

A Statistical Framework for Spatial Comparative Genomics

Thesis Proposal

Rose Hoberman

Carnegie Mellon University, August 2005

Thesis Committee

Dannie Durand (chair)

Andrew Moore

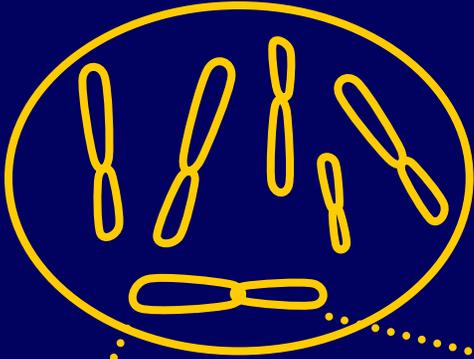
Russell Schwartz

Jeffrey Lawrence (Univ. of Pittsburgh, Dept. of Biological Sciences)

David Sankoff (Univ. of Ottawa, Dept. of Math & Statistics)



Genome: the complete set of genetic material of an organism or species



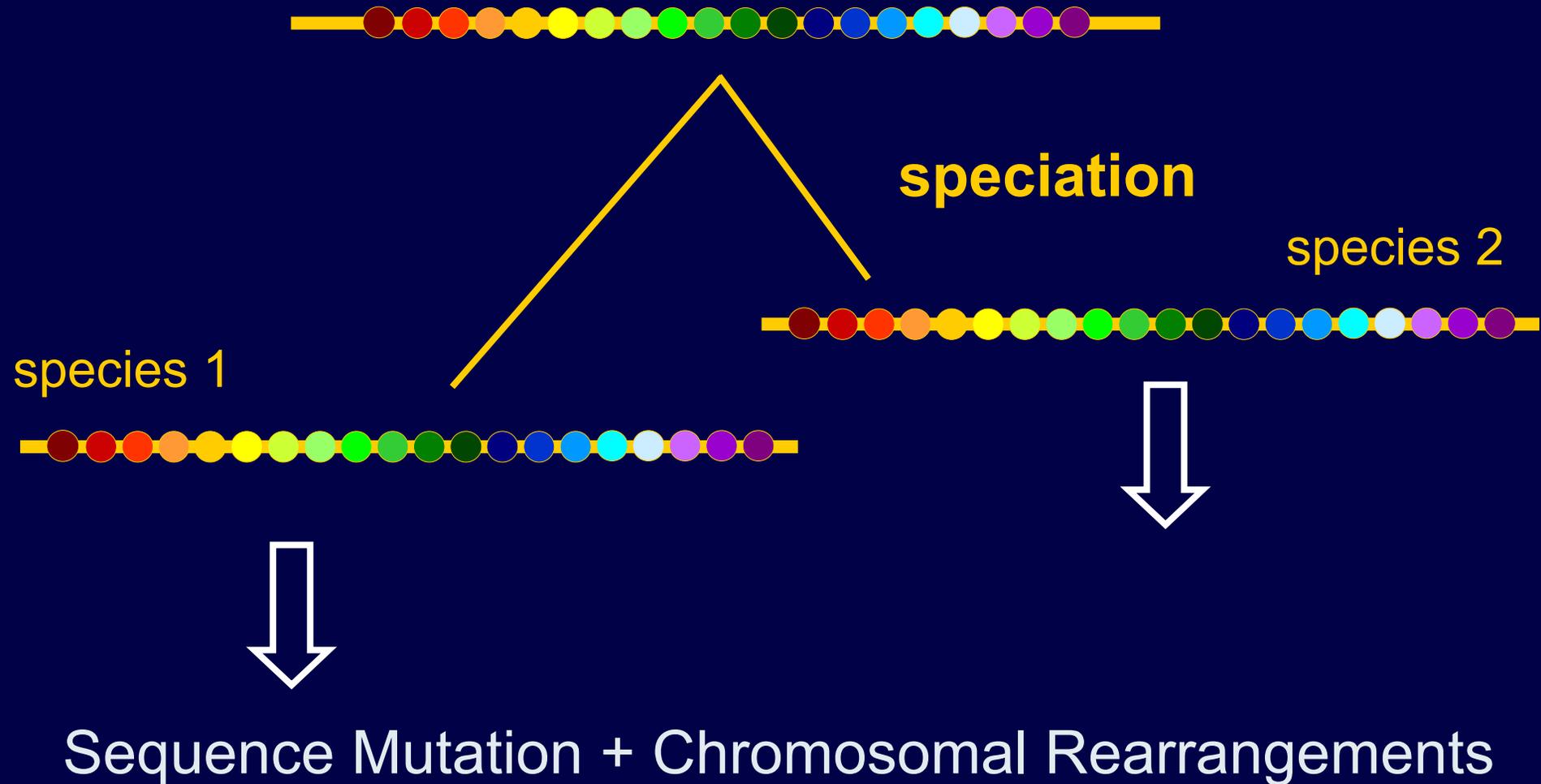
Noncoding DNA:
Large stretches of DNA
with unknown function.

CGGACACTTCGTCTTCAGACCCTTAGCTAGACCTTTAGGAGGATTAAAAATGAGGGAGAGGGGCGGGCCCCCGCCCCCGCCCCCCCCCCCCC
CCCTGTGAAGCAGAAGTCTGGGAATCGATCTGGAAATCCTCCTAATTTTACTCCCTCTCCCCGCCCGGGGGCGGGGGCGGGGGGGGGGGGG

Regulatory regions:
Regions of DNA
where regulatory
proteins bind

Genes: DNA sequences that code for
a specific functional product,
most commonly proteins.

Genome Evolution



Chromosomal Rearrangements

Species 1



Duplications

Species 2

Inversions

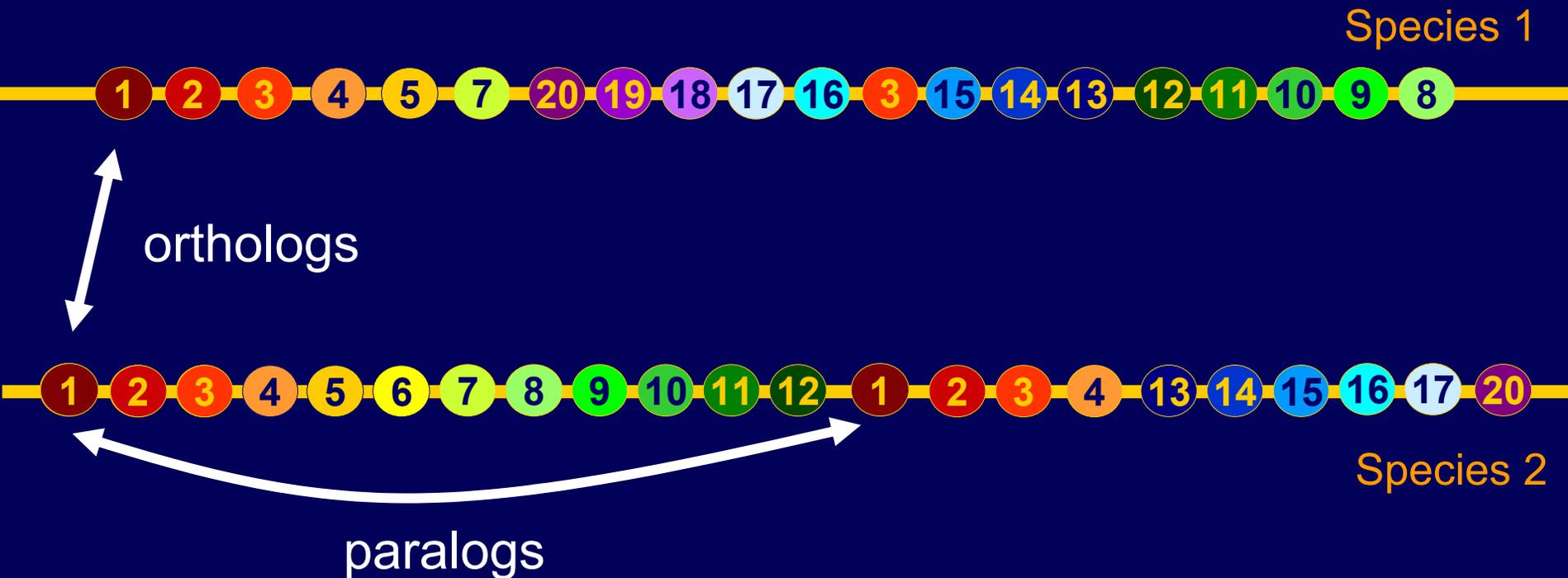
Loss

My focus: Spatial Comparative Genomics

Understanding genome structure, especially how the spatial arrangement of elements within the genome changes and evolves.

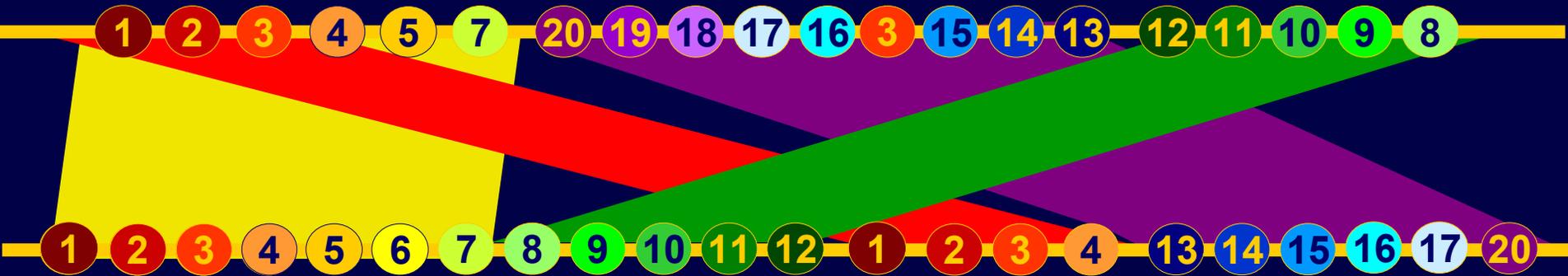
Terminology

- Homologous: related through common ancestry
 - Orthologous: related through speciation
 - Paralogous: related through duplication



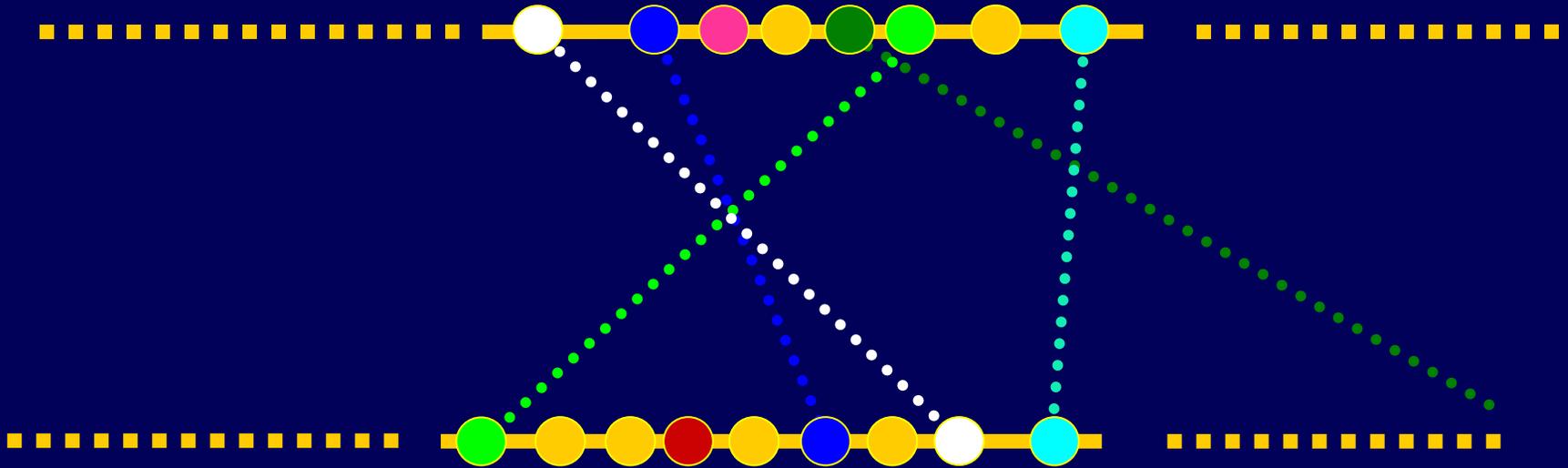
An Essential Task for Spatial Comparative Genomics

Identify *homologous* blocks, chromosomal regions that correspond to the same chromosomal region in an ancestral genome



My thesis: how to find and statistically
validate homologous blocks

More distantly related segments:



Gene Clusters: similar gene content, but neither gene content nor order is strictly conserved

Gene Clusters are Used in Many Types of Genomic Analysis

Inferring **functional coupling** of genes in bacteria (Overbeek et al 1999)

