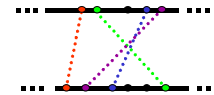


## Gene clusters in comparative genomics: Accident or design?

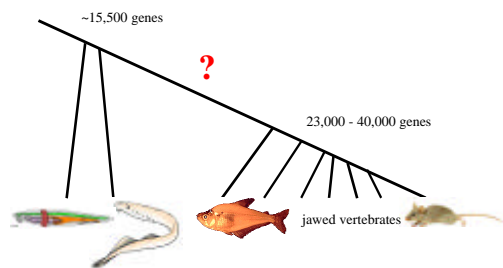
Dannie Durand  
Computer Science, Biological Sciences  
Carnegie Mellon University

## Outline

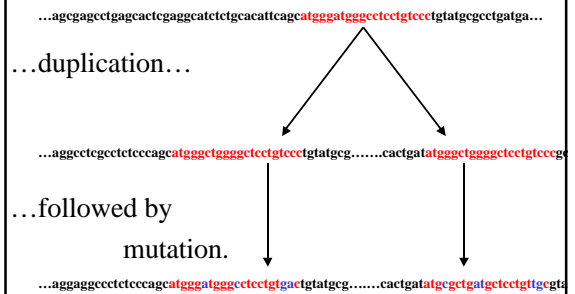
- Vertebrate genome evolution
- Tests for gene clustering
- An application: Evolution of the insulin/IGF1 signalling pathway
- Open problems



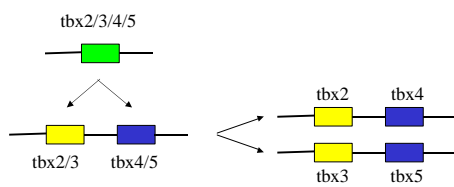
## Evolution of vertebrate genomes



## New genes come from ...

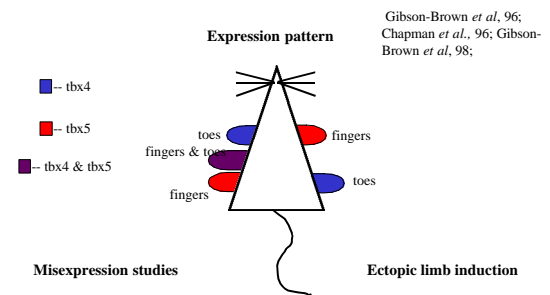


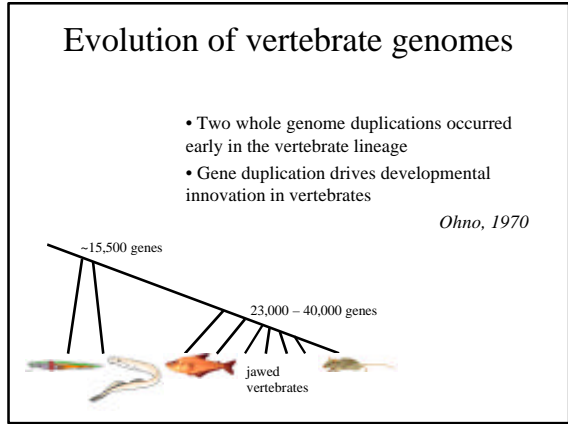
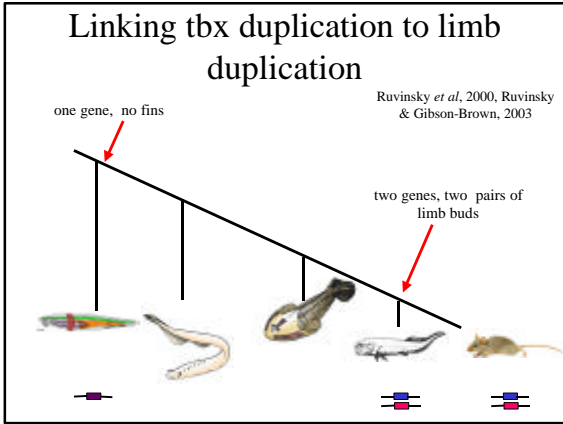
## An example: linking gene duplication to limb development



Agulnik *et al.*, 1996

## Tbx4 and Tbx5 have a role in limb specification and patterning





### Predictions of Ohno's Hypothesis

- Temporal:
  - An excess of duplicated genes originated before the emergence of bony fish (~500MYA)
- Spatial:
  - Regions that share *significant* similarity in gene content and order.

