Significance Tests for Max-Gap Gene Clusters

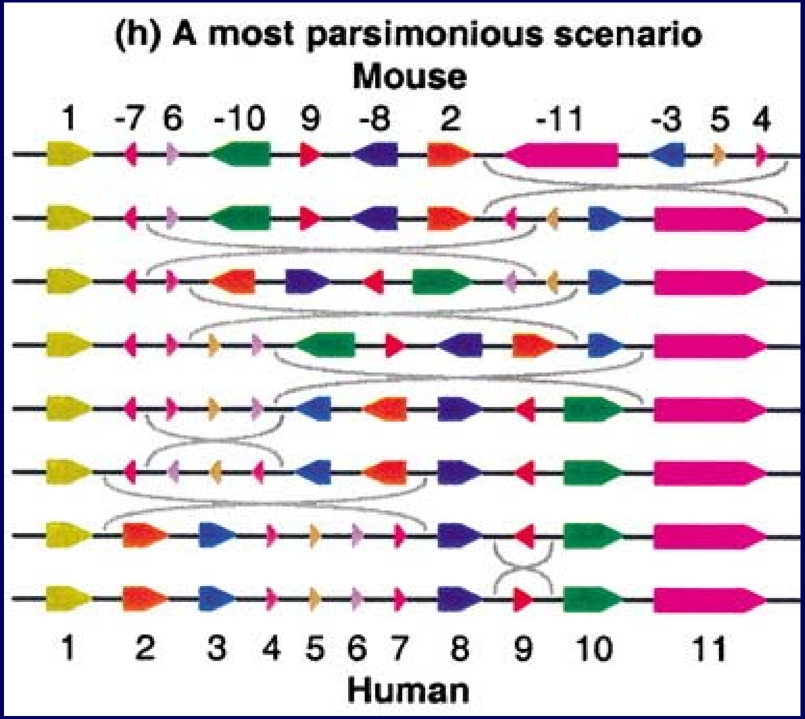
Rose Hoberman

joint work with Dannie Durand and David Sankoff

Identification of homologous chromosomal segments is a key task in comparative genomics

- Genome evolution
  - Reconstruct history of chromosomal rearrangements
  - Infer ancestral genetic map
  - Phylogeny reconstruction

…



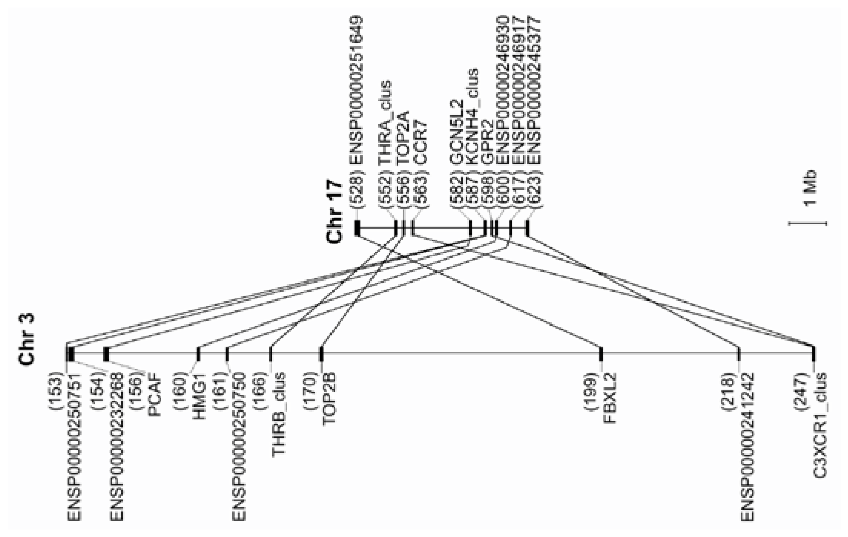
Pevzner, Tesler. Genome Research 2003

Identification of homologous chromosomal segments is a key task in comparative genomics

...

- Genome self-comparisons
  - evidence for ancient whole-genome duplications

...

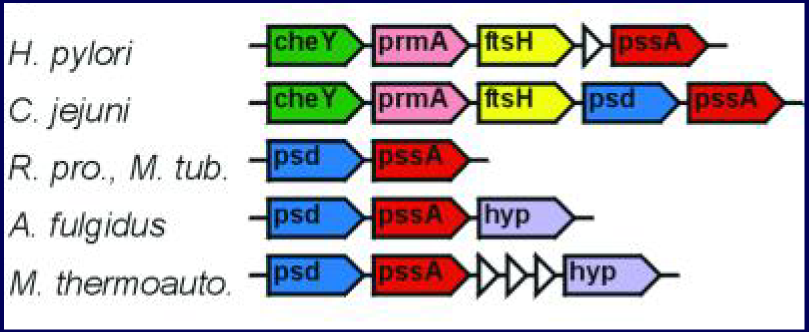


McLysaght, Hokamp, Wolfe. Nature Genetics, 2002.

Identification of homologous chromosomal segments is a key task in comparative genomics

…

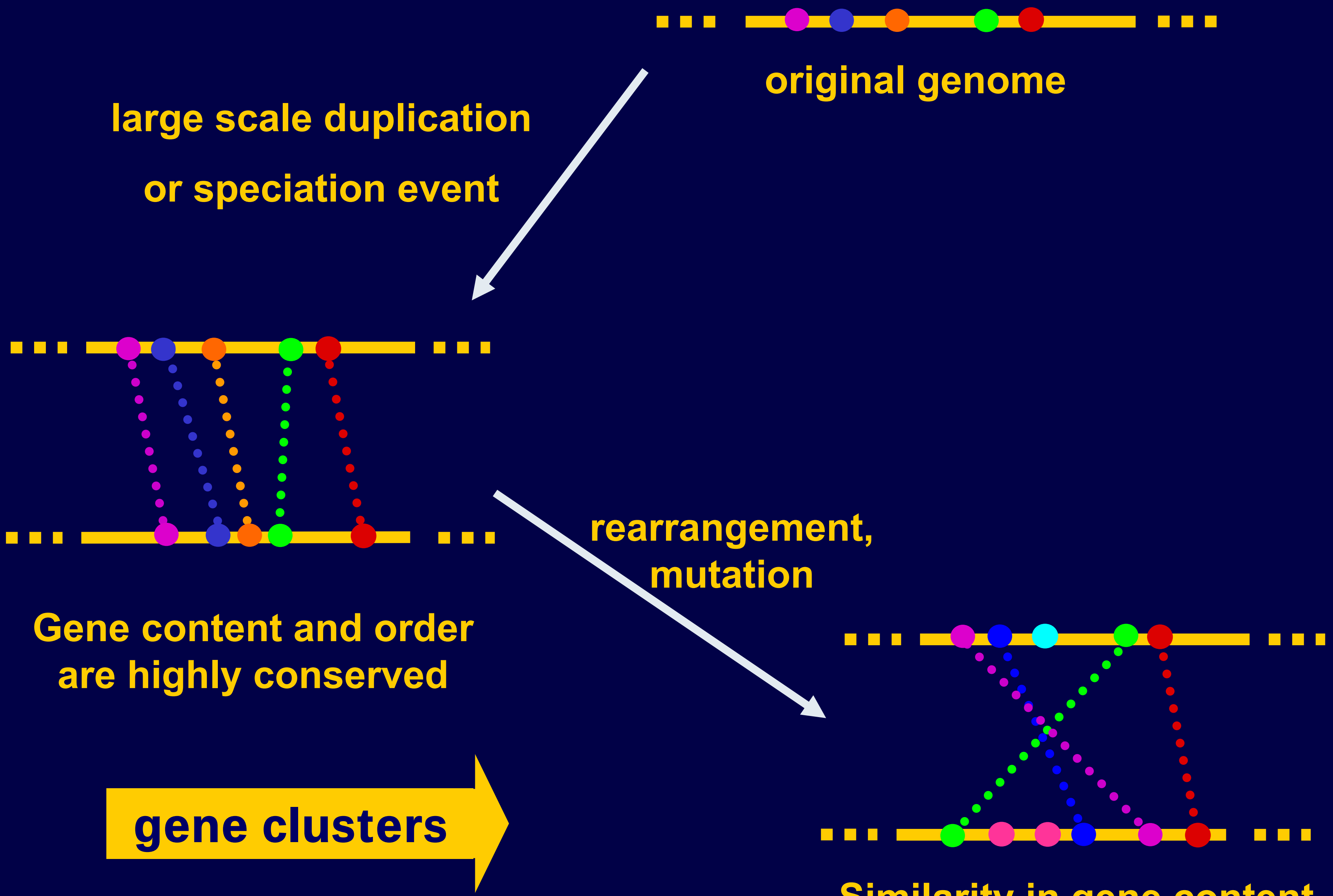
- Understand gene function and regulation in bacteria
  - Predict operons
  - Identify horizontal transfers
  - Infer functional associations



Snel, Bork, Huynen. PNAS 2002

What do such homologous segments look like?