Recent polyploidy in Arabidopsis (Blanc et al 2003)

Sequence of the **human genome** (Venter et al 2001)

Duplications in Arabidopsis through comparison with rice (Vandepoele et al 2002)

Duplications in **Eukaryotes** (Vision et al 2000)

Identification of **horizontal transfers** (Lawrence and Roth 1996)

Evolution of **gene order** conservation in **prokaryotes** (Tamames 2001)

Ancient **yeast duplication** (Wolfe and Shields 1997)

Genomic duplication during early **chordate evolution** (McLysaght et al 2002)

Comparing rates of **rearrangements** (Coghlan and Wolfe 2002)

Genome **rearrangements** after duplication in yeast (Seoighe and Wolfe 1998)

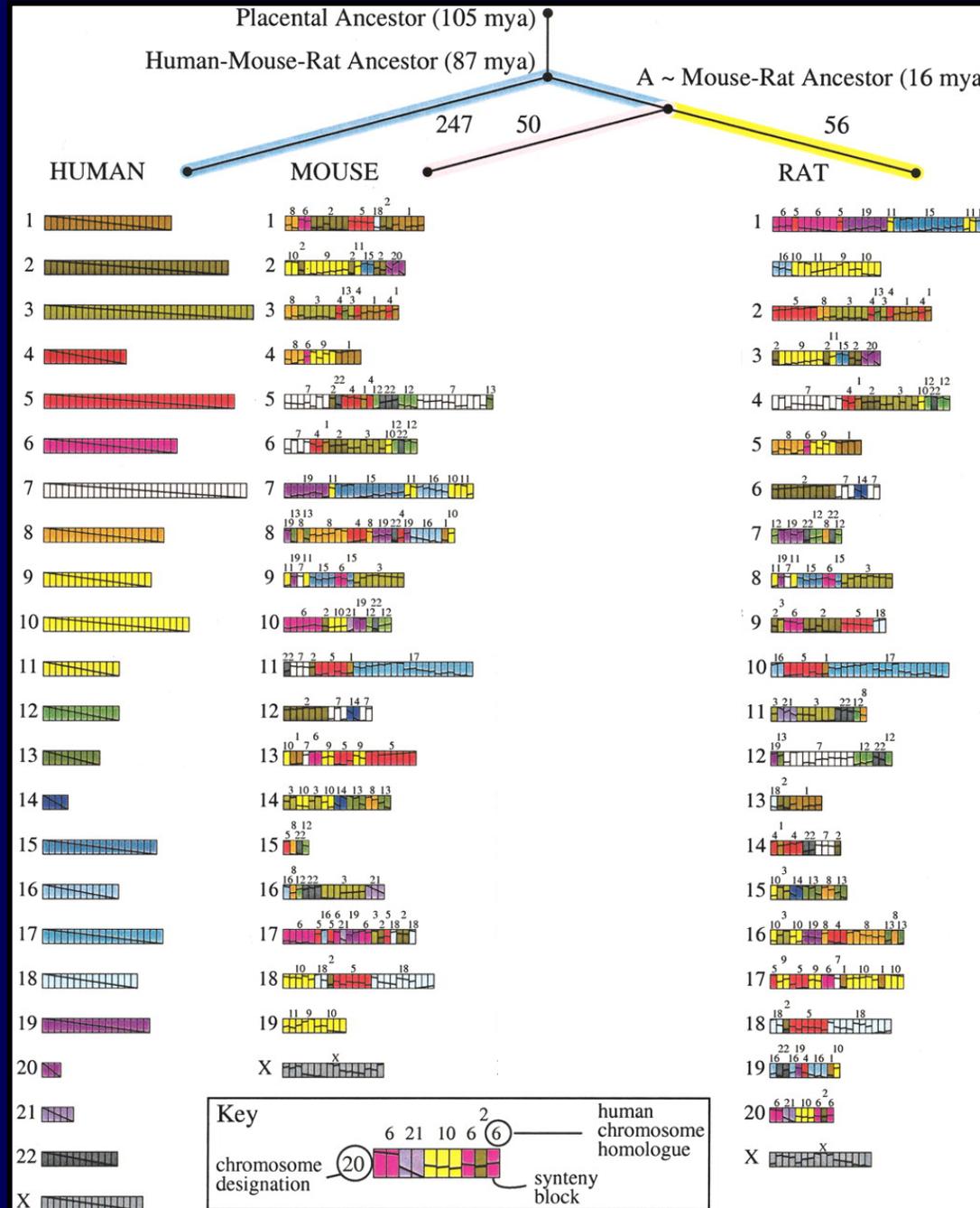
Operon prediction in newly sequenced bacteria (Chen et al 2004)

Breakpoints as **phylogenetic features** (Blanchette et al 1999)

...

Spatial Comparative Genomics

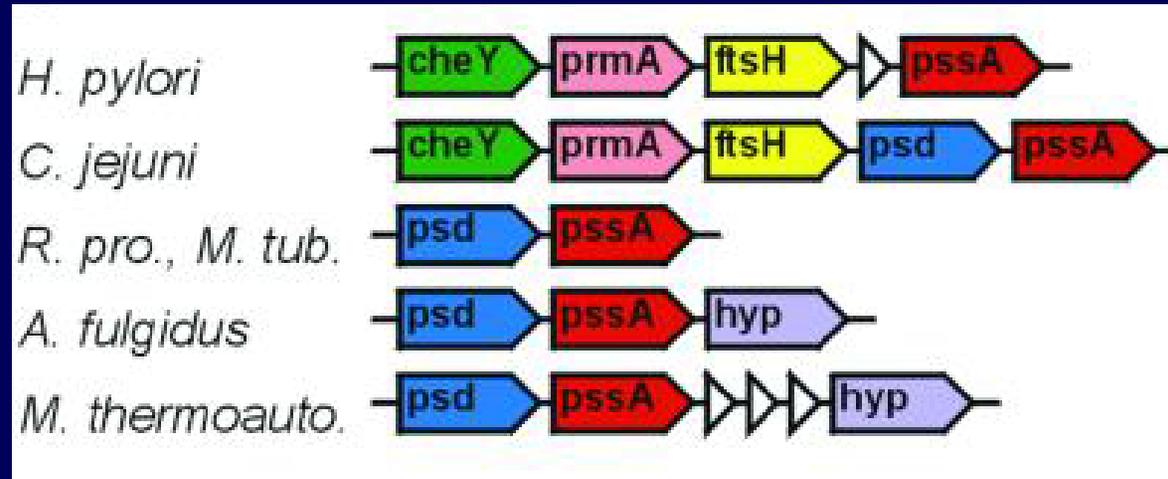
- reconstruct the history of chromosomal rearrangements
- infer an ancestral genetic map
- build phylogenies
- transfer knowledge



Spatial Comparative Genomics

Function

Snel, Bork, Huynen. PNAS 2002



- Consider evolution as an enormous experiment
- Unimportant structure is randomized or lost
- Exploit evolutionary patterns to infer functional associations

Outline

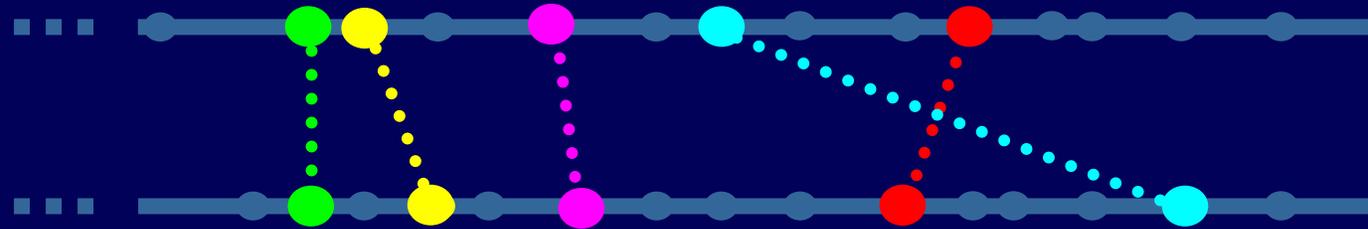
- Introduction and Applications
- Formal framework for gene clusters
 - Genome representation
 - Gene homology mapping
 - Cluster definition
- Introduction to Statistical Issues
- Preliminary work: Testing cluster significance
- Proposed work

Basic Genome Model



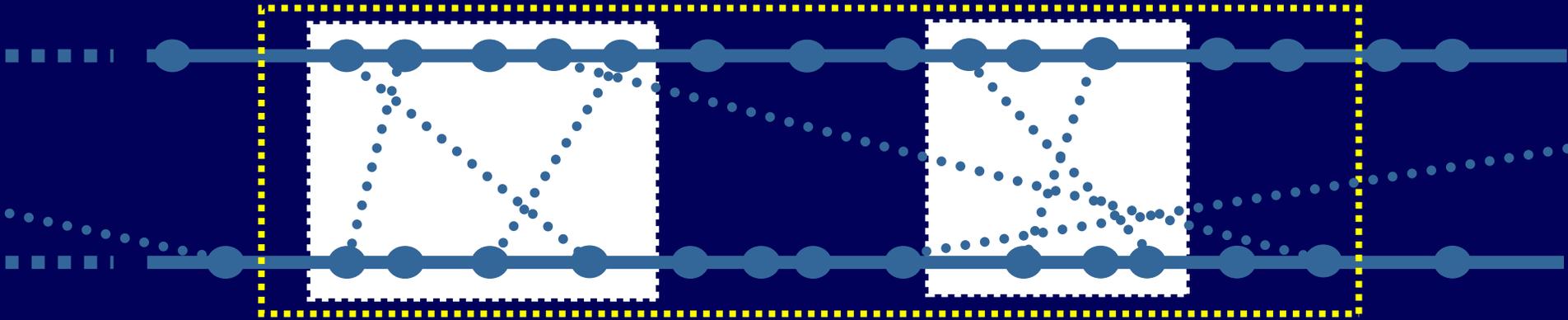
- a sequence of unique genes
- distance between genes is equal to the number of intervening genes
- gene orientation unknown
- a single, linear chromosome

Gene Homology



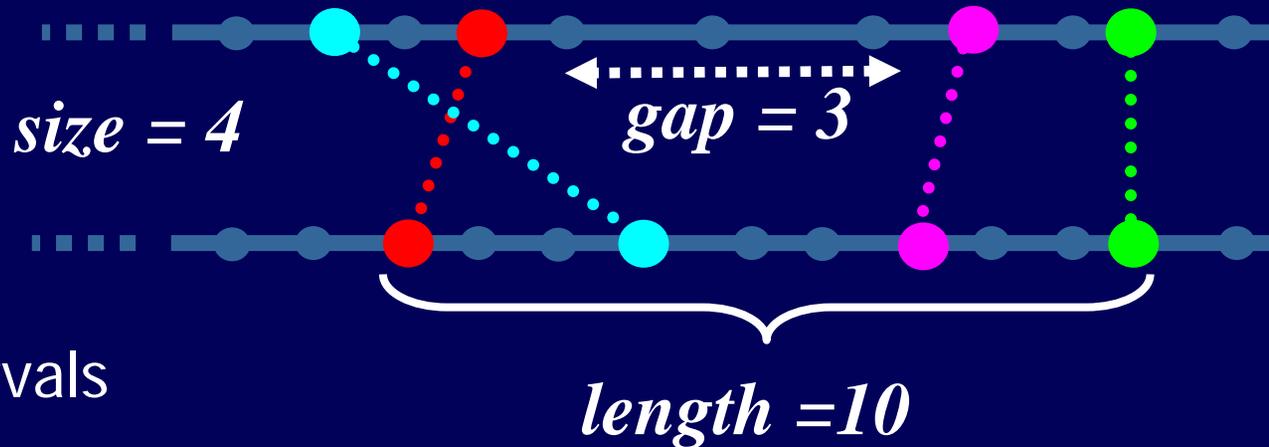
- Identification of homologous gene pairs
 - generally based on sequence similarity
 - still an imprecise science
 - preprocessing step
- Assumptions
 - matches are binary (similarity scores are discarded)
 - each gene is homologous to at most one other gene in the other genome

Where are the gene clusters?



- Intuitive notions of what clusters look like
 - Enriched for homologous gene pairs
 - Neither gene content nor order is perfectly preserved
- Need a more rigorous definition

Cluster Definitions



■ Descriptive:

- common intervals
- r-window
- max-gap
- ...

