Identifying conserved spatial patterns in genomes

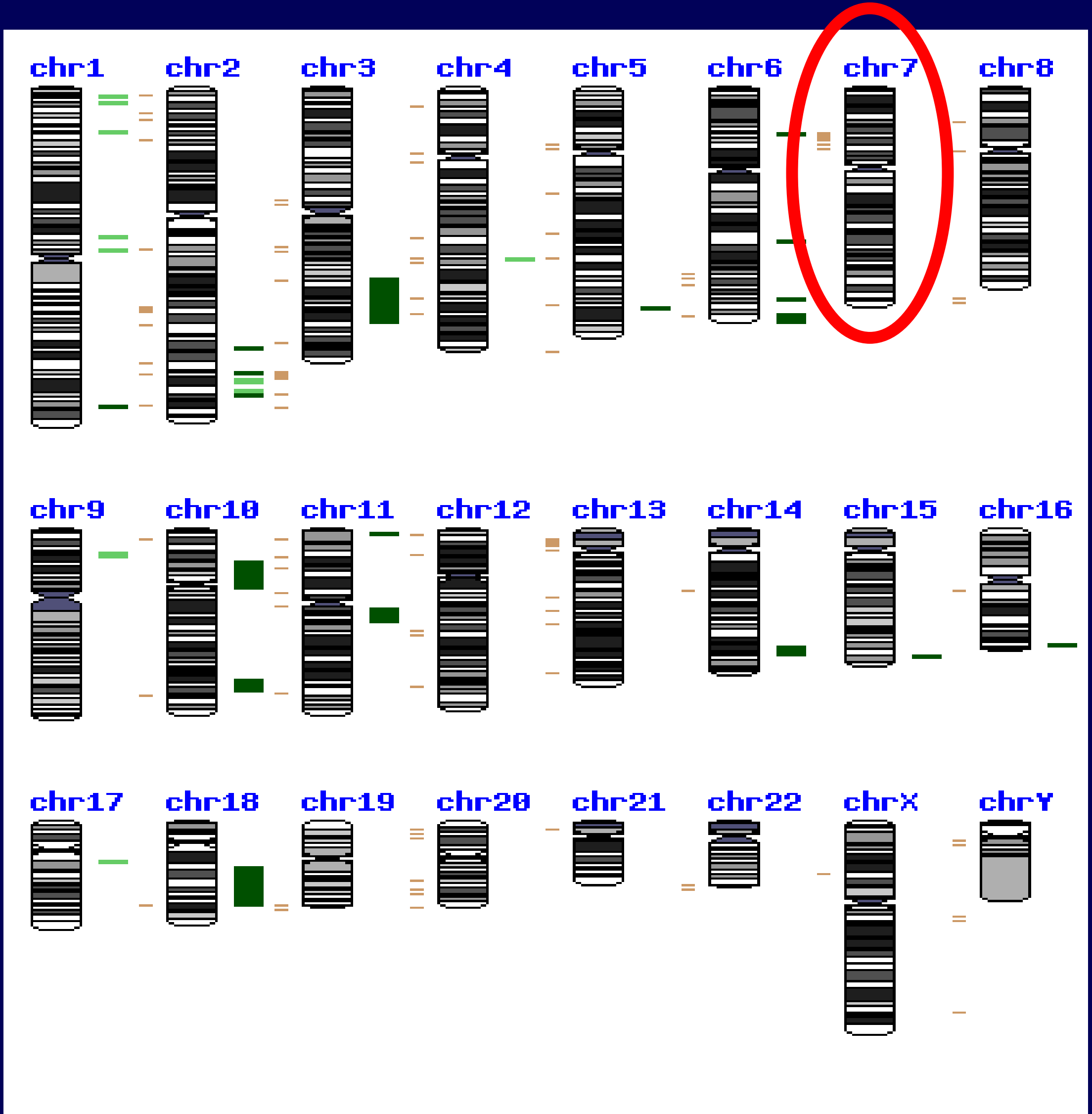
Rose Hoberman

Dannie Durand
Depts. of Biological Sciences
and Computer Science, CMU

David Sankoff
Dept. of Math and Statistics
University of Ottawa

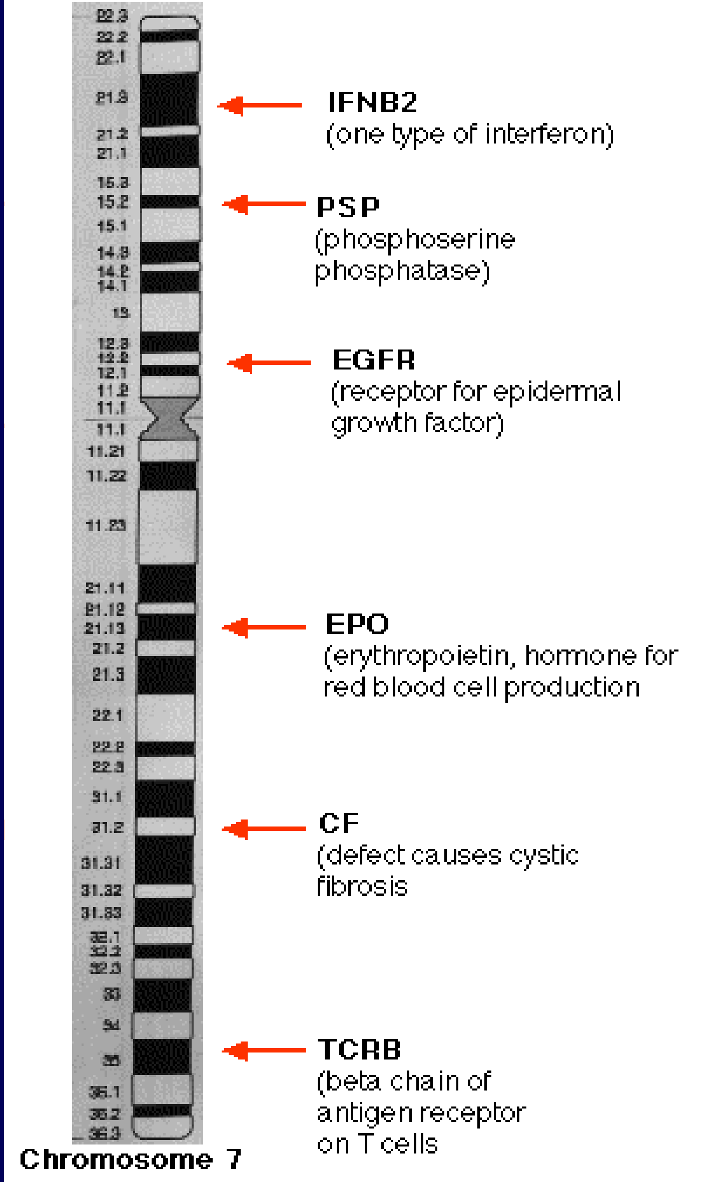
Student Seminar Series
Jan 20, 2006

The Genome



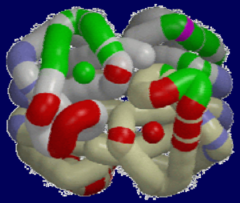
The complete genetic material of an organism or species

Key genomic component: genes


IFNB2
(one type of interferon)
PSP
(phosphoserine phosphatase)
EGFR
(receptor for epidermal growth factor)
EPO
(erythropoietin, hormone for red blood cell production
CF
(defect causes cystic fibrosis
TCRB
(beta chain of antigen receptor on T cells
Chromosome 7

A gene is a DNA subsequence

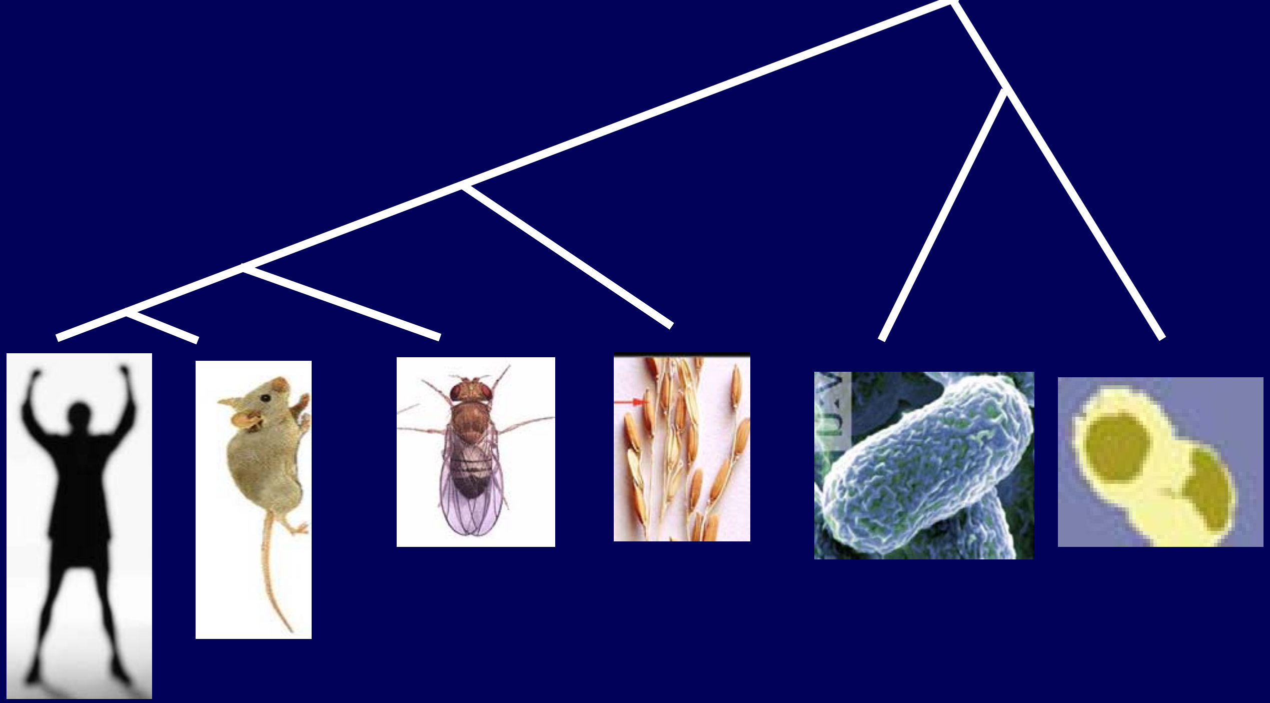
ACCCTTAGCTAGACCTTTAGGAGG...