Why is identifying them a difficult problem?



original genome

large scale duplication

or speciation event

Gene content and order
are highly conserved

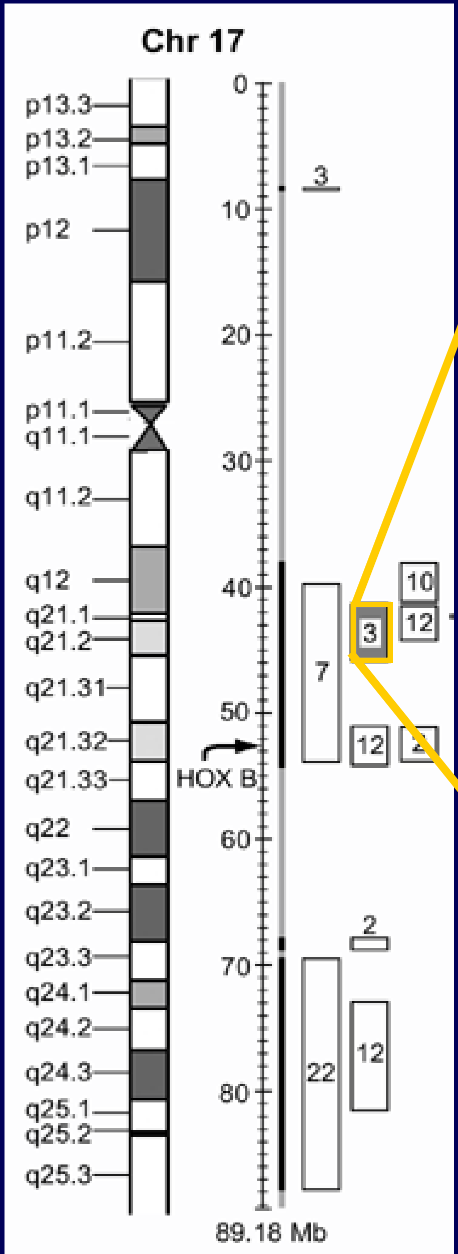
rearrangement,
mutation

gene clusters

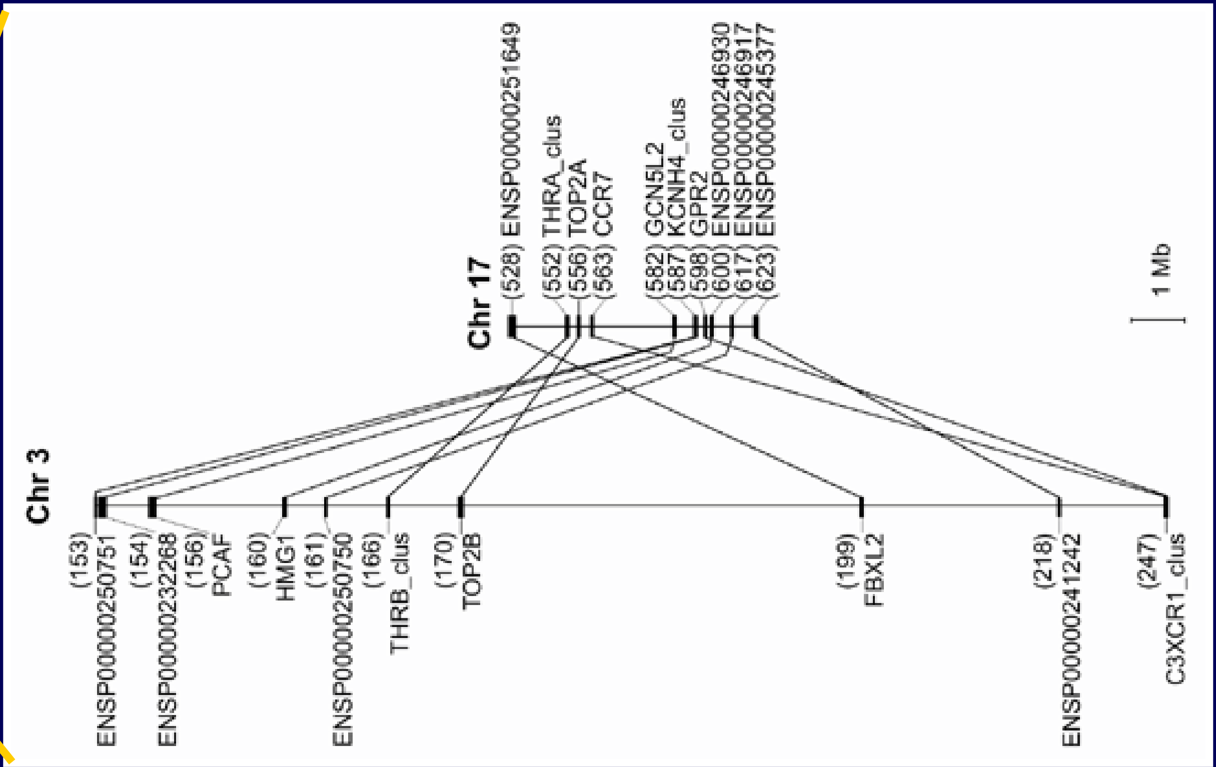
Similarity in gene content

Neither gene content nor order is strictly preserved

Whole Genome Comparison of Human with Human



McLysaght, Hokamp, Wolfe. Nature Genetics, 2002.



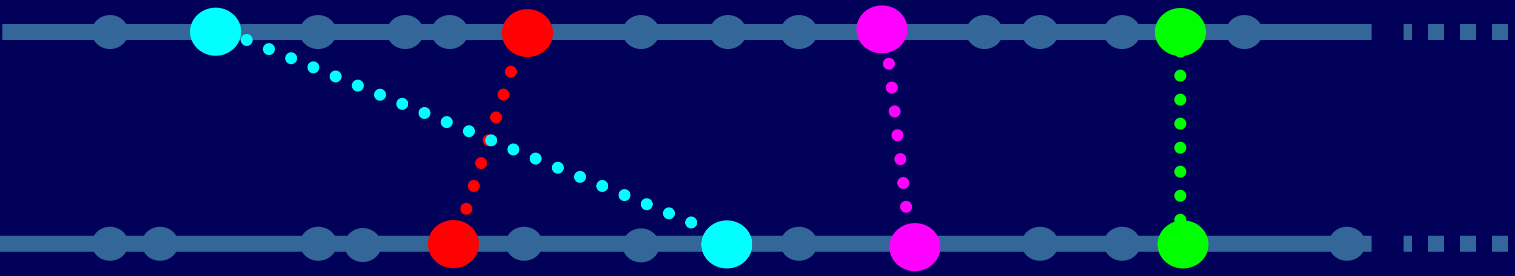
Could this pattern have occurred by chance?

Approach

- Genome as a sequence of genes (or markers)
  - a single chromosome
  - genes are unique
  - each gene has at most one match in the other genome

- Hypothesis testing
  - Alternate hypothesis: common ancestry
  - Null hypothesis: random gene order

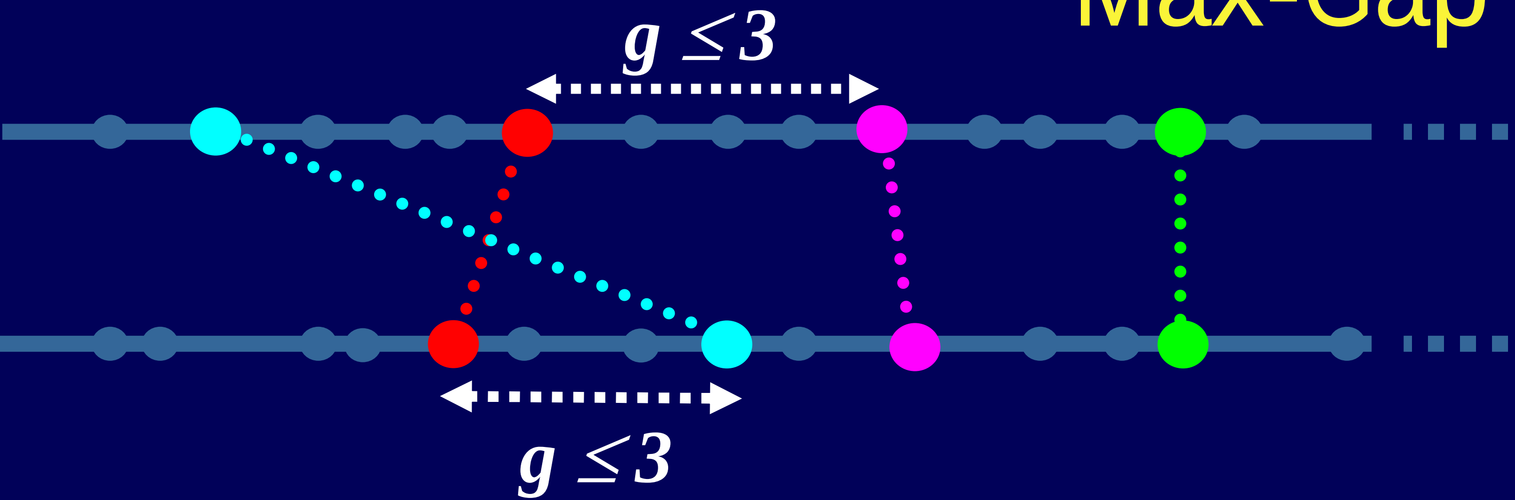
Gene Clusters



Similar gene content

Neither gene content nor order is strictly preserved

Max-Gap Clusters



- The *gap* between genes is the number of intervening genes
- A set of genes form a max-gap cluster if the gap between adjacent genes is never greater than *g* on either genome

Max-Gap Clusters are Commonly Used in Genomic Analyses

Blanc et al 2003, recent polyploidy in Arabidopsis

Venter et al 2001, sequence of the human genome

Overbeek et al 1999, inferring functional coupling of genes in bacteria

Vandepoele et al 2002, duplications in Arabidopsis through comparison with rice

Vision et al 2000, duplications in Eukaryotes

Lawr

Tama

Wolf

McLy

Cogh

Seoighe and Wolfe 1998, genome rearrangements after duplication in yeast

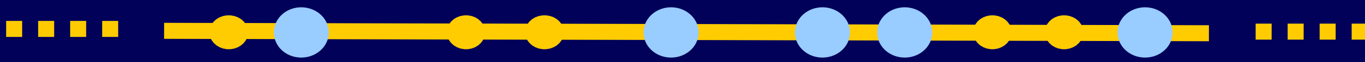
Chen et al 2004, operon prediction in newly sequenced bacteria

Blanchette et al 1999, breakpoints as phylogenetic features

...