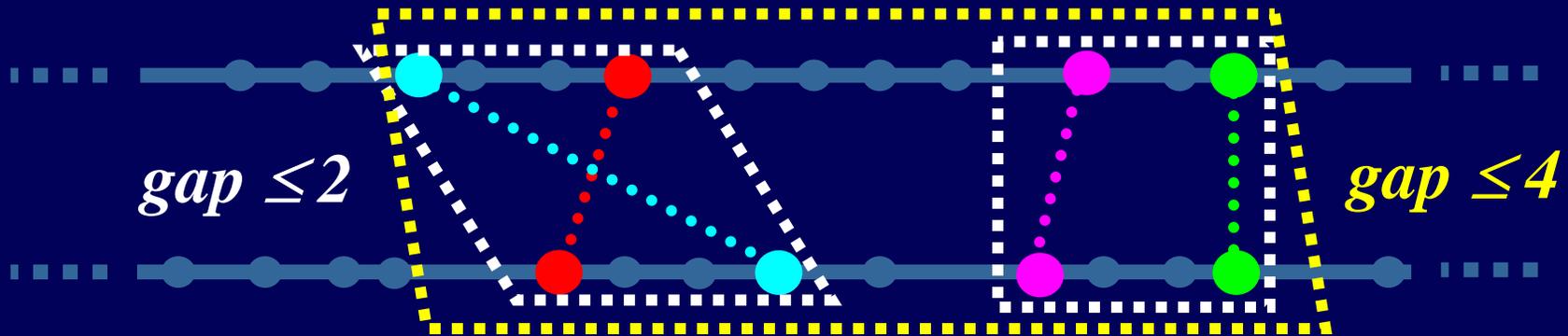
■ Constructive:

- LineUp
- CloseUp
- FISH
- ...

■ Cluster properties

- order
- size
- length
- density
- gaps

Max-Gap: a common cluster definition



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

Why Max-Gap?

- Allows extensive rearrangement of gene order
- Allows limited gene insertion and deletions
- Allows the cluster to grow to its natural size

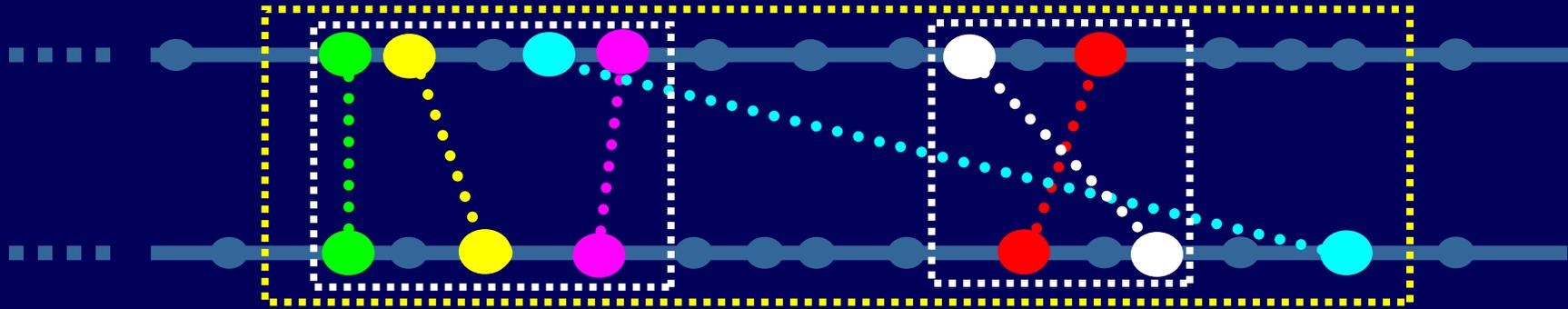
It's the most widely used
in genomic analyses

**no formal statistical model for
max-gap clusters**

Outline

- Introduction and Applications
- Formal framework for gene clusters
- Introduction to statistical issues
- Preliminary work: Testing cluster significance
- Proposed work

Detecting Homologous Chromosomal Segments



1. Formally define a "gene cluster" ...*modeling*
2. Devise an algorithm to identify clusters ...*algorithms*
3. Verify that clusters indicate common ancestry ...*statistics*

How can we verify that a gene cluster indicates common ancestry?

- True histories are rarely known
- Experimental verification is often not possible
- Rates and patterns of large-scale rearrangement processes are not well understood

Statistical Testing Provides Additional Evidence for Common Ancestry

Statistical Testing

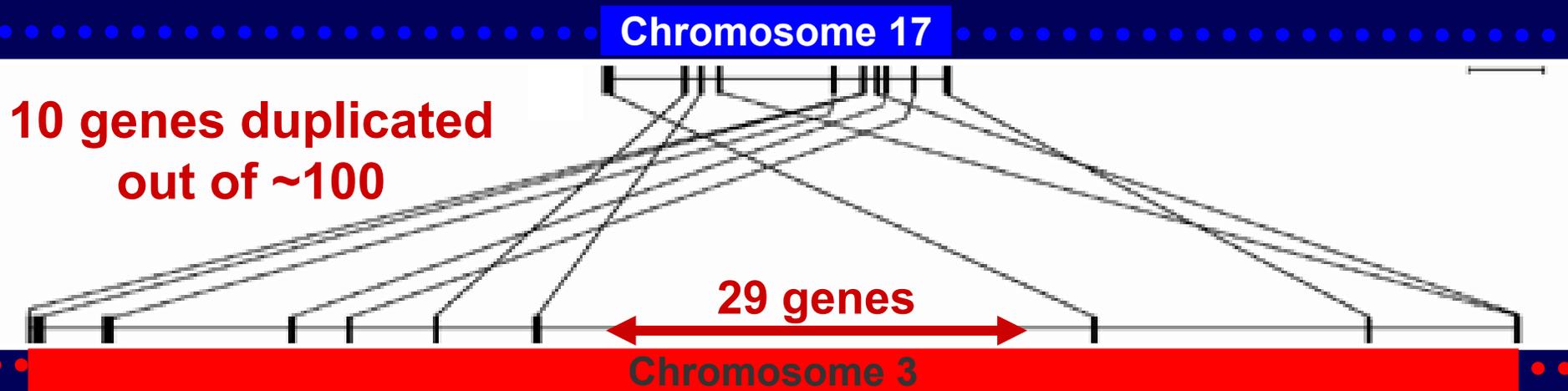
- Goal: distinguish ancient homologies from chance similarities
- Hypothesis testing
 - **Alternate hypothesis**: shared ancestry
 - **Null hypothesis**: random gene order
- Determine the probability of seeing a cluster by chance under the null hypothesis

An example...

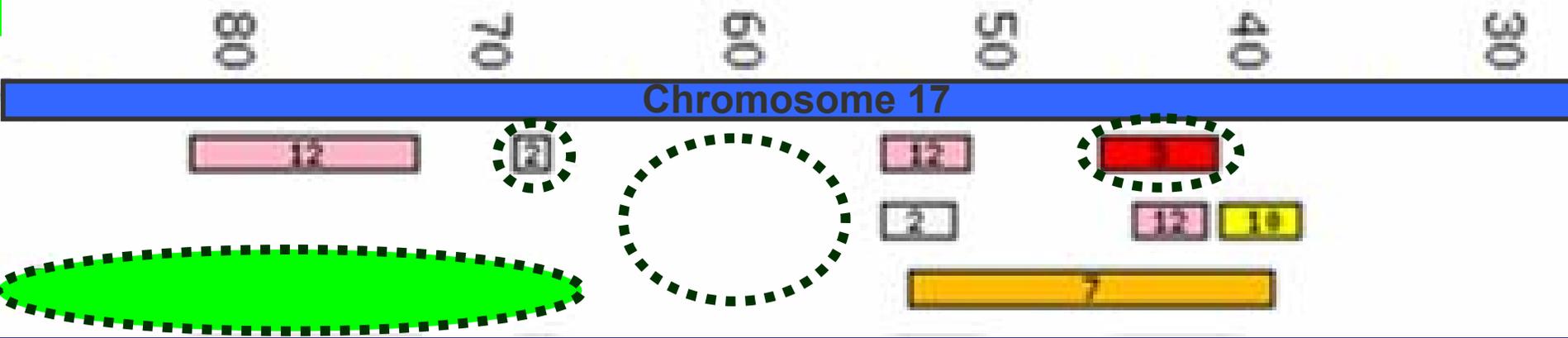
Whole Genome Self-Comparison

McLysaght, Hokamp, Wolfe. Nature Genetics, 2002.

- Compared all human chromosomes to all other chromosome to find gene clusters
- Identified 96 clusters of size 6 or greater



Could two regions display this degree of similarity simply by chance?



Clusters with similarity to human chromosome 17

1. Are larger clusters more likely to occur by chance?
2. Are there other duplicated segments that their method did not detect?

Cluster Significance: Related Work

- Randomization tests
 - most common approach
 - generally compare clusters by size
- Very simple models
 - Excessively strict simplifying assumptions
 - Overly conservative cluster definitions

Citations in proposal

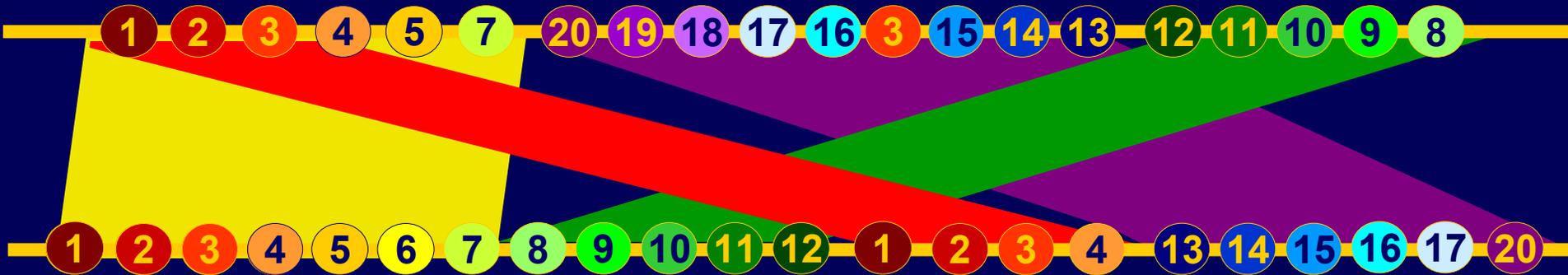
Cluster Significance: Related Work

- Calabrese *et al*, 2003
 - statistics introduced in the context of developing a heuristic search for clusters
- Durand and Sankoff, 2003
 - definition: m homologs in a window of size r
- My thesis
 - max-gap definition

Outline

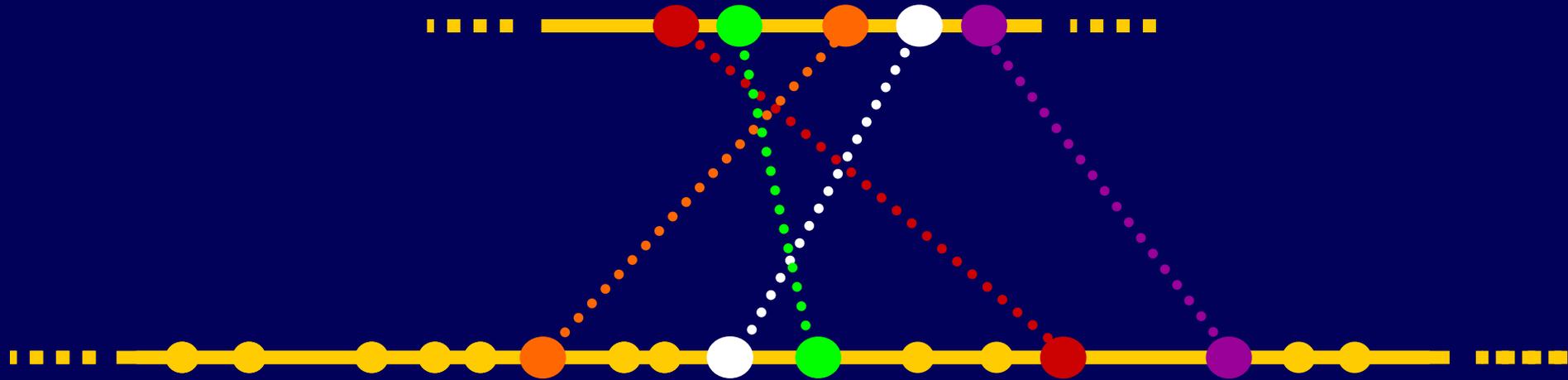
- Introduction and Applications
- Formal framework for gene clusters
- Introduction to statistical issues
- Preliminary work: max-gap cluster statistics
 - reference set
 - whole-genome comparison
- Proposed work

Cluster statistics depend on how the cluster was found



Whole genome comparison: find all (maximal) sets of genes that are clustered together in both genomes.

Cluster statistics depend on how the cluster was found



Reference set: does a *particular* set of genes cluster together in one genome?

- complete cluster: contains all genes in the set
- incomplete cluster: contains only a subset

Preliminary results: Max-Gap Cluster Statistics

■ Reference set

- *complete clusters*

- complete clusters with length restriction

- incomplete clusters

■ Whole genome comparison

- upper bound

- lower bound

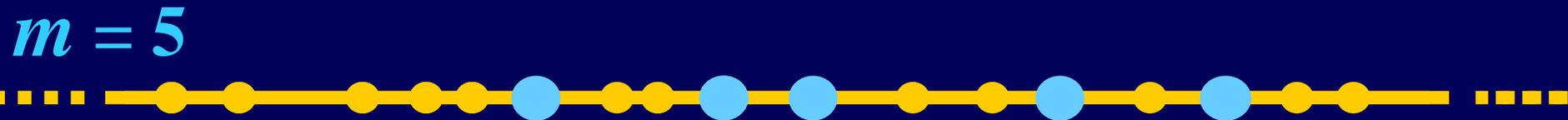
Hoberman, Sankoff, and Durand. *Journal of Computational Biology* 2005.

Hoberman and Durand. *RECOMB Comparative Genomics* 2005.

Hoberman, Sankoff, and Durand. *RECOMB Comparative Genomics* 2004.

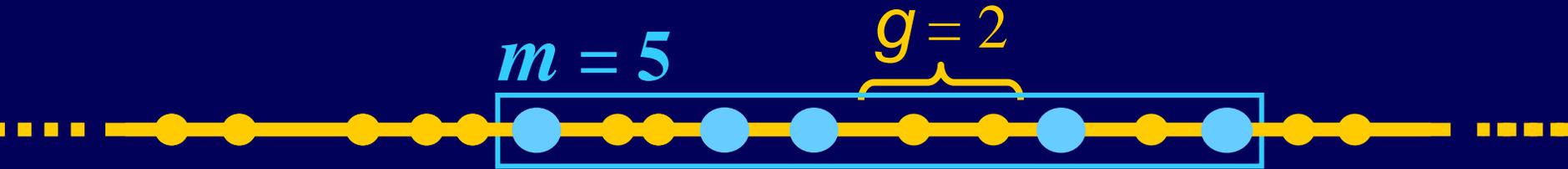
Reference set, complete clusters

Given: a genome: $G = 1, \dots, n$ unique genes
a set of m genes of interest (in blue)



Do all m *blue genes* form a significant cluster?

Reference set, complete clusters



- Test statistic: the maximum gap observed between adjacent blue genes
- P-value: the probability of observing a maximum gap $\leq g$, under the null hypothesis

Compute probabilities by counting

The problem
is how to
count this

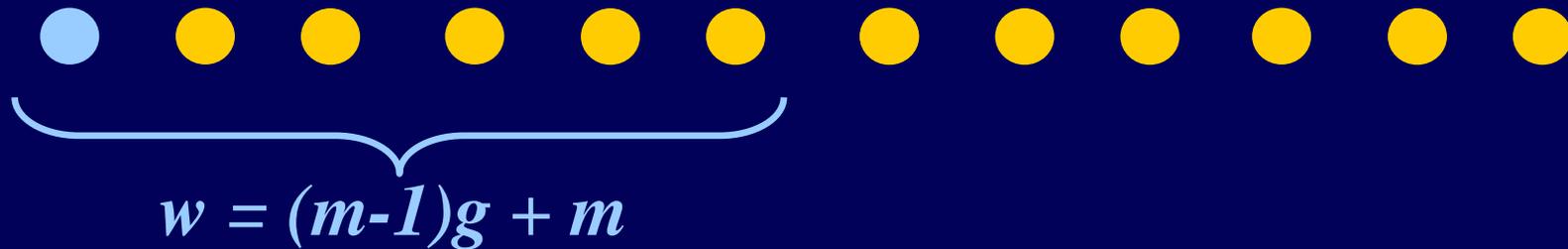
$$\text{P-val} = \frac{N(m, g, n)}{\binom{n}{m}}$$

**All possible
unlabeled permutations**

**Permutations
where the
maximum gap $\leq g$**

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

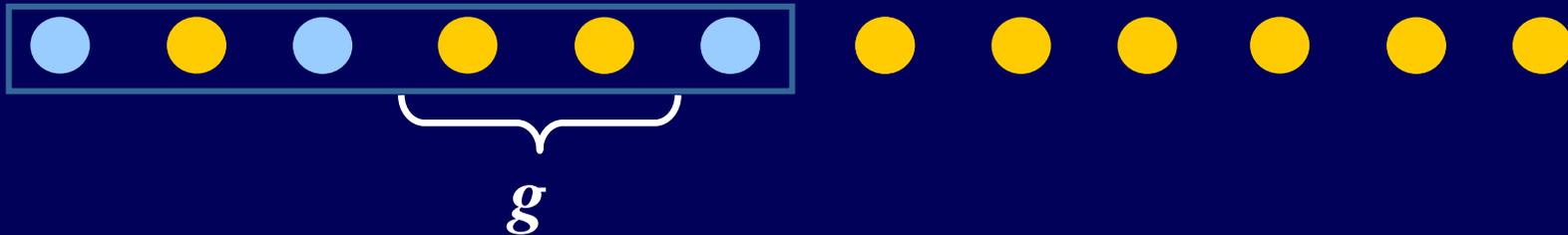
number of ways to start a cluster, e.g. ways to place the first gene and still have $w-1$ slots left



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

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

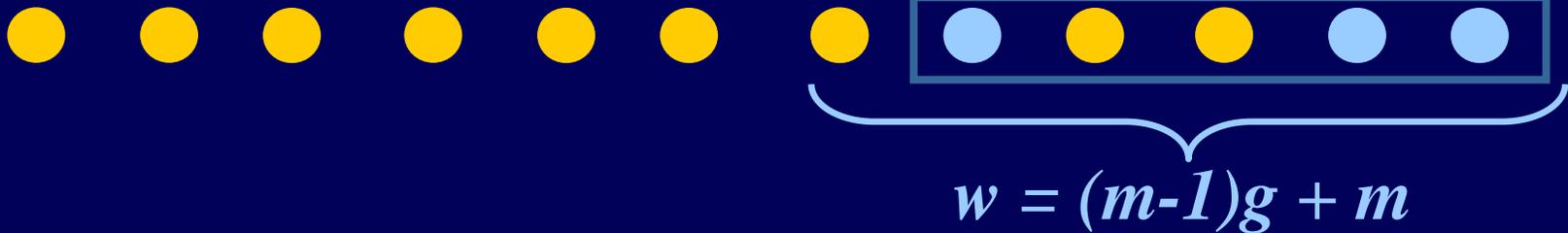


$$N(m, g, n) = (n - w + 1)(g + 1)^{m-1} + E$$

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

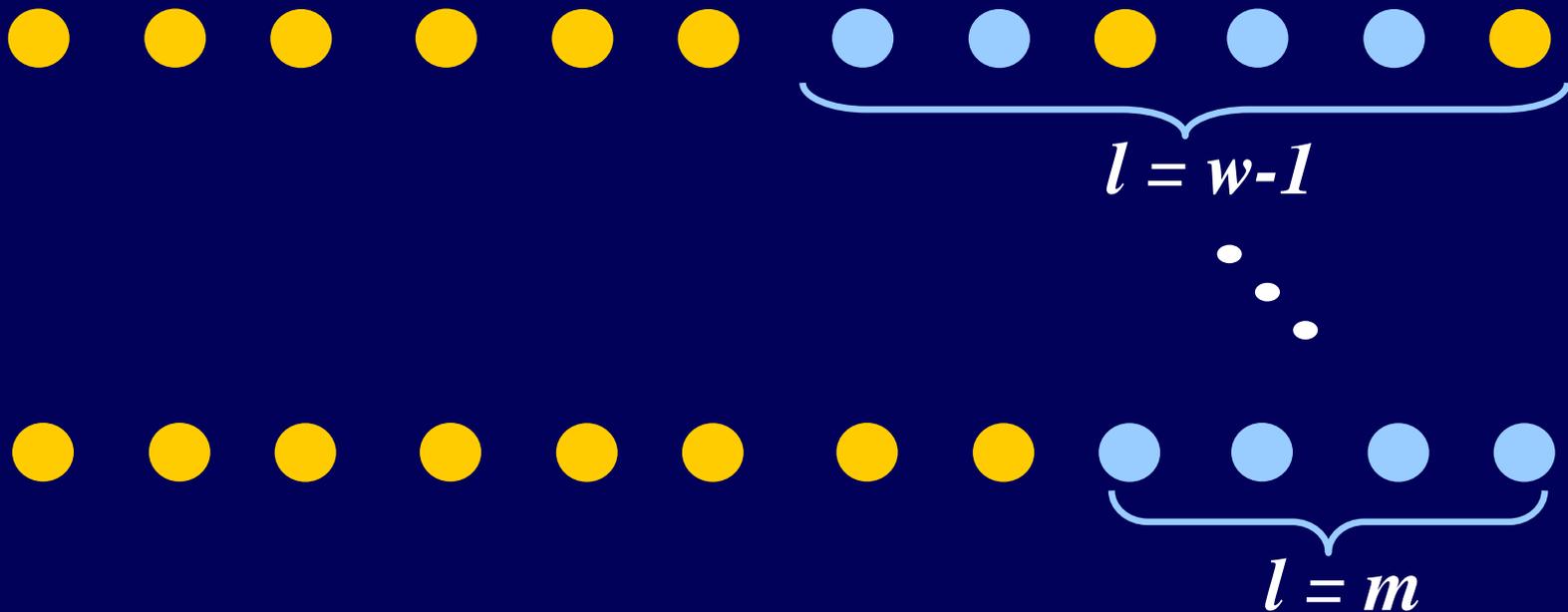
edge effects



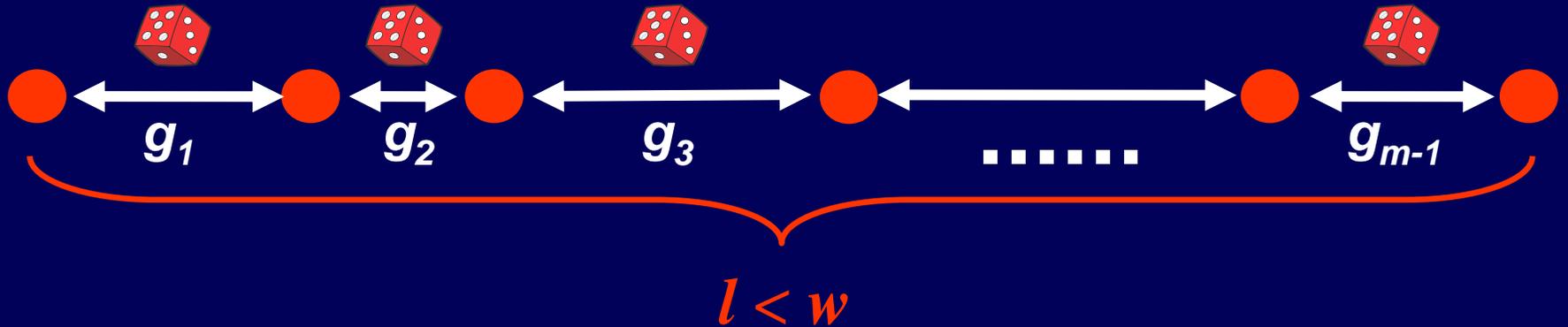
Counting clusters at the end of the genome

Gaps are constrained: $0 \leq g_i \leq g \quad \forall i$

And *sum of gaps* is constrained: $\sum_{i=1}^{m-1} g_i = l$



$d(m, g, l)$ = the number of ways to form a max-gap cluster of size m and length l



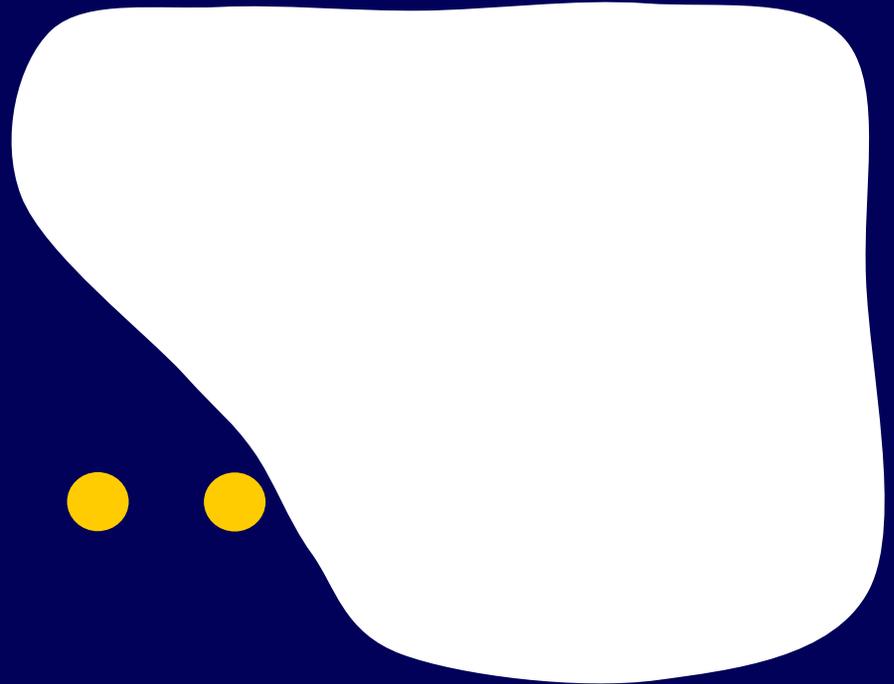
A known solution:

$$d(m, g, l) = \sum_{i=0}^{\lfloor (l-m)/(g+1) \rfloor} (-1)^i \binom{m-1}{i} \binom{l-i(g+1)-2}{m-2}$$

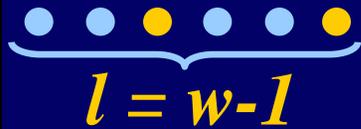
Counting clusters at the end of the genome

Gaps are constrained: $0 \leq g_i \leq g \quad \forall i$

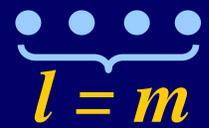
And *sum of gaps* is constrained: $\sum_{i=1}^{m-1} g_i < w - m$



Cluster Length



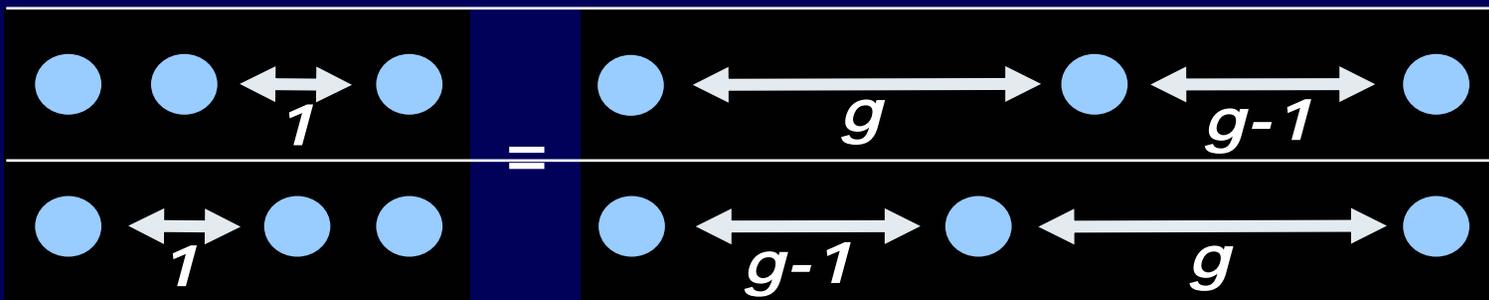
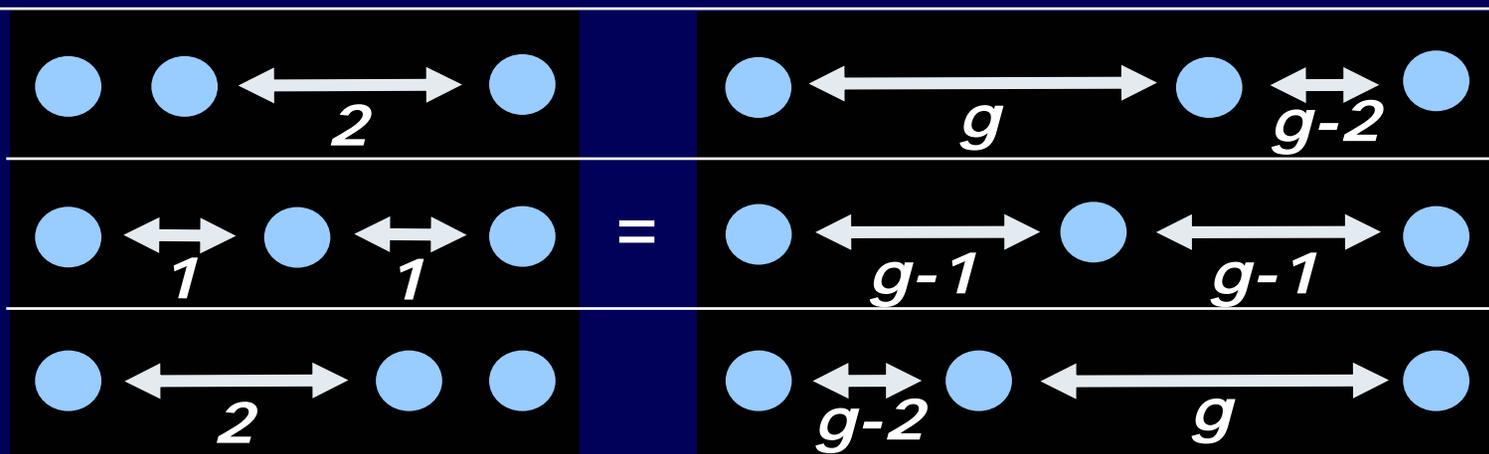
m	m+1	.	w-2	w-1	w
$d(m,g,m) + d(m,g,m+1) +$.		$+ d(m,g,w-2) + d(m,g,w-1)$	
$d(m,g,m) + d(m,g,m+1) +$			$+ d(m,g,w-2)$		
$d(m,g,m) + d(m,g,m+1) +$.			
$d(m,g,m) + d(m,g,m+1) + \dots$					
$d(m,g,m) + d(m,g,m+1)$					
$d(m,g,m)$					



Line of Symmetry

$l =$

Exploiting Symmetry

 $l =$ m  w $m+1$  $w-1$ $m+2$  $w-2$

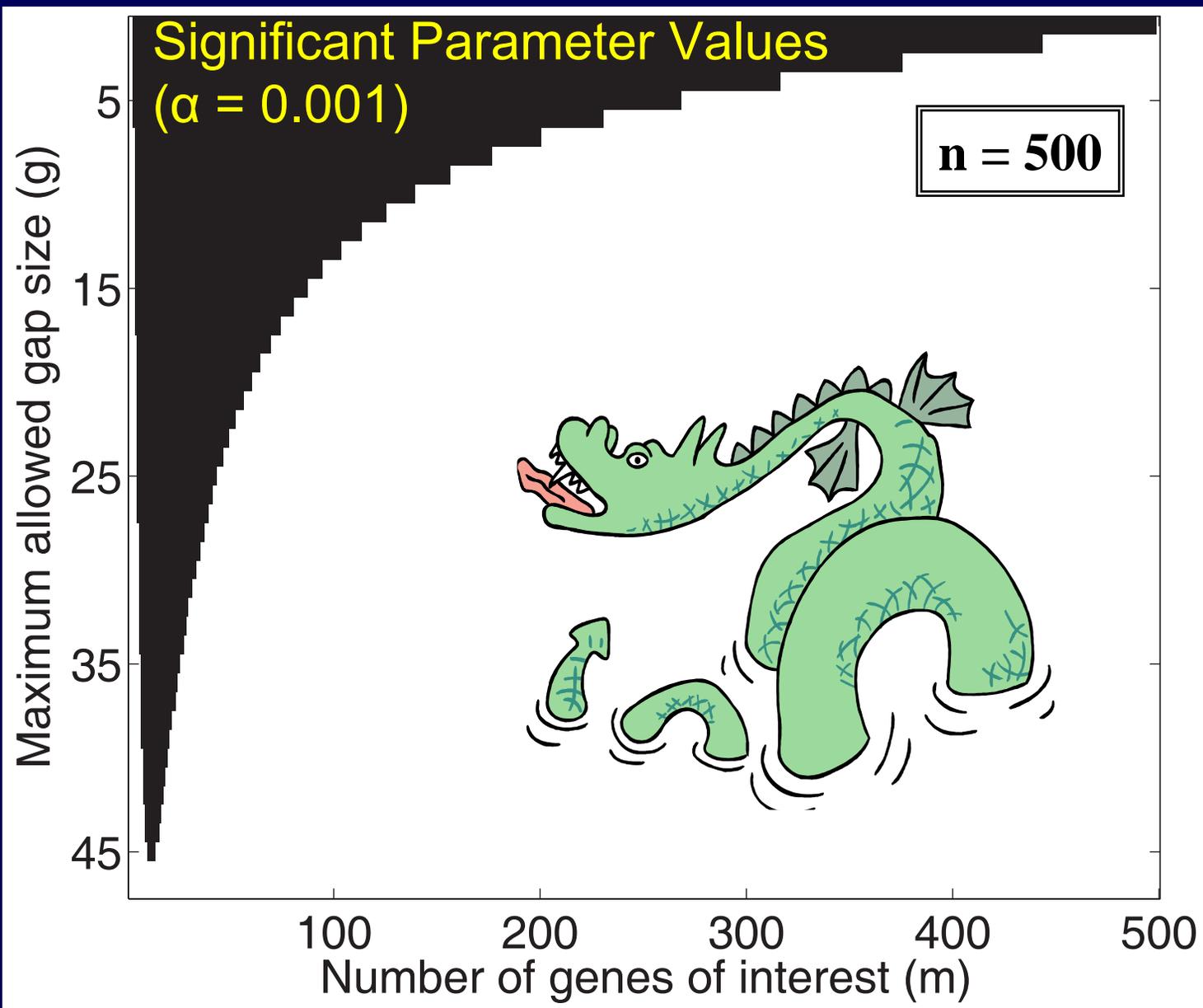
Adding edge effects...

$$N(m, g, n) =$$

$$\left[\underbrace{n - w + 1}_{\text{Starting positions}} + \underbrace{\frac{w - m}{2}}_{\text{Starting positions near end}} \right] \cdot \underbrace{(g + 1)^{m-1}}_{\text{Ways to place remaining } m-1}$$

assuming $w - 1 \leq n$

Using statistics to choose parameter values



Preliminary Results: Max-Gap Cluster Statistics

■ Reference set

- complete clusters
- complete clusters with length restriction
- incomplete clusters

➤ Whole genome comparison

- upper and lower bounds

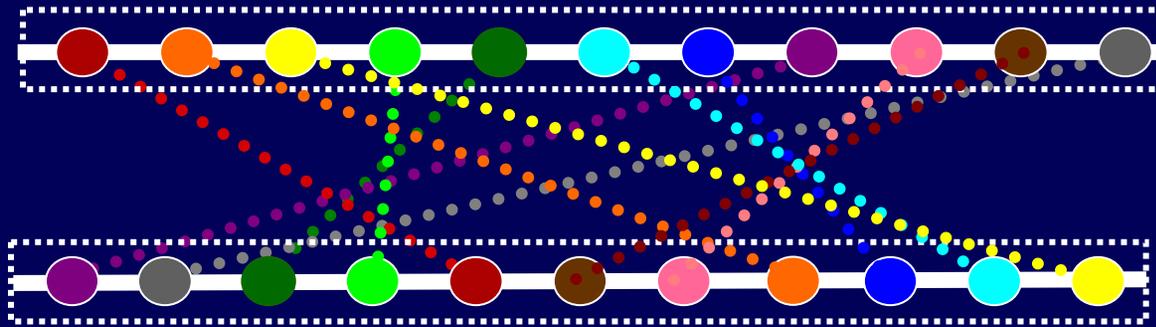
Hoberman, Sankoff, Durand. Journal of Computational Biology 2005.

Hoberman and Durand. RECOMB Comparative Genomics 2005.

Hoberman, Sankoff, Durand. RECOMB Comparative Genomics 2004.

Whole genome comparison

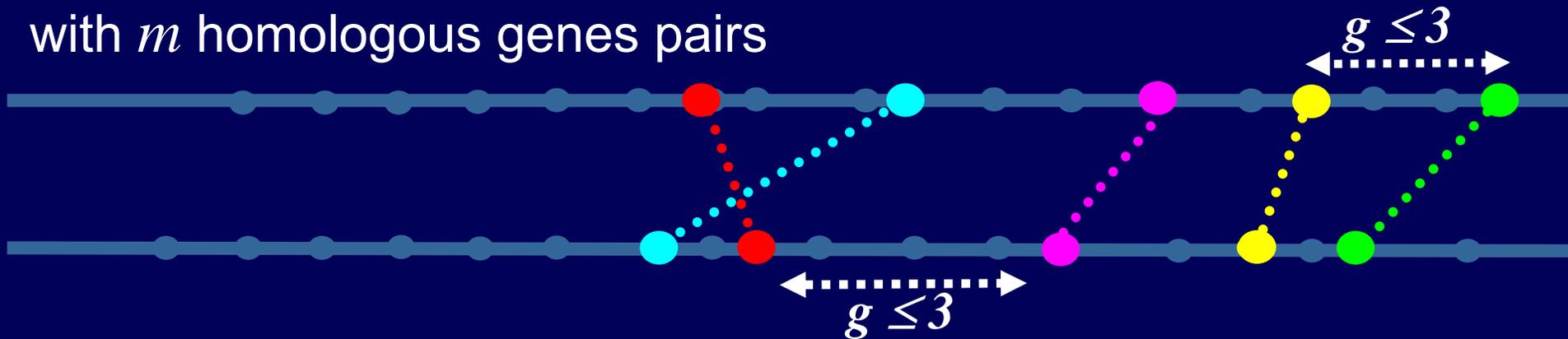
A surprising result



If gene content is identical,
the probability of a max-gap cluster is 1
(regardless of the allowed gap size)

Whole Genome Comparison: $m \leq n$

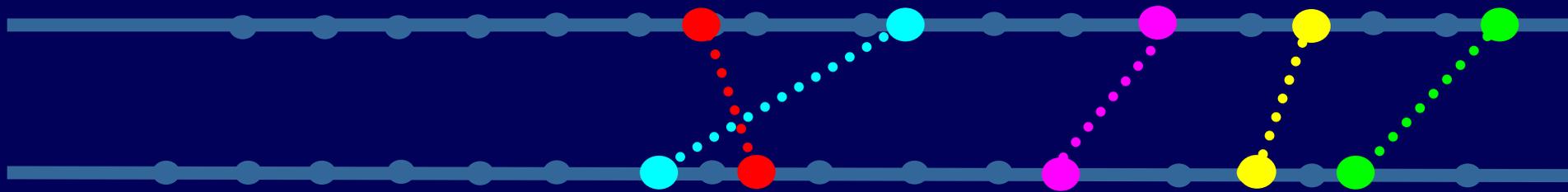
Two genomes of n genes with
with m homologous genes pairs



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

A cluster is **maximal** if it is not a subset of a larger cluster

A constructive approach



All configurations
of two genomes

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

Configurations
that contain a cluster
of exactly size h

??

Constructive Approach

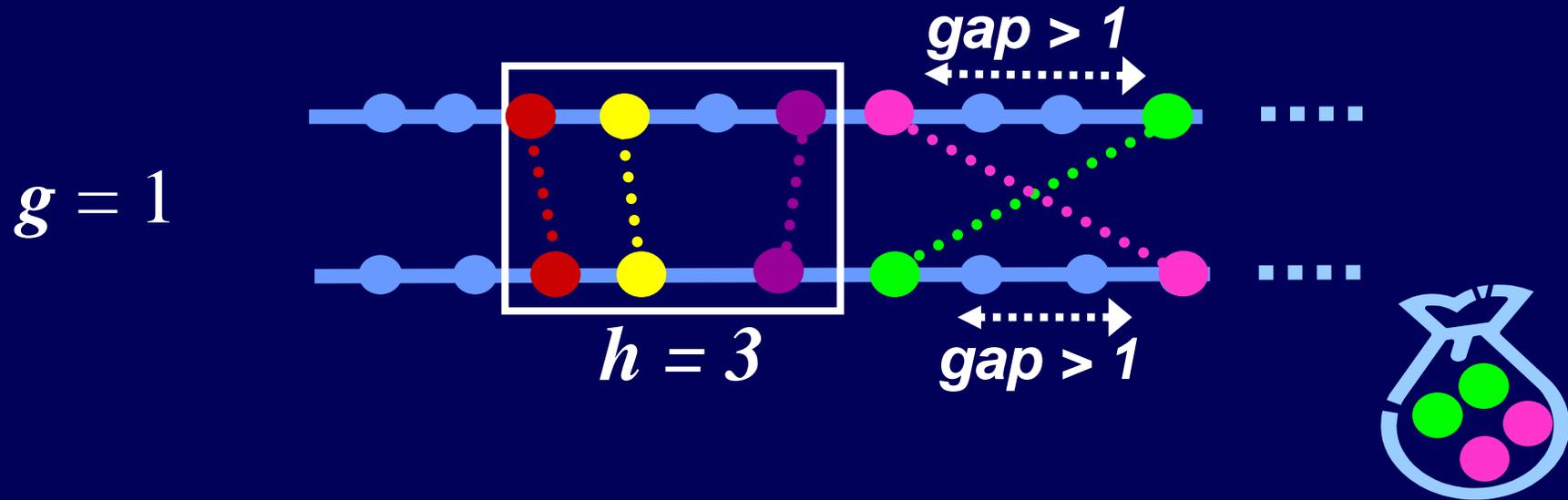
Number of configurations that contain a cluster of exactly size h

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

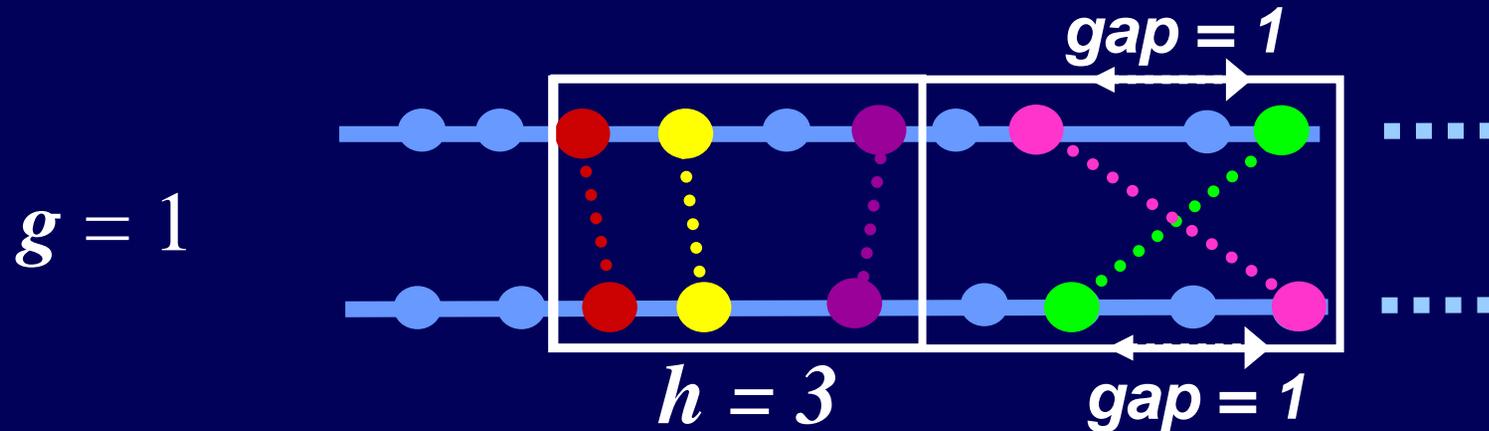

$$N(h, g, n)^2 \cdot h! \cdot ??$$

A tricky case...



Where can we place the pink and green genes so
With this placement, the cluster cannot be extended
that they do not extend this cluster of size three?

A tricky case...



Moving genes further away from the cluster may make them more likely to extend the cluster

My whole-genome comparison results

I derived upper and lower bounds on the probability of observing a cluster containing h homologs, via whole genome comparison

- Lower bound: guarantees no tricky cases
- Upper bound: a few tricky cases sneak in

Hoberman, Sankoff, Durand. Journal of Computational Biology 2005.

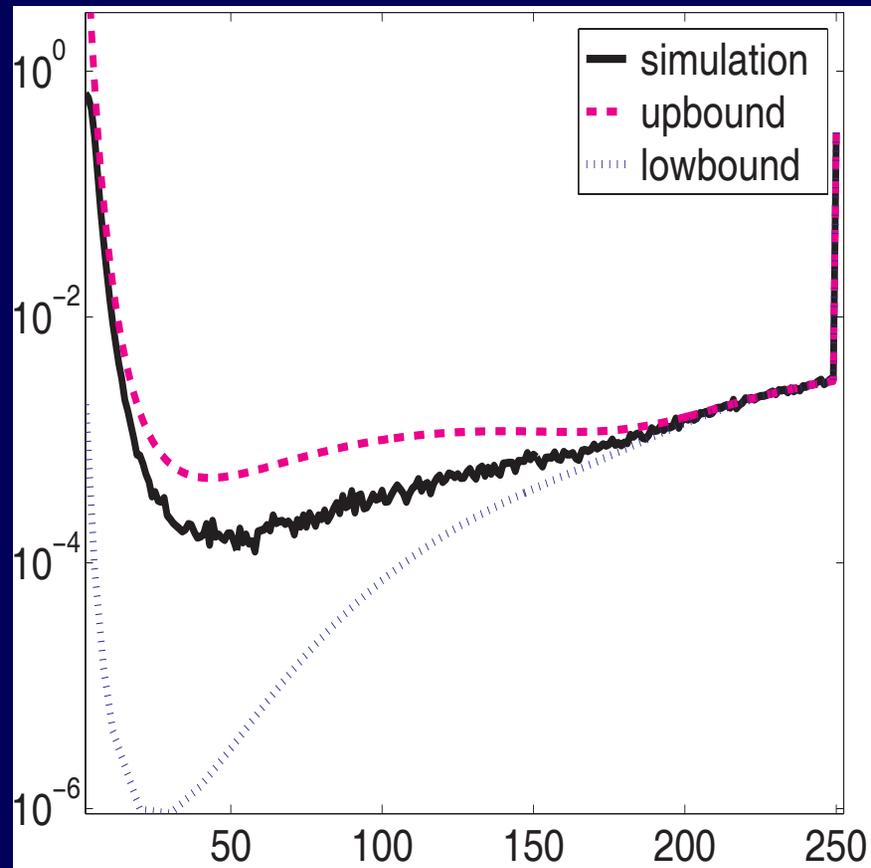
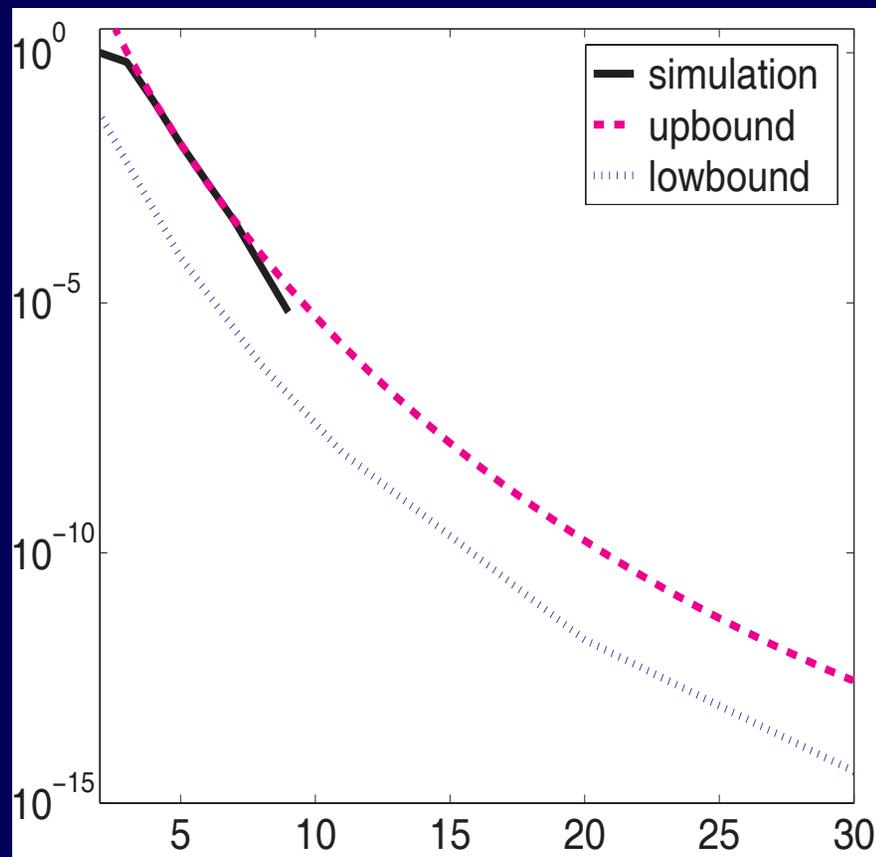
Whole-genome comparison cluster statistics

$n=1000, m=250$

$g=10$

$g=20$

Cluster Probability

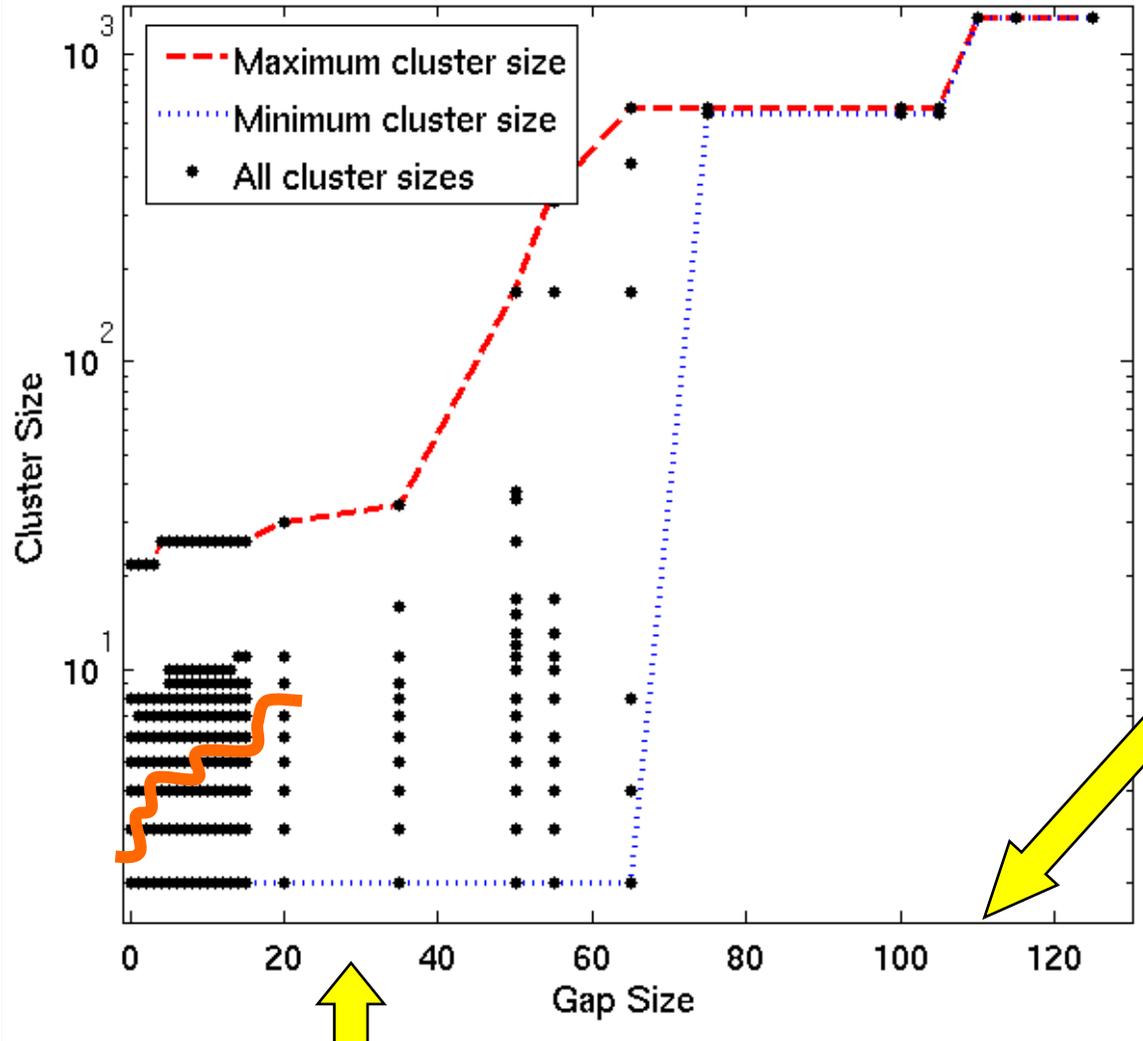


Cluster size

E. coli vs B. Subtilis

Algorithm:
Bergeron
et al, 02

Statistics:
Hoberman
et al, 05



Typical
operon
sizes

Complete
cluster
doesn't
form until
g=110

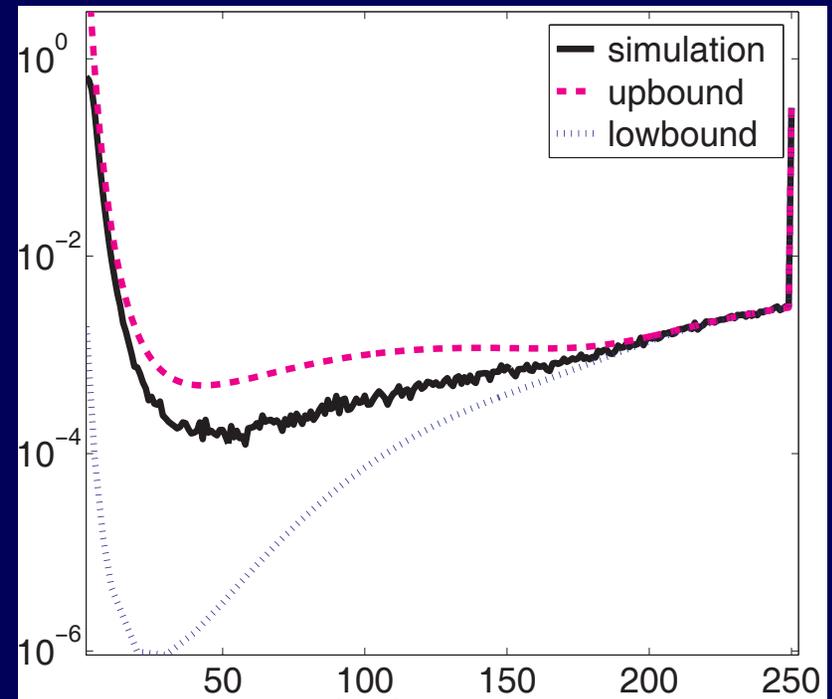
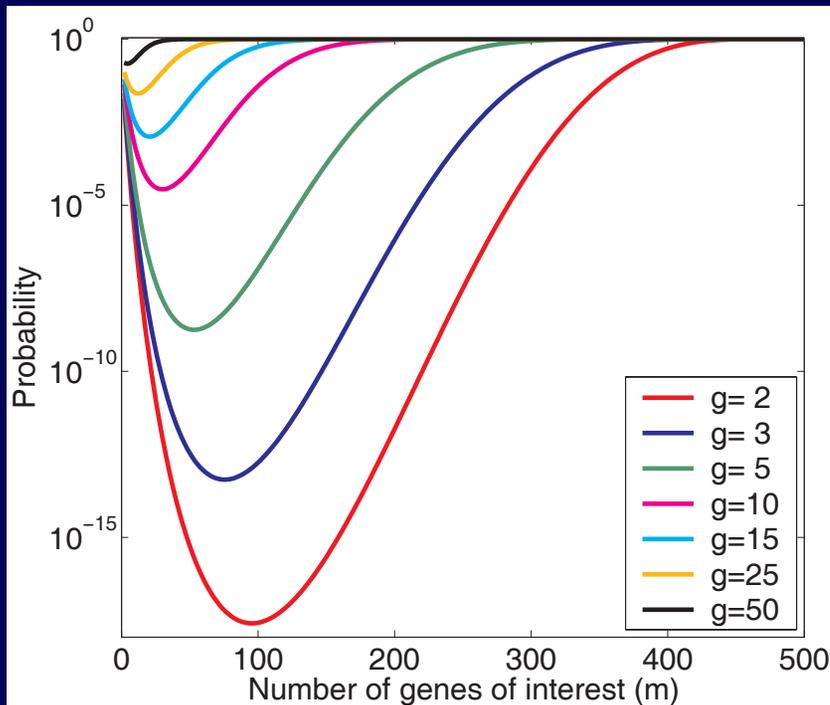
clusters above the orange line
are significant at the 0.01 level
Under null hypothesis, by g=25 all
genes should form a single cluster

Summary of preliminary work

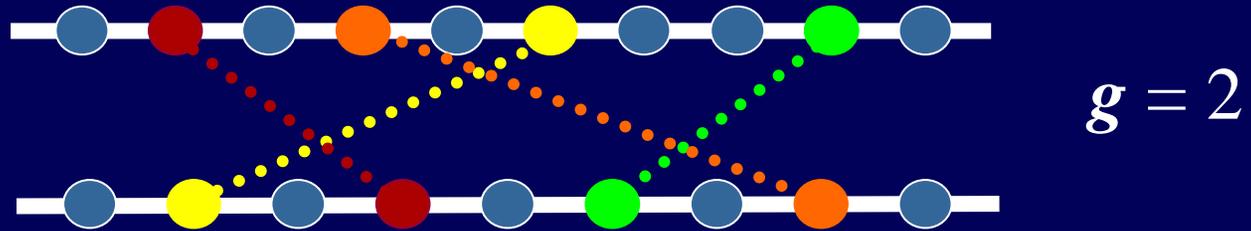
- Developed statistical tests using a combinatoric approach
 - reference region
 - whole genome comparison
- Some surprising results
- Results raise concerns about current methods used in comparative genomics studies

Larger clusters do not always imply greater significance

A max-gap cluster containing many genes may be *more* likely to occur by chance than one containing few genes



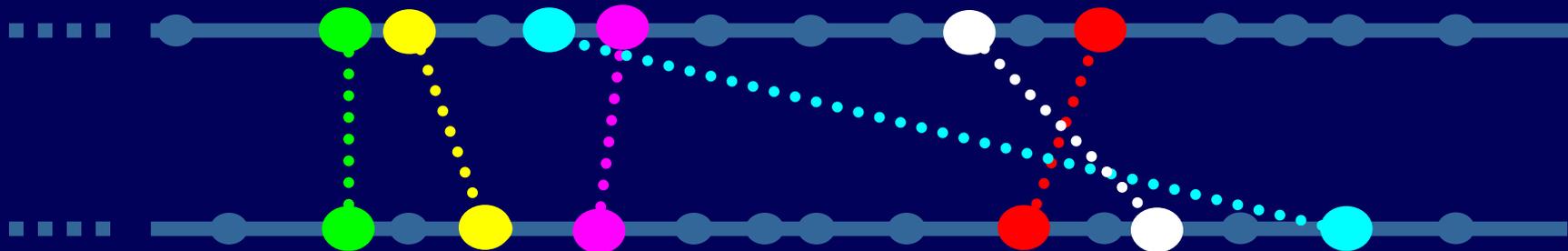
Algorithms and Definition Mismatch



- Greedy, bottom-up algorithms will not find all max-gap clusters
- There is an efficient divide-and-conquer algorithm to find maximal max-gap clusters (Bergeron *et al*, WABI, 2002)

Extending the Model

- Directions for generalization
 - Circular chromosomes
 - Multiple chromosomes
 - Genome self-comparison
 - Gene order and orientation
 - Gene families



Outline

- Introduction and Applications
- Formal framework for gene clusters
- An introduction to statistical issues
- Preliminary work: Testing cluster significance
- **Proposed work**

Proposed Work Outline

- Generalizing the model

- At least one of the following:

1. Joint detection of orthologous genes and chromosomal regions
2. Finding and assessing clusters in multiple genomes
3. Detecting selection for spatial organization

- Validation

Joint Identification of Orthologous Genes and Chromosomal Regions

- The identification of orthologous genes is a prerequisite for a marker-based approach
- Orthology identification
 - is often difficult to determine from gene sequence alone
 - is an important unsolved research problem
 - can be improved by incorporating genomic context

Problem: for more diverged genomes,
unambiguous orthologs will be sparse
and clusters will be more rearranged

Identify homologous genes



Solution: Identify orthologs and gene clusters
simultaneously

- Work that combines sequence similarity and genomic context
 - Bansal, Bioinformatics 99
 - Kellis et al, J Comp Biol 04
 - Bourque et al, RECOMB Comp Genomics 05
 - Chen et al, ACM/IEEE Trans Comput Biol and Bioinf 05

- Limitations
 - No flexible cluster definitions
 - No statistical approaches
 - Little real evaluation

Possible computational approaches:

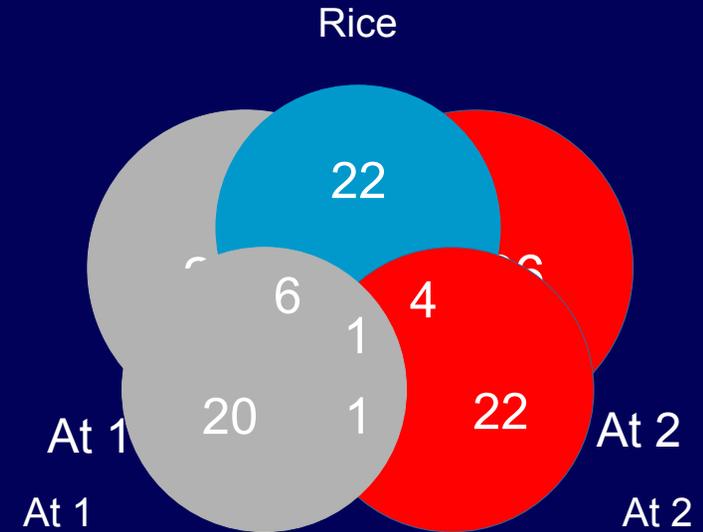
- Expectation Maximization (EM)
 - treat ortholog assignment as a hidden variable
- Maximal bipartite matching
 - use an objective function that incorporates both sequence similarity and spatial clustering

Proposed Work

- Generalizing the model
- At least one of the following:
 1. Joint detection of orthologous genes and chromosomal regions
 2. Finding and assessing clusters in multiple genomes
 3. Detecting selection for spatial organization
- Validation

Comparing Multiple Genomes Simultaneously

Comparison of multiple genomes offers significantly more power to detect highly diverged homologous segments



Arabidopsis thaliana



Rice

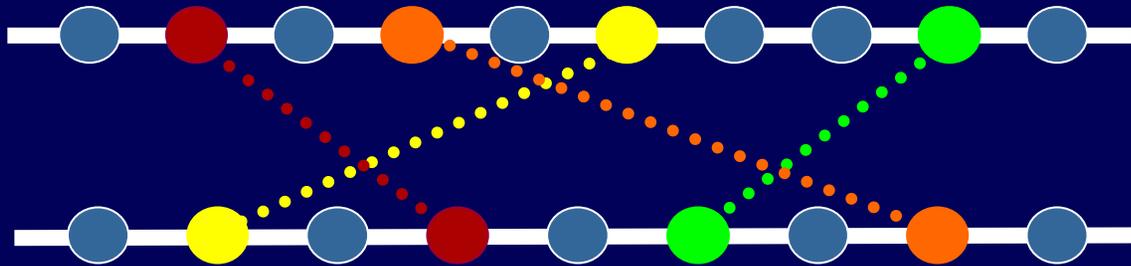


Arabidopsis thaliana

Vandepoele et al, 2002

Current Approaches

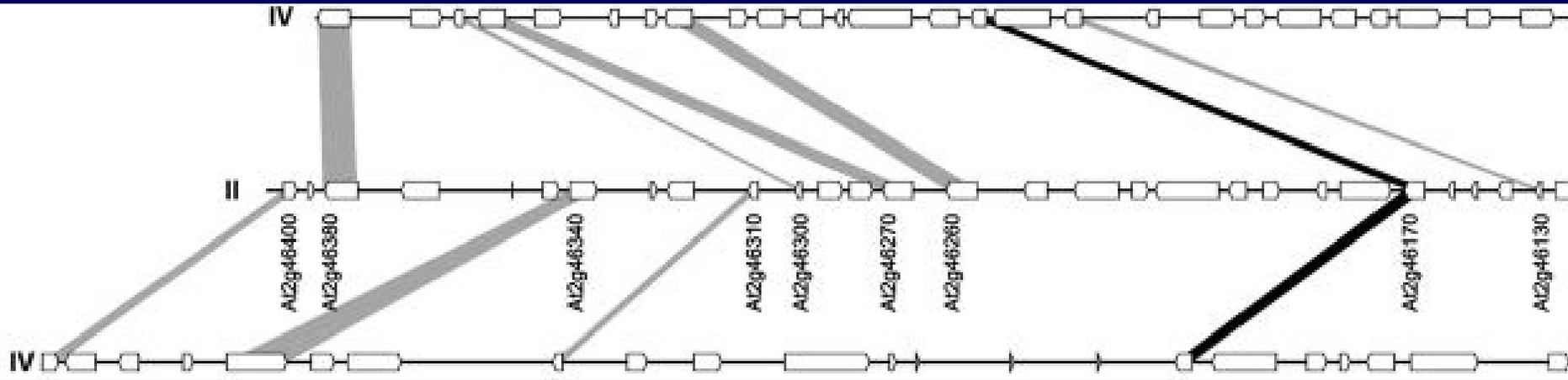
1. Identify clusters based on conserved pairs of genes, using heuristics



Limitation: A highly rearranged cluster may have no pairs in proximity

Current Approaches

2. Identify clusters with conserved gene order,

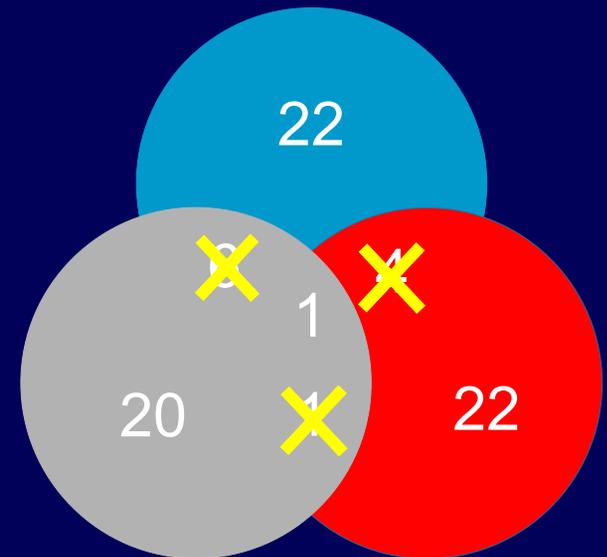


Limitation: rearranged clusters will not be detected

Current Approaches

3. Search for max-gap clusters, but require the cluster to be found in its entirety in all genomes

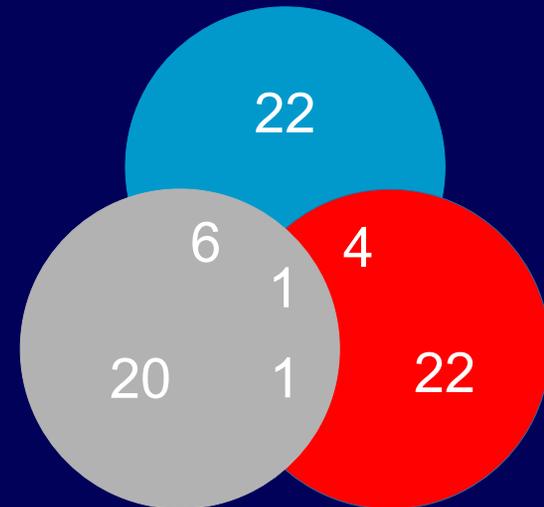
Will lead to a *reduction* in power as more genomes are added



...No formal statistics

Initial Investigations

- Modeling: Maximum gap between genes with a match in *any* of the regions must be small
- Algorithms: how to find such clusters
- Statistics: choice of test statistic; *i.e.*, how to weight genes that occur in only a subset of the regions



Proposed Work

- Generalizing the model
- At least one of the following:
 1. Joint detection of orthologous genes and chromosomal regions
 2. Finding and assessing clusters in multiple genomes
 3. Detecting selection for spatial organization
- Validation

Tests for Selective Pressure on Spatial Organization

Preliminary work:

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

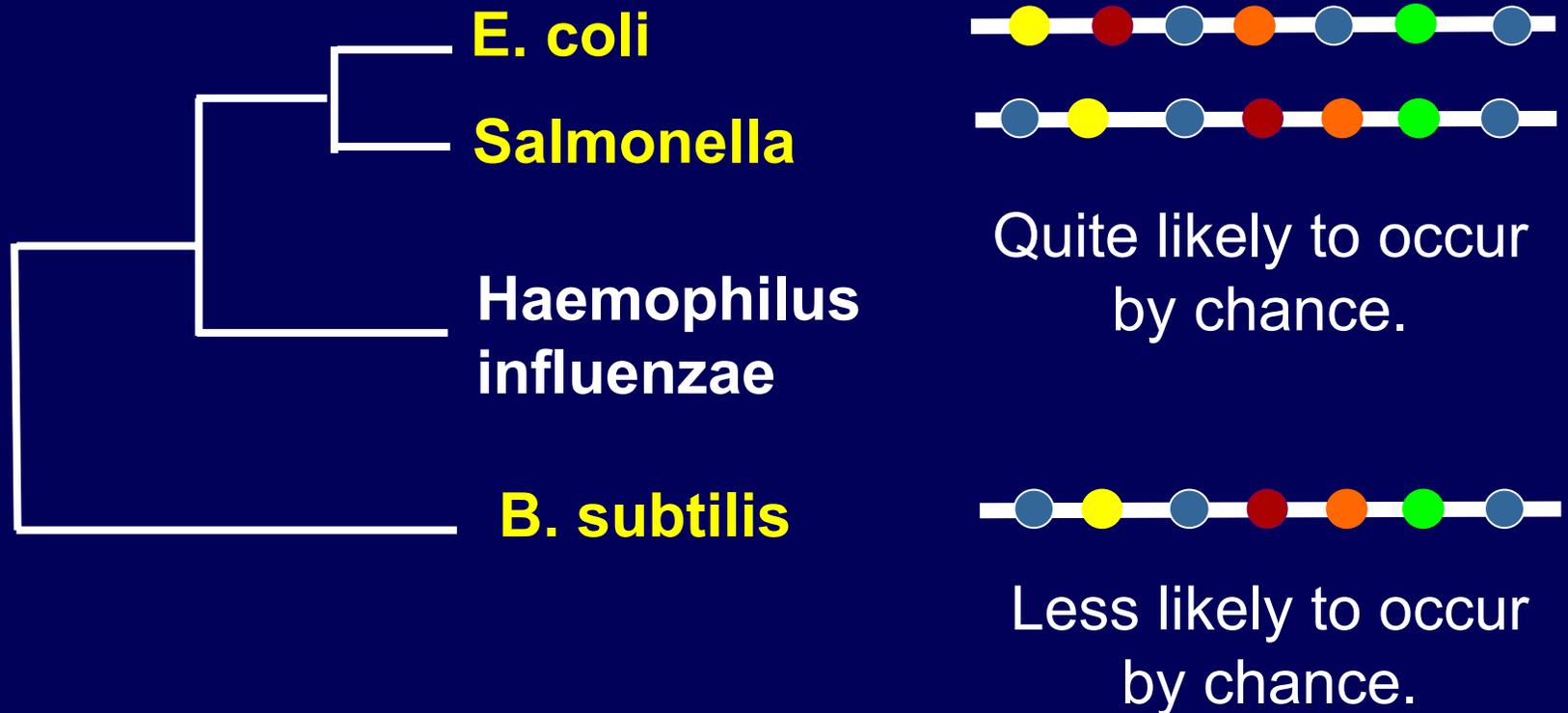
Proposed work:

- Null hypothesis: common ancestry
- Alternate hypothesis: functional selection



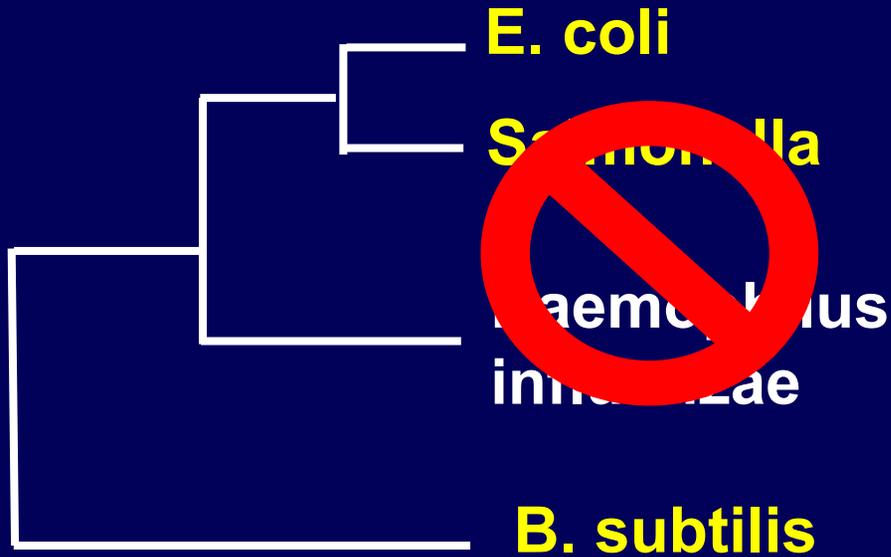
Probability of finding a cluster under the null hypothesis now depends on the phylogenetic distance between the species

Tests for selective pressure must consider phylogenetic distance



Current Approaches

1. Discard closely related genomes, and test against random gene order



Current Approaches

2. Some formal statistical tests, but based on gene pairs only.

Limitation: considering only pairs of genes could result in a loss of power

Detecting Selective Pressure on Spatial Organization

Initial Explorations

- Searching for evidence of selective pressure to maintain non-operon structure in bacteria
- Locations of clusters with respect to
 - origin and terminus
 - left and right arm of chromosome
 - functional classification

Proposed Work

- Generalizing the model
- At least one of the following:
 1. Joint detection of orthologous genes and chromosomal regions
 2. Finding and assessing clusters in multiple genomes
 3. Detecting selection for spatial organization
- **Validation**

How Should Gene Cluster Statistics be Validated?

- No established benchmarks
 - True evolutionary histories are rarely known
 - Rearrangement processes are not yet understood
- We'd like to evaluate
 - Discriminatory power
 - Parameter selection strategies
- Possible strategies depend on specific problem
 - Synthetic data
 - Hand-curated ortholog databases
 - Databases of experimentally verified operons

timeline

S	O	N	D	J	F	M	A	M	J	J	A	S	O	N	D
9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12
2005				2006											
Loose ends	Model Extensions					Selected Problem(s) And Validation			Writing						
	Initial Investigations														

Acknowledgements

- My Thesis Committee
- Barbara Lazarus Women@IT Fellowship
- The Sloan Foundation
- The Durand Lab

Advantages of an analytical approach

- Analyzing incomplete datasets
- Principled parameter selection
- Efficiency
- Understanding statistical trends
- Insight into tradeoffs between definitions

The Max-Gap Definition is the Most Widely 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

Lawrence and Roth 1996, identification of horizontal transfers

Tamames 2001, evolution of gene order conservation in prokaryotes

Wolfe and Shields 1997, ancient yeast duplication

McLysaght02, genomic duplication during early chordate evolution

Coghlan and Wolfe 2002, comparing rates of rearrangements

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

...