

# Measures of Conserved Synteny

Work was funded by the National  
Science Foundation's Interdisciplinary  
Grants in the Mathematical Sciences

All work is joint with  
John Postlethwait

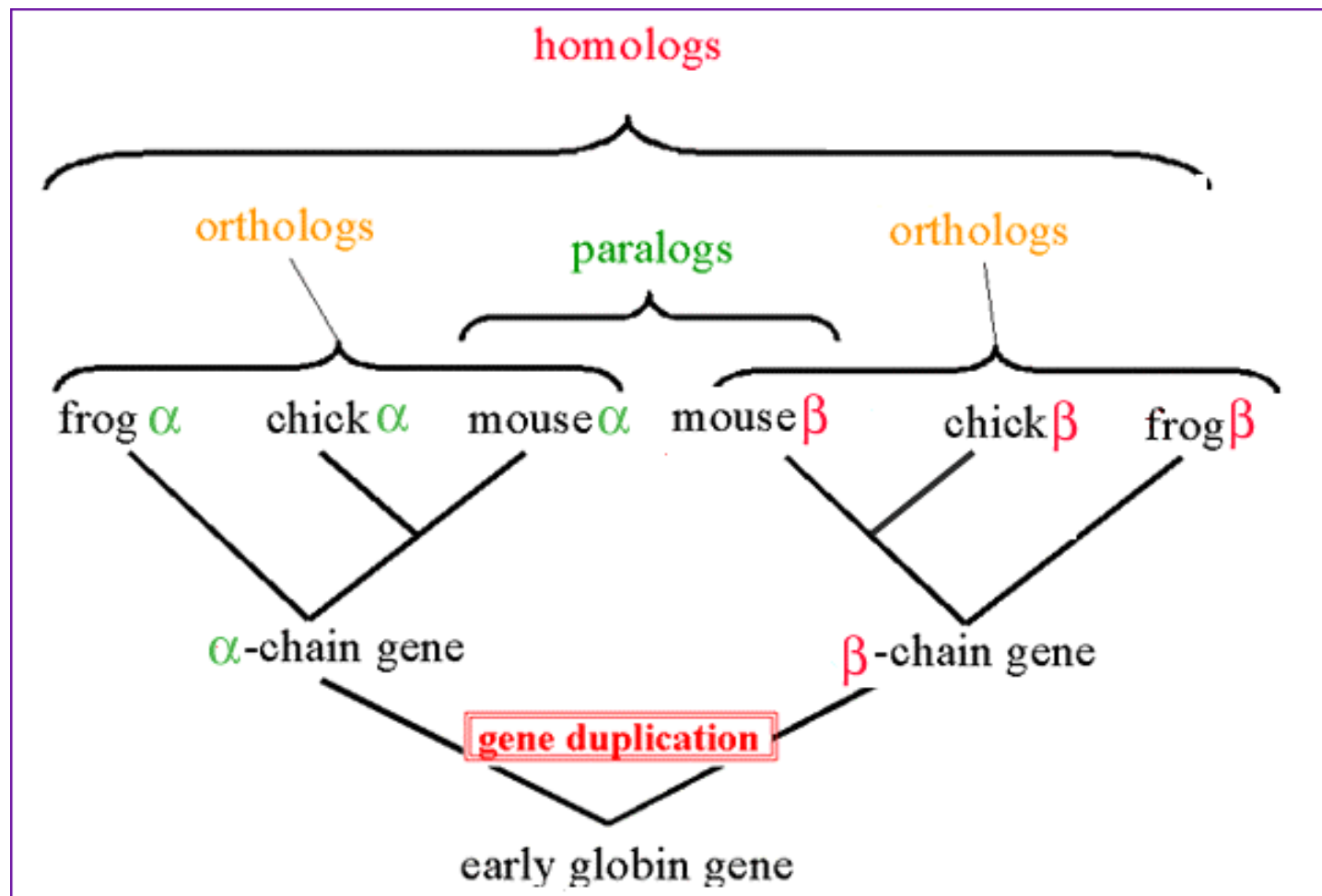
# THE QUESTION

As posed by **John Postlethwait**:

*When we look at maps of conserved synteny between two species, we see that the orthologous genes are not randomly scattered between the pairs of chromosomes. Some chromosome pairs contain many orthologs and some contain none. If we have mapped some ( $n$ ) of the orthologs between two species and observe  $k$  conserved syntenies, can we estimate the number of conserved syntenies,  $l$ , that will exist after we have mapped all the orthologs shared by both species?*

