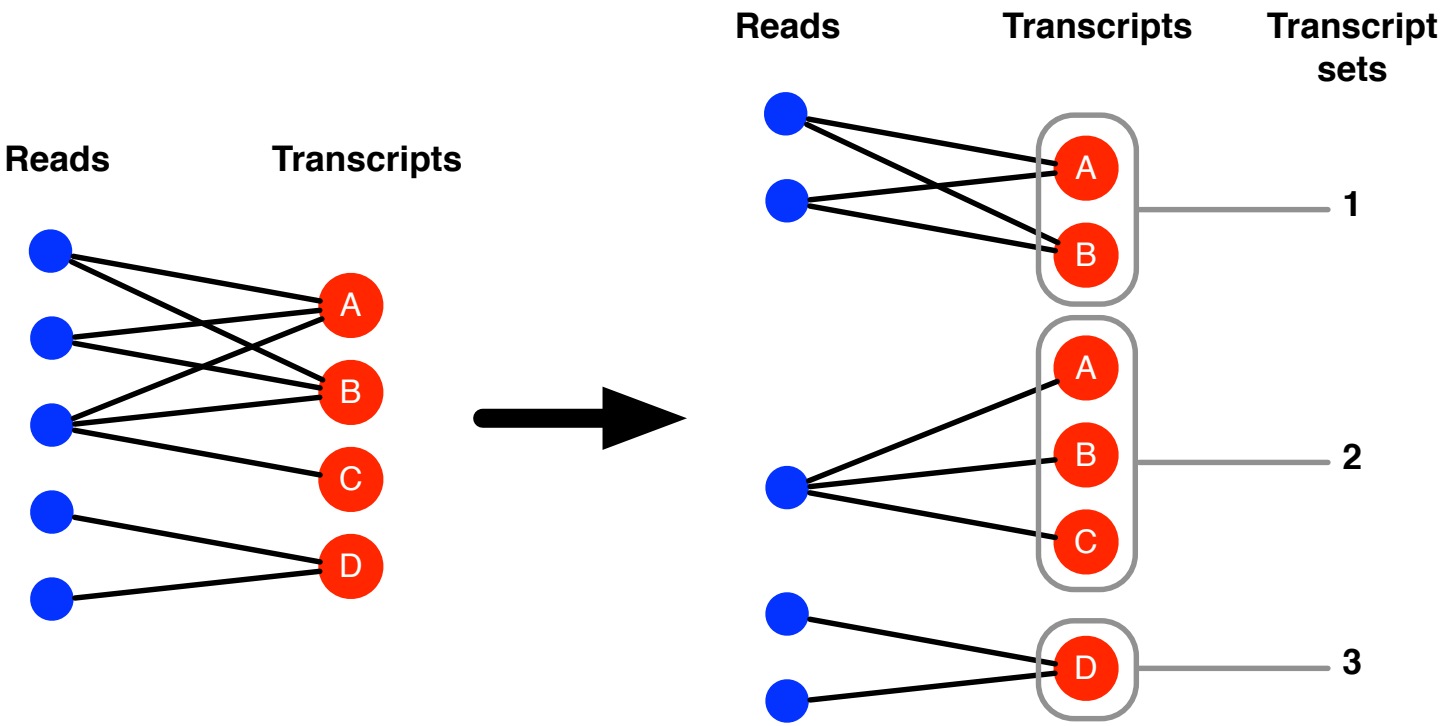
Multi-mapping reads

- Align reads back to reference transcript sequences with Bowtie (Langmead et al. 2009), allowing multiple alignments per read
- Multi-mapping structure between reads and transcripts



Multi-mapping reads

- Obtain transcript sets, such that each read maps to only 1 set
- Transcripts may belong to more than one set
- Read counts per set can be observed
- Transcripts can be isoforms sharing exons or from multiple genes



Poisson model for transcript set reads counts

Model reads per transcript set instead of per gene (Turro et al. 2011).