- ***no* analytical statistical model for max-gap clusters**
- **statistical significance assessed through randomization**

Statistics for max-gap gene clusters



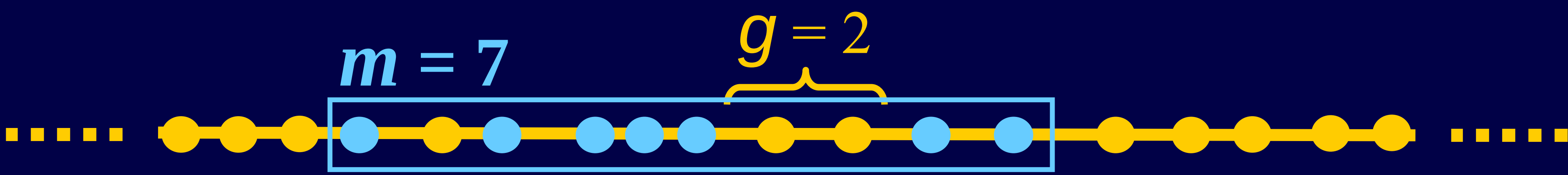
1. Reference set:

Inputs

1. a genome: ***G = 1, ..., n*** of unique genes
2. a set of ***m special*** genes

2. Whole Genome Comparison

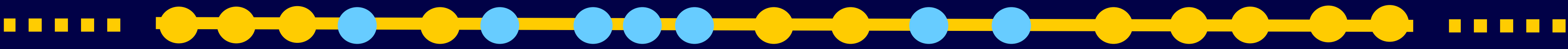
Significance of a *complete cluster*



- **Test statistic:** the maximum gap observed between adjacent blue genes

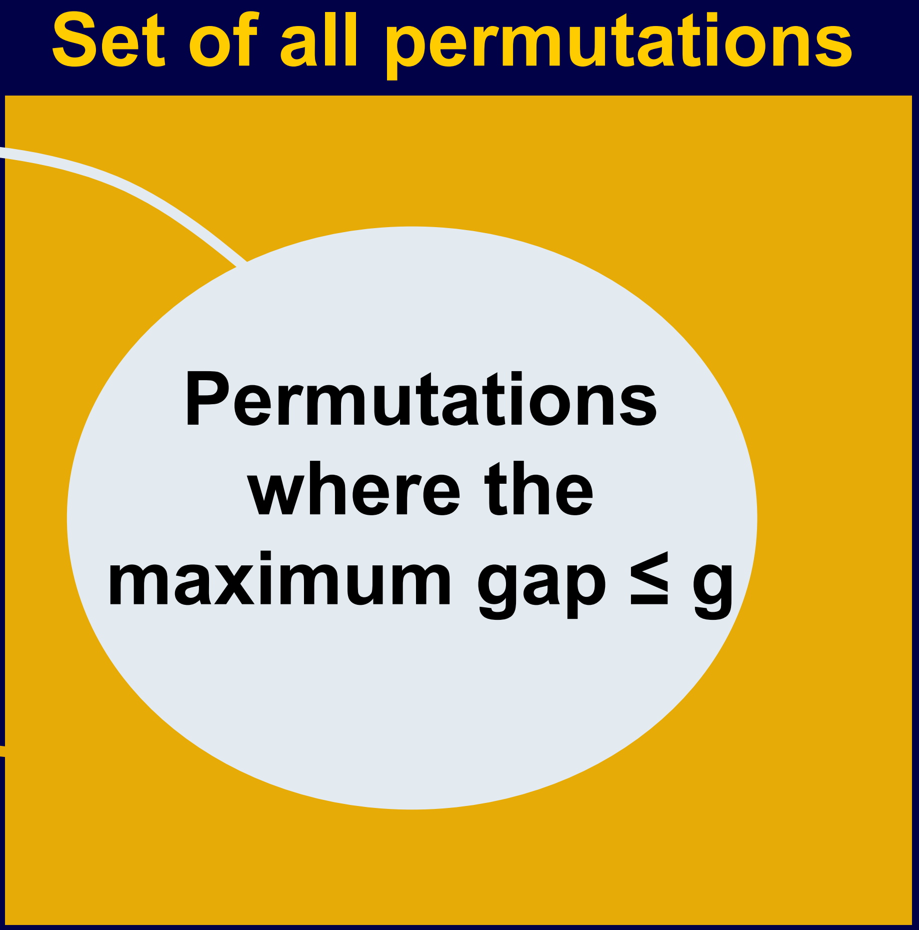
- **P-value**: the probability of observing a maximum gap ≤ g, under the null hypothesis

Compute probabilities by counting



The problem is how to count this

$$\text{P-val} = \frac{N(m, g, n)}{\binom{n}{m}}$$

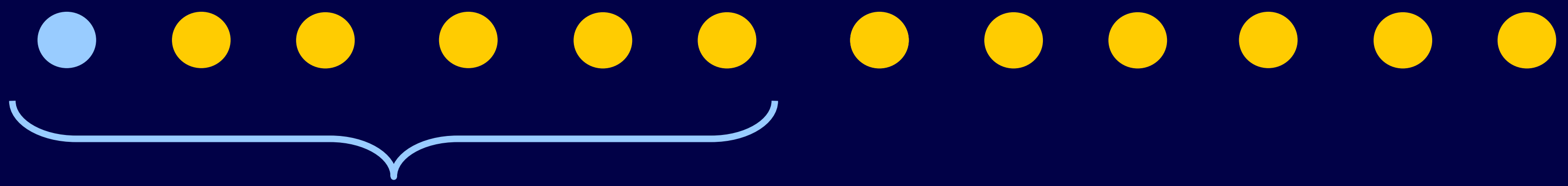


Set of all permutations

Permutations where the maximum gap ≤ g

$$N(m,g,n) = (n - w + 1)(g+1)^{m-1} + E$$

number of ways to start a cluster, e.g. ways to place the first gene and still have w-1 slots left

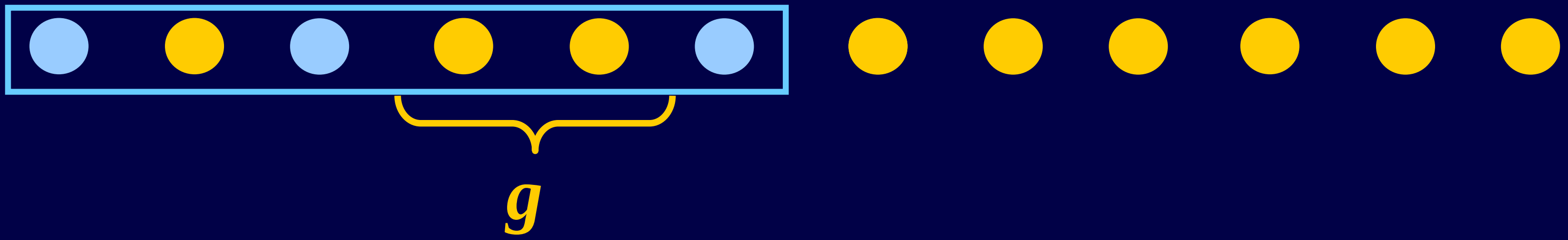


$w = (m\text{-}1)g + m$

$$N(m,g,n) = (n - w + 1)(g + 1)^{m-1} + E$$

number of ways to start a cluster, e.g. ways to place the first gene and still have ***w*-1** slots left

ways to place the remaining ***m*-1** blue genes, so that no gap exceeds *g*

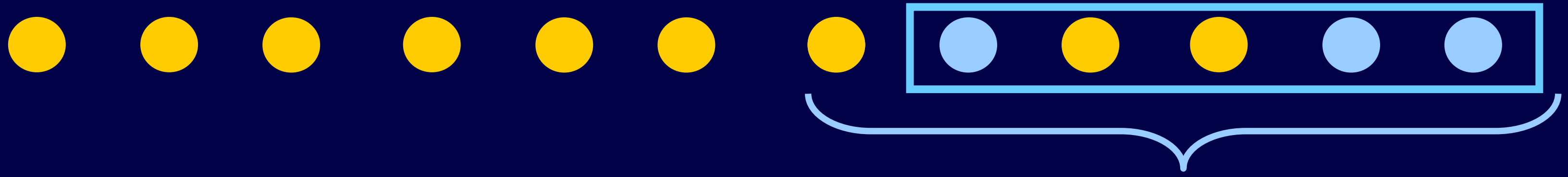


$$N(m,g,n) = (n-w+1)(g+1)^{m-1} + E$$

number of ways to start a cluster, e.g. ways to place the first gene and still have *w-1* slots left