# DEFINITION

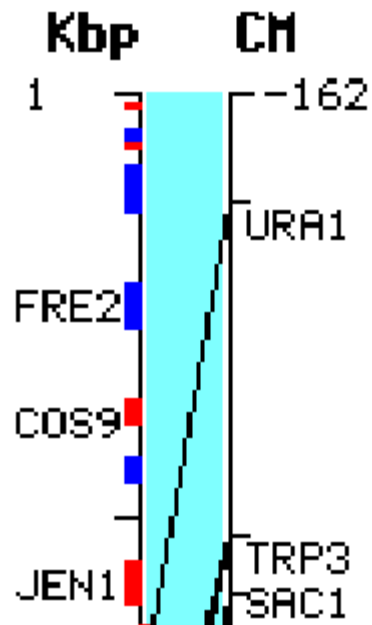
- *Orthologs* – genes that have evolved directly from the most recent common ancestor's gene



From the  
NCBI  
WEBSite

# DEFINITION

- ***Synten*** – the occurrence of two or more genes on the same chromosome within one species



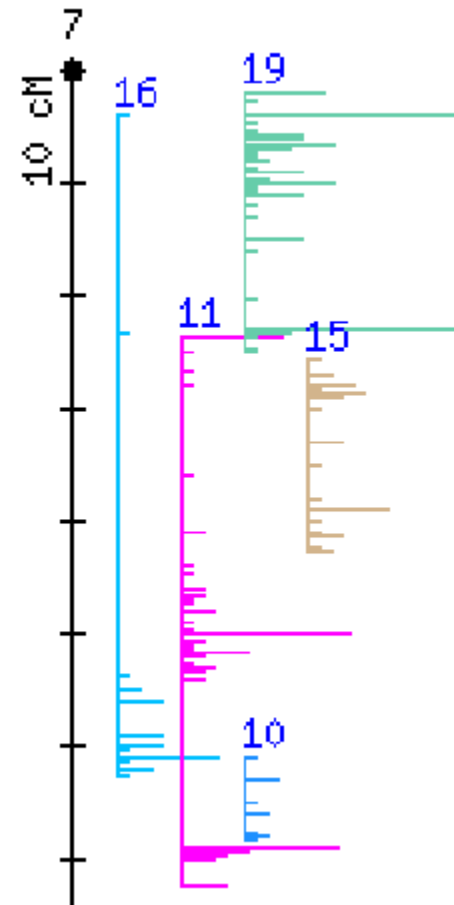
Uracil and Tryptophan  
requiring genes are  
syntenic on yeast  
Chromosome XI

# DEFINITION

- ***Conserved Synteny*** – a chromosome pair, one from each of two species, that share at least one ortholog

Mouse chromosome 7 shares genes with human chromosomes 10, 11, 15, 16, 19 but not with human chromosomes 1 and 2 (it appears).