### Genome duplication and divergence

An Example:

A B C D E F   P Q R S T U V

A B C D E F   P Q R S T U V

*Whole genome duplication*

### Genome duplication and divergence

An Example:

A B C D E F   P Q R S T U V

A B   P Q R   S T U V

C D E F

*Reciprocal translocation*

### Genome duplication and divergence

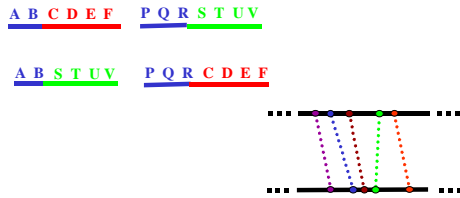
An Example:

A B C D E F   P Q R S T U V

A B S T U V   P Q R C D E F

*Reciprocal translocation*

## Conserved Segments



Distinct chromosomal regions with *identical gene content and order.*

## Genome duplication and divergence

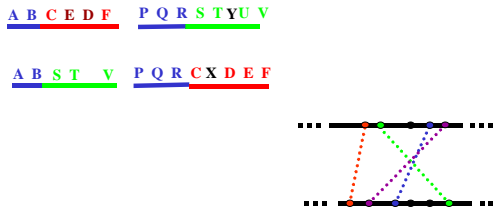
An Example:



Local mutations

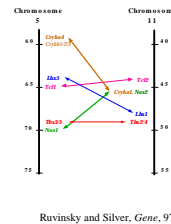
- Small inversions
- Deletions
- Insertions

## Gene Clusters



Distinct chromosomal regions with similar gene content. *Gene content and order are not preserved.*

## Local Spatial Evidence



Ruvinsky and Silver, *Gene*, 97

Vertebrate gene clusters discussed in the literature	
TBOX	Cryb, Lhx, Nos, Tbx, Tcf, Prkar
MHC	Abc, C3/4/5, Col, Hsp, Notch, Pbx, Psmb, Ror, Ten
HOX	Achr, Ccnd, Cdc, Cdk, Dlx, En, Evx, Gli, Hh, Hox, If, Inhb, Nhr, Npy/Ppy, Wnt
FGR	Adr, Ank, Egr, Fgfr, Pa, Vmat, Lpl
MATN	Eya, Hck, Matn, Myb, Myc, Sdc, Src
...	...

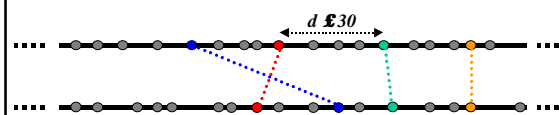
## Spatial Evidence :

### Genome scale study of paralogous regions

- Identify paralogous genes
  - Stringent standards reduced data set from 20,800 to 9,500 aa sequences
- Find candidate duplicate regions (“paralogons”)
- Test significance of candidate paralogons.

McLysaght, Hokamp, Wolfe, *Nat Gen*, 2002

## Paralagon Identification



- “Compared blast hits with those of neighboring proteins, scanning them for matches within the same remote chromosomal location”
- “Gap size: maximum number ( $d$ ) of unduplicated genes allowed between two duplicated genes in each paralagon.”

McLysaght, Hokamp, Wolfe, *Nat Gen*, 2002

## Spatial Evidence : Genome scale study of paralogous regions

- Identify paralogous genes
  - Stringent standards reduced data set from 20,800 to 9,500 aa sequences
- Find candidate paralogous regions (“paralogons”)
- Test significance of candidate paralogons.

McLysaght, Hokamp, Wolfe, *Nat Gen*, 2002

## Significance Testing

McLysaght, Hokamp, Wolfe, *Nat Gen*, 2002

### Monte Carlo hypothesis testing

- Shuffled the mapping from sequences to loci
- Searched for paralogons
- Compared the number of paralogons containing  $m$  paralogons in the observed data and the shuffled data

## Significance Testing

McLysaght, Hokamp, Wolfe, *Nat Gen*, 2002

$m$	$O$	$S$	$Z$
3	260	159.1	8.2
4	93	30.1	11.2
5	55	6.9	17.8
$\geq 6$	96	2.6	57.5