Genes encode proteins,

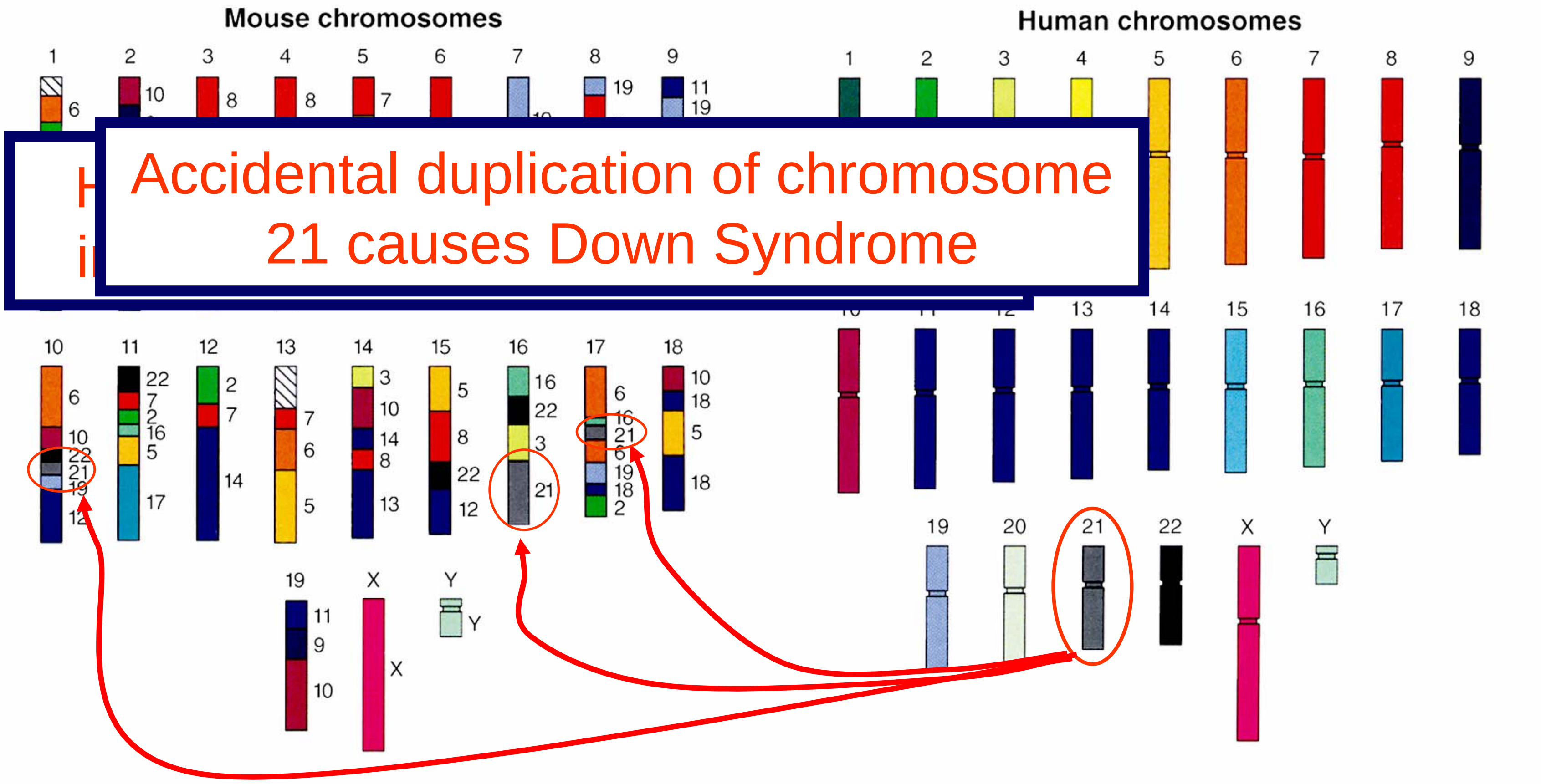
the building blocks of the cell

Comparing Genomes



	Human	Mouse	Fly	Rice	*E. Coli*	*Chlamydia*
Chromosomes	23	20	4	12	1	1
Genes	20-25k	20-25k	13.6k	~40k	3200	936

Mouse and Human Genetic Similarities



Accidental duplication of chromosome 21 causes Down Syndrome

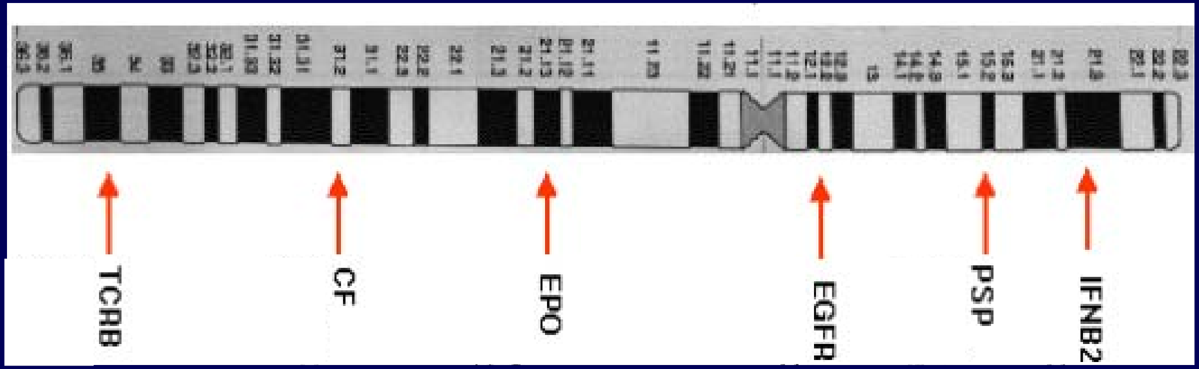
Courtesy Lisa Stubbs
Oak Ridge National Laboratory

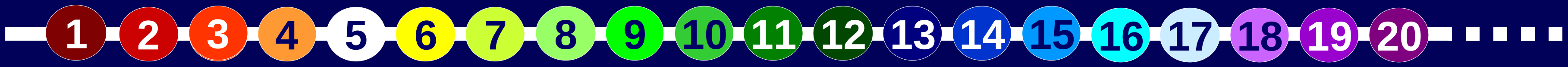
YGA 98-075R2

Outline

- Evolution of genome organization
- Why identify related genomic regions?
- How do we find them?
  - **Identification:** Formal cluster definition
  - **Validation:** Testing cluster significance

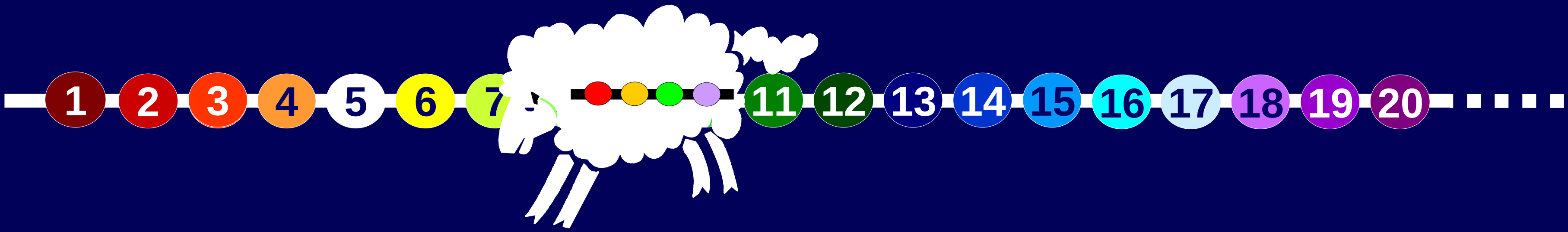
A simple model of a chromosome



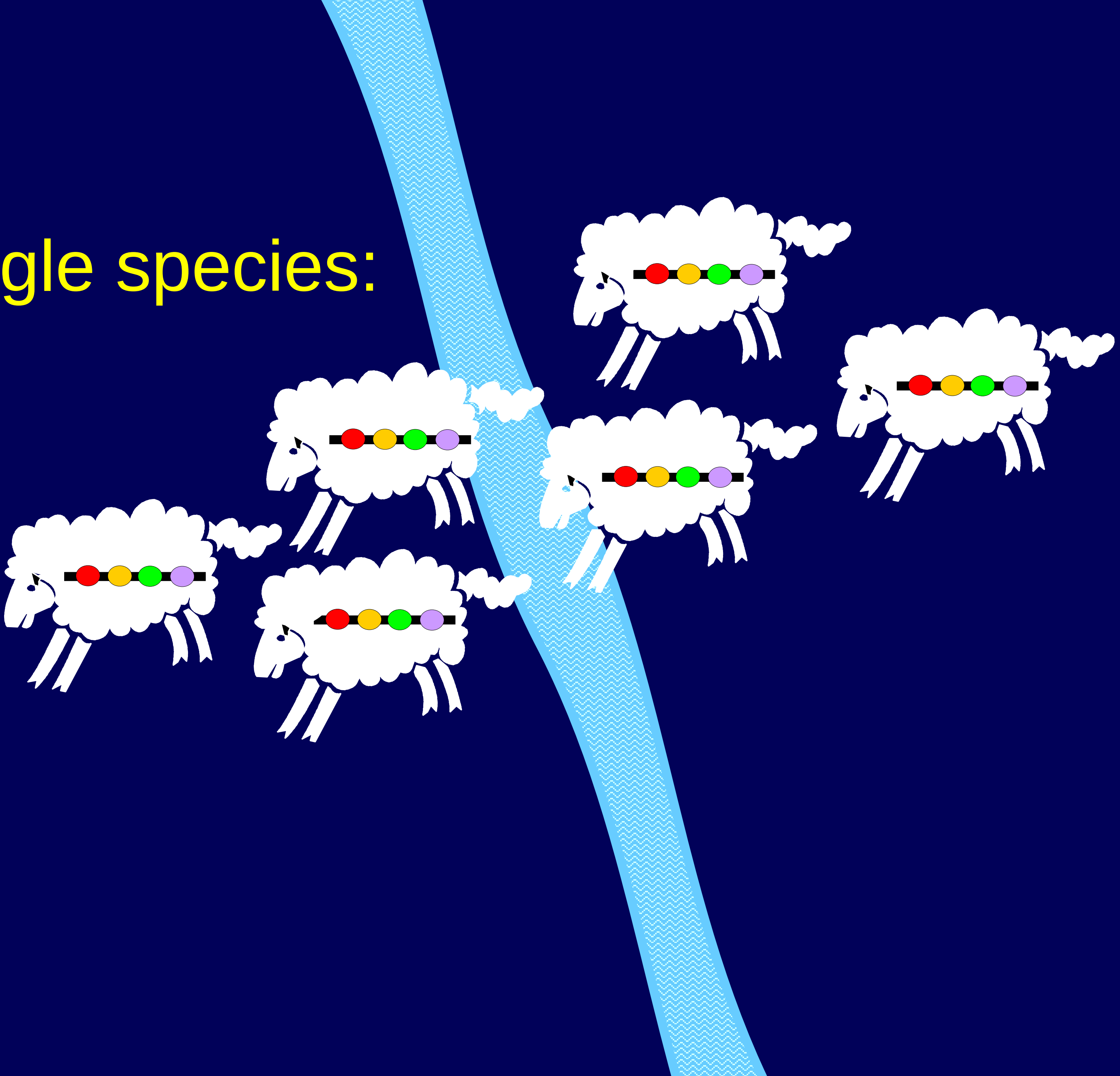


an ordered list of genes

What are the processes of genomic change?



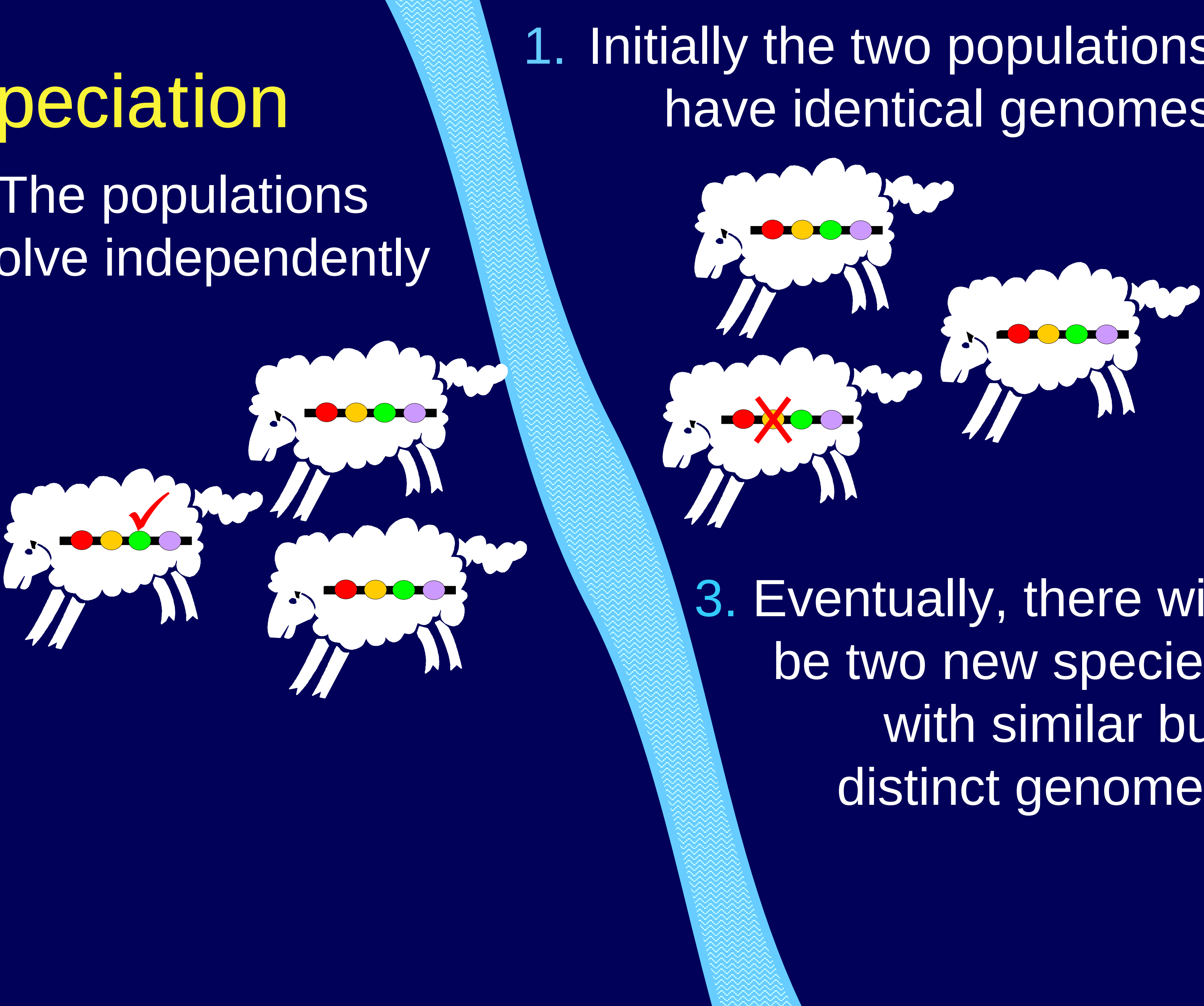
A single species:



Speciation

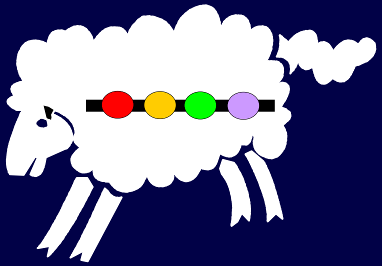
1. Initially the two populations have identical genomes

2. The populations evolve independently

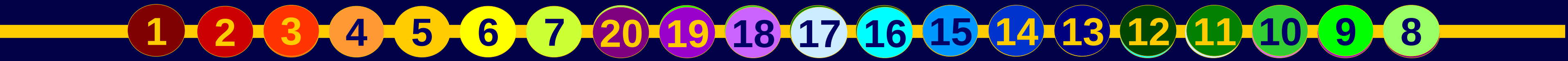


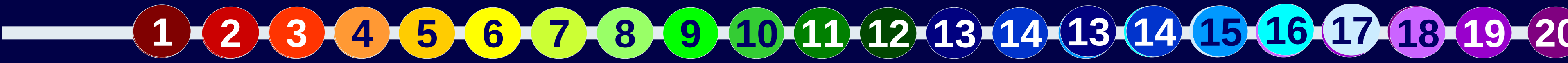
3. Eventually, there will be two new species with similar but distinct genomes

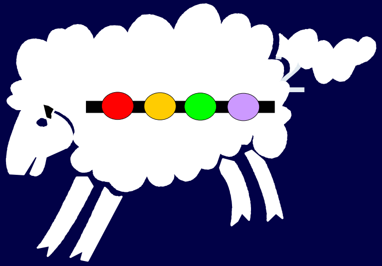
Types of Genomic Rearrangements



Inversions





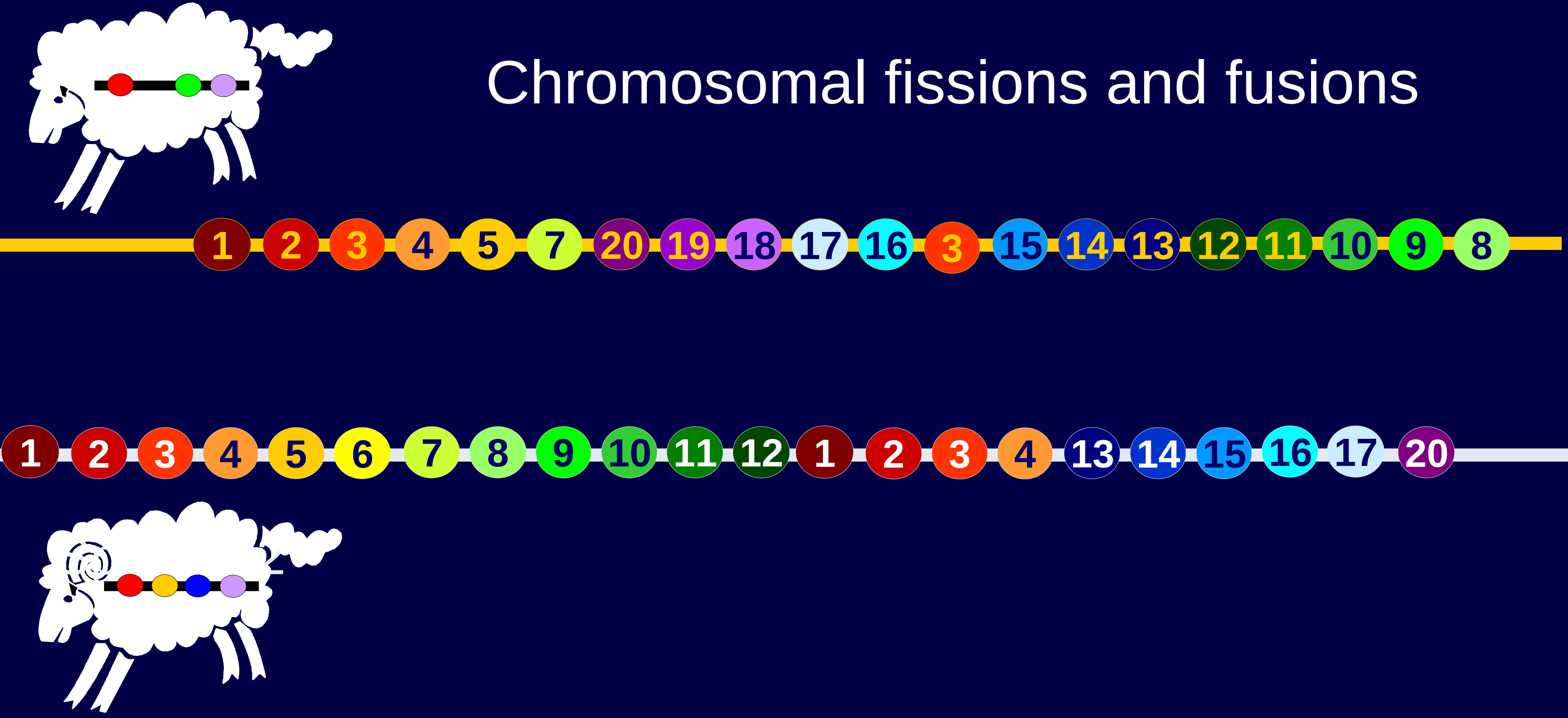


Duplications/Insertions

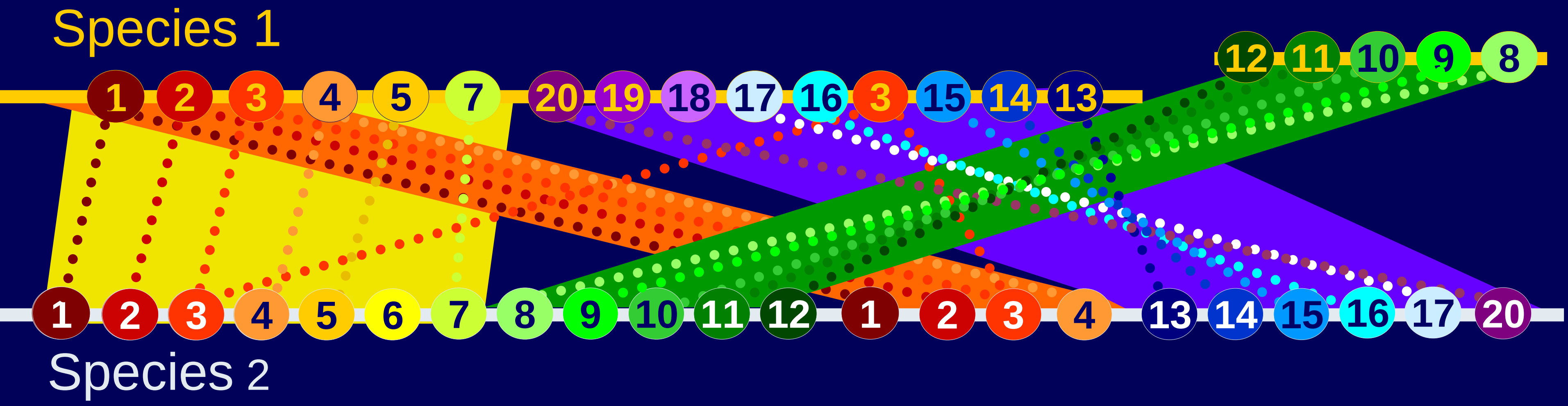
Loss

Types of Genomic Rearrangements

Chromosomal fissions and fusions



Genome Comparison



Our goal: identify chromosomal regions that descended from the same region in the genome of the common ancestor

Outline

- Evolution of genome organization
- Why identify related genomic regions?
- How do we find them?
  - **Identification:** Formal cluster definition
  - **Validation:** Testing cluster significance

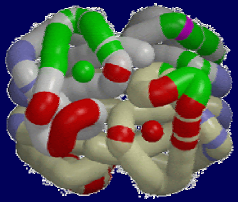
Genome Annotation Problem

Given the set of genes in the genome, label each with its function

Gene

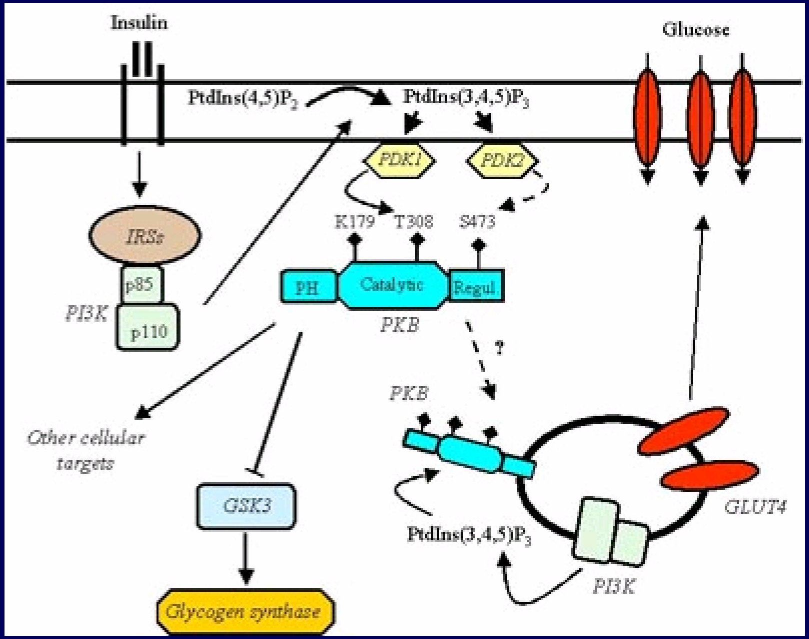
ACCCTTAGCTAGACCTTTAGGAGGTGCAGGA

Protein



Cellular Pathway:

Glucose Metabolism



There are many aspects of gene function

- Gene: trpA
- Biochemical Function: cleaves a double bond
- Cellular Process: amino-acid biosynthesis
- Protein-protein interactions: binds trpB

There are many aspects of gene function

- Gene: a typical gene
- Biochemical Function: ?
- Biological Process: ?
- Protein-protein interactions: ?

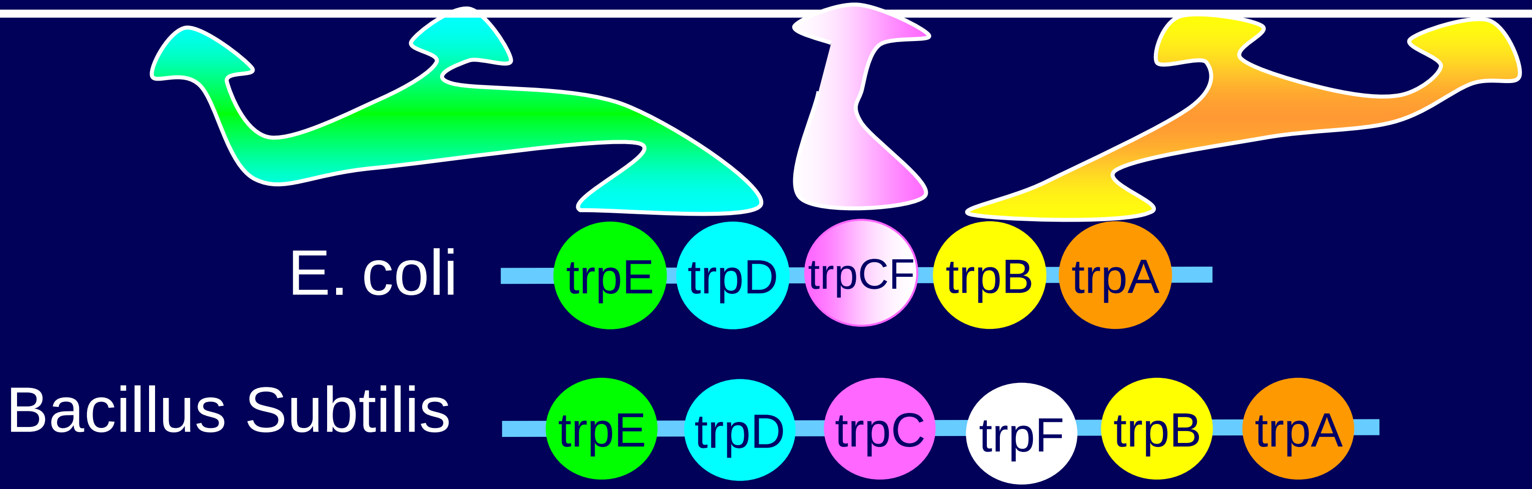
40-60% of genes in most genomes have unknown function

Comparisons of spatial organization within genomes can yield gene function predictions

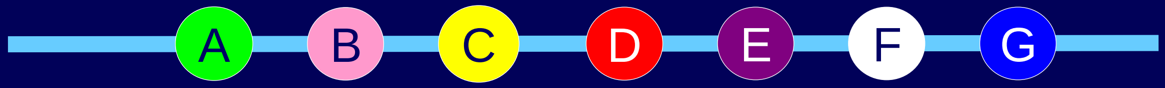
In bacteria, genes in the same pathway often occur together in the genome

Tryptophan Synthesis Pathway

Chorismate → Anthranilate → N-5’-P-ribosyl-anthranilate → 1-2 Carboxy-phenylaminodeoxy-ribulose-5’P → 3-Indole Glycerol-P → Tryptophan



Conserved spatial organization between distantly related species suggests functional associations between the genes



A Glucose metabolism

B Glucose metabolism

C ?

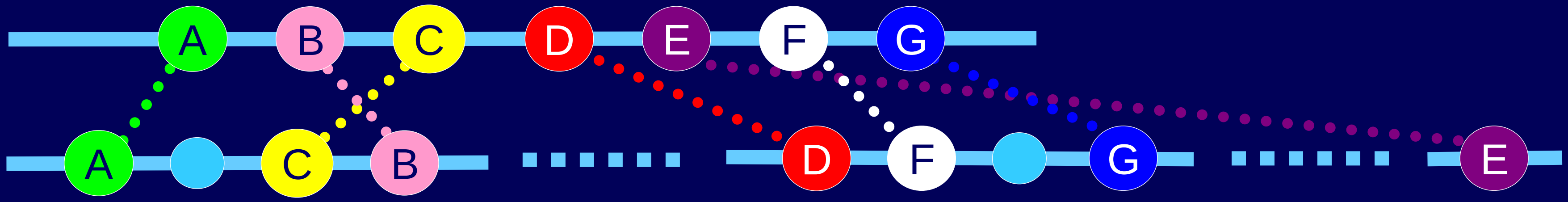
D Tryptophan synthesis

E ?

F ?

G Tryptophan synthesis

Conserved spatial organization between distantly related species suggests functional associations between the genes



A Glucose metabolism

B Glucose metabolism

C *Prediction:* Glucose metabolism

D Tryptophan synthesis

E *?*

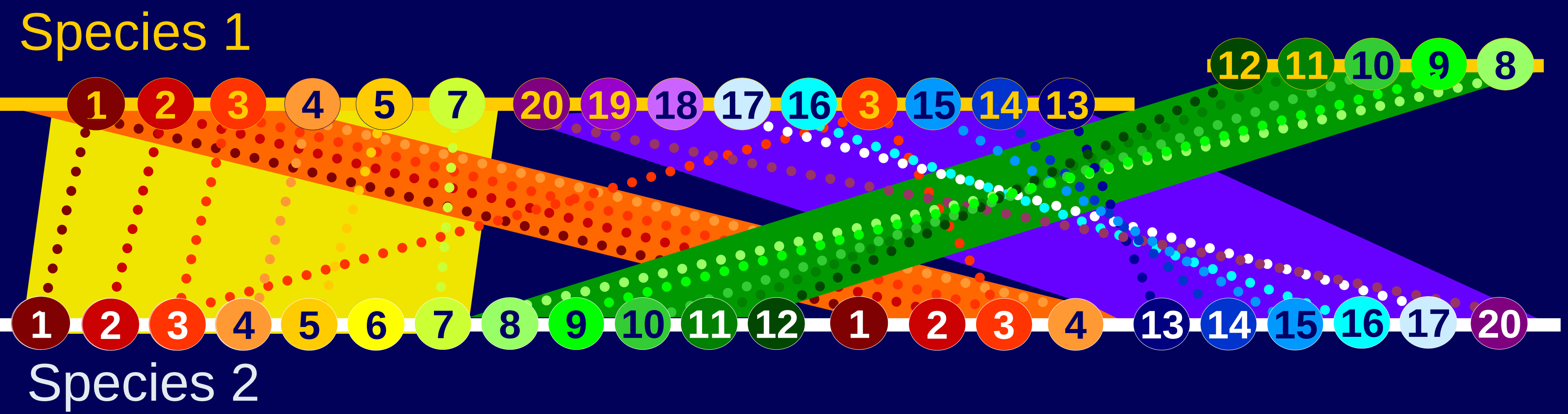
F *Prediction:* Tryptophan synthesis

G Tryptophan synthesis

Outline

- Evolution of genome organization
- Why identify related genomic regions?
- How do we find them?
  - Identification: Formal cluster definition
  - Validation: Testing cluster significance

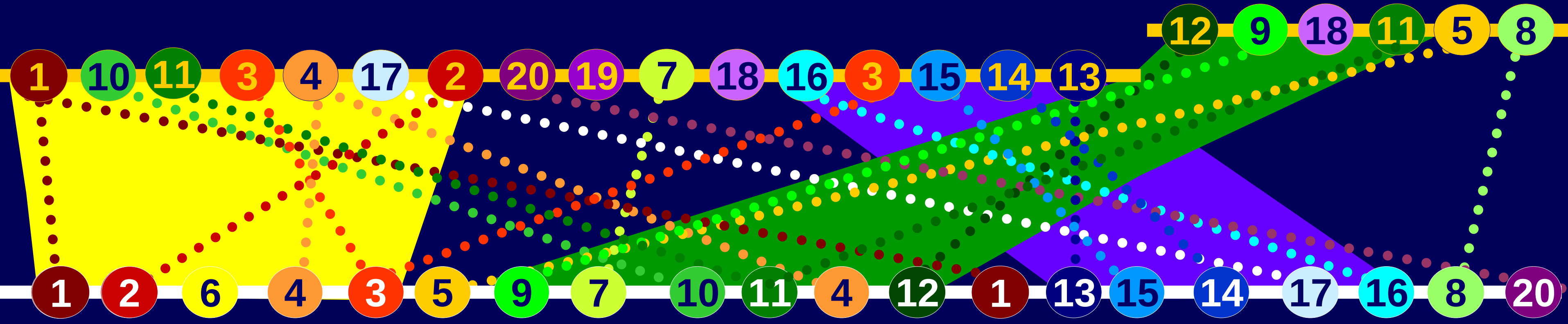
Closely related genomes



Related regions, regions that descended from the same region in the genome of the common ancestor, are easy to identify

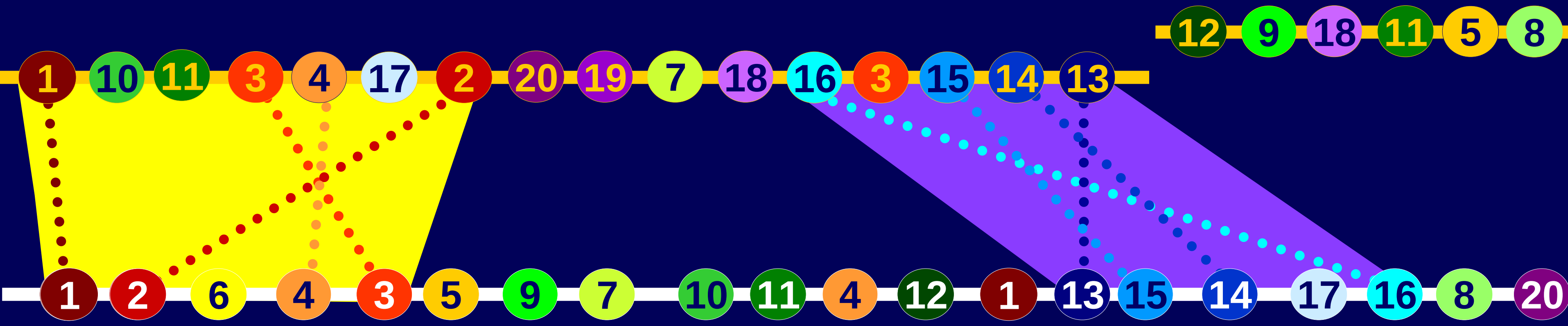
A hundred million years…

More Diverged Genomes



- Related regions are harder to detect, but there is still spatial evidence of common ancestry
  - Similar gene content
  - Neither gene content nor order is perfectly preserved

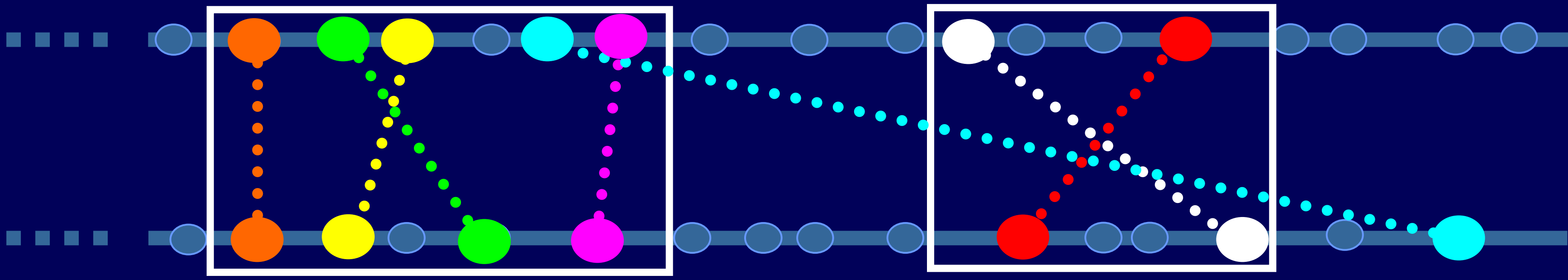
The signature of diverged regions



Gene clusters

- Similar gene content
- Neither gene content nor order is perfectly

A Framework for Identifying Gene Clusters



1. Find corresponding genes — given as input
2. Formally define a "gene cluster" — review the most common definition
3. Devise an algorithm to identify clusters
4. Statistically verify clusters — my work

Clusters are signatures of distantly related regions.

Without functional constraints...

- After sufficient time has passed, gene order will become randomized
- Uniform random data tends to be “clumpy”
  - some genes will end up proximal in both genomes simply by chance

Not all clusters have biological significance.

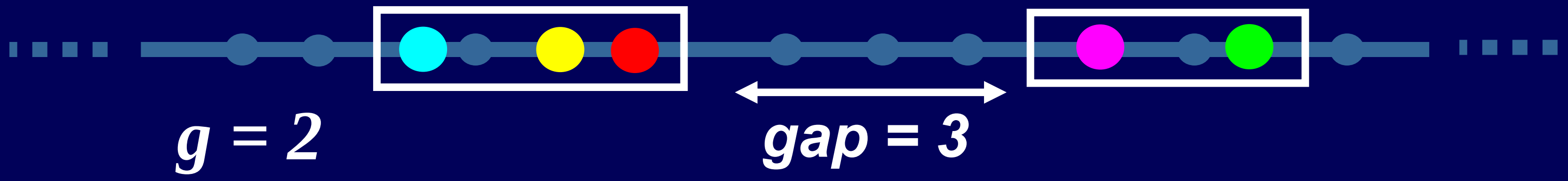
Cluster Validation via Hypothesis Testing

- Null hypothesis: random gene order
- Reject gene clusters that could have arisen under the null model
- Clusters that cannot be rejected are likely to be functionally constrained

Outline

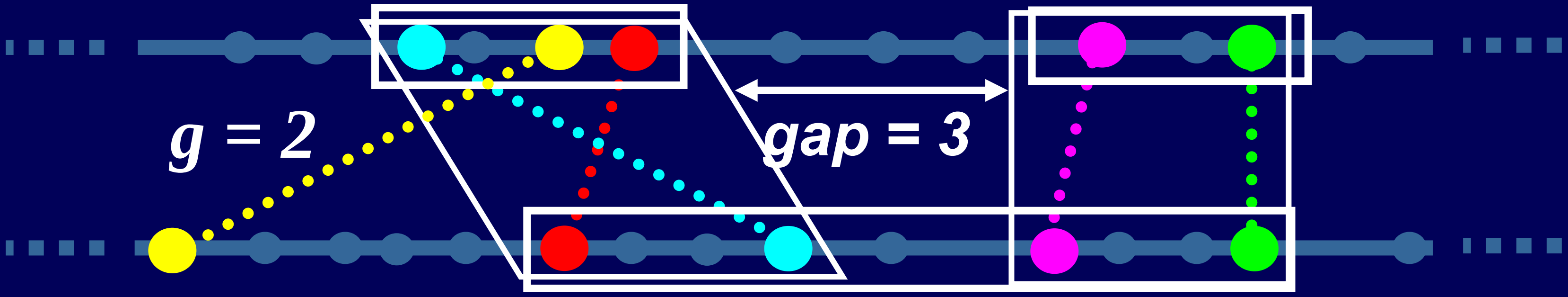
- Evolution of genome organization
- Why find related genomic regions?
- How do we find them?
  - Identification: max-gap cluster definition
  - Validation: Testing cluster significance

A max-gap chain



- The distance or “gap” between genes is equal to the number of intervening genes
- A set of genes in a genome form a max-gap chain if
    - the gap between adjacent genes is never greater than *g* (a user-specified parameter)

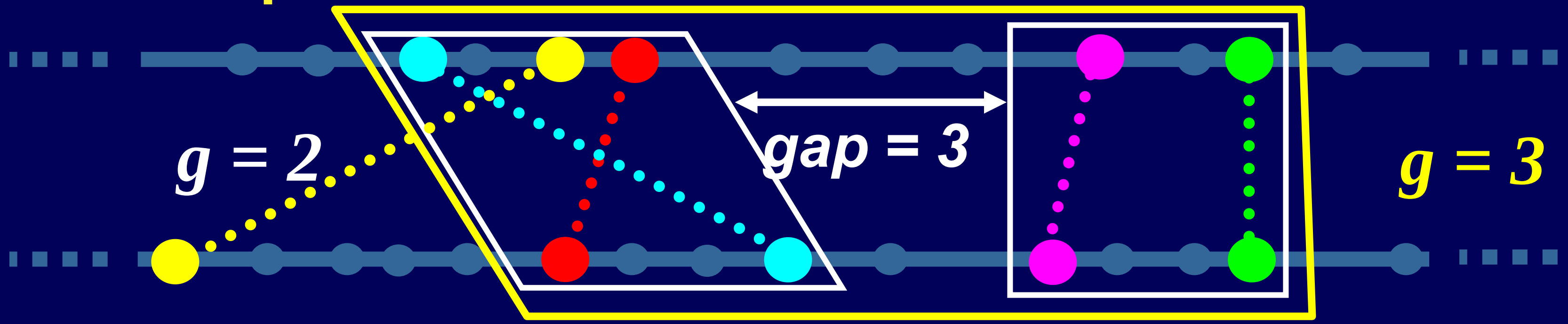
Max-Gap cluster definition



A set of genes form a max-gap cluster of two genomes if

1. the genes forms a max-gap chain in each genome
2. the cluster is maximal (i.e. not contained within a larger cluster)

Max-Gap cluster definition



A set of genes form a max-gap cluster of two genomes if

1. the genes forms a max-gap chain in each genome
2. the cluster is maximal (i.e. not contained within a larger cluster)

The max-gap definition is the most widely used cluster definition in genomic analyses

- Allows extensive rearrangement of gene order
- Allows limited gene insertion and losses

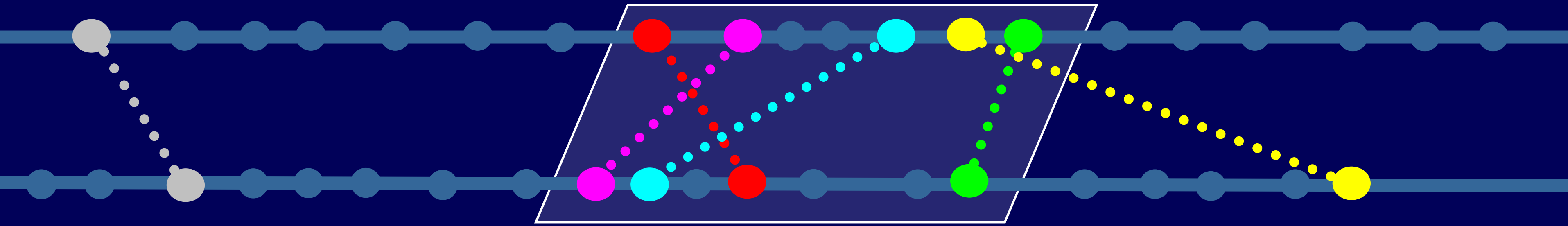
There is no formal statistical model for max-gap clusters

Outline

- Evolution of genome organization
- Why find related genomic regions?
- How do we find them?
  - Identification: max-gap cluster definition
  - Validation: Testing cluster significance

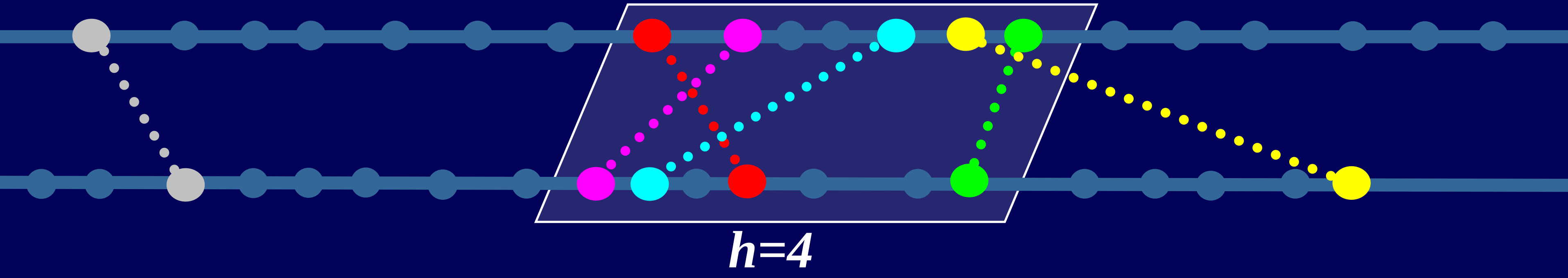
The Questions

Suppose two whole genomes were compared, and this max-gap cluster was identified:



- Is this cluster biologically meaningful?
- Could it have occurred in a comparison of random genomes?

The Inputs



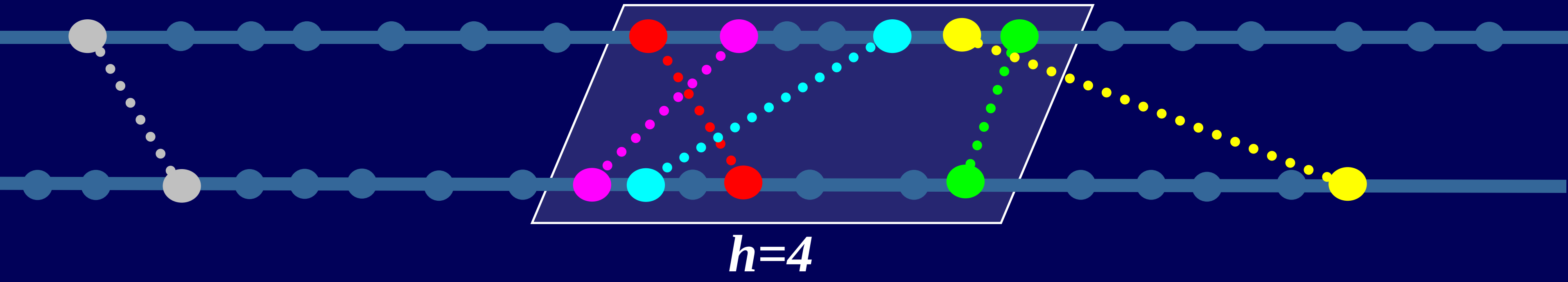
n: number of genes in each genome

m: number of matching genes pairs

g: the maximum gap allowed in a cluster

h: number of matching genes in the cluster

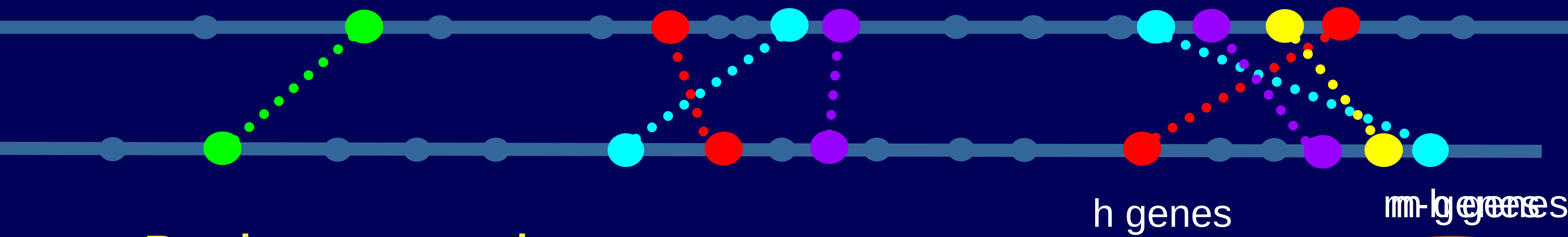
The Problem



What is the probability of observing a max-gap cluster

- containing exactly h matching gene pairs
- assuming the genomes are randomly ordered

Probability of a cluster of size h



Basic approach
Enumerate all ways to:

1. Create chains of h genes in both genomes * 2. Place m-h remaining genes so they do *not* extend the cluster

---

3. Normalize to get a probability

Probability of observing a cluster of size h

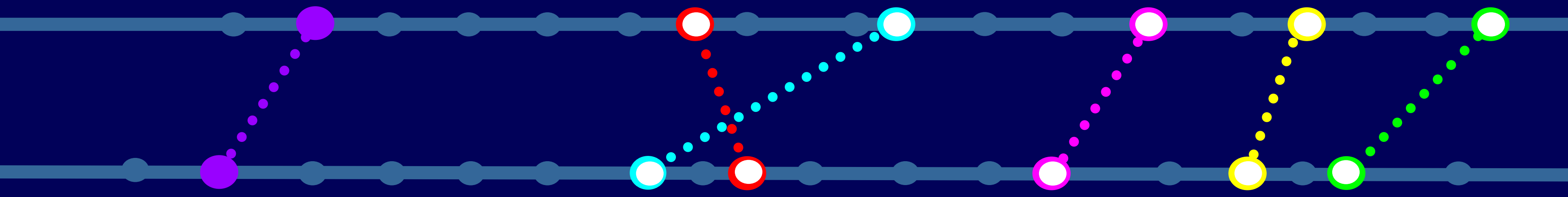
number of ways to place h genes so they form a chain in both genomes

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

$$\frac{F(h,g,n)\;G(m-h,g,n)}{\binom{n}{m}^2 m!^2}$$

All configurations of m gene pairs in two genomes of size n

Total number of configurations of m gene pairs in two genomes of size n



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

m genes

Probability of observing a cluster of size h

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

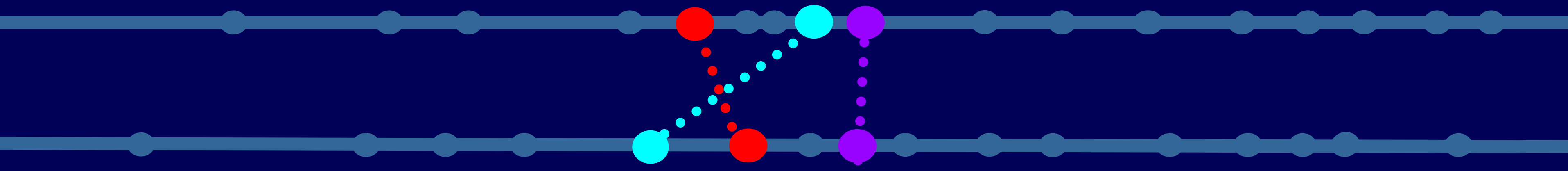
number of ways to place m-h remaining genes so they do *not* extend the cluster

$$\frac{\boxed{F(h,g,n)}\;G(m-h,g,n)}{\binom{n}{m}^{2} m!^{2}}$$

All configurations of m gene pairs in two genomes of size n

h genes m-h genes m genes

Number of ways to place h genes in two genomes so they form a cluster



$$\underbrace{\binom{m}{h}}\ \underbrace{\left[n-L+1+\frac{L-h}{2}\right]\cdot(g+1)^{h-1}}\ \underbrace{h!^2}$$

Choose *h* genes to compose the cluster

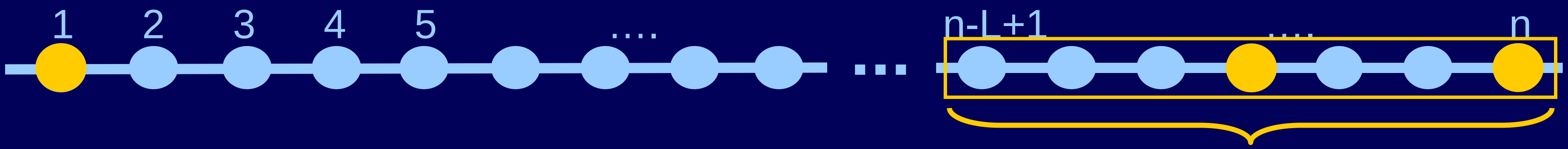
Select *h* spots in each genome, *so they form a max-gap chain*

Assign each gene to a selected spot in each genome

The number of ways to create a chain of h genes

$$F(h,g,n) = \underline{(n - L + 1)}(g+1)^{h-1} + E$$

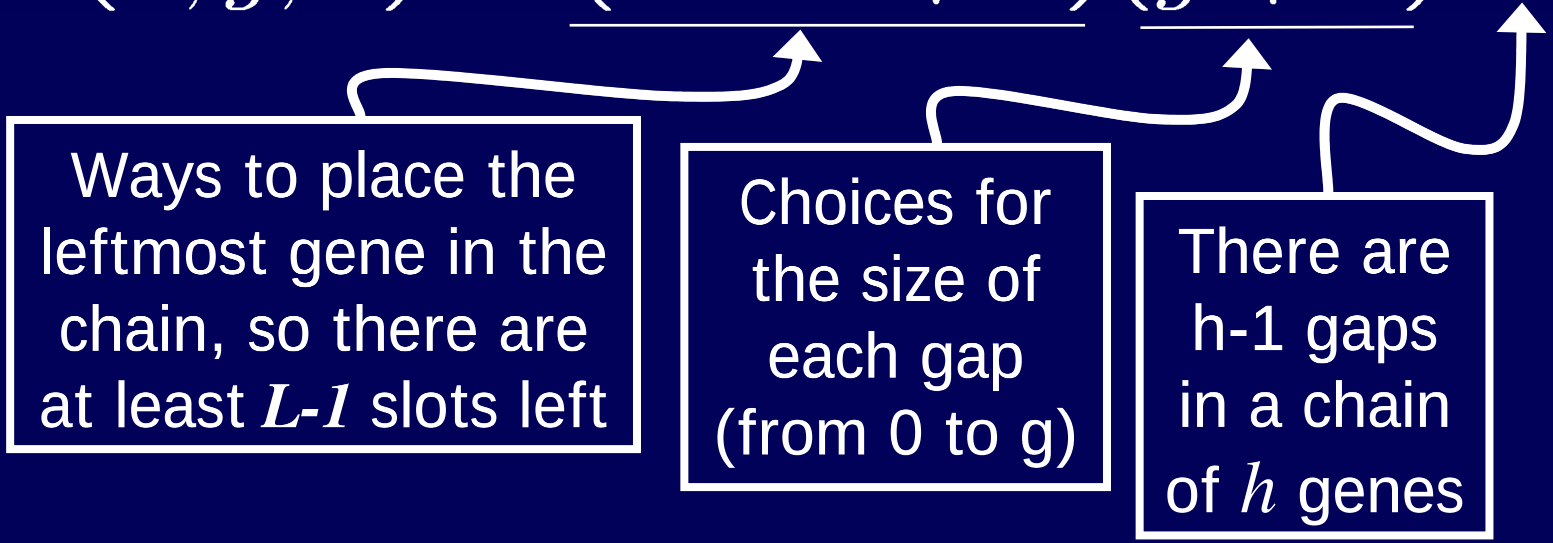
Ways to place the leftmost gene in the chain, so there are at least *L-1* places left



The maximum length of the chain is: ***L* = (*h*-1)*g* + *h***

The number of ways to create a chain of h genes

$$F(h,g,n) = \underline{(n-L+1)}\,\underline{(g+1)}^{\underline{h-1}} + E$$



Ways to place the leftmost gene in the chain, so there are at least *L-1* slots left

Choices for the size of each gap (from 0 to g)

There are h-1 gaps in a chain of h genes

The number of ways to create a chain of h genes

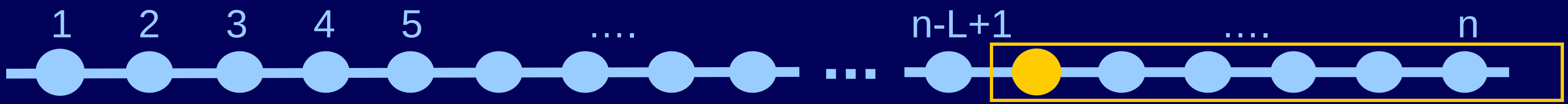
$$F(h,g,n) = \underline{(n-L+1)}\,\underline{(g+1)}^{\underline{h-1}} + E$$

Ways to place the leftmost gene in the chain, so there are at least *L-1* slots left

Choices for the size of each gap (from 0 to g)

There are h-1 gaps in a chain of h genes

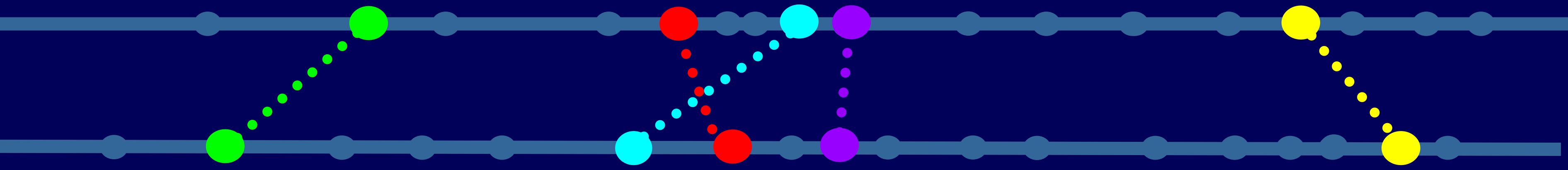
Chains near the end of the genome

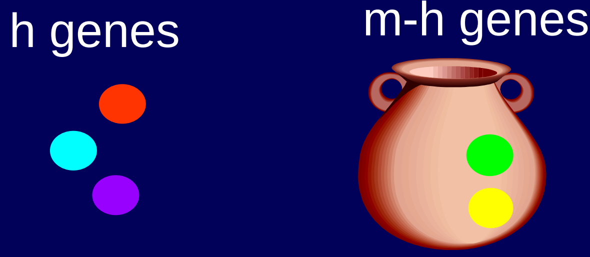


Number of ways to position *h* genes in a genome of n genes so they form a max-gap chain

$$F(h,g,n) = \left[\underbrace{n - L + 1}_{\text{Starting positions}} + \underbrace{\frac{L-h}{2}}_{\text{Starting positions near end}}\right] \cdot \underbrace{(g+1)^{h-1}}_{\text{Ways to place remaining h-1 genes}}$$

Probability of a cluster of size h





Basic approach
Enumerate all ways to:

1. Create chains of h genes in both genomes

*

2. Place m-h remaining genes so they do *not* extend the cluster

Probability of observing a cluster of size h

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

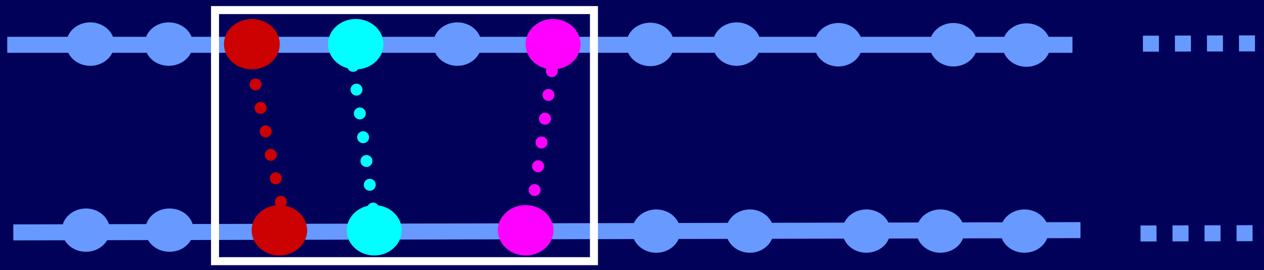
number of ways to place m-h remaining genes so they do *not* extend the cluster

$$\frac{F(h,g,n)\;\boxed{G(m-h,g,n)}}{\binom{n}{m}^2 m!^2}$$

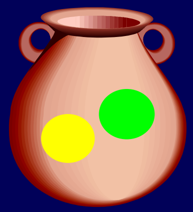
All configurations of m gene pairs in two genomes of size n

Counting the number of ways to place m-h genes outside the cluster

$g = 1$



$h = 3$

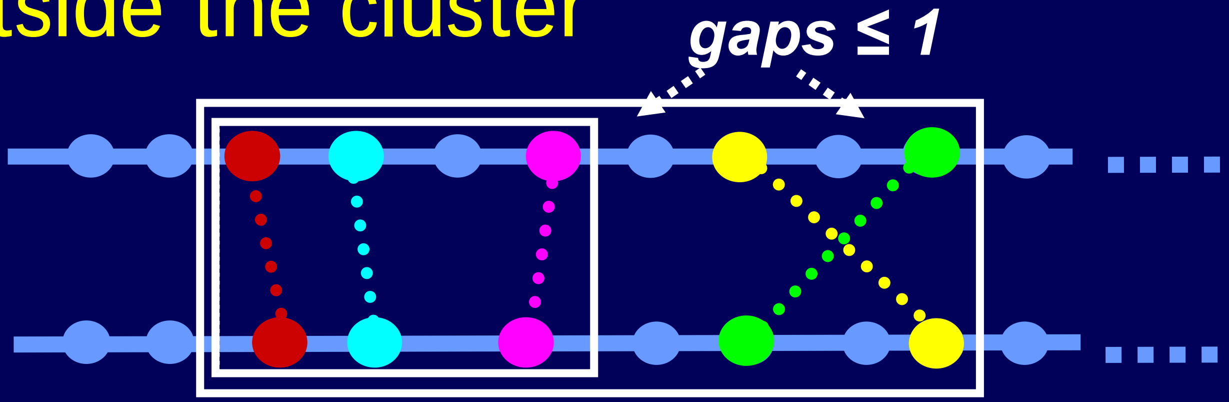


Approach:

- design a rule specifying *where* the genes can be placed so that the cluster is not extended
- count the positions

Counting the number of ways to place m-h genes outside the cluster

$g = 1$

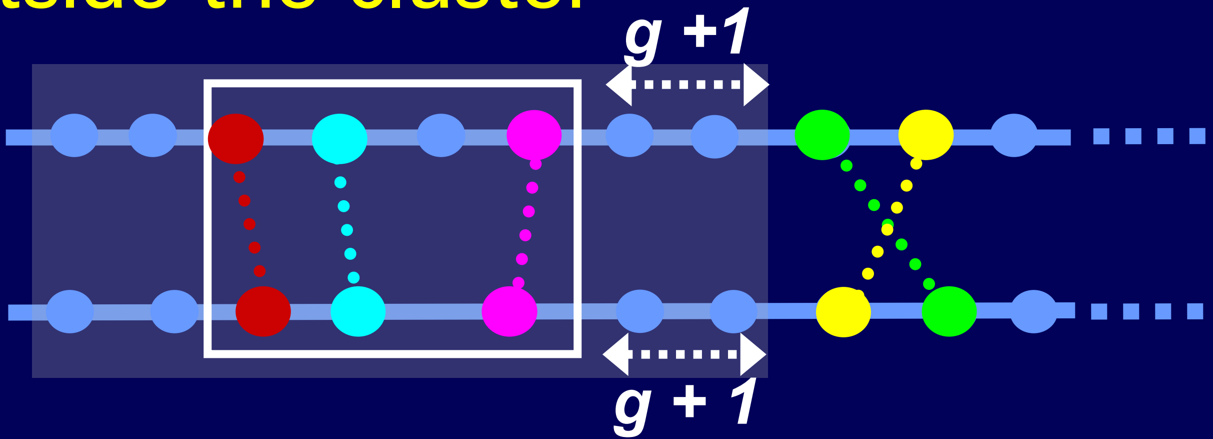


Rule 1: A gene can go anywhere except in the cluster (the white box).

Too lenient

Counting the number of ways to place m-h genes outside the cluster

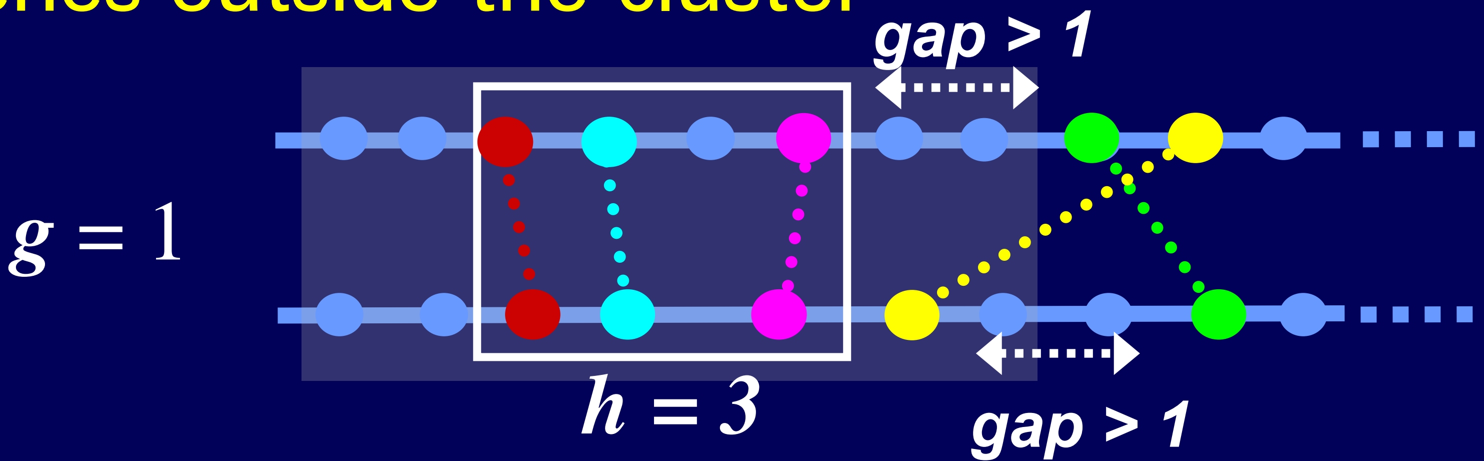
$g = 1$



Rule 2: Every gene must be at least g+1 positions from the cluster (outside the grey box).

Too strict

Counting the number of ways to place m-h genes outside the cluster

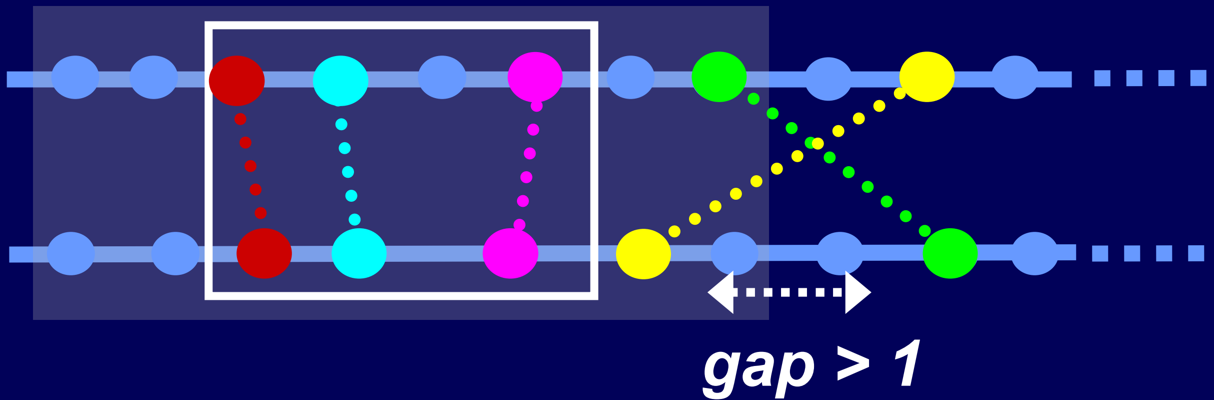


Rule 2: Every gene must be at least g+1 positions from the cluster (outside the grey box).

Too strict

Counting the number of ways to place m-h genes outside the cluster

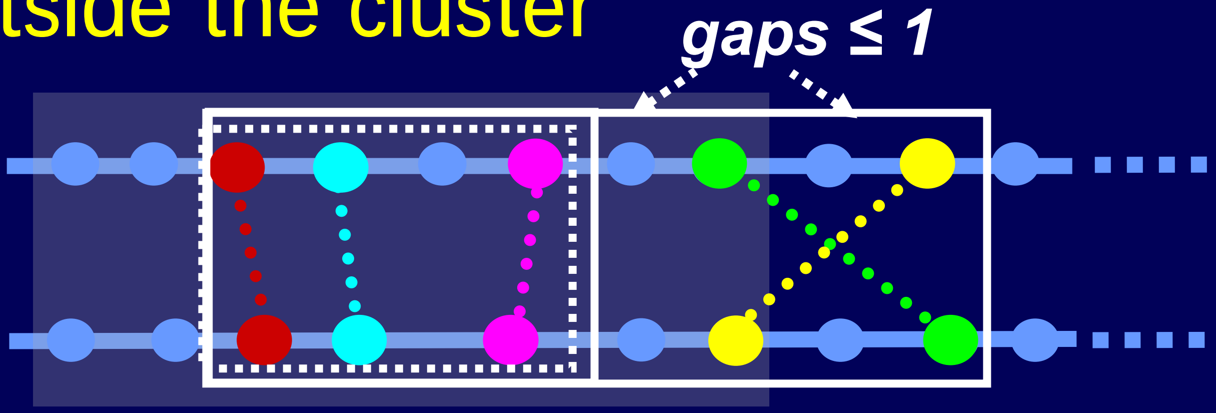
$g = 1$



Rule 3: At most one member of each gene pair can be in the grey box.

Too lenient

Counting the number of ways to place m-h genes outside the cluster



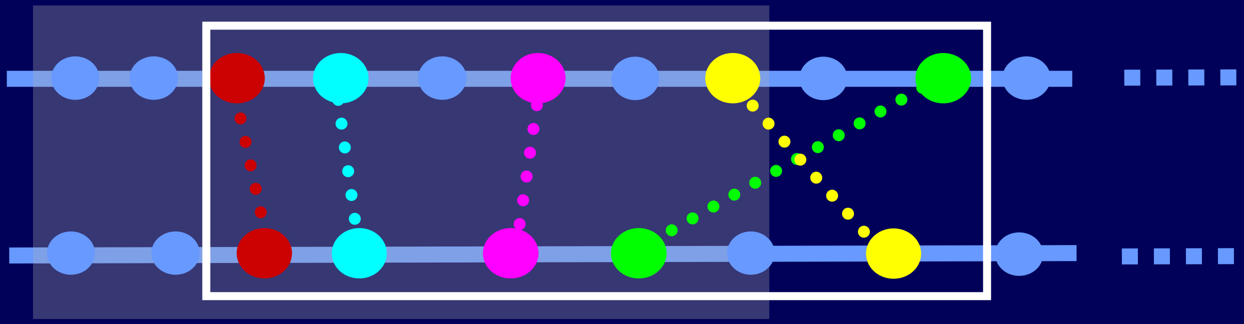
$g = 1$

Rule 3: At most one member of each gene pair can be in the grey box.

Too lenient

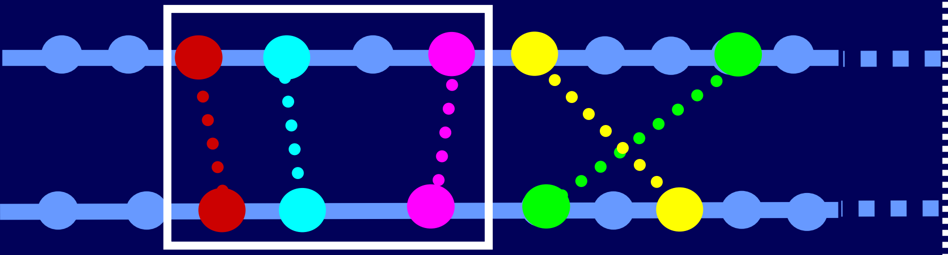
Counting the number of ways to place m-h genes outside the cluster

$g = 1$



- Acceptable positions for a gene depend on the positions of the remaining genes
- Use strict and lenient rules to calculate upper and lower bounds on G

Estimating G

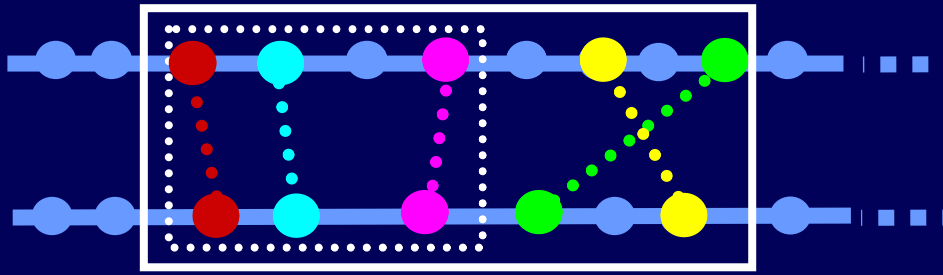


⬆ Lower bound:

Fails to enumerate this configuration

⬇ Upper bound:

Erroneously enumerates this configuration



Probability of observing a cluster of size h

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

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

$$\frac{F(h,g,n)\;G(m-h,g,n)}{\binom{n}{m}^{2} m!^{2}}$$

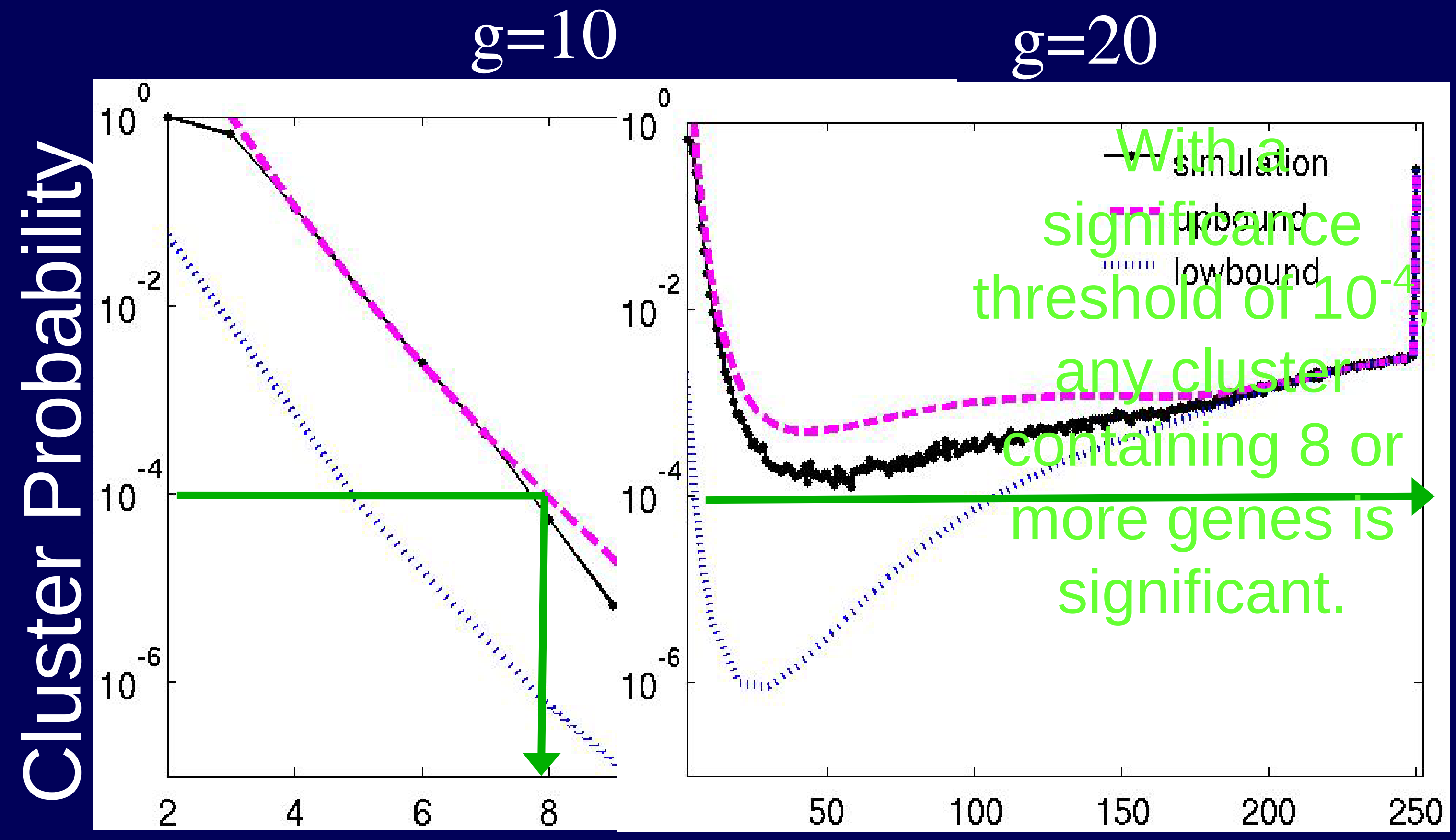
Hoberman, Sankoff, Durand
Journal of Computational Biology, 2005

What can we learn from this statistical result?

- Are we less likely to observe a large cluster (containing more gene pairs) than a small cluster?
- How large does a cluster have to be before we are surprised to observe it?
- How do we choose the maximum allowed gap value? Larger values will
  - yield more clusters
  - more of these will be false positives

Whole-genome comparison cluster statistics

n=1000, m=250



With a significance threshold of 10^{-4}, any cluster containing 8 or more genes is significant.

h (cluster size)

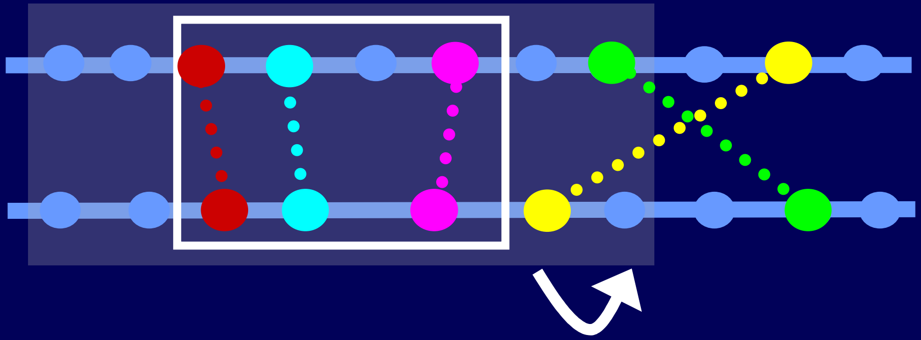
Conclusion

Statistical analysis of max-gap gene clusters

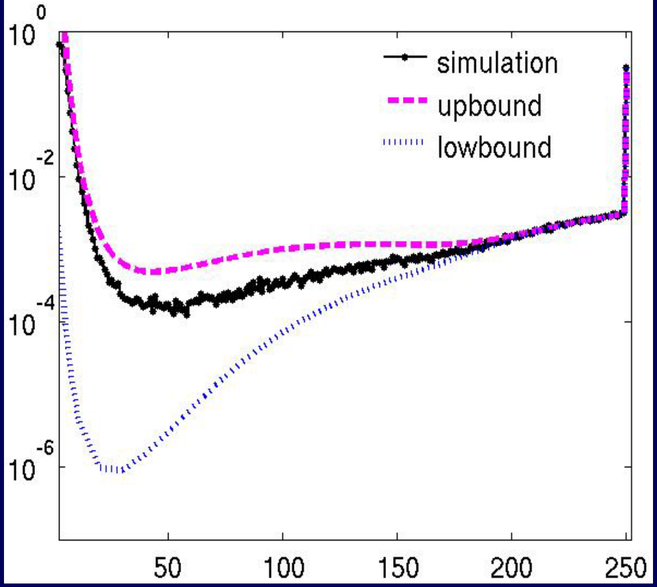
1. Provides a principled approach for choosing a gap size that will yield significant clusters
2. Allows statistically significant max-gap clusters to be identified
3. Provides insight on criteria for cluster definitions

Odd properties of max-gap clusters

1. Moving a gene further away may make a cluster more likely



2. A larger cluster may be less significant



Acknowledgements

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

Thanks

Questions?