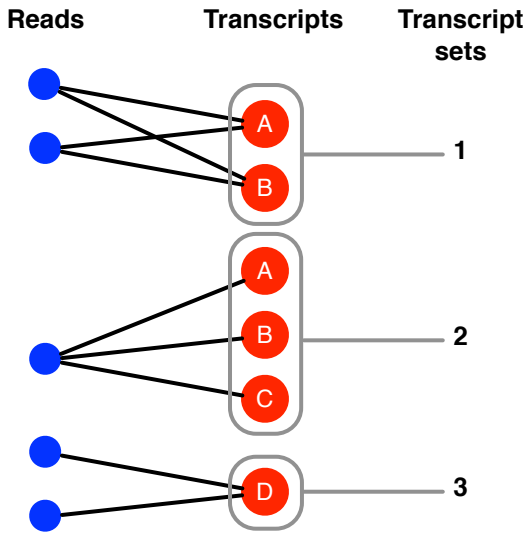
$$\text{Define } M_{it} = \begin{cases} 1 & \text{if transcript } t \text{ in set } i, \\ 0 & \text{otherwise.} \end{cases}$$

Now model for reads counts is:

$$k_i \sim Poisson(bs_i \sum_t M_{it}\mu_t),$$

where s_i is the effective length shared by transcripts in set i.

Latent variables for read counts



Transcripts

	A	B	C	D	
$M =$	1	1	0	0	1
	1	1	1	0	2
	0	0	0	1	3

Transcript sets

Observed set counts

$$\mathbf{k} = \begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix} \qquad X = \begin{pmatrix} X_{11} & X_{12} & 0 & 0 \\ X_{21} & X_{22} & X_{23} & 0 \\ 0 & 0 & 0 & X_{34} \end{pmatrix}$$

Transcript sets

Unobserved transcript counts

$$\mathbf{r} = \begin{pmatrix} r_1 & r_2 & r_3 & r_4 \end{pmatrix}$$

$$X_{it} \sim Poisson(bs_i M_{it} \mu_t),$$

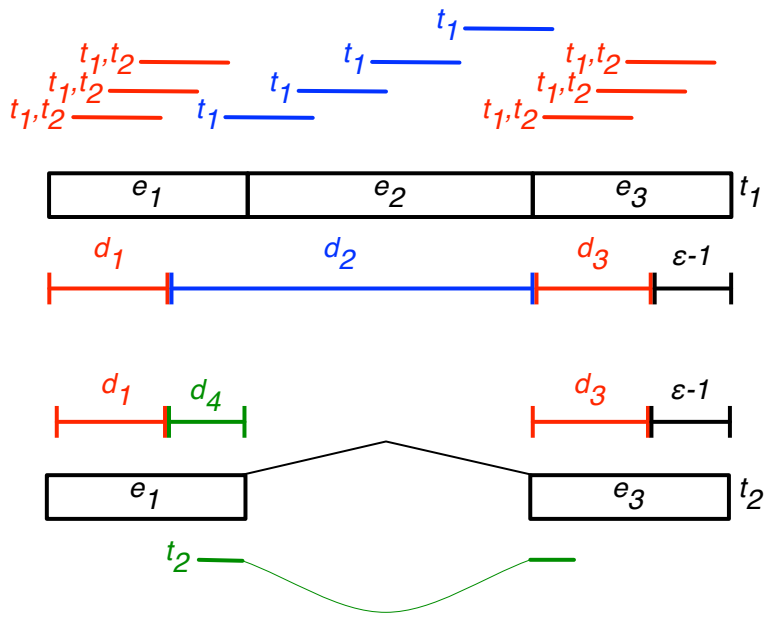
$$k_i \sim Poisson(bs_i \sum_t M_{it} \mu_t),$$

$$r_t \sim Poisson(b\mu_t \sum_i M_{it} s_i) = Poisson(bl_t \mu_t),$$

$$\{X_{i1}, \ldots, X_{in}\} | \{\mu_1, \ldots, \mu_n\}, k_i \sim Mult(k_i, \frac{M_{i1}\mu_1}{\sum_t M_{it}\mu_t}, \ldots, \frac{M_{in}\mu_n}{\sum_t M_{it}\mu_t}).$$