ways to place the remaining ***m-1*** blue genes, so that no gap exceeds *g*

edge effects



$w = (m-1)g + m$

Adding edge effects…

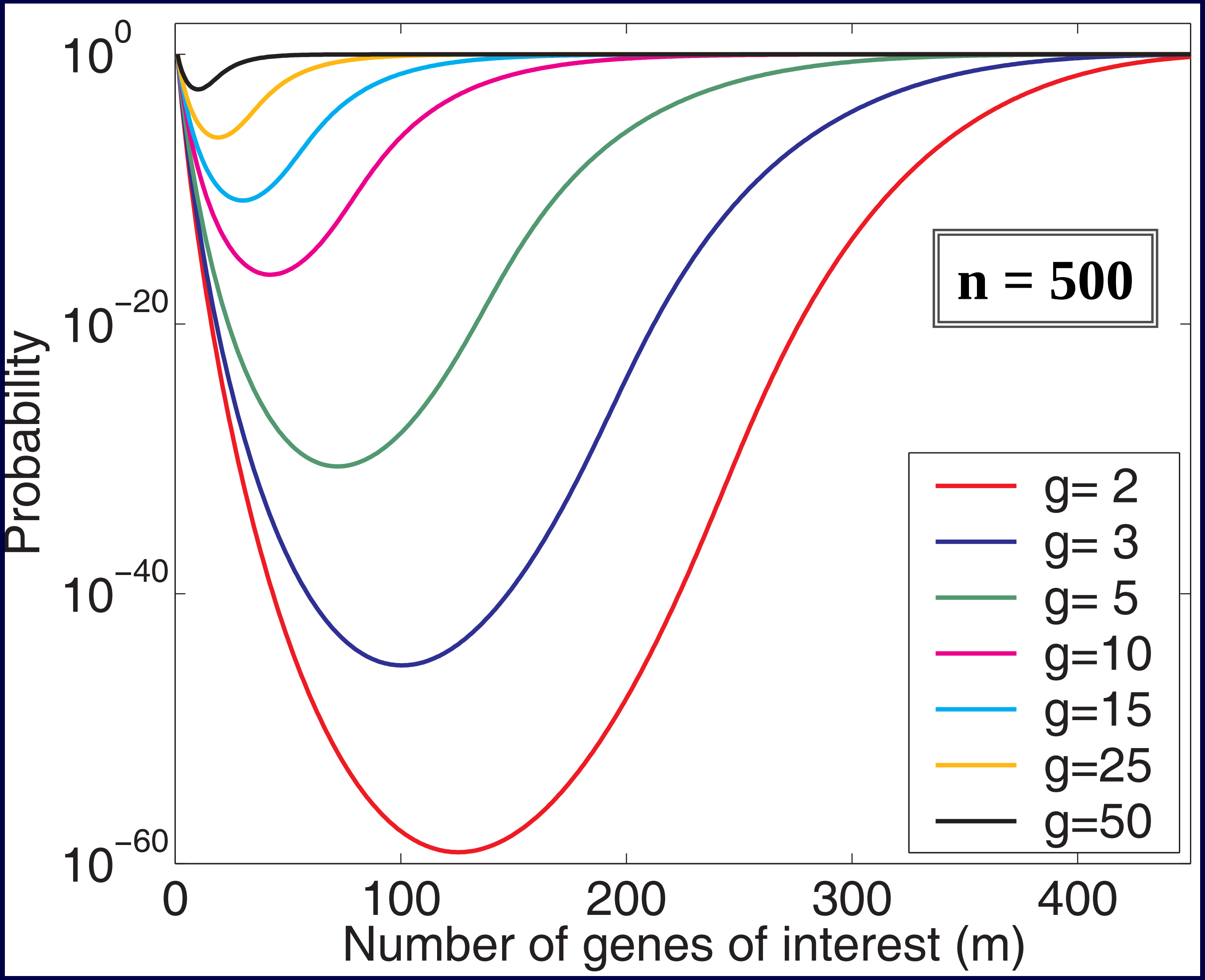
$$N(m,g,n) = \begin{cases} \left[n - w + 1 + \dfrac{w-m}{2} \right] \cdot (g+1)^{m-1}, & \text{if } w \leq n+1 \\ \displaystyle\sum_{i=0}^{\lfloor (n-m)/(g+1) \rfloor} (-1)^i \binom{m-1}{i} \binom{n-i(g+1)}{m} & \text{otherwise.} \end{cases}$$

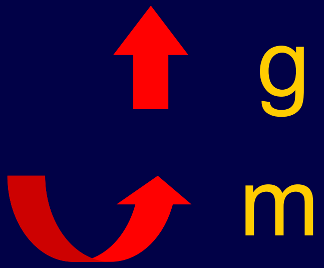
Hoberman, Sankoff, Durand. JCB 2005.

I used this equation to calculate probabilities

for various parameter values ➲

Probability of Observing a Complete Cluster





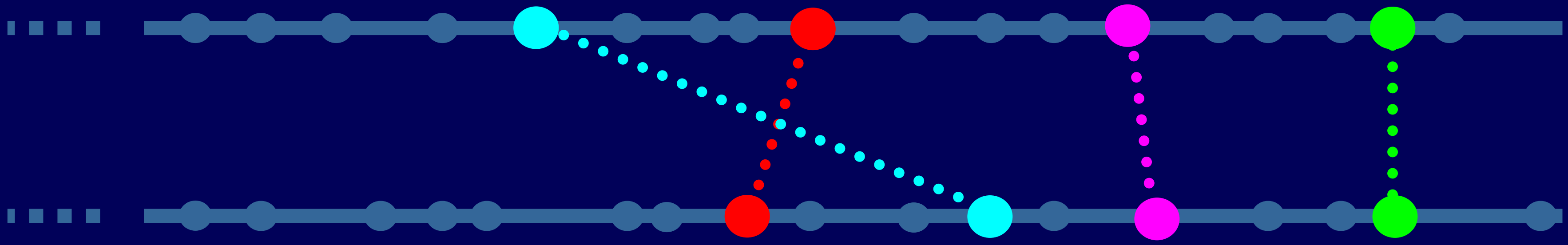
Statistics for max-gap gene clusters

- Reference set

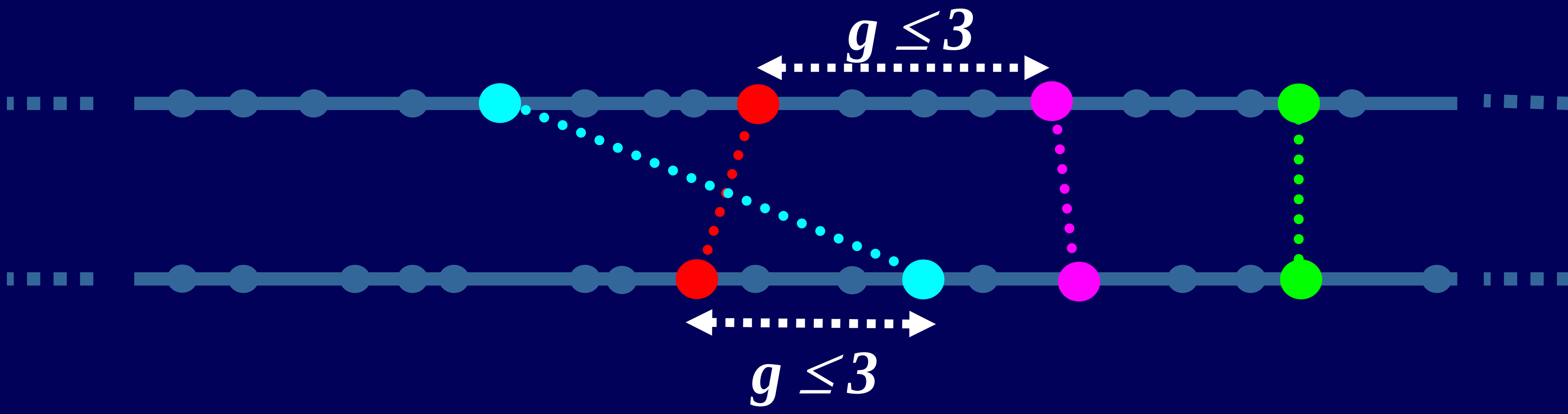
- Whole Genome Comparison

Inputs

1. two genomes of n genes
2. m homologous genes pairs
3. a maximum gap size g



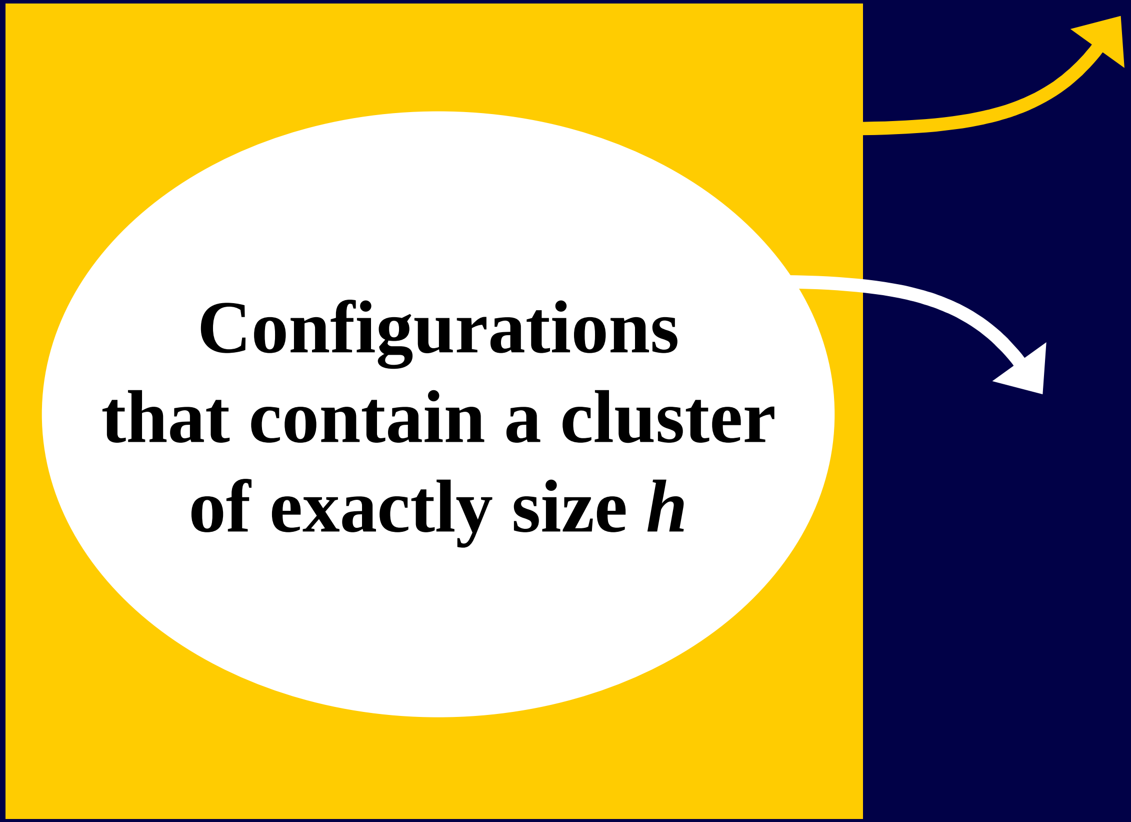
Whole Genome Comparison



- What is the probability of observing a maximal max-gap cluster of size exactly h, if both genomes are randomly ordered?