$O$ : Observed data

$S$ : Shuffled data

$Z$ : Zscore (*st.dev.*)

## Conclusions

McLysaght, Hokamp, Wolfe, *Nat Gen*, 2002

- Human genome was generated by a pattern of large scale duplication.
- Observed paralogons are
  - consistent with one whole genome duplication
  - do not show strong support for two rounds
- “...any paralogon with  $m \geq 6$  was very likely to have been formed by a single duplication of a chromosomal region”

## Questions about Gene Clusters

- Is a particular cluster statistically significant?  
Evidence concerning a particular region.
- Is the observation of  $k$  clusters in a whole genome comparison significant?  
Evidence concerning the processes that contributed to the evolution of the genome.

## Outline

- Vertebrate genome evolution
- Tests for gene clustering
- An application: Evolution of the insulin/IGF1 signalling pathway
- Open problems

## Tests for Gene Clustering

Durand and Sankoff, *JCB*, 2003

### Individual clusters:

Is a *particular* gene cluster significant?

### Aggregate clusters counts:

Is it significant to observe  $k$  clusters?

### Orthologous comparisons:

Two different genomes

### Paralogous comparisons:

Genome self-comparison

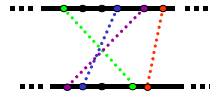
### Null hypothesis:

Random gene order

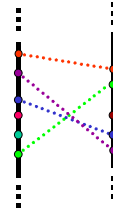
### Alternate hypotheses:

Evolutionary history

Functional selection



## Individual Gene Clusters

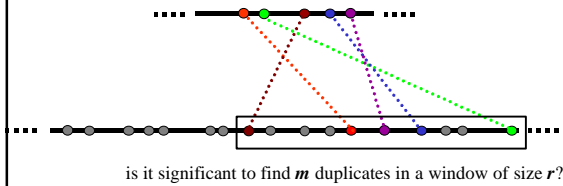


- Reference region
- Window sampling
- Whole genome scans

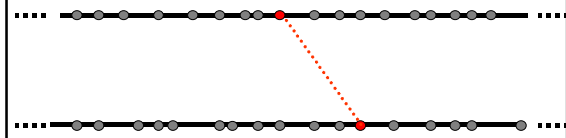
The significance of a cluster depends on how you found it

## Reference Region: A simple two parameter model

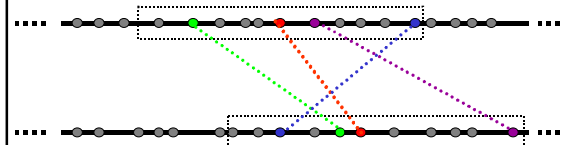
Given: a region of interest containing  $m$  genes,



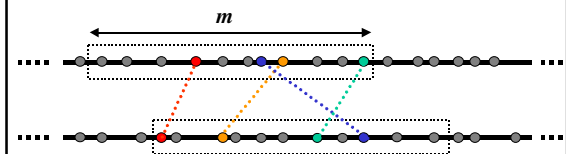
## Window sampling



## Window sampling

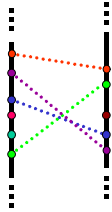


## Whole genome scans



For each run of  $m$  consecutive genes, what is the probability of finding the same genes in a window of size  $r$  elsewhere?

## Individual Gene Clusters



- Reference region  $O(n)$
- Window sampling  $O(1)$
- Whole genome scans  $O(n^2)$

The significance of a cluster depends on how you found it

## Individual cluster significance

- ↓ Possibilities considered
- Conserved order
- Gene families
- Partial clusters
- Multiple copies

## Reference Region

**Genome:**  $G = 1, \dots, n$ , no gene families (initially).

**Event:** observation of  $m$  genes in a window of size at most  $r$  in  $G$ .

$$q(n, m, r) = \frac{\left[ \binom{n-r}{m-1} \right] + \binom{r}{m}}{\binom{n}{m}}$$

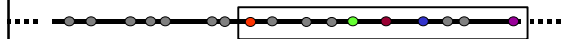
**Example:**  $M = \{ \text{red, green, blue, purple} \}$ ,  $m = 5$ ,  $r = 10$



## Reference Region conserved order

$$q(n, m, r) = \frac{1}{m!} \frac{\left[ \binom{n-r}{m-1} \right] + \binom{r}{m}}{\binom{n}{m}}$$