Same model structure for isoforms and haplo-isoforms

A

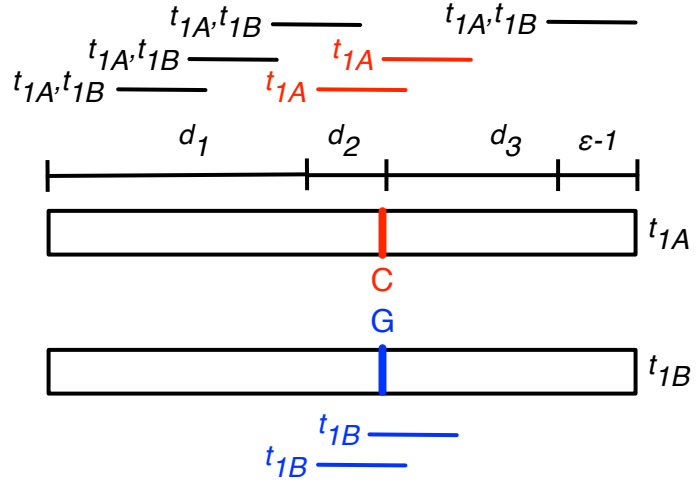


$$M = \begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \mathbf{k} = \begin{pmatrix} 6 \\ 4 \\ 1 \end{pmatrix} \qquad \begin{matrix} t_1 & t_2 \end{matrix}$$

$$\mathbf{s} = \begin{pmatrix} d_1 + d_3 \\ d_2 \\ d_4 \end{pmatrix} = \begin{pmatrix} e_1 + e_3 - 2(\epsilon - 1) \\ e_2 + \epsilon - 1 \\ \epsilon - 1 \end{pmatrix}$$

$$l_1 = s_1 + s_2 = e_1 + e_2 + e_3 - (\epsilon - 1)$$
$$l_2 = s_1 + s_3 = e_1 + e_3 - (\epsilon - 1)$$

B



$$M = \begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \mathbf{k} = \begin{pmatrix} 4 \\ 2 \\ 2 \end{pmatrix} \qquad \begin{matrix} t_{1A} & t_{1B} \end{matrix}$$

Heterozygotes can be treated like alternative exons!

Model fitting with MMSEQ

Bayesian models capture the scientific method:

- We have prior beliefs on a quantity of interest
- We perform an experiment and observe data
- We update our beliefs based on the observations

Place a Gamma prior on μ_t (vague belief about μ_t before seeing the data)