Compute probabilities by counting

All configurations of two genomes



Configurations that contain a cluster of exactly size *h*

$$\binom{n}{m}^2 m!$$

??

Constructive Approach

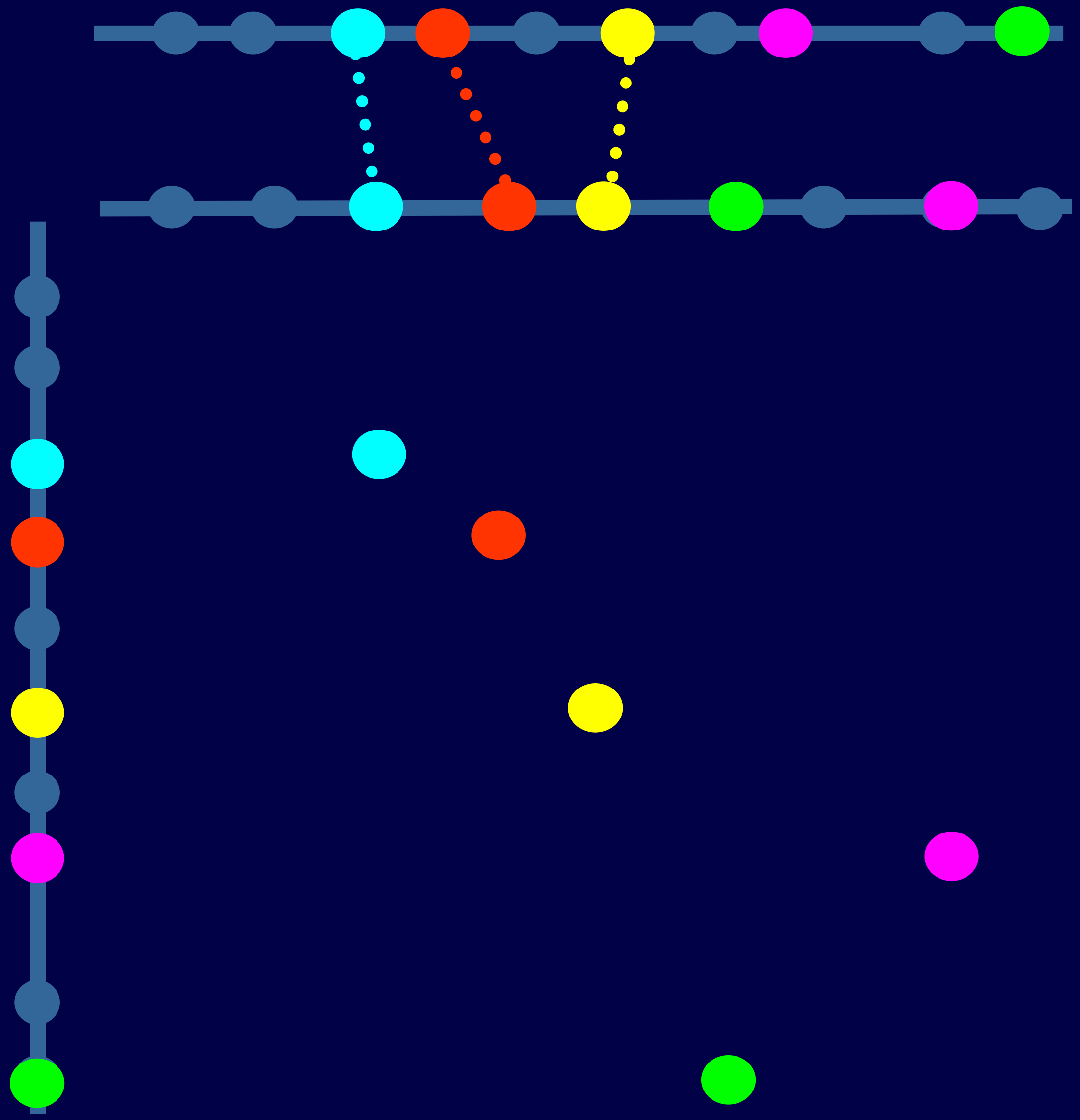
Number of configurations that contain a cluster of exactly size *h*