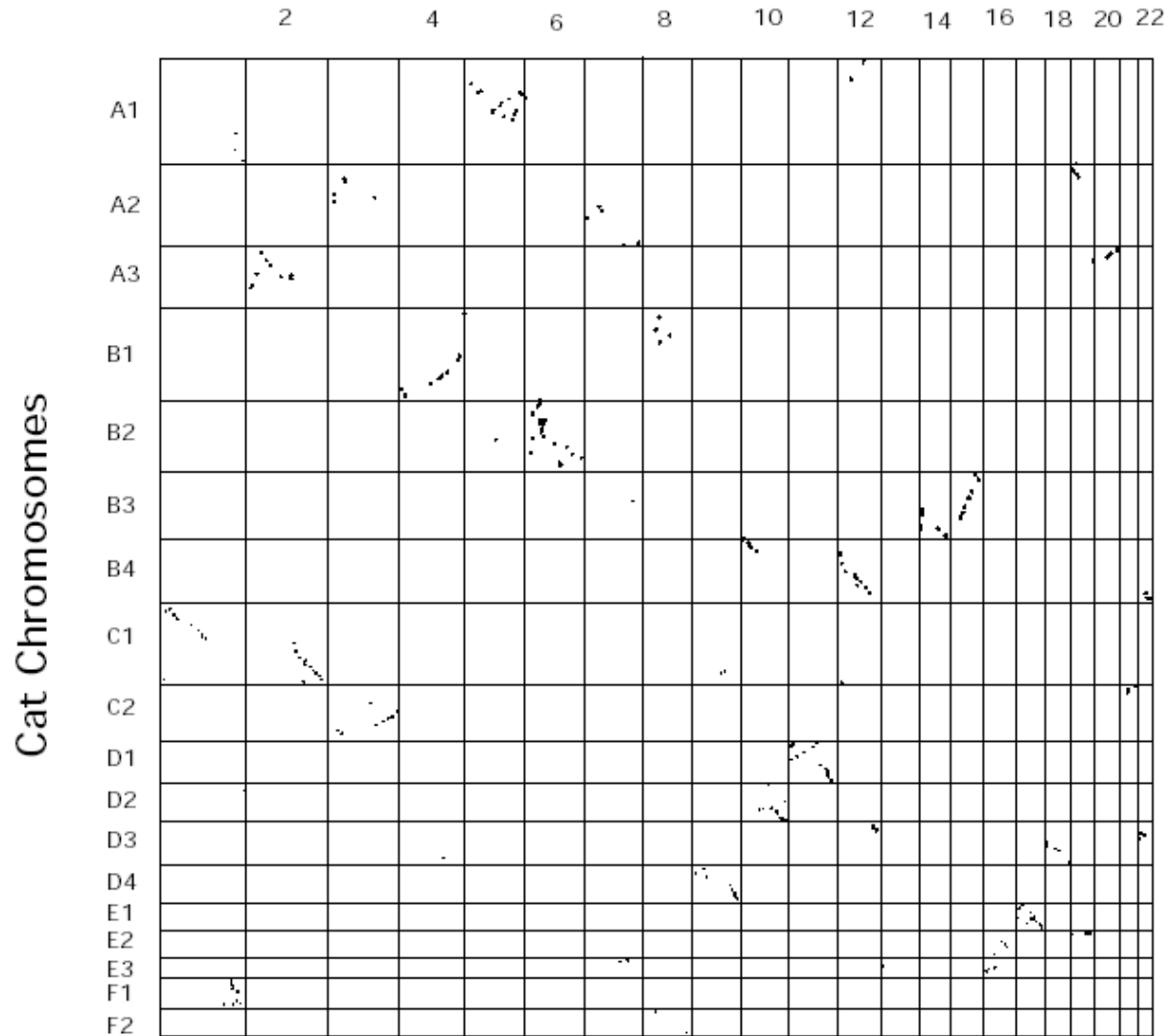
Thus, there is conserved synteny between mouse chromosome 7 and human chromosome 10, for instance.



from  
Jackson  
Laboratory

# DATA: from Murphy et al. 2000

Human Chromosomes



# OXFORD GRIDS — from Jackson Laboratory

	MOUSE																		
HUMAN	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
<b>1</b>	104	1	91	105	2	1	0	6	0	0	2	1	8	1	0	0	0	1	0
<b>2</b>	79	48	0	0	7	30	0	0	0	0	9	17	1	0	0	0	10	1	0
<b>3</b>	1	0	19	1	0	29	0	1	62	0	1	0	0	12	0	32	1	0	0
<b>4</b>	0	0	28	0	82	4	0	21	0	0	0	0	2	0	0	0	1	0	1
<b>5</b>	1	0	0	0	0	0	0	0	0	0	45	0	56	0	13	0	4	35	0
<b>6</b>	5	0	0	6	1	0	1	0	8	31	0	0	24	0	0	0	115	0	0
<b>7</b>	0	1	0	0	72	71	1	0	0	0	14	14	6	0	0	0	1	0	0
<b>8</b>	6	0	8	10	1	0	0	26	0	0	0	0	0	21	33	3	0	0	0
<b>9</b>	0	45	0	56	1	0	0	0	0	0	0	0	12	0	0	0	0	0	12
<b>10</b>	0	12	0	0	0	3	12	1	0	13	0	0	1	17	0	1	0	3	49
<b>11</b>	0	34	0	1	0	0	89	0	52	0	1	0	1	0	0	0	0	0	48
<b>12</b>	0	1	1	0	23	69	0	0	0	41	0	0	0	0	47	0	0	0	0
<b>13</b>	3	0	2	0	8	0	0	14	0	0	0	0	0	24	0	0	0	0	0
<b>14</b>	0	0	1	0	0	0	1	0	1	1	0	52	0	30	0	0	0	0	0
<b>15</b>	0	30	0	0	0	0	33	0	37	0	1	0	1	0	0	0	0	0	0
<b>16</b>	0	1	1	0	0	0	23	60	0	0	8	0	0	0	0	17	20	0	0
<b>17</b>	0	1	0	0	0	1	0	1	0	0	203	2	0	1	0	0	0	0	0
<b>18</b>	3	0	0	0	2	0	0	1	0	0	0	0	0	0	0	1	4	34	0
<b>19</b>	0	0	0	0	0	1	103	31	13	29	0	1	0	0	0	0	17	1	0
<b>20</b>	0	80	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
<b>21</b>	0	0	0	0	0	0	0	0	0	24	0	0	0	0	0	33	6	0	0
<b>22</b>	1	0	0	0	6	3	0	3	0	13	9	0	0	0	41	33	0	0	0