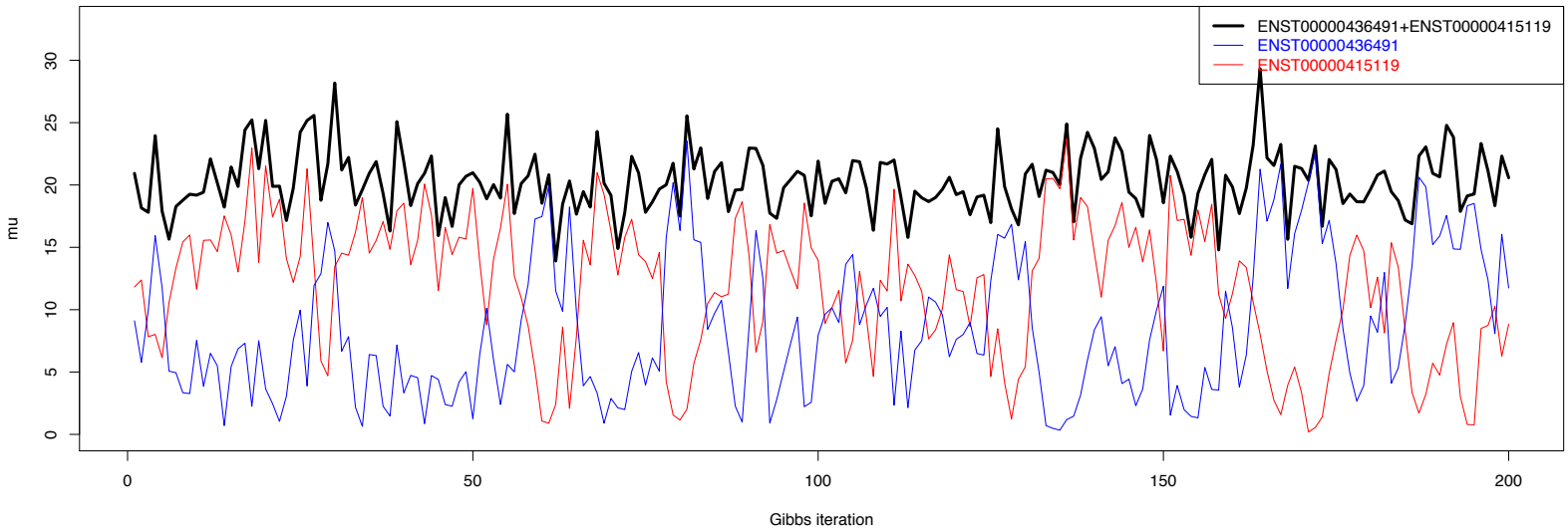
$$X_{it}|\mu_t \sim Poisson(bs_i M_{it}\mu_t),$$
$$\mu_t \sim Gamma(\alpha, \beta).$$

Sample iteratively from the full conditional distributions:

$$\{X_{i1}, \ldots, X_{it}\}|\{\mu_1, \ldots, \mu_t\}, k_i \sim Mult(k_i, \frac{M_{i1}\mu_1}{\sum_t M_{it}\mu_t}, \ldots, \frac{M_{in}\mu_n}{\sum_t M_{it}\mu_t}),$$
$$\mu_t|\{X_{1t}, \ldots, X_{mt}\} \sim Gam(\alpha + \sum_i X_{it}, \beta + bl_t).$$

Transcript amalgamation

- Some transcripts identical
- High uncertainty individually but good precision for sum
- → sum over anticorrelated posterior traces

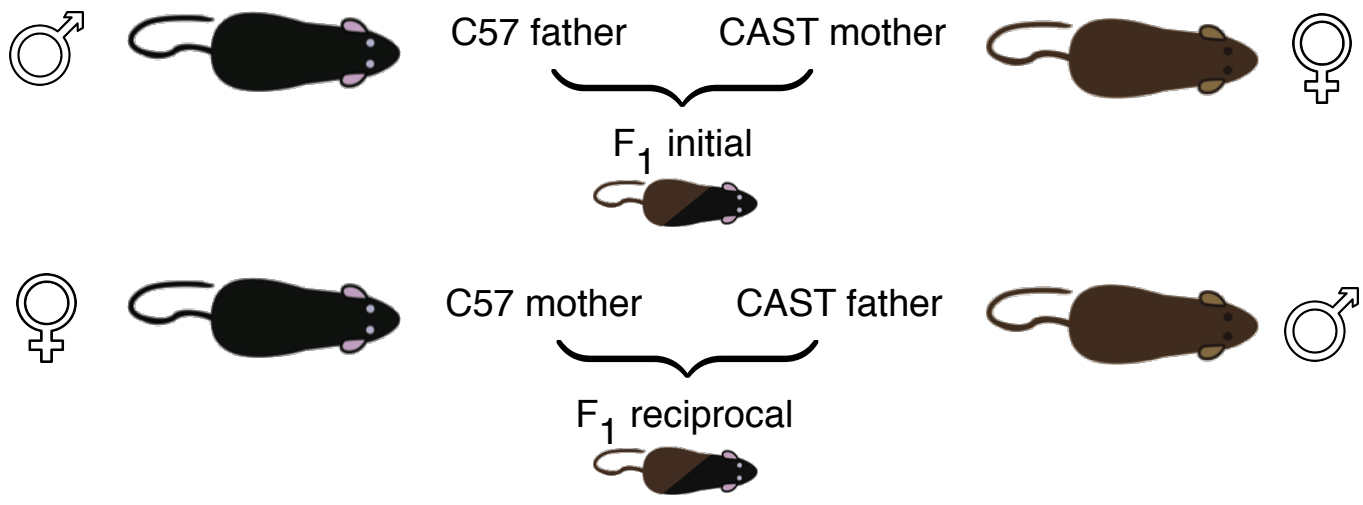


Also sum over isoform-level traces to obtain *gene-level* estimates of expression **and** associated standard deviation.

Detecting imprinting in F_1 hybrid mice

Application

- Imprinting: dominant expression according to sex of parent from whom the haplotype was inherited
- Model system: F_1 initial and reciprocal crosses of C57 and CAST inbred mice (Gregg et al. 2010)



- Can we use MMSEQ to detect imprinting?

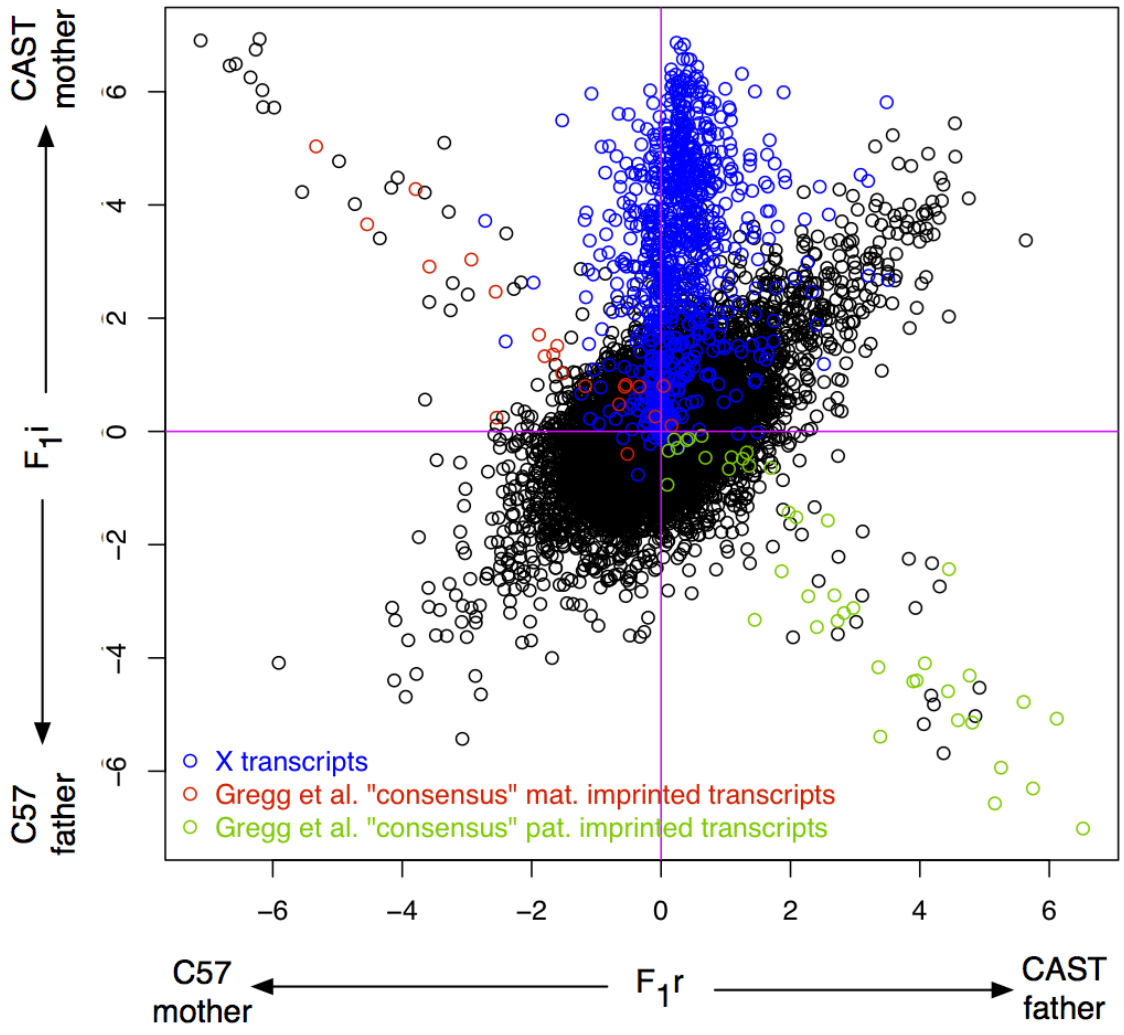
Detecting imprinting in F_1 hybrid mice

- Estimate strain and isoform-specific expression in $F_1 i$ and $F_1 r$ crosses
- Fold change for $F_1 i$ mice = $\dfrac{\text{paternal expression (C57i)}}{\text{maternal expression (CASTi)}}$
- Fold change for $F_1 r$ mice = $\dfrac{\text{paternal expression (CASTr)}}{\text{maternal expression (C57r)}}$
- A transcript is imprinted if:

$$\frac{\text{paternal expression (C57i)}}{\text{maternal expression (CASTi)}} \simeq \frac{\text{paternal expression (CASTr)}}{\text{maternal expression (C57r)}}$$

Detecting imprinting in F_1 hybrid mice

- Top-left: maternal imprinting
- Bottom-right paternal imprinting
- Diagonal: *cis* regulation
- $F_1 i$ male and $F_1 r$ female

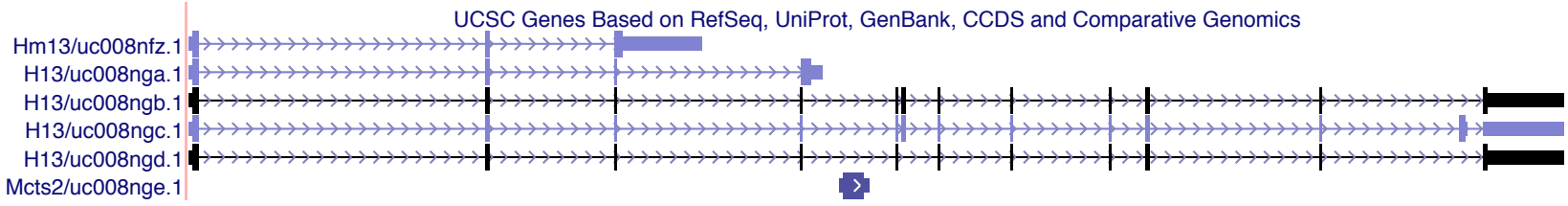


Gregg et al."Consensus imprinted": > 2 hets in favour of same parental sex, one of which has $p < 0.05$ (χ^2 test).

Detecting imprinting in F_1 hybrid mice

- Unclear how to interpret within-gene contradictions in SNP-by-SNP analyses.
- Haplo-isoform deconvolution allows us to automatically detect isoform imbalances in opposite directions within same gene:

	Mother		Father	
	$CAST_i$	$C57_r$	$C57_i$	$CAST_r$
uc008nfz.1	1.26	1.63	9.61	9.17
uc008nga.1	1.29	3.58	7.68	7.51
uc008ngb.1	12.97	9.81	0.94	0.39
uc008ngc.1	13.63	10.51	1.10	1.08
uc008ngd.1	0.22	0.18	0.30	0.13
uc008nge.1	2.01	4.20	11.29	14.66



Closing remarks

- Poisson distribution captures the unavoidable variance due to counting independent events
- The mapping of a read or read pair to a feature can be ambiguous
- Deconvolution methods help quantify expression of different isoforms and even haplotype-specific isoforms
- This is not possible with microarrays
- Yet there can be considerable uncertainty in expression parameters and posterior anti-correlation between transcripts
- Next time we will discuss statistical analysis of data from multiple samples