**Example:**  $M = \{ \text{red, green, blue, purple} \}$ ,  $m = 5$ ,  $r = 10$

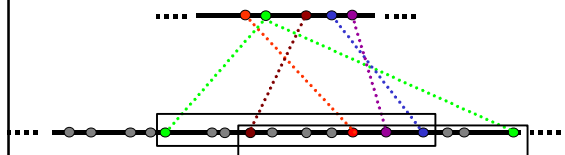


## Individual cluster significance

- ↓ Possibilities considered
- ↑ Conserved order
- Gene families
- Partial clusters
- Multiple copies

## Gene Families

**Given:** a reference region containing  $m$  genes,



is it significant to find  $m$  paralogs in a window of size  $r$ ?

## Gene families: Expected number of clusters

**Given**

- a set  $M$  of  $m$  pre-specified genes,
- each gene,  $j$ , has,  $f(j)$  copies in  $G$

there are

$$\Phi(M) = \prod_{j \in M} f(j)$$

sets of genes that match  $M$ .

Expected number of clusters,

$$S_F(n, m, r) = \Phi(M) \cdot q(n, m, r)$$

For fixed gene family size,  $f$ ,

$$S_F(n, m, r) = f^m \cdot q(n, m, r)$$

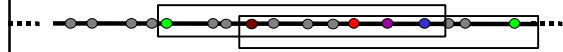
## Gene Families: Probability of at least one cluster

**Event  $E_i$ :** observation of the  $i$ th set of paralogs homologous to  $M$ .

$$P(\bigcup E_i) = \sum P(E_i) - \sum P(E_{i,j}) + \sum P(E_{i,j,k}) \dots$$

$$\approx f^m q(n, m, r) - f^{m-1} \binom{f}{2} q(n, m+1, 2r-m+1)$$

$$M = \{ \bullet \bullet \bullet \bullet \bullet \}$$

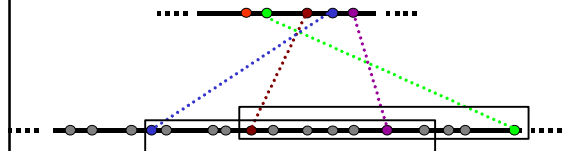


## Individual cluster significance

- ↓ Possibilities considered
- ↑ Conserved order
- ↓ Gene families
- Partial clusters
- Multiple copies

## Partial clusters

**Given:** a reference region containing  $m$  genes,

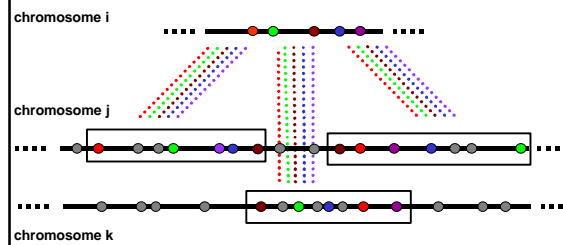


Three out of five genes are found in a window of size 10

## Individual cluster significance

- ↓ Possibilities considered
- ↑ Conserved order
- ↓ Gene families
- ↓ Partial clusters
- Multiple copies

## Multiple copies



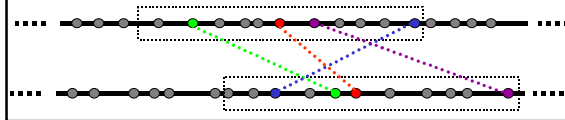
## Individual cluster significance

- ↓ Possibilities considered
- ↑ Conserved order
- ↓ Gene families
- ↓ Partial clusters
- ↑ Multiple copies

## Window sampling No gene families

Given windows  $W1$  and  $W2$  of size  $r$  drawn from genomes  $G1$  and  $G2$ , the probability that they share at least  $m$  genes is:

$$q(n, m, r) = \sum_{i=m}^r \binom{r}{i} \binom{n-r}{r-i}$$

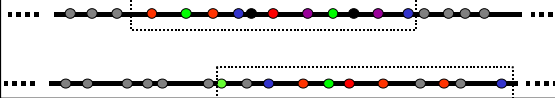


## Window sampling With gene families

Given windows  $W1$  and  $W2$  of size  $r$  drawn from genomes  $G1$  and  $G2$ , the probability that they share at least  $m$  distinct gene families is:

$$\sum_{k=m}^r P_1(k, n, r) P_2(m, k, n, r)$$

where  $P1()$  is the probability that there are  $k$  distinct gene families in  $W1$  and  $P2()$  is the probability that at least  $m$  of those  $k$  families are represented in  $W2$ .