# FREQUENTIST MODEL

## Notation:

- Let  $m = r \times c$  be the total number of cells in the Oxford Grid
- Label the cells  $1, 2, \dots, m$
- Let the probability an ortholog fall into cell  $i$  be  $p_i$
- Let  $k$  be the number of cells observed to contain orthologs
- Let  $l$  be the number of cells that will eventually contain orthologs when the entire genomes of the two species are mapped and analyzed
- Let  $n$  be the number of orthologs mapped

$$\text{data} = \{n_1 \text{ in cell } c_1, n_2 \text{ in cell } c_2, \dots, n_k \text{ in cell } c_k\}$$

$$f(\text{data}|l) = \sum_{\text{choices for the occupied cells}} f(\text{data}|l, \text{ choice of cells}) f_1(\text{ choice of cells}|l)$$

$$= \frac{\binom{m-k}{l-k}}{\binom{m}{l}} f(\text{data}|l, \text{ choice of cells})$$

$$= \frac{\binom{m-k}{l-k}}{\binom{m}{l}} \int f(\text{data}|l, \text{ choice of cells}, p_{c_1}, \dots, p_{c_l}) \times f_2(p_{c_1}, \dots, p_{c_l}|l, \text{ choice of cells}) dp_{c_1} \cdots dp_{c_{l-1}}$$

$$= \frac{\binom{m-k}{l-k}}{\binom{m}{l}} \int \binom{n}{n_1 \dots n_k} p_{c_1}^{n_1} \cdots p_{c_k}^{n_k} (l-1)! dp_{c_1} \cdots dp_{c_{l-1}}$$

$$= \frac{\binom{m-k}{l-k}}{\binom{m}{l} \binom{n+l-1}{l-1}} = \frac{\binom{l}{k}}{\binom{m}{k} \binom{n+l-1}{l-1}}$$

- The denominator is the number of ways to choose the  $l$  chromosome pairs (cells) to contain the conserved syntenies times the number of ways to distribute  $n$  orthologs among those cells
- The numerator is the number of ways to choose the  $(l-k)$  unseen conserved syntenies from the  $(m-k)$  possibilities
- The maximum likelihood estimator for  $l$  depends on  $m$  (the number of possible conserved syntenies) only through the constraint that  $l$  is no more than  $m$

# BAYESIAN MODEL

If we assume a uniform prior distribution on  $l$ , then the posterior distribution on  $l$  is proportional to the likelihood function of  $l$  given the data. The proportionality constant is given by:

$$\frac{1}{C} = \sum_{l=k}^m f(\text{data} | l) = \sum_{l=k}^m \frac{\binom{m-k}{l-k}}{\binom{m}{l} \binom{n+l-1}{l-1}}$$

And a 95% credibility interval runs from the number of observed syntenies,  $k$ , to  $L$  where  $L$  is determined by:

$$.95 \leq C \sum_{l=k}^L \frac{\binom{m-k}{l-k}}{\binom{m}{l} \binom{n+l-1}{l-1}}$$

# RESULTS

---

	<b>Mouse</b>	<b>Rat</b>	<b>Cattle</b>	<b>Human</b>	<b>Cat</b>
<b>Mouse (19)</b>	_____				
<b>Rat (20)</b>	<b>[58, 62, 68]</b> <b>(752)</b>	_____			
<b>Cattle (29)</b>	<b>[104,138,154]</b> <b>(416)</b>	<b>[94,149,174]</b> <b>(252)</b>	_____		
<b>Human (22)</b>	<b>[154,164,170]</b> <b>(3521)</b>	<b>[99,113,122]</b> <b>(776)</b>	<b>[72,84,93]</b> <b>(776)</b>	_____	
<b>Cat (18)</b>				<b>[39,44,50]</b> <b>(324)</b>	_____

---

Entries of the form [observed number of conserved syntenies, maximum likelihood estimate, 95% upper bound], (number of mapped orthologs). Data from the Mouse Genome Database at Jackson Laboratory (July 28, 2001), BovBase at the Roslin Institute and LocusLink from NIH (June, 2001), and the Murphy *et al.* 2000 article.