number of ways to place *h* genes so they form a cluster in both genomes

number of ways to place *m-h* remaining genes so they do not extend the cluster

$$N(h,g,n)^2 \cdot h! \cdot \;??$$

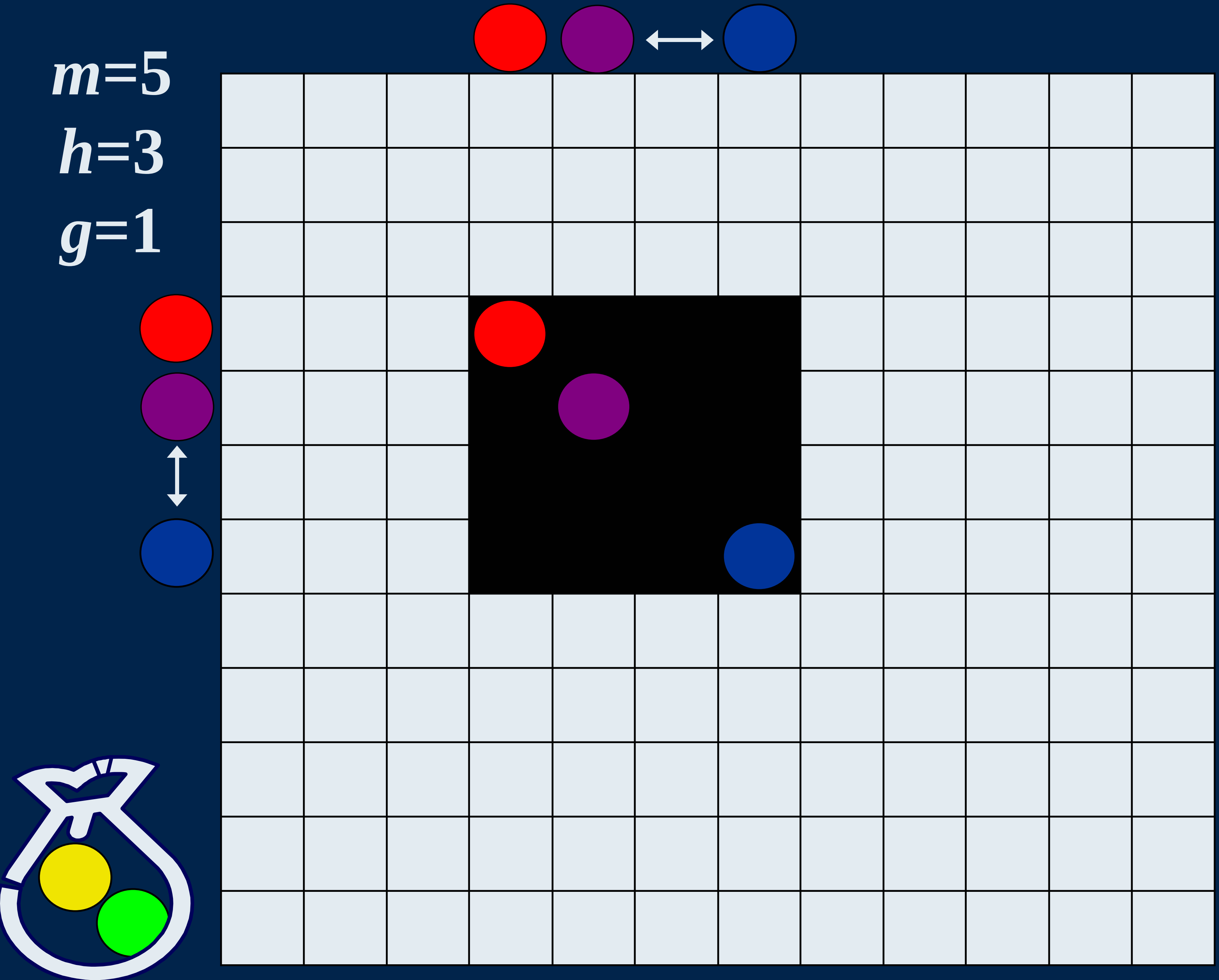
Switching Representations



$m=5$

$h=3$

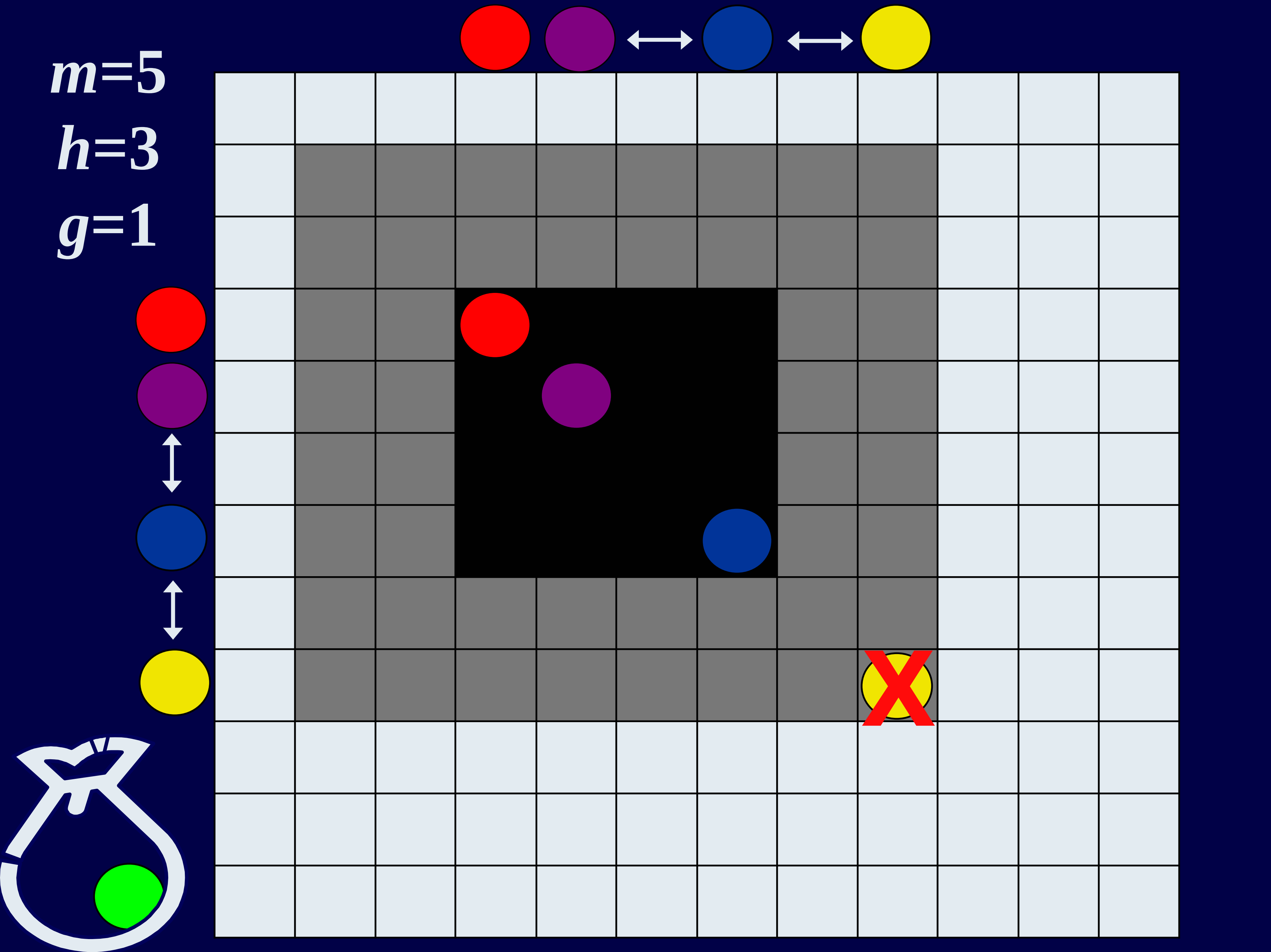
$g=1$



$m=5$

$h=3$

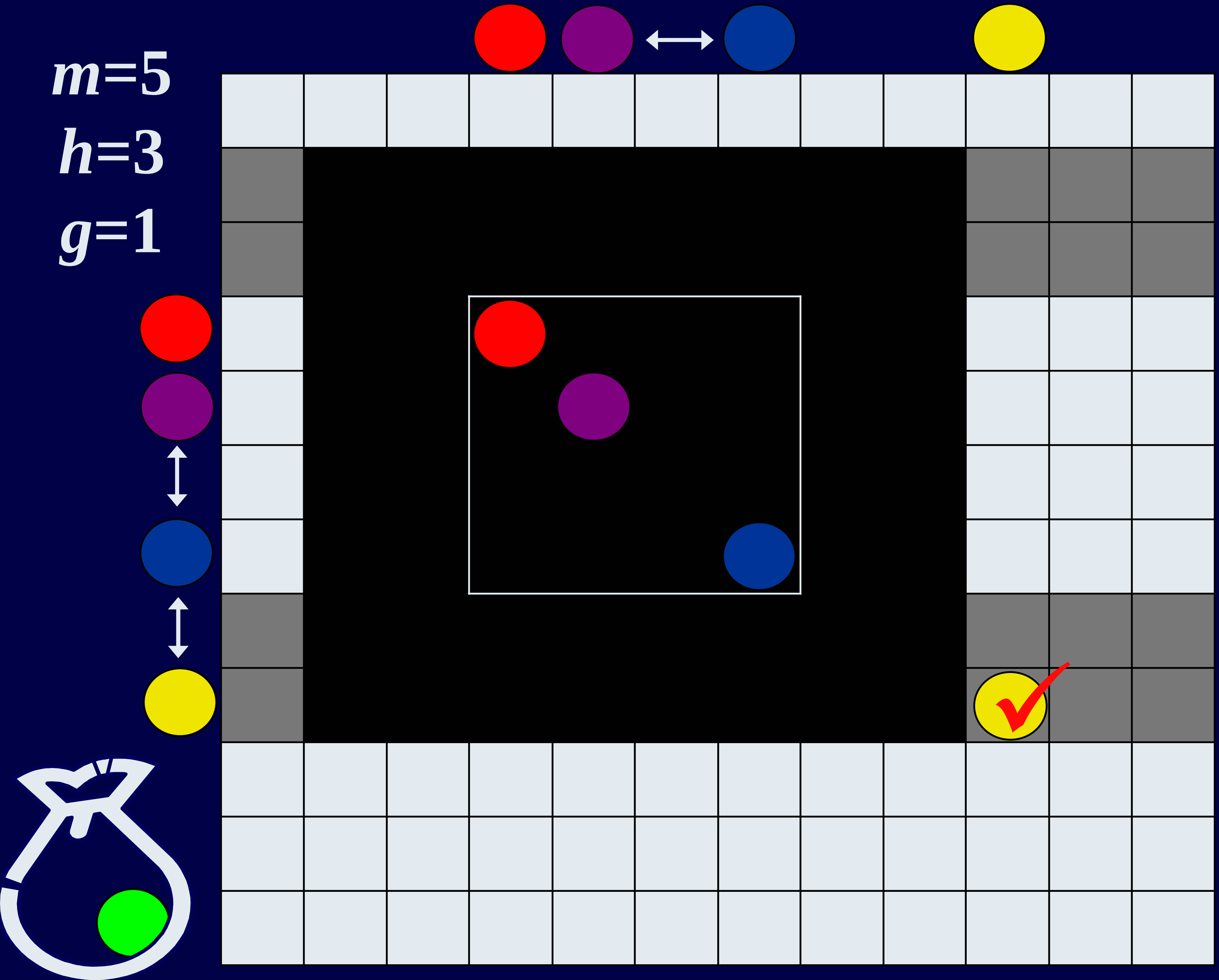
$g=1$



$m=5$

$h=3$

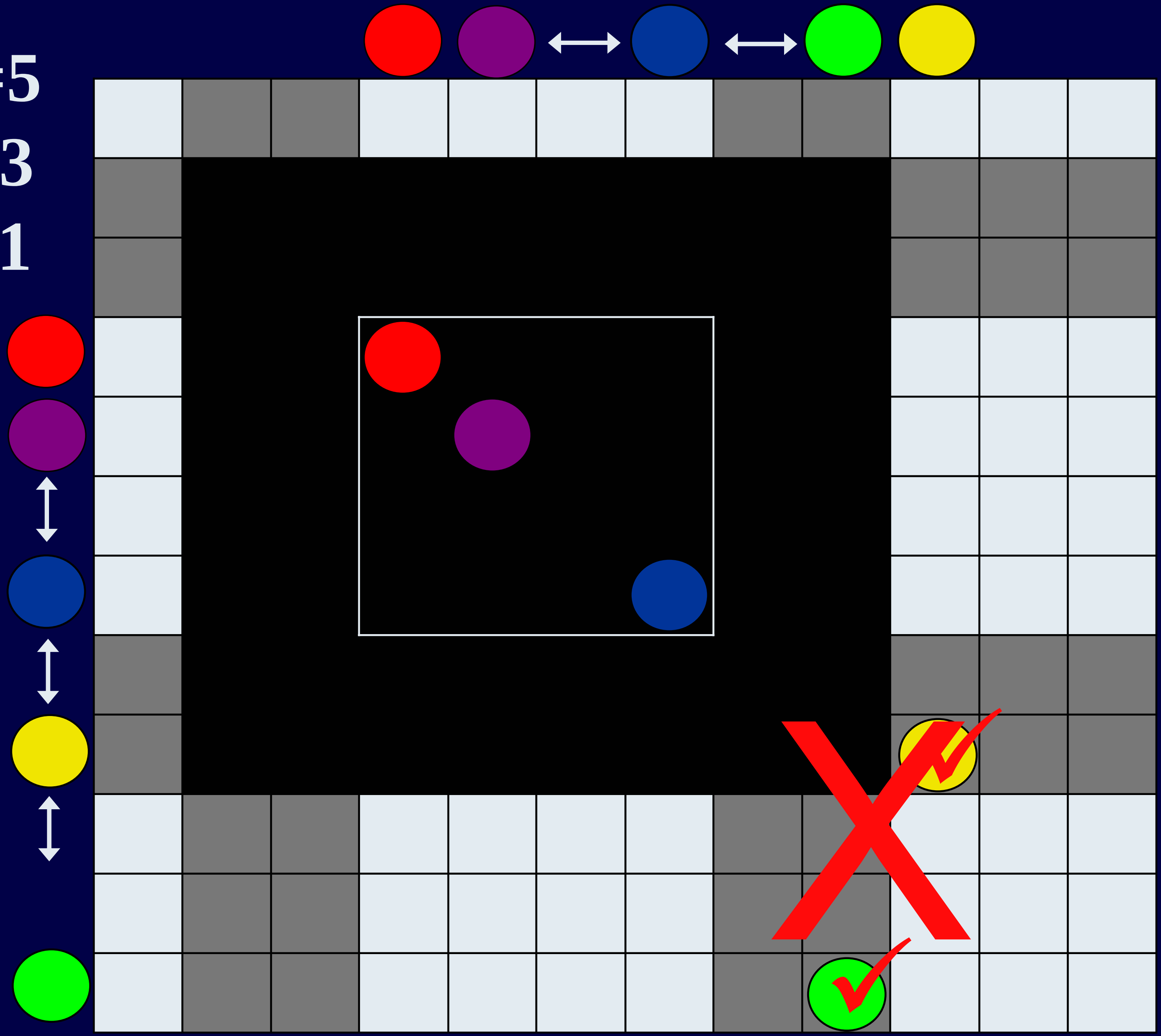
$g=1$



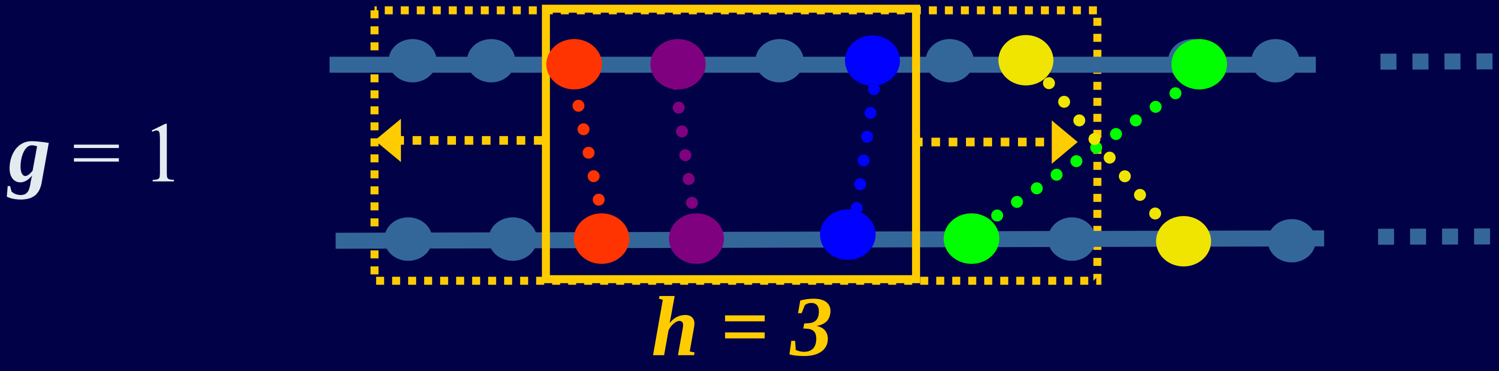
$m=5$

$h=3$

$g=1$



Why is counting hard in this case?

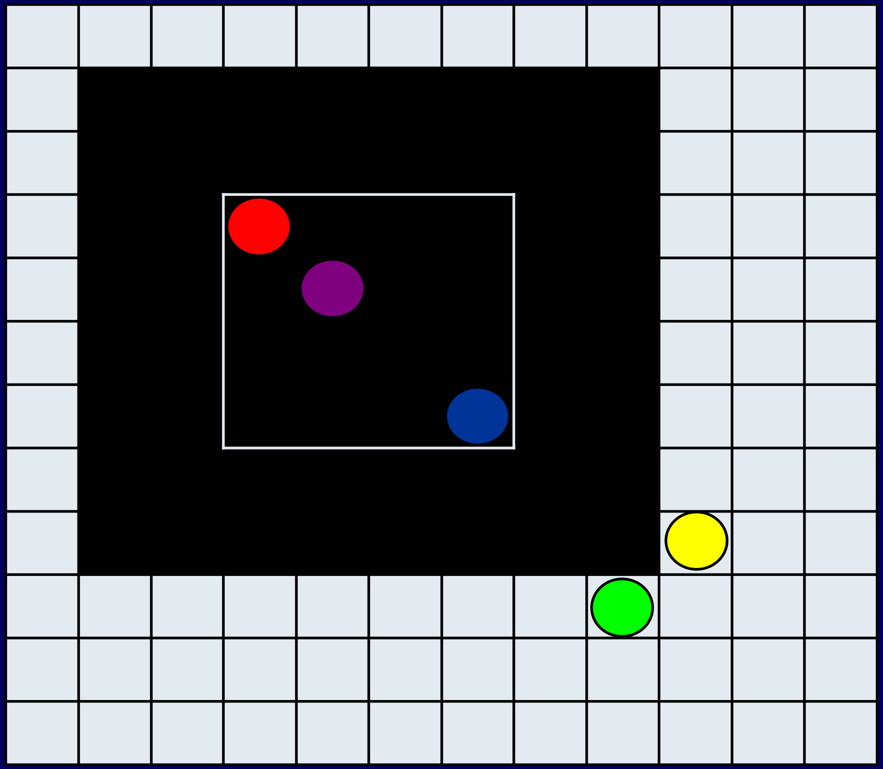


- There are no other homologs within *g* of this cluster on both genomes, yet this cluster is *not* maximal

- Greedy agglomerative algorithm doesn’t find all max-gap clusters

- There is an efficient divide-and-conquer algorithm to find maximal max-gap clusters (Bergeron, Corteel, Raffinot 2002)

Bounding the Cluster Probabilities

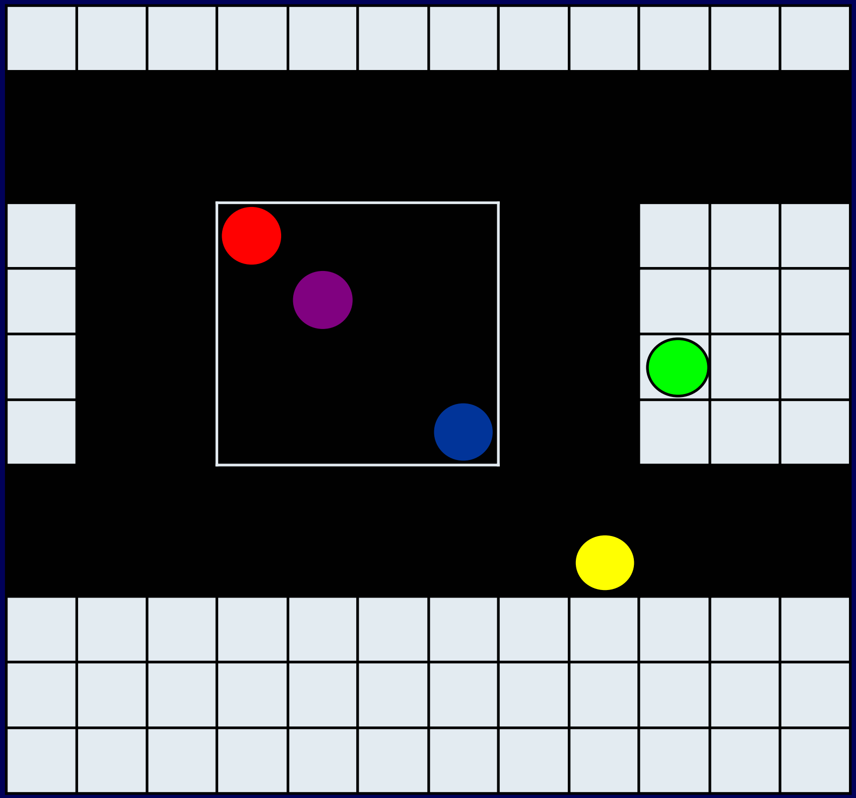


Upper bound:

Erroneously enumerates this configuration as a maximal cluster of size three

Lower bound:

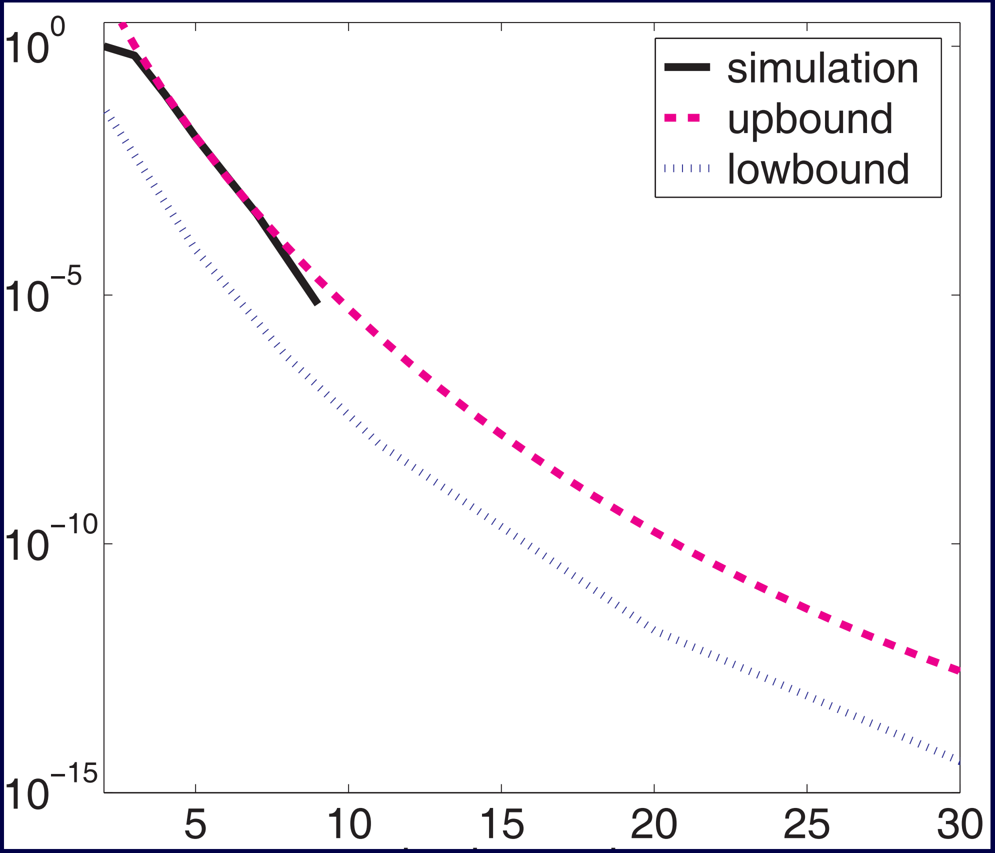
Fails to enumerate this permutation as containing a maximal cluster of size three



Whole-genome comparison

n=1000, m=250, g=10

Probability of observing a maximal max-gap cluster of size *h* by chance



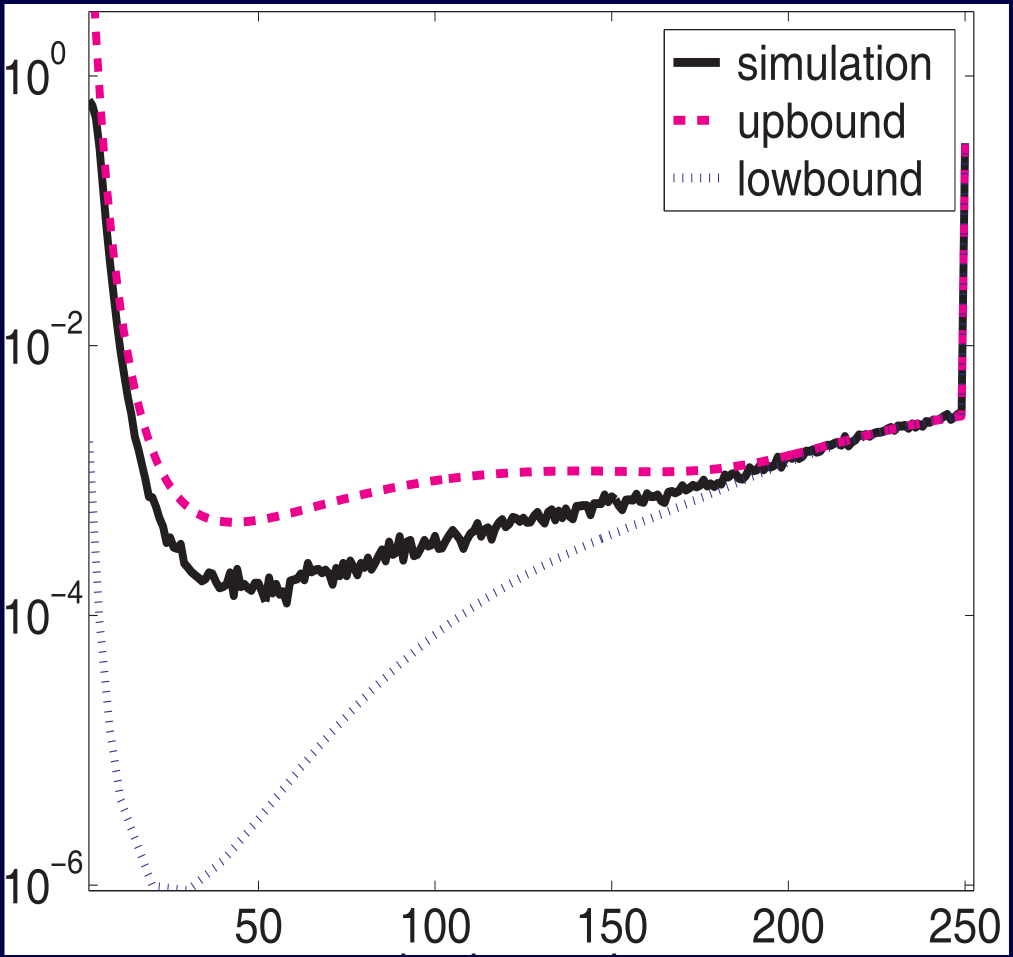
Cluster size

Whole-genome comparison

Probability of observing a maximal max-gap cluster of size *h* by chance

…is no longer strictly decreasing!

n=1000, m=250, g=20



Cluster size

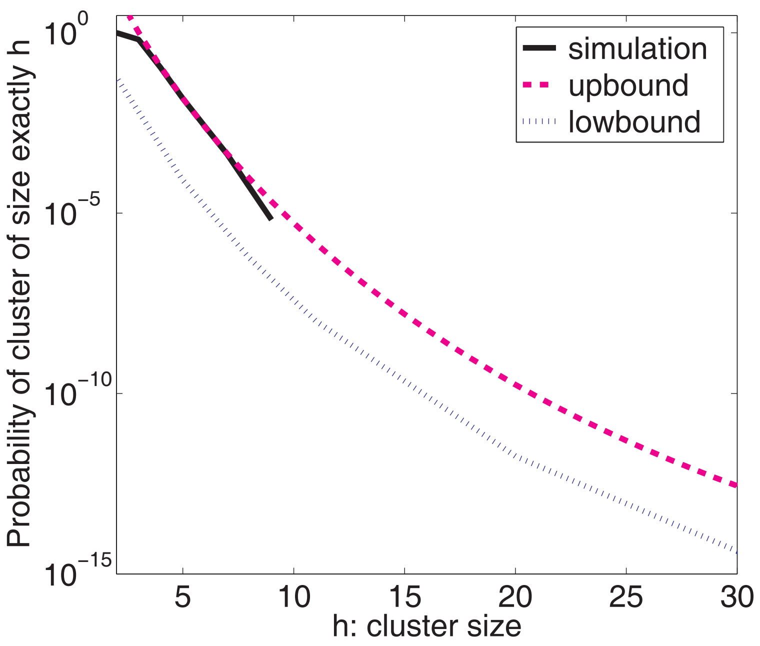
Conclusions

Presented statistical tests for max-gap clusters

- **Evaluate the significance of observed clusters**

– Choose parameters effectively

– Understand trends

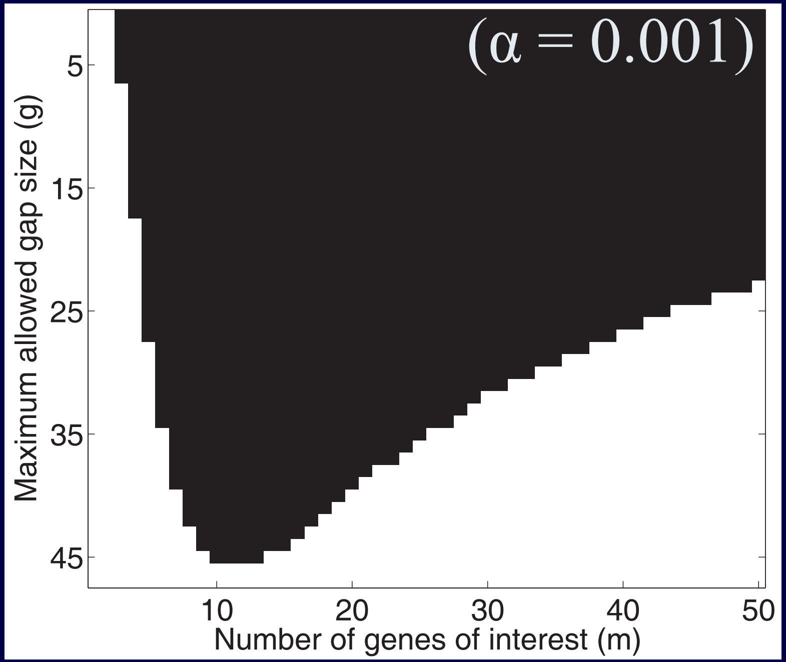


Conclusions

Presented statistical tests for max-gap clusters

- Evaluate the significance of observed clusters
- **Choose parameters effectively**
- Understand trends

Significant Parameter Values

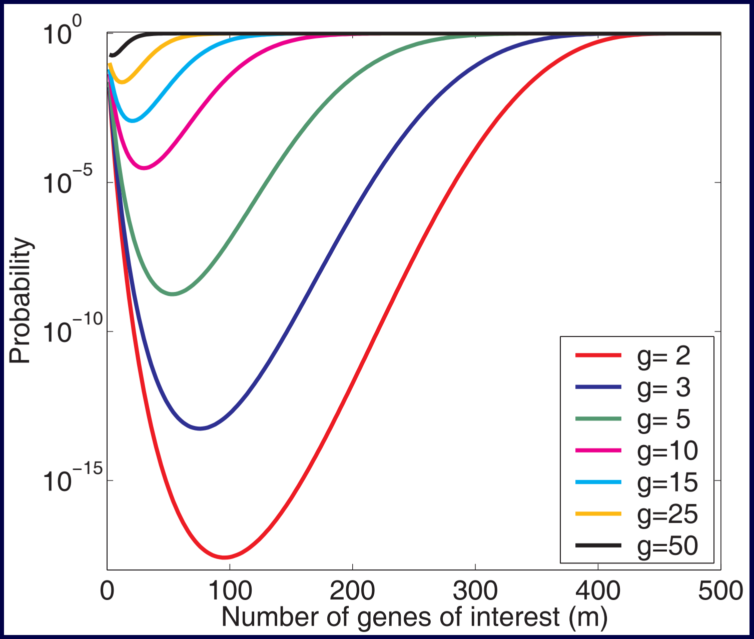


Conclusions

Presented statistical tests for max-gap clusters

- Evaluate the significance of clusters of a pre-specified set of genes
- Choose parameters effectively

➢ **Understand trends**



Thank You