## Tests for individual gene clusters

**Event:** Given a set  $M$  of  $m$  pre-specified genes,  $h < m$  are found in  $k$  windows of size at most  $r$  in a random genome with gene families.

### Tests:

- Expected number of clusters
- Probability of observing a least one cluster

Also: tests for clusters found through

- Window sampling
- Whole genome scans

## Example: Significance of TBX cluster

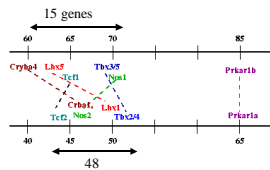
Expected number of gene partial clusters:

$$S_{FH}(n, h, m, r) = \binom{m}{h} (f-1)^h \cdot q(n-r, h, m, r)$$

Reference chromosome 5

$n = 2888$ ,  $f = 3$   
 $m = 15$ ,  $h = 6$ ,  $r = 48$

$S_{FH}(2888, 15, 48, 3) = 1.0 \cdot 10^{-3}$



Ruvinsky and Silver, 1997

## Example: Significance of TBX cluster

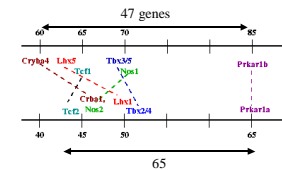
Expected number of partial gene clusters:

$$S_{FH}(n, h, m, r) = \binom{m}{h} (f-1)^h \cdot q(n-r, h, m, r)$$

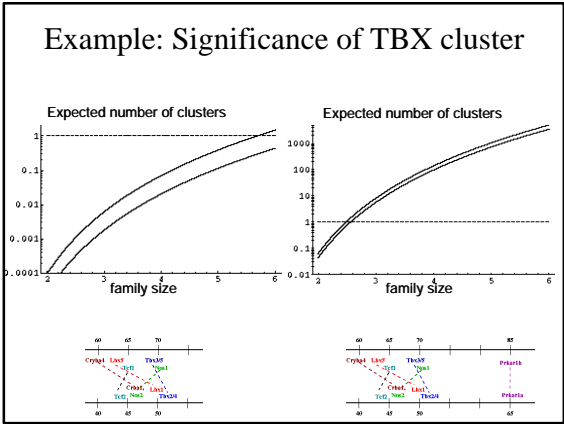
Reference chromosome 5

$n = 2888$ ,  $f = 3$   
 $m = 47$ ,  $h = 7$ ,  $r = 65$

$S_{FH}(2888, 47, 65, 3) = 1.0 \cdot 10^{-3}$



Ruvinsky and Silver, 1997



### Tests for *whole genome comparison*

**Aggregate clusters counts:**  
Is it significant to observe  $k$  clusters?

**Tests:**

- Whole genome comparison:  
Expected number of paired clusters
- Window sampling:  
Given a particular sample of window pairs,  
– Expected number of paired clusters in sample.  
– Probability of observing a least one cluster in sample

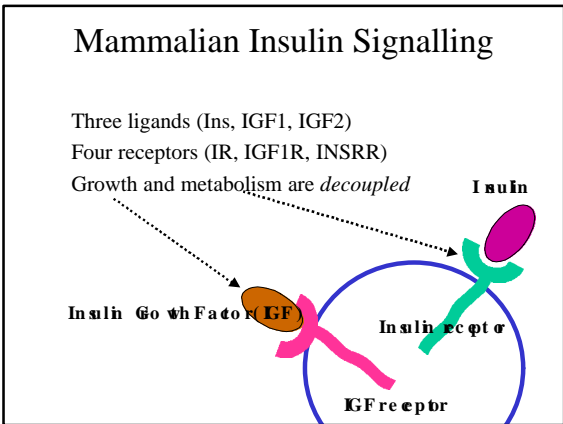
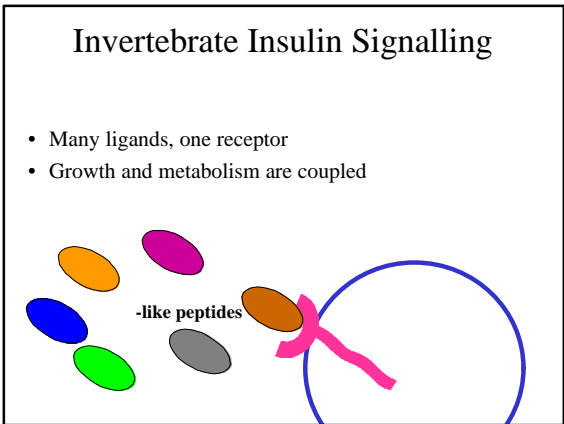
- ### Outline
- Vertebrate genome evolution
  - Tests for gene clustering
  - **An application: Evolution of the insulin/IGF1 signalling pathway**
  - Open problems

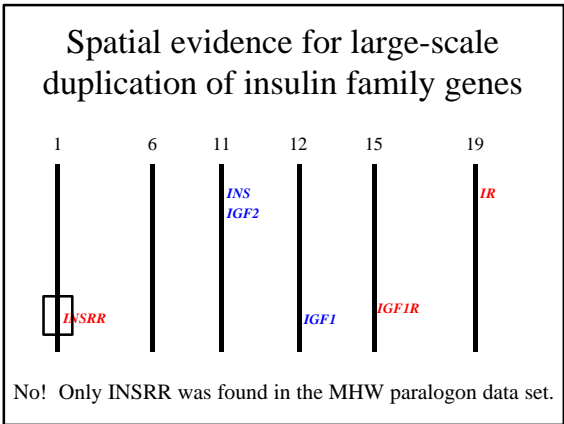
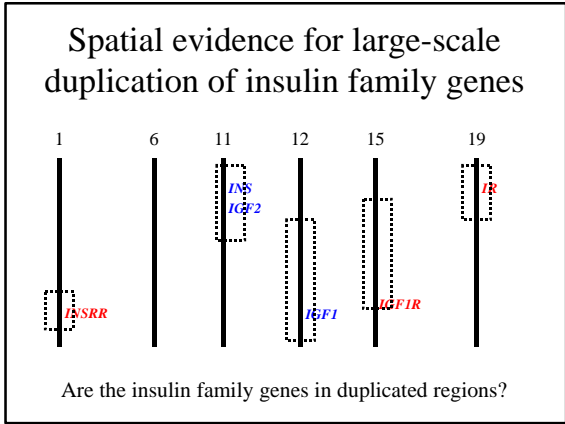
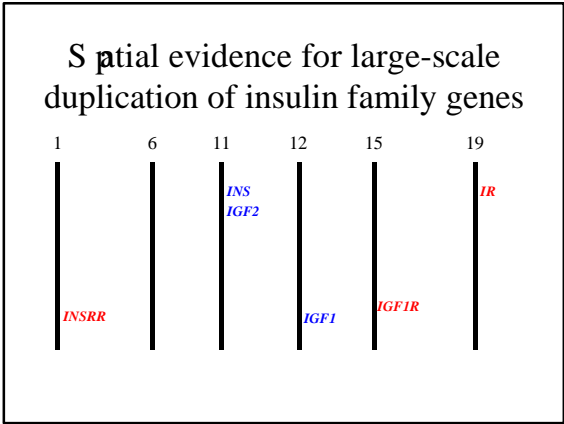
### An example: evolution of insulin

The insulin family regulates metabolism, growth, cell differentiation, aging.

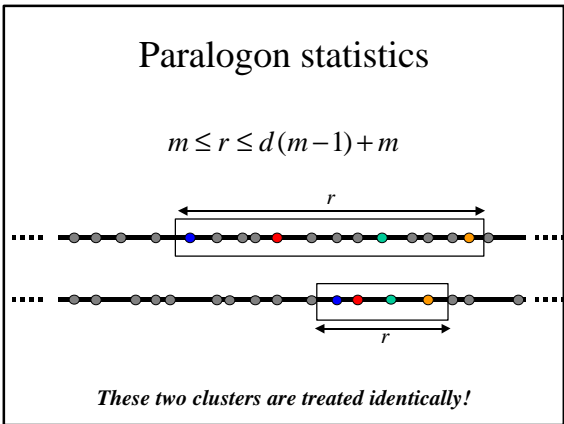
- Reduced insulin signalling in increases life-span in fly, worm and mouse.
- We must understand the comparative evolution of the insulin family in vertebrates and invertebrates to know the relevance of invertebrate models for human aging.

Tatar, Bartke and Antebi, *Science*, 2003





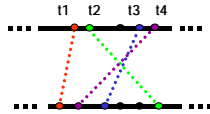
- ### Spatial evidence for large-scale duplication of insulin family genes
- Current spatial evidence does not *support* large-scale duplication in vertebrate insulin evolution.
  - Current spatial evidence does not *falsify* large-scale duplication in vertebrate insulin evolution.
    - State of human gene finding
    - Many genes were excluded from analysis
    - Only paralogs of size  $\geq 6$  are included in data set.
- Statistical analysis target at regions surrounding insulin genes is needed!*



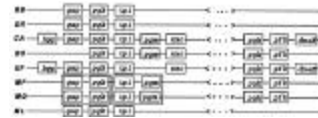
- ### Outline
- Vertebrate genome evolution
  - Tests for gene clustering
  - An application: Evolution of the insulin/IGF1 signalling pathway
  - Open problems

## Including additional information

- Gene orientation
- Intron/exon structure, flanking regions,...
- Age



## Gene Clustering for Functional Inference in Bacterial Genomes



The Use of Gene Clusters to Infer Functional Coupling, Overbeek et al., *PNAS* 96: 2896-2901, 1999.

Incorporating evolutionary history in the null hypothesis

## Modeling overlapping clusters



Individual clusters:

Better estimates of the probability of observing at least one cluster

Aggregate cluster counts:

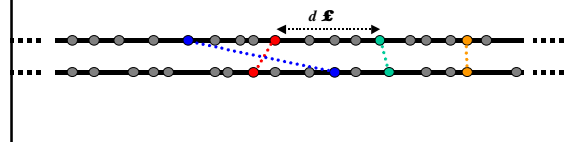
$p$ -values

Distribution of  $P(\text{observing } k \text{ clusters})$

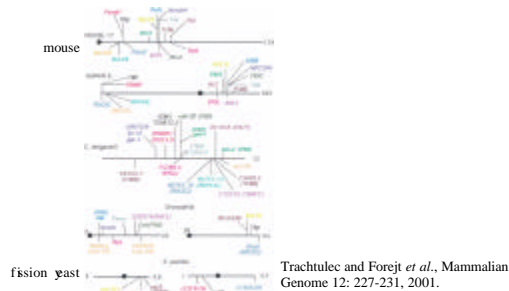
## How to choose the window size to fit the data?



versus



## A window sampling model for three or more clusters



Trachtulec and Forejt et al., *Mammalian Genome* 12: 227-231, 2001.

## Better models of gene families

- Exact gene families sizes  
Each gene,  $j$  has  $f(j)$  copies in  $G$

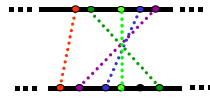
*Hard to calculate*

- Fixed gene family sizes  
 $j$  has  $c_j$  copies in  $G$

*Need more accurate approximation that is tractible*

## Determining homology

### 1. Distinguishing orthologs and paralogs



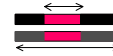
### 2. Identifying pairs of genes that share a common ancestor across their entire length.

## Identifying Homologous Genes

### 1. Identify genes with significant sequence similarity

```
... atgcaggaatccocagtgaatgcaaggagtccocagagcgtgccaggcgtgtct...
... cgaagacttcgggtcacgtatgagaggctccocagtgtagaggtcggcagacgt...
```

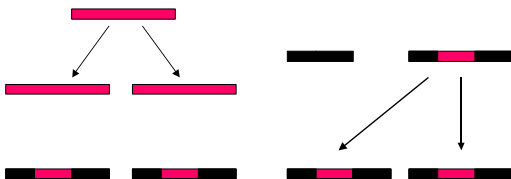
### 2. Length limitations: require that $\geq$



## Alternate Hypotheses

Whole gene copy

Domain copy



## Acknowledgements

David Sankoff,  
University of Ottawa

Narayanan Raghupathy  
CMU

Marc Tatar  
Brown University

Limb Development:

- Jeremy Gibson-Brown
- Ilya Ruvinsky
- Sergei Agulnik
- Silver Lab  
(Princeton)
- Papaioannou Lab  
(Columbia)

NHGRI, David and Lucille Packard Foundation