# ASSUMPTIONS/PROBLEMS

- Orthologs found and mapped are a random sample of all orthologs between two species
- Raw number of conserved syntenies do not offer a good comparison of genomic distances between pairs of species due to variation in chromosome numbers

# SYNTENIC CORRELATION

## Notation:

- Let  $(i, j)$  denote chromosome  $i$  in species A and chromosome  $j$  in species B
- Let  $n_{i,j}$  be the number of orthologs mapped to  $(i, j)$
- Let  $n_{*,j}$  be the number of orthologs mapped to chromosome  $j$  in species B and anywhere in species A
- Let  $n_{i,*}$  be the number of orthologs mapped to chromosome  $i$  in species A and anywhere in species B
- Let  $e_{i,j} = (n_{*,j} n_{i,*})/n$  be the expected number of orthologs

$$\rho = \frac{\sum_{i=1}^r \sum_{j=1}^c (n_{i,j} - e_{i,j})^2}{n \min\{r - 1, c - 1\} e_{i,j}}$$

# PROPERTIES

- Proposed by Cramer, 1946
- Syntenic correlation always exists if  $0^2/0$  is interpreted as 0
- $\rho$  always lies between 0 and 1
- $\rho$  is 0 if and only if the orthologs are perfectly independently scattered
- $\rho$  is 1 if and only if, for one of the two species, knowing which chromosome an ortholog belongs to in that species determines which chromosome the ortholog belongs to in the other species

# RESULTS

---

	Mouse	Rat	Cattle	Human	Cat
<b>Mouse (19)</b>	_____	<b>0.69</b> $\bar{\pm}$ 0.04	<b>0.36</b> $\bar{\pm}$ 0.07	<b>0.31</b> $\bar{\pm}$ 0.01	
<b>Rat (20)</b>		_____	<b>0.39</b> $\bar{\pm}$ 0.10	<b>0.32</b> $\bar{\pm}$ 0.04	
<b>Cattle (29)</b>			_____	<b>0.64</b> $\bar{\pm}$ 0.05	
<b>Human (22)</b>				_____	<b>0.66</b> $\bar{\pm}$ 0.05
<b>Cat (18)</b>					_____

---

Entries above the diagonal are the syntenic correlations with 95% confidence intervals obtained through resampling procedures. Note that the syntenic correlation for humans and cats is not statistically different than that for mice and rats even though the times since divergence are approximately 92 my vs. 40.7 my (Kumar and Hedges, 1999).

## QUOTE from Sir Ronald Fisher:

*...the value of  $\chi^2$  indicates the fact, but does not measure the degree of association. ... To measure the degree of association it is necessary to have some hypothesis as to the nature of the departure from independence to be measured.*

# SYNTENIC ASSOCIATION

## Notation:

- Let  $m_A$  ( $m_B$ ) be the maximum number of orthologs mapped to any chromosome in species A (B)
- Let  $m_{*,j}$  ( $m_{i,*}$ ) be the maximum number of orthologs mapped from species B (A) chromosome  $j$  ( $i$ ) to any chromosome in species A (B)
- Let  $n$  be the total number of orthologs mapped between species A and B

$$\lambda = \frac{\sum_{i=1}^r m_{i,*} + \sum_{j=1}^c m_{*,j} - (m_A + m_B)}{2n - (m_A + m_B)}$$

# PROPERTIES

- Proposed by Guttman (1941) and Goodman and Kruskal (1954)
- Measures the proportion of errors made in assigning a gene to a chromosome in one species that can be eliminated by knowing which chromosome the gene belongs to in the other
- Makes sense when there is more than one conserved synteny
- $\lambda$  always lies between 0 and 1
- $\lambda$  is 0 if knowing which chromosome contains a gene in one species is of no help predicting the gene's chromosome in the other
- $\lambda$  is 1 if all the orthologs are concentrated in cells no two of which are in the same row or column

# EXAMPLES

Perfect dependence

		A			
		1	2	3	
B					
1		150	0	0	$\rho = 1$
2		0	100	0	
3		0	0	150	$\lambda = 1$

Perfect independence

		A			
		1	2	3	
B					
1		10	10	20	$\rho = 0$
2		40	40	80	
3		50	50	100	$\lambda = 0$

# EXAMPLES

	A		
B	1	2	3
1	100	0	0
2	50	0	0
3	0	150	0
4	0	0	100

$$\rho = 1$$

$$\lambda = 9/10$$

	A		
B	1	2	3
1	100	60	40
2	50	35	15
3	50	15	35

$$\rho = 0.0227$$

$$\lambda = 0$$

# EXAMPLES

		A		
B		1	2	3
1	150	50	0	
2	0	0	50	
3	0	0	150	

$\rho = \frac{1}{2}$  (only depends on form, not on the entries)

$\lambda = \frac{3}{4}$  (depends on the entries)

Ancestor Guess:

