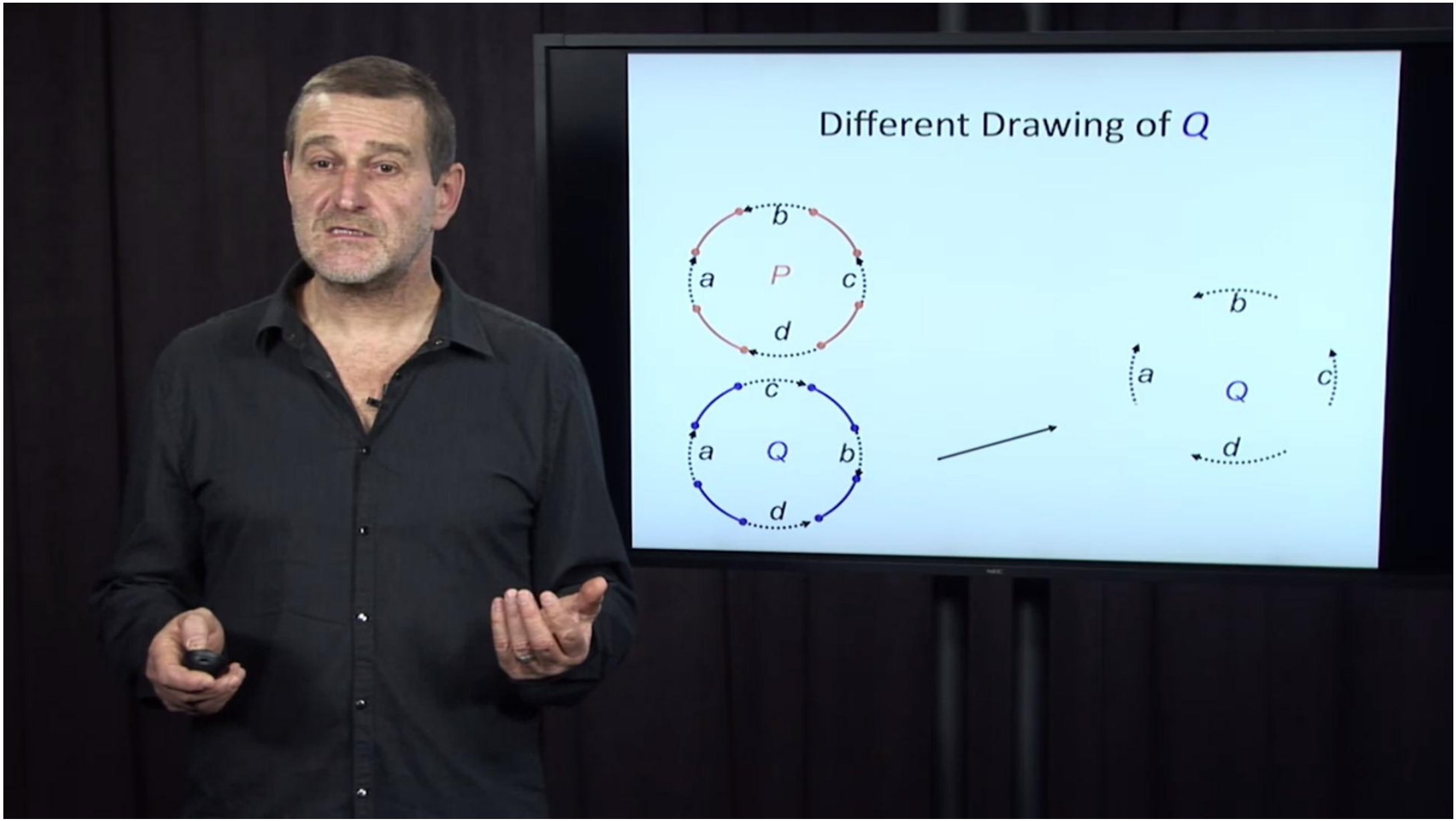
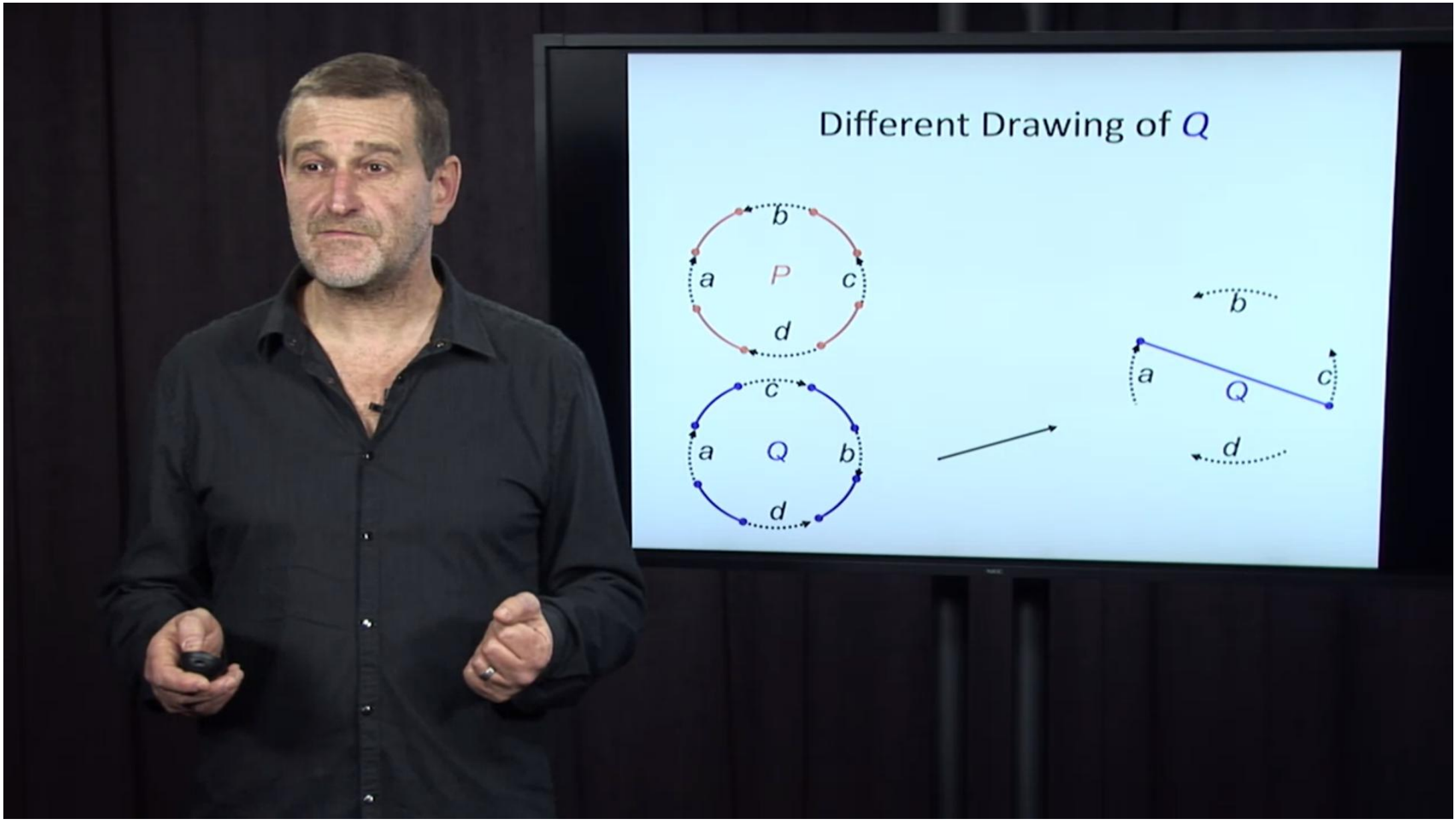
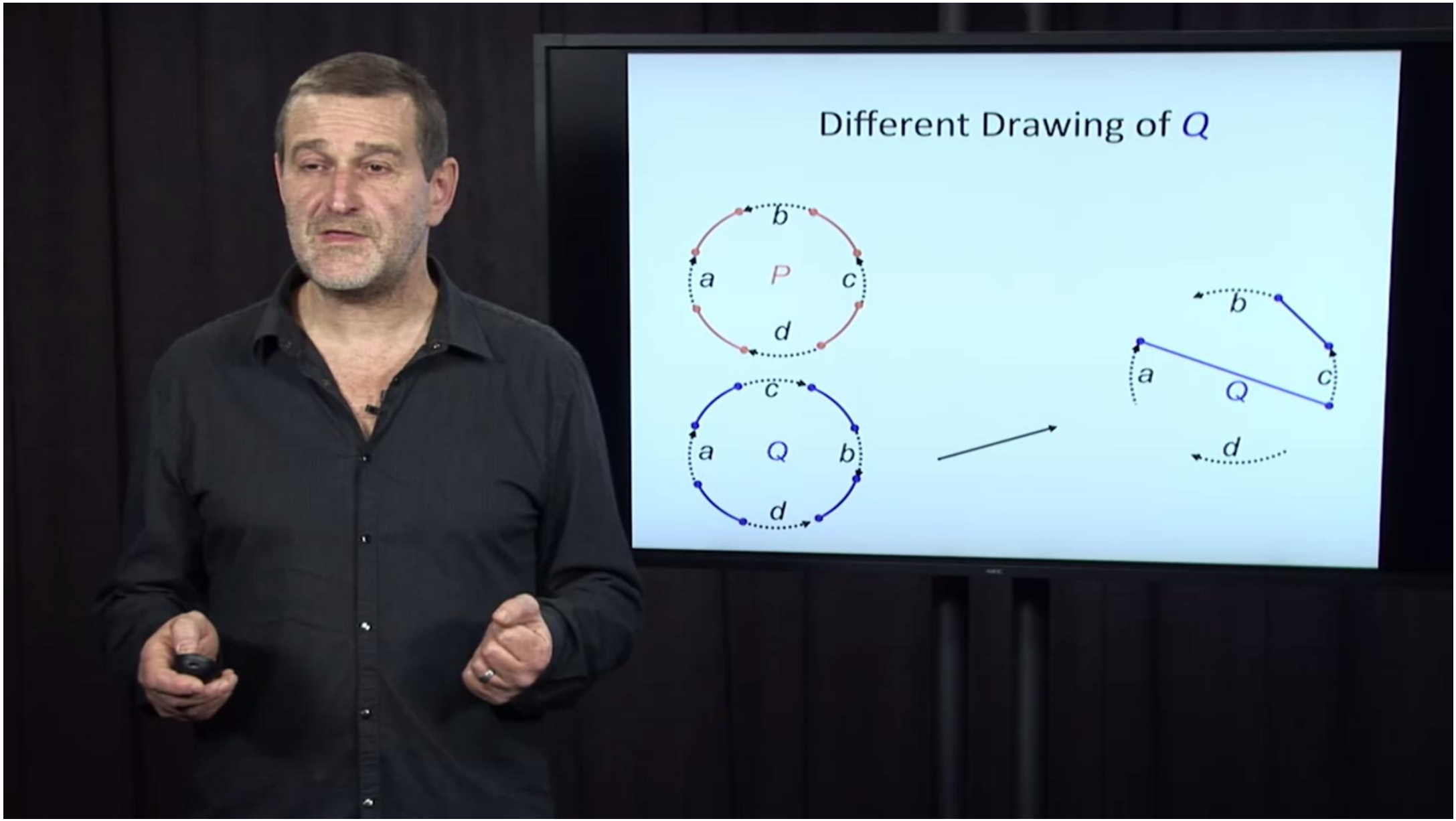
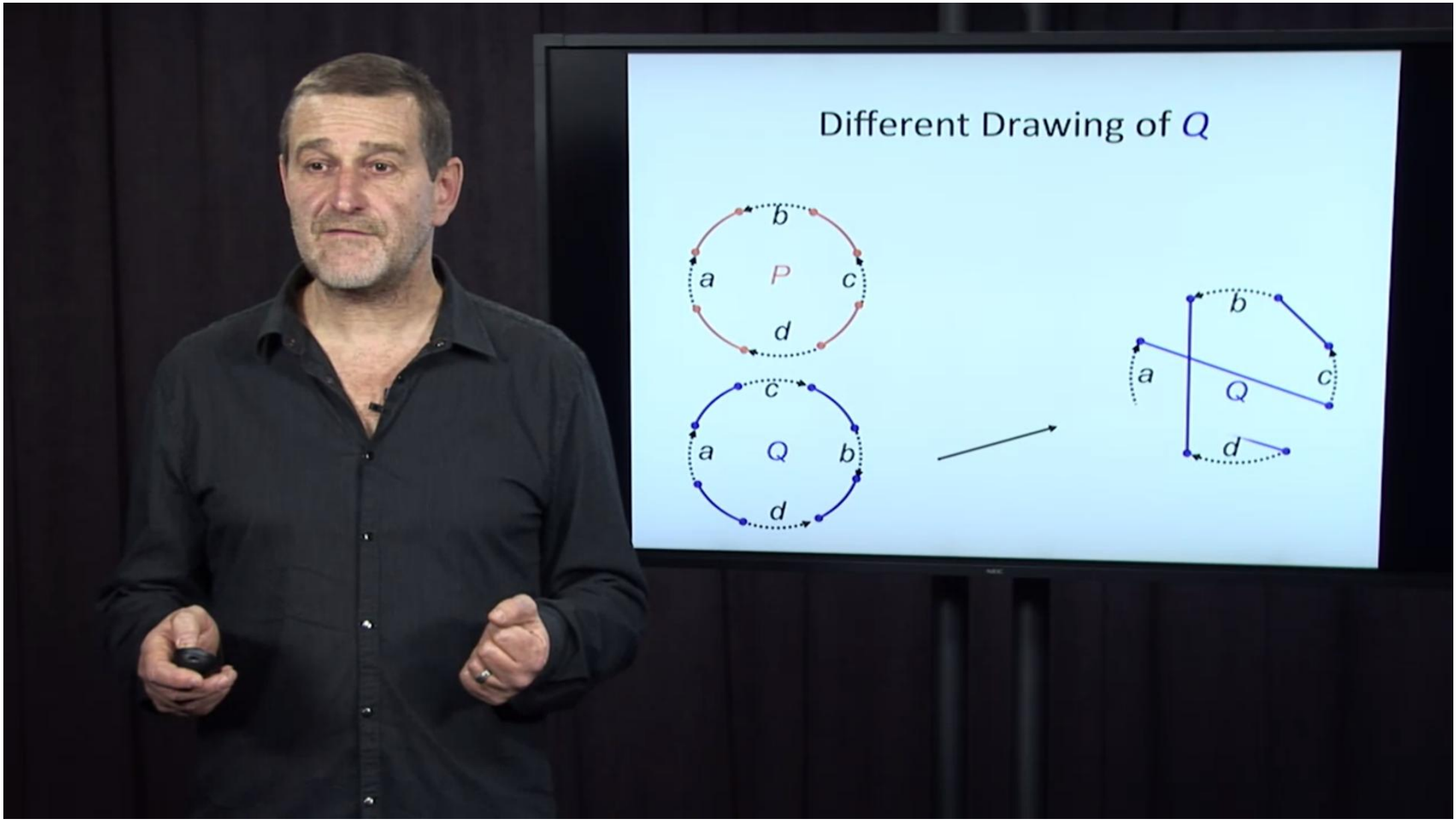
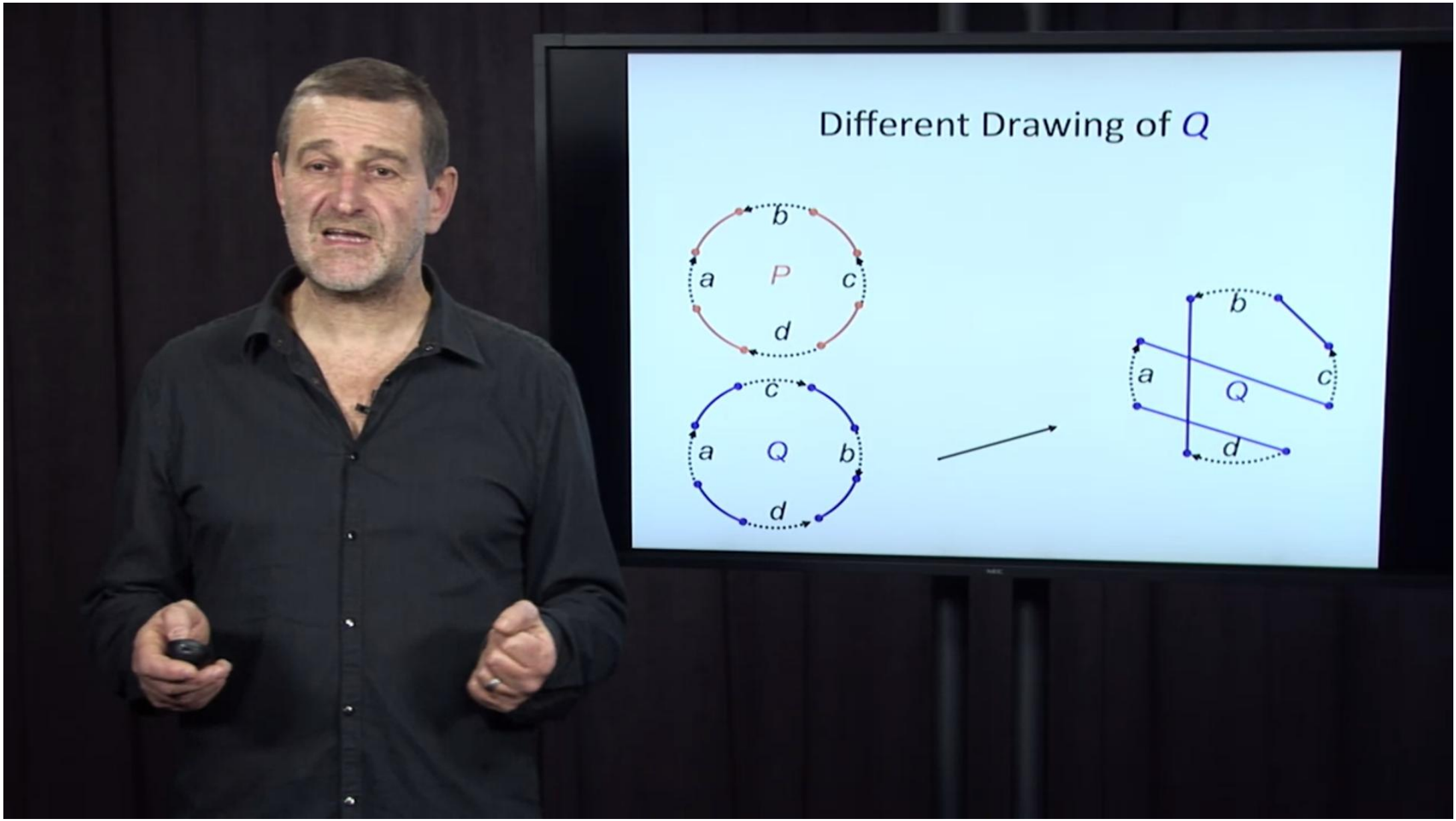

Comparing Genomes P and Q
b
a
P
c
d
c
a
Q
b
d

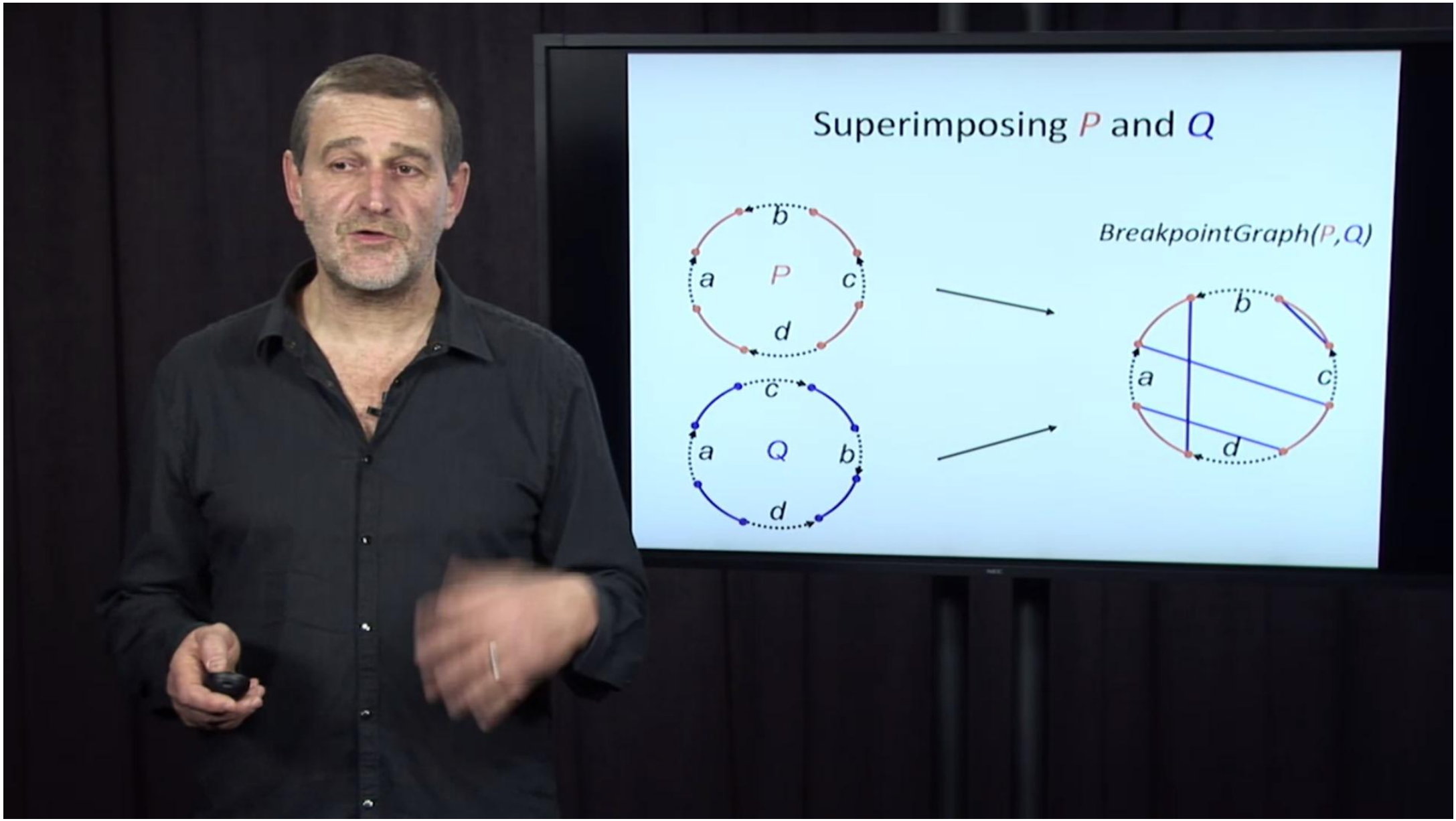

Different Drawing of Q
b
a
P
c
d
c
a
Q
b
d
b
a
Q
c
d


Different Drawing of Q
b
a
P
c
d
c
a
Q
b
d
b
a
Q
c
d

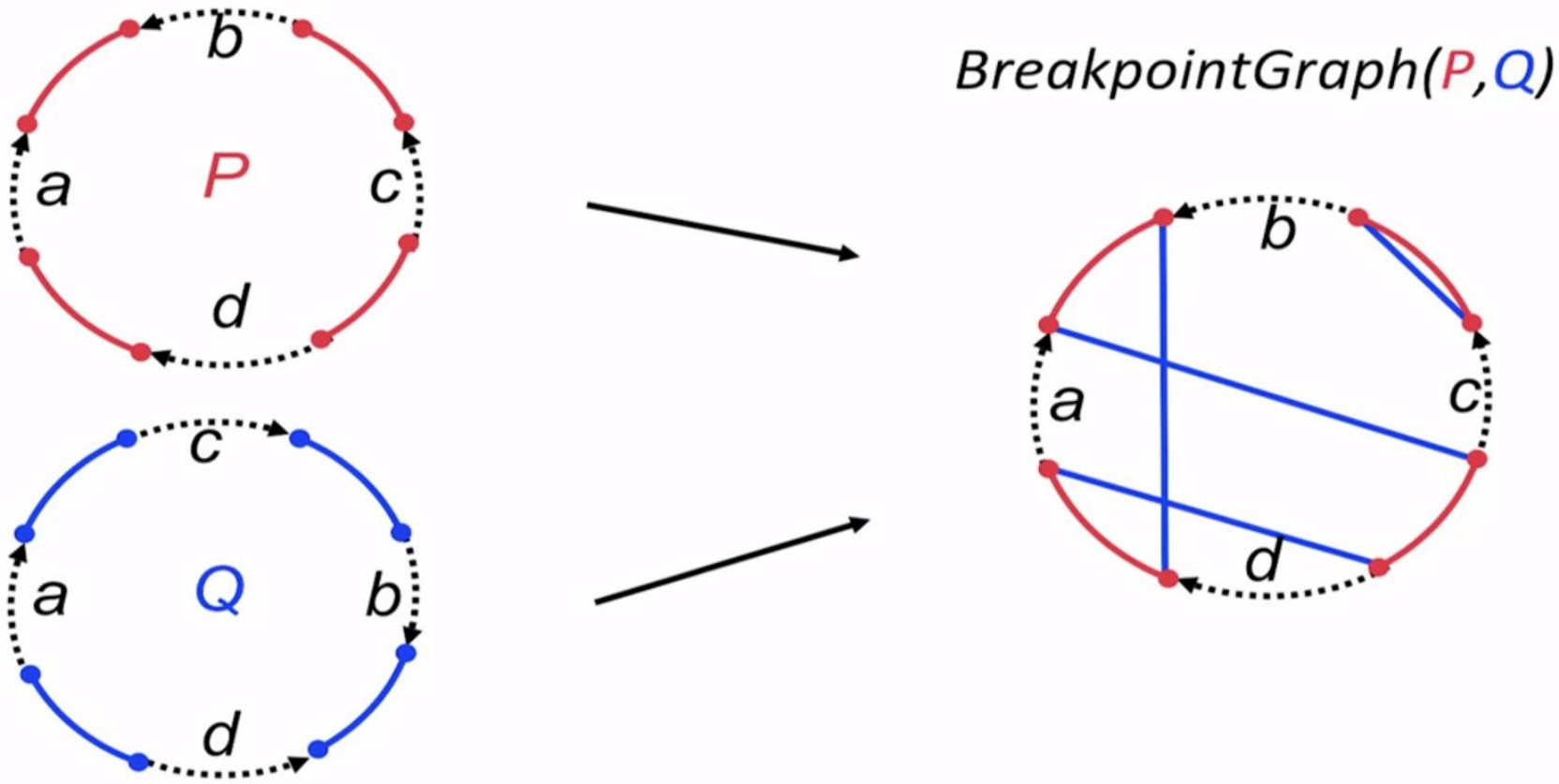

Different Drawing of Q
b
a
P
c
d
c
a
Q
b
d
b
a
Q
c
d


Different Drawing of Q
b
a
P
c
d
c
a
Q
b
d
b
a
c
Q
d

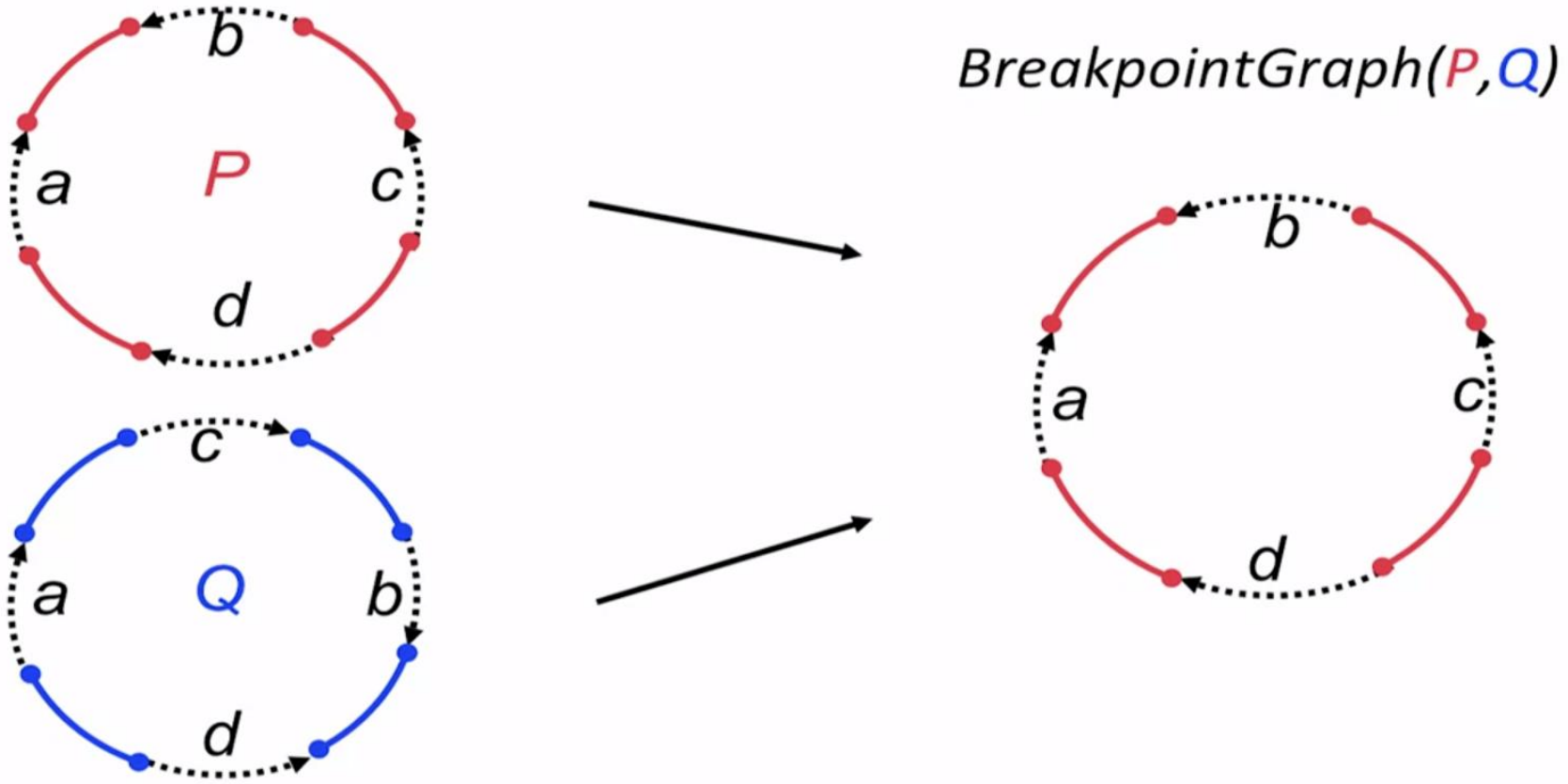

Different Drawing of Q
b
a
P
c
d
c
a
Q
b
d
b
a
Q
c
d


Superimposing P and Q
b
a
P
c
d
c
a
Q
b
d
BreakpointGraph(P,Q)
b
a
c
d

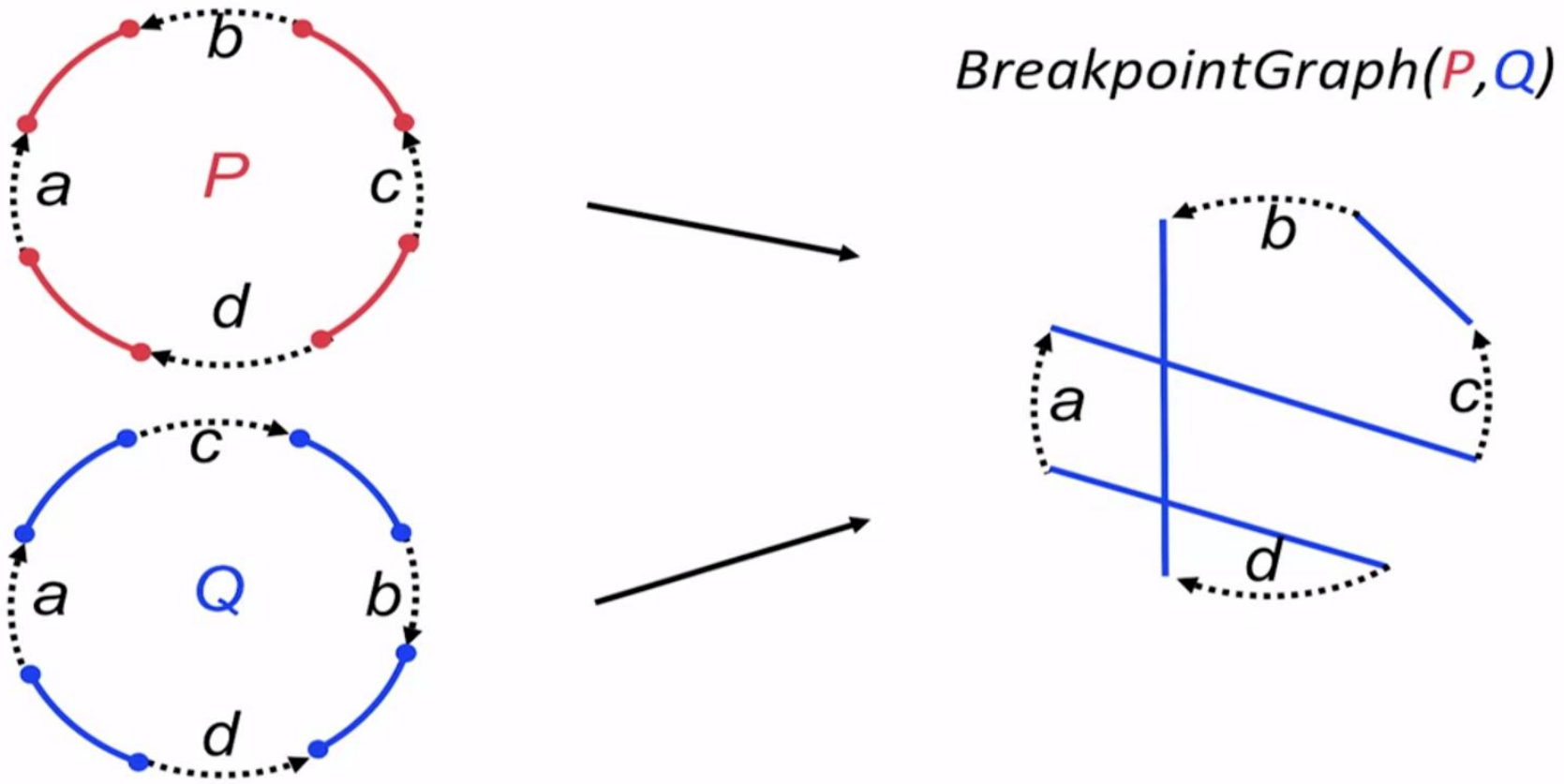
Superimposing *P* and *Q*



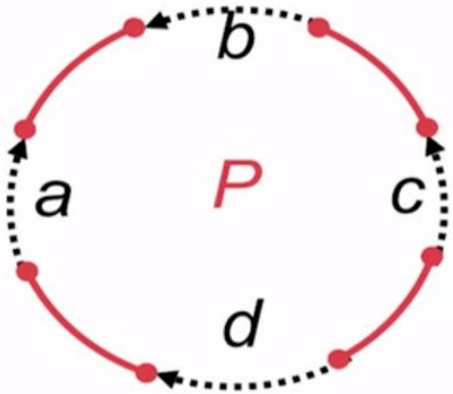
Red and Black Edges in Breakpoint Graph Form..

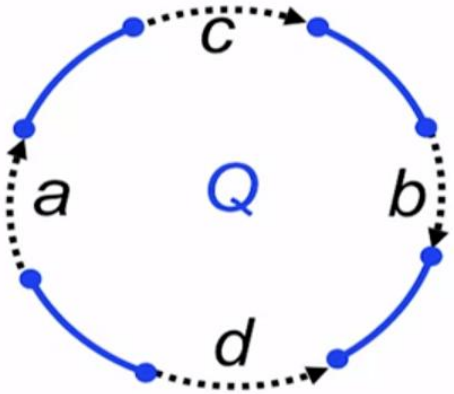


Blue and Black Edges in Breakpoint Graph Form…

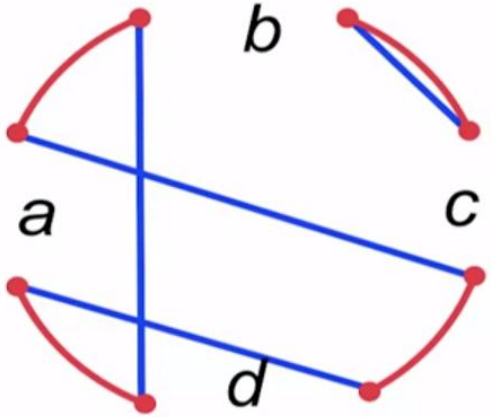


What About Red and Blue Edges???

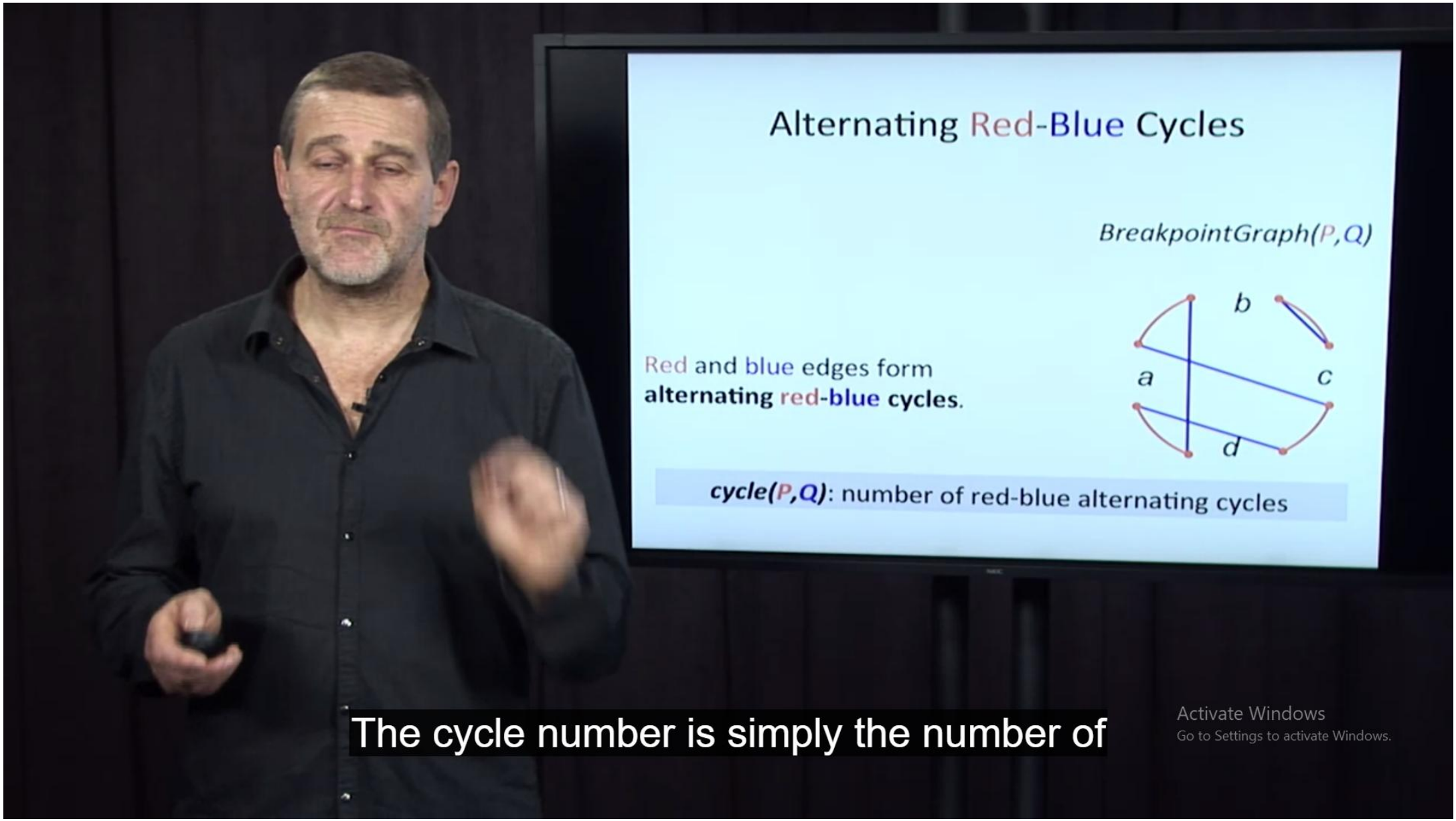


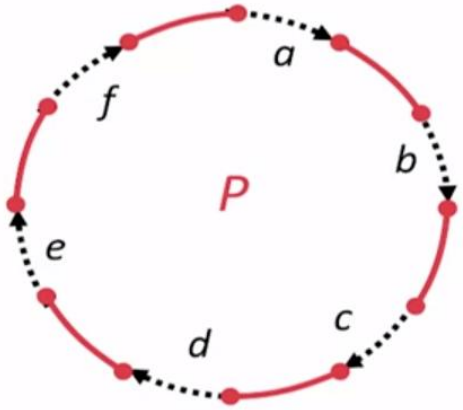


BreakpointGraph(P,Q)

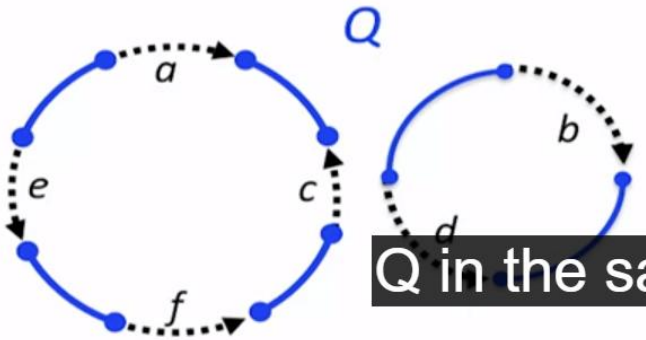



Alternating Red-Blue Cycles
BreakpointGraph(P,Q)
Red and blue edges form
alternating red-blue cycles.
b
a
c
d
Why?

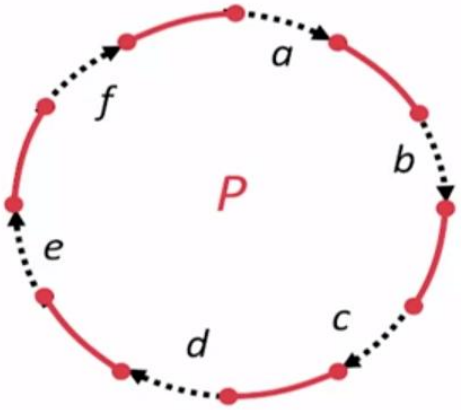




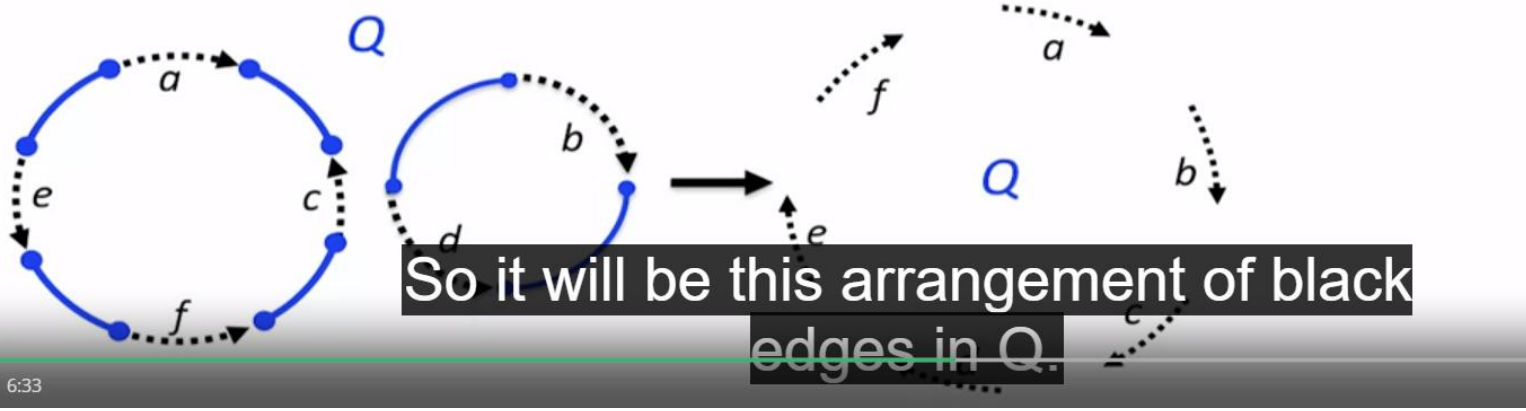
What is the cycle number *cycle(P,Q)*?



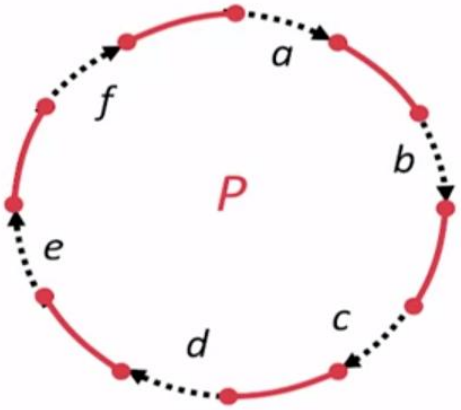
Q in the same order they are present in genome P.



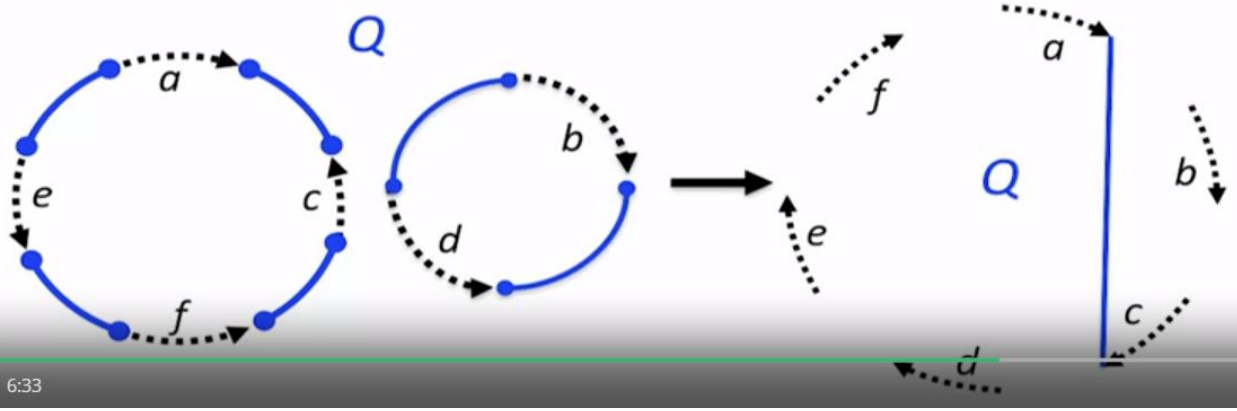
Arranging blacks edges of *Q*
in the same order as black edges in *P*

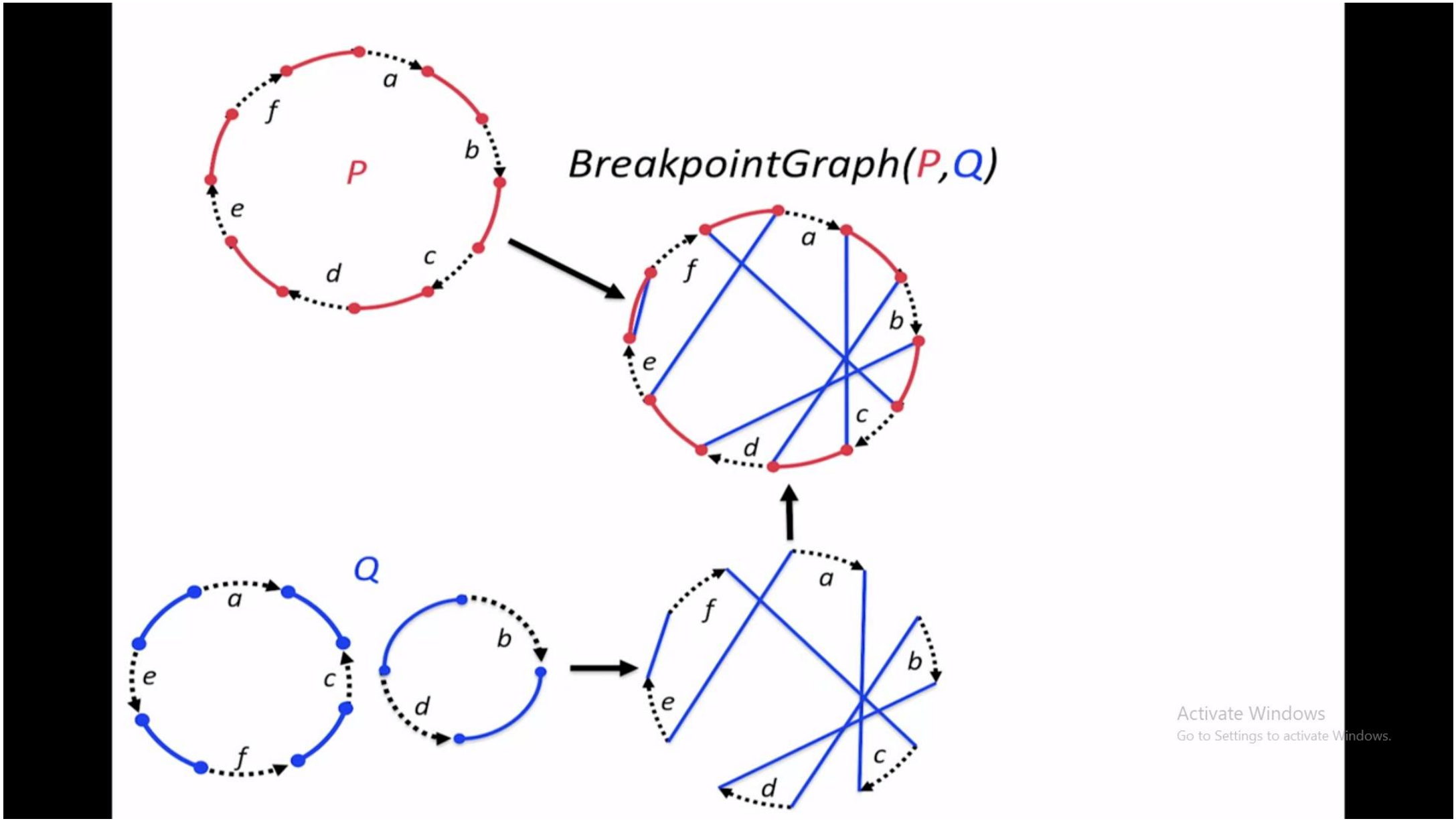


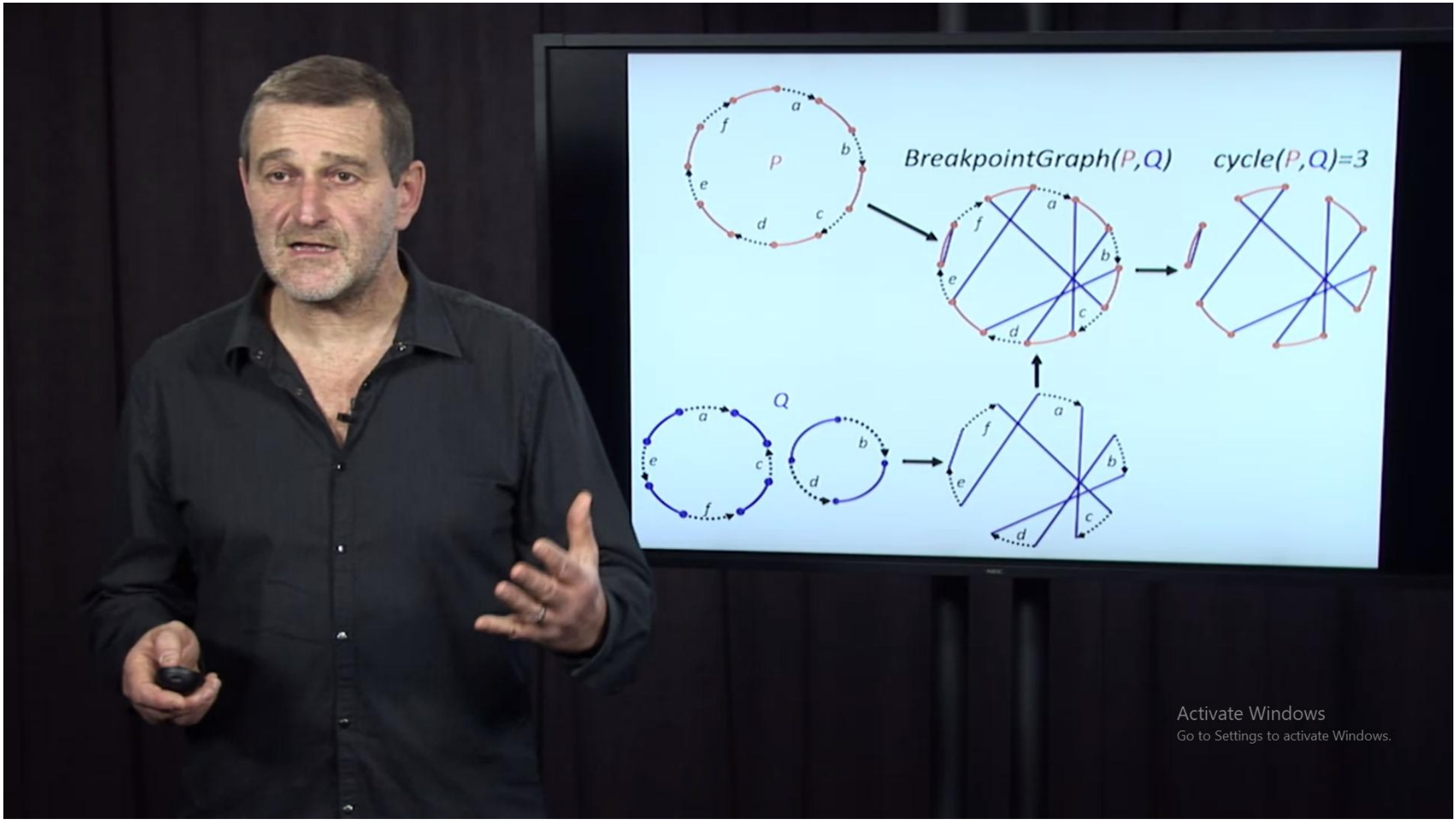
So it will be this arrangement of black edges in Q.

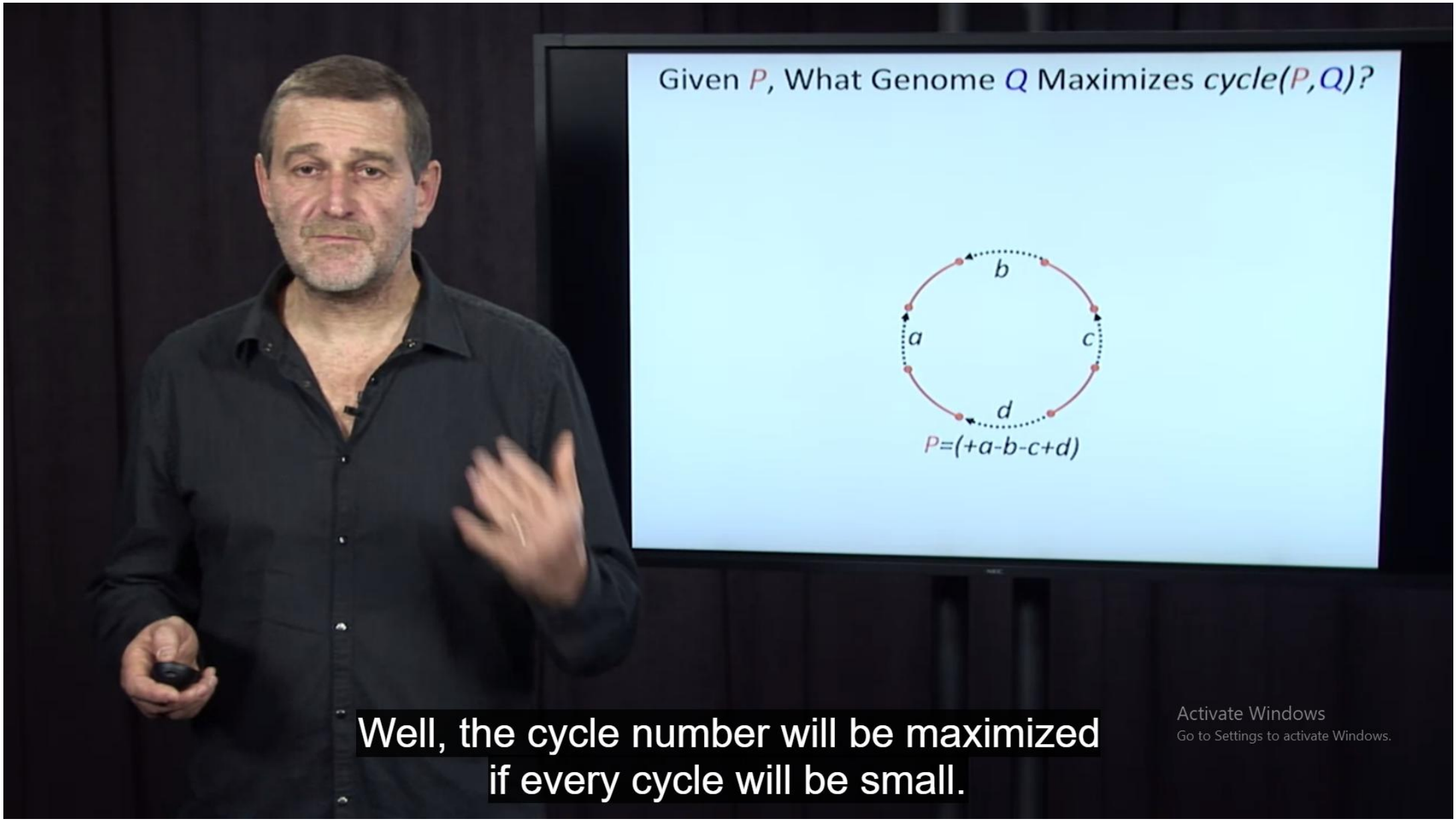


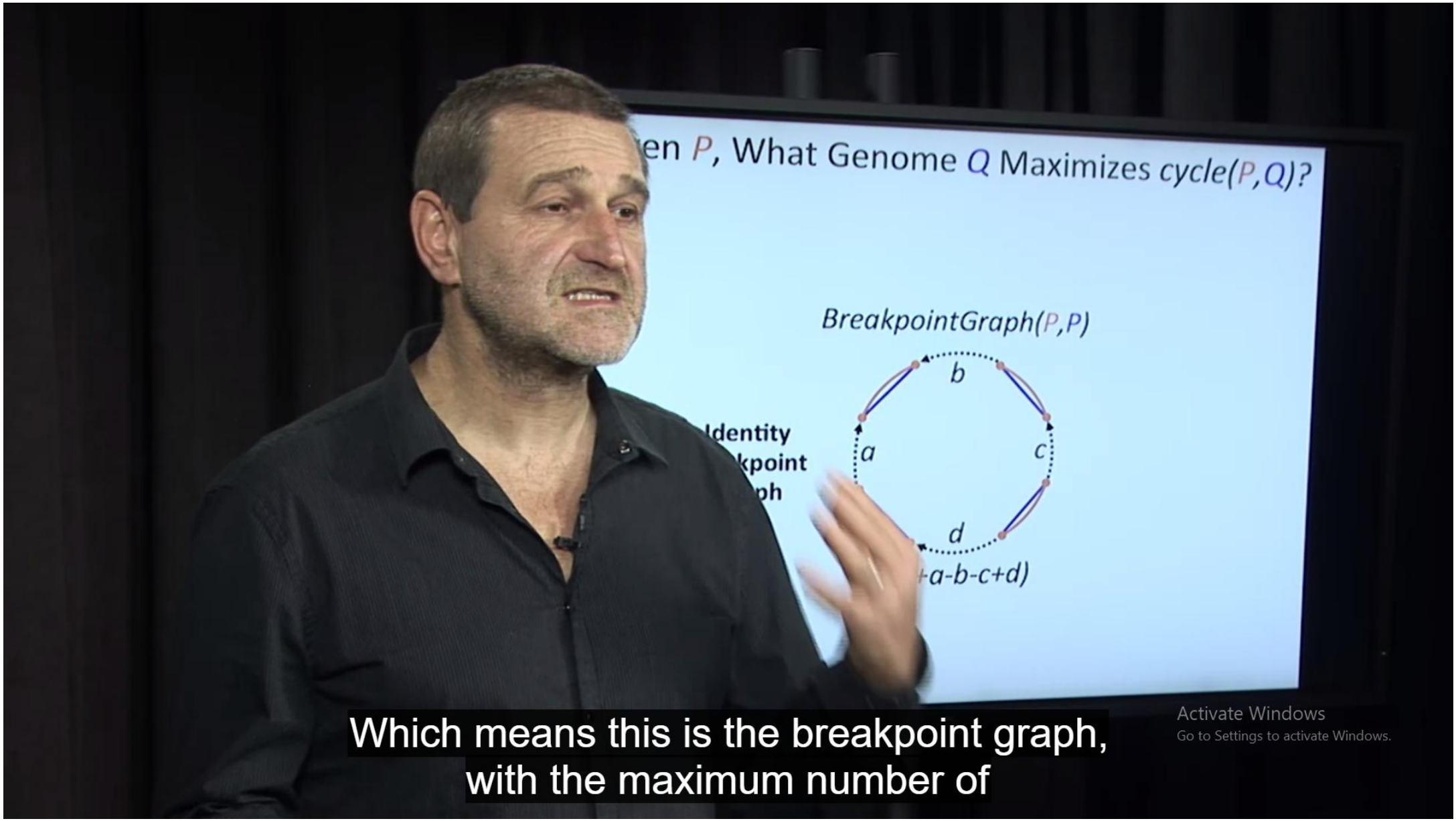
Arranging blacks edges of *Q*
in the same order as black edges in *P*





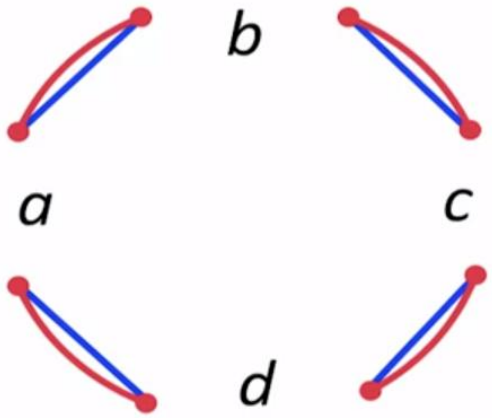



Given P, What Genome Q Maximizes cycle(P,Q)?
b
a
c
d
P=(+a-b-c+d)
Well, the cycle number will be maximized
if every cycle will be small.
Activate Windows
Go to Settings to activate Windows.


en P, What Genome Q Maximizes cycle(P,Q)?
BreakpointGraph(P,P)
b
a
c
d
Identity
kpoint
ph
-a-b-c+d)
Which means this is the breakpoint graph,
with the maximum number of
Activate Windows
Go to Settings to activate Windows.

Given *P*, What Genome *Q* Maximizes *cycle(P,Q)*?

BreakpointGraph(P,P)



Identity Breakpoint Graph

cycle(P,P)=#blocks=4

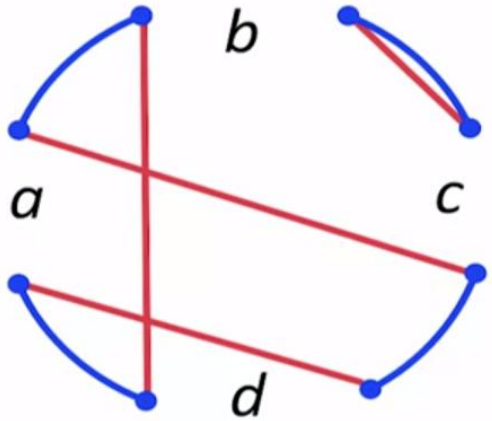
P=(+a-b-c+d)

of course simply the number of blocks in genome P.

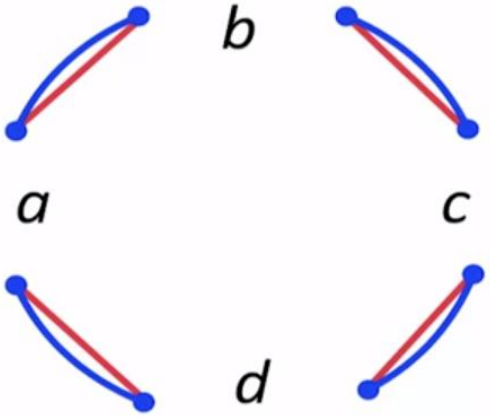
Genome Rearrangements Affect Red-Blue Cycles

Each transformation $P \rightarrow Q$ corresponds to a transformation:

BreakpointGraph(P,Q)



BreakpointGraph(Q,Q)

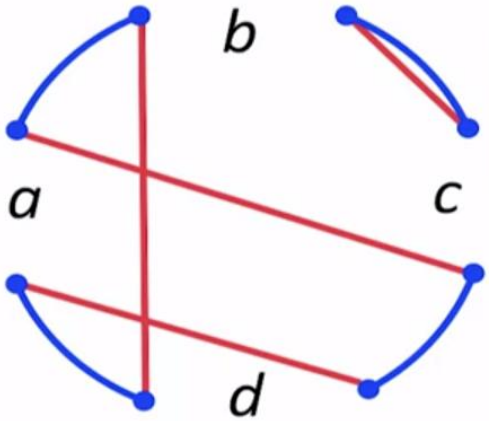


So it's important to realize that genome rearrangements affect

Genome Rearrangements Affect Red-Blue Cycles

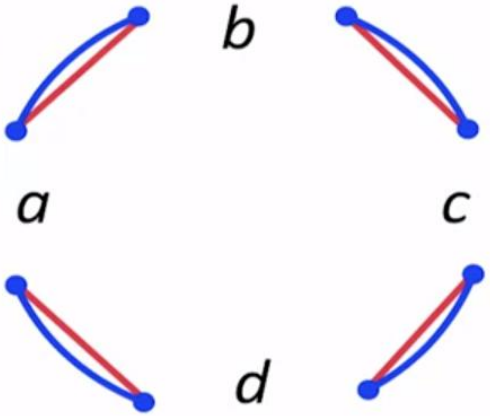
Each transformation $P \rightarrow Q$ corresponds to a transformation:

BreakpointGraph(P,Q)



Series of 2-breaks transforming *P* into *Q*

BreakpointGraph(Q,Q)

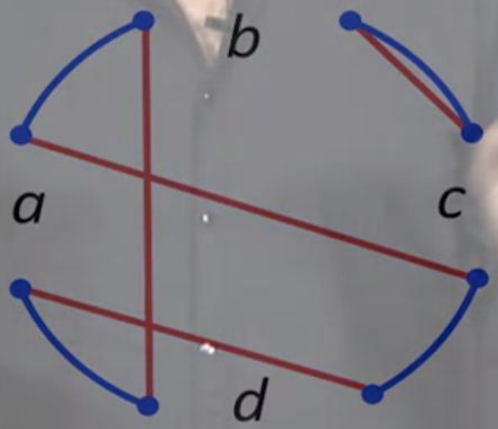


the fact that the series of 2-break transforming P into Q is unknown,

Genome Rearrangements Affect Red-Blue Cycles

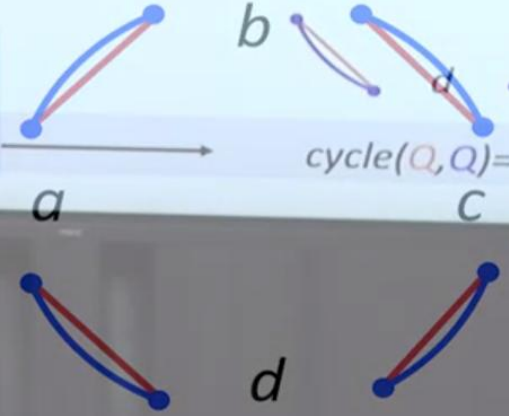
Each transformation $P \rightarrow Q$ corresponds to a transformation:

BreakpointGraph(P,Q)



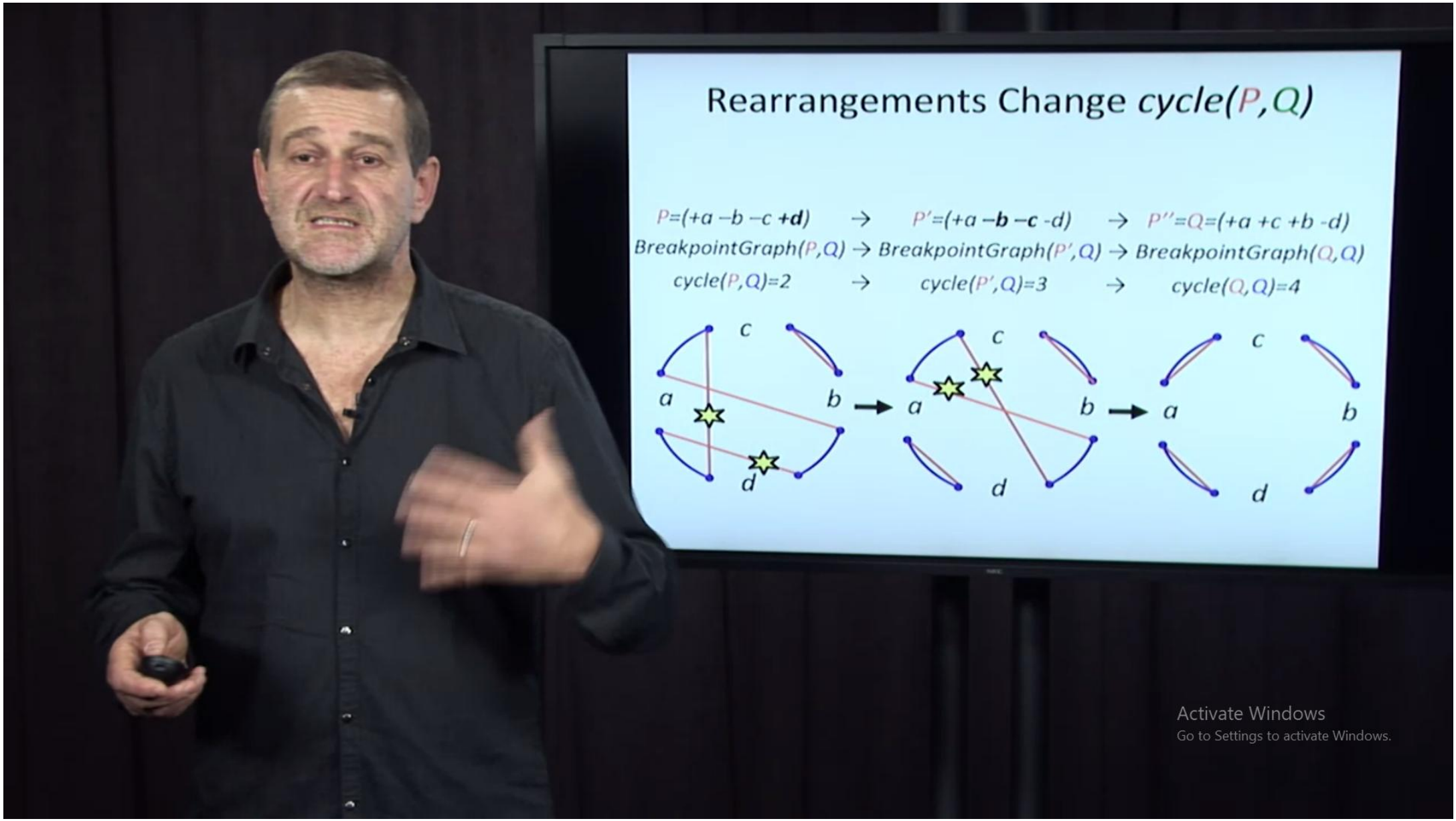
Series of 2-breaks transforming P into Q

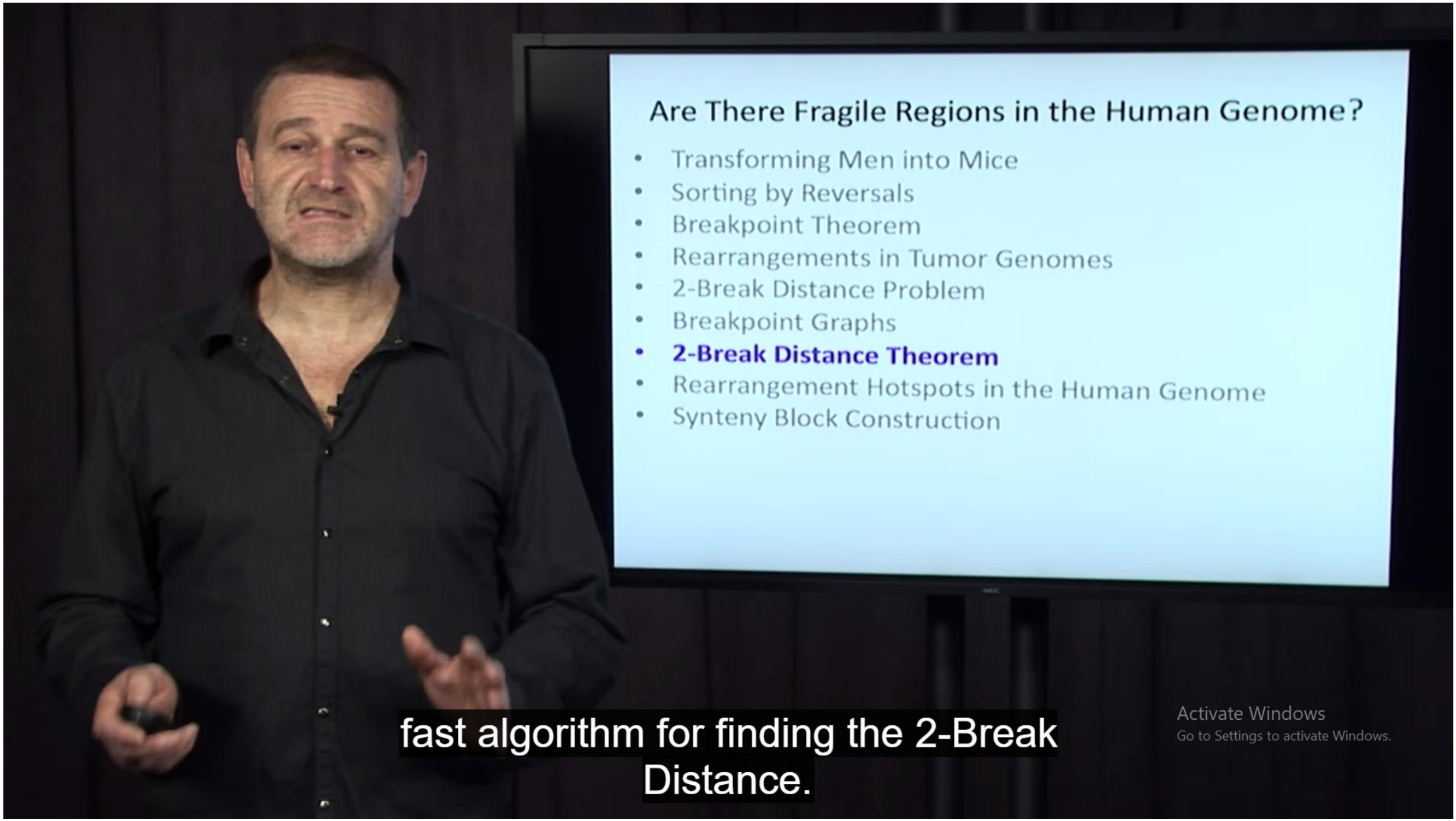
BreakpointGraph(Q,Q)

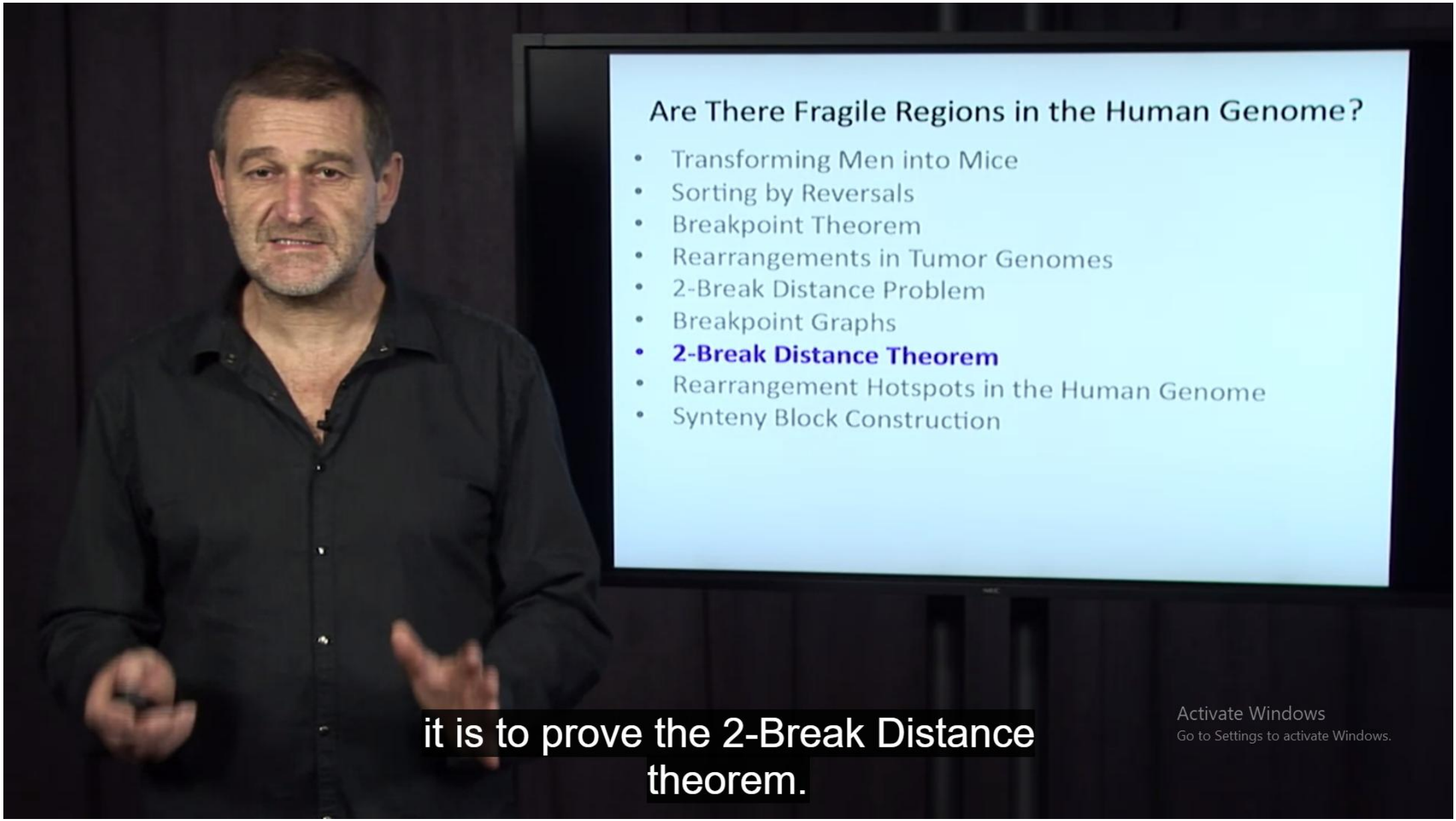


cycle(P,Q)=2

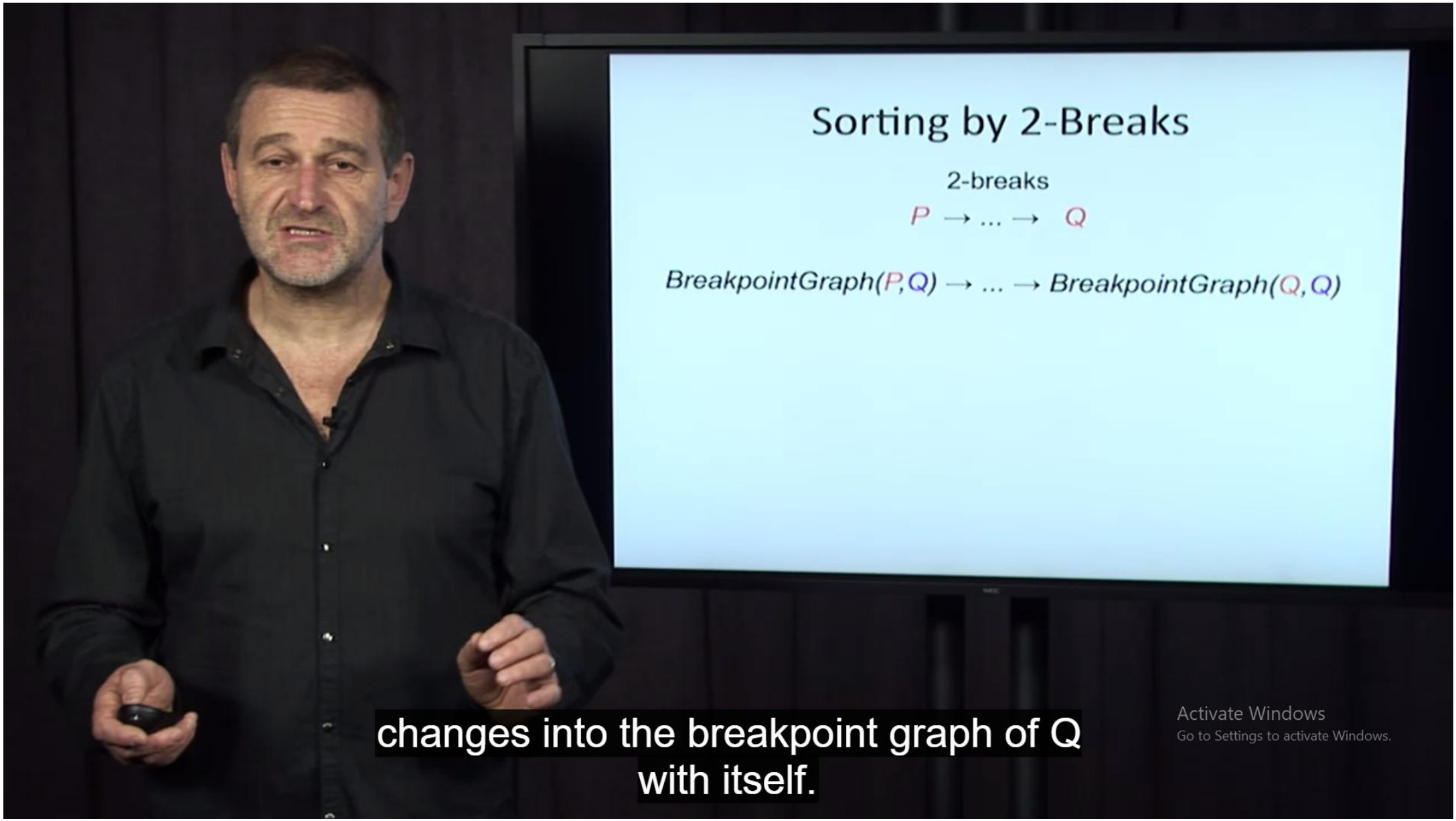
cycle(Q,Q)= 4


Rearrangements Change cycle(P,Q)
P=(+a −b −c +d) → P′=(+a −b −c -d) → P″=Q=(+a +c +b -d)
BreakpointGraph(P,Q) → BreakpointGraph(P′,Q) → BreakpointGraph(Q,Q)
cycle(P,Q)=2 → cycle(P′,Q)=3 → cycle(Q,Q)=4
c
a
b
d
c
a
b
d
c
a
b
d
Activate Windows
Go to Settings to activate Windows.


Are There Fragile Regions in the Human Genome?
• Transforming Men into Mice
• Sorting by Reversals
• Breakpoint Theorem
• Rearrangements in Tumor Genomes
• 2-Break Distance Problem
• Breakpoint Graphs
• 2-Break Distance Theorem
• Rearrangement Hotspots in the Human Genome
• Synteny Block Construction
fast algorithm for finding the 2-Break Distance.
Activate Windows
Go to Settings to activate Windows.


Are There Fragile Regions in the Human Genome?
• Transforming Men into Mice
• Sorting by Reversals
• Breakpoint Theorem
• Rearrangements in Tumor Genomes
• 2-Break Distance Problem
• Breakpoint Graphs
• 2-Break Distance Theorem
• Rearrangement Hotspots in the Human Genome
• Synteny Block Construction
it is to prove the 2-Break Distance theorem.
Activate Windows
Go to Settings to activate Windows.


Sorting by 2-Breaks
2-breaks
P → ... → Q
Let's consider an arbitrary transformation
of genome P into genome Q.
Activate Windows
Go to Settings to activate Windows.


Sorting by 2-Breaks
2-breaks
P → ... → Q
BreakpointGraph(P,Q) → ... → BreakpointGraph(Q,Q)
changes into the breakpoint graph of Q
with itself.
Activate Windows
Go to Settings to activate Windows.

Sorting by 2-Breaks

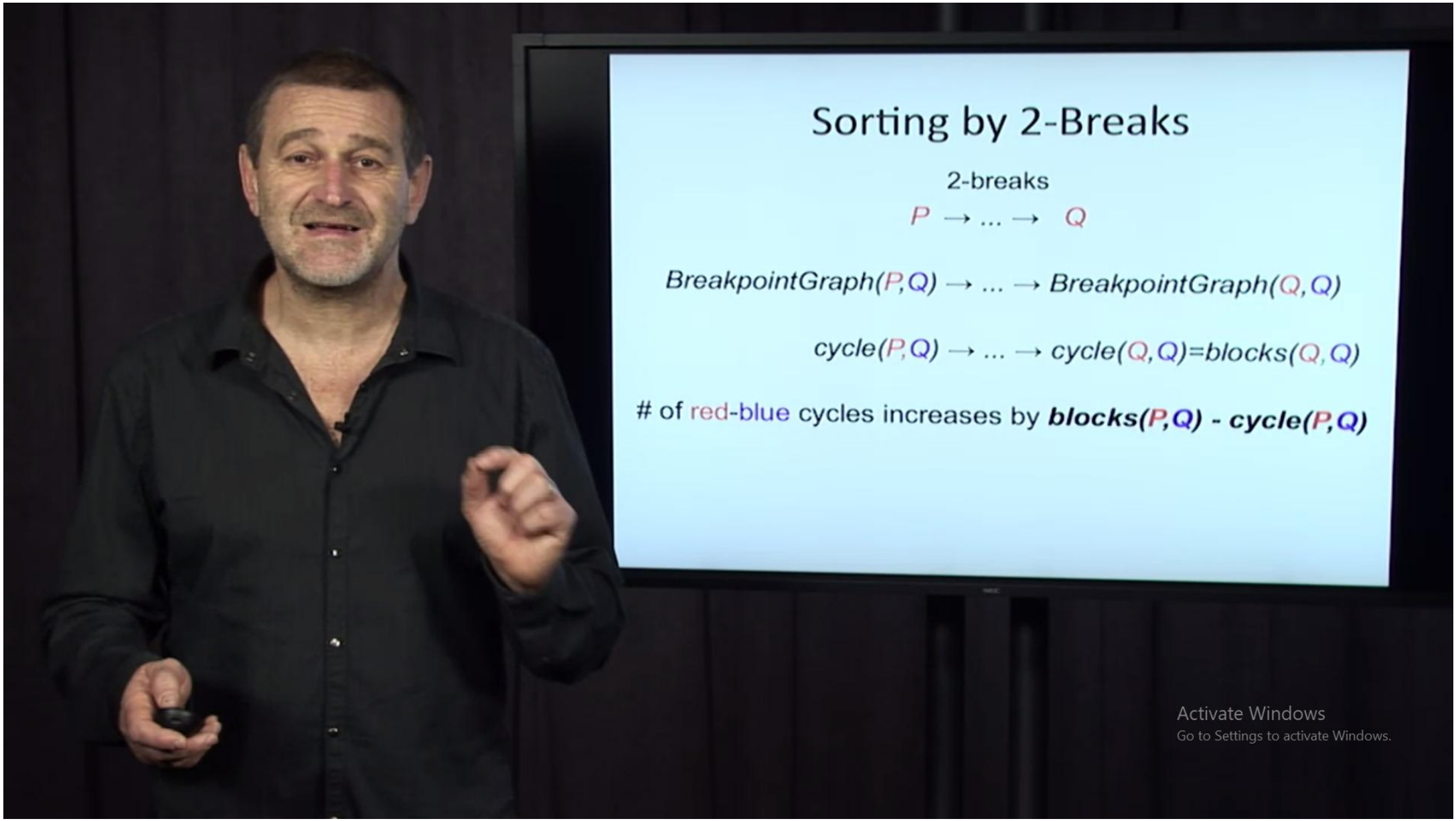
2-breaks

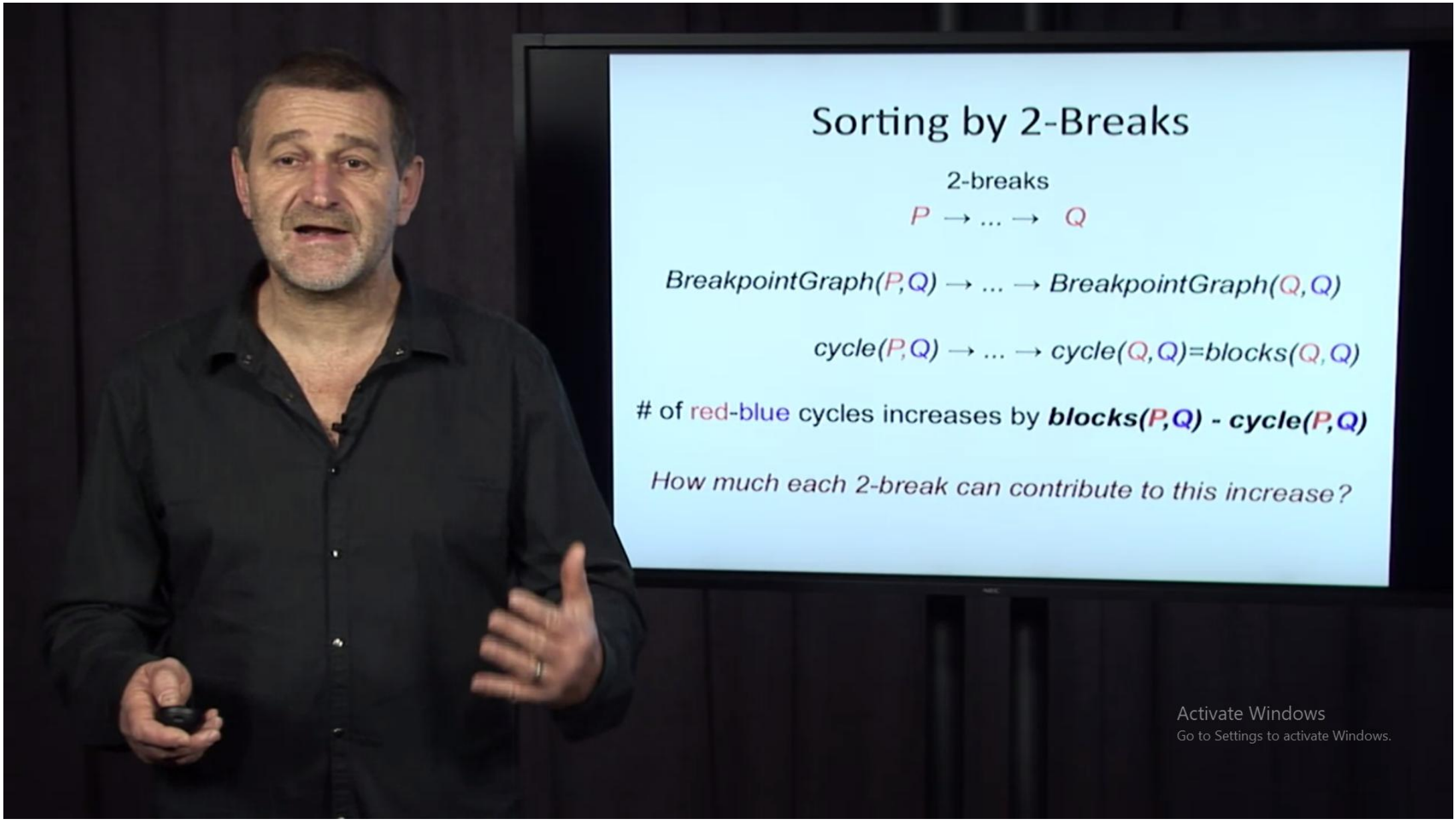
$P \rightarrow \ldots \rightarrow Q$

$BreakpointGraph(P,Q) \rightarrow \ldots \rightarrow BreakpointGraph(Q,Q)$

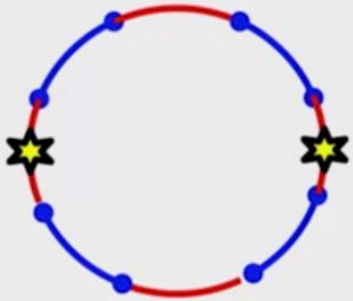
$cycle(P,Q) \rightarrow \ldots \rightarrow cycle(Q,Q) = blocks(Q,Q)$

Which means that the cycle number of P and Q is changing into the

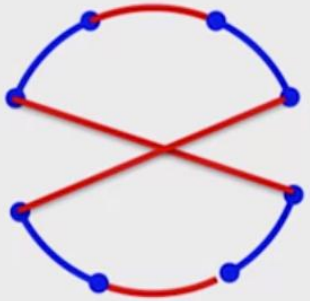




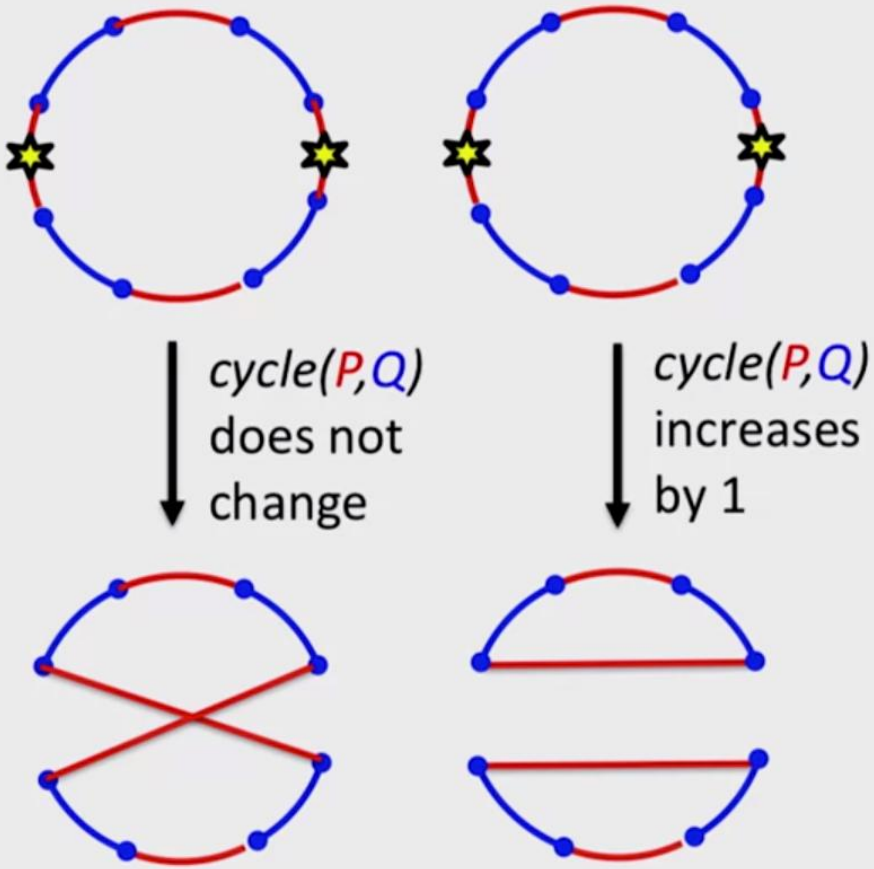
A 2-Break May Change *cycle(P,Q)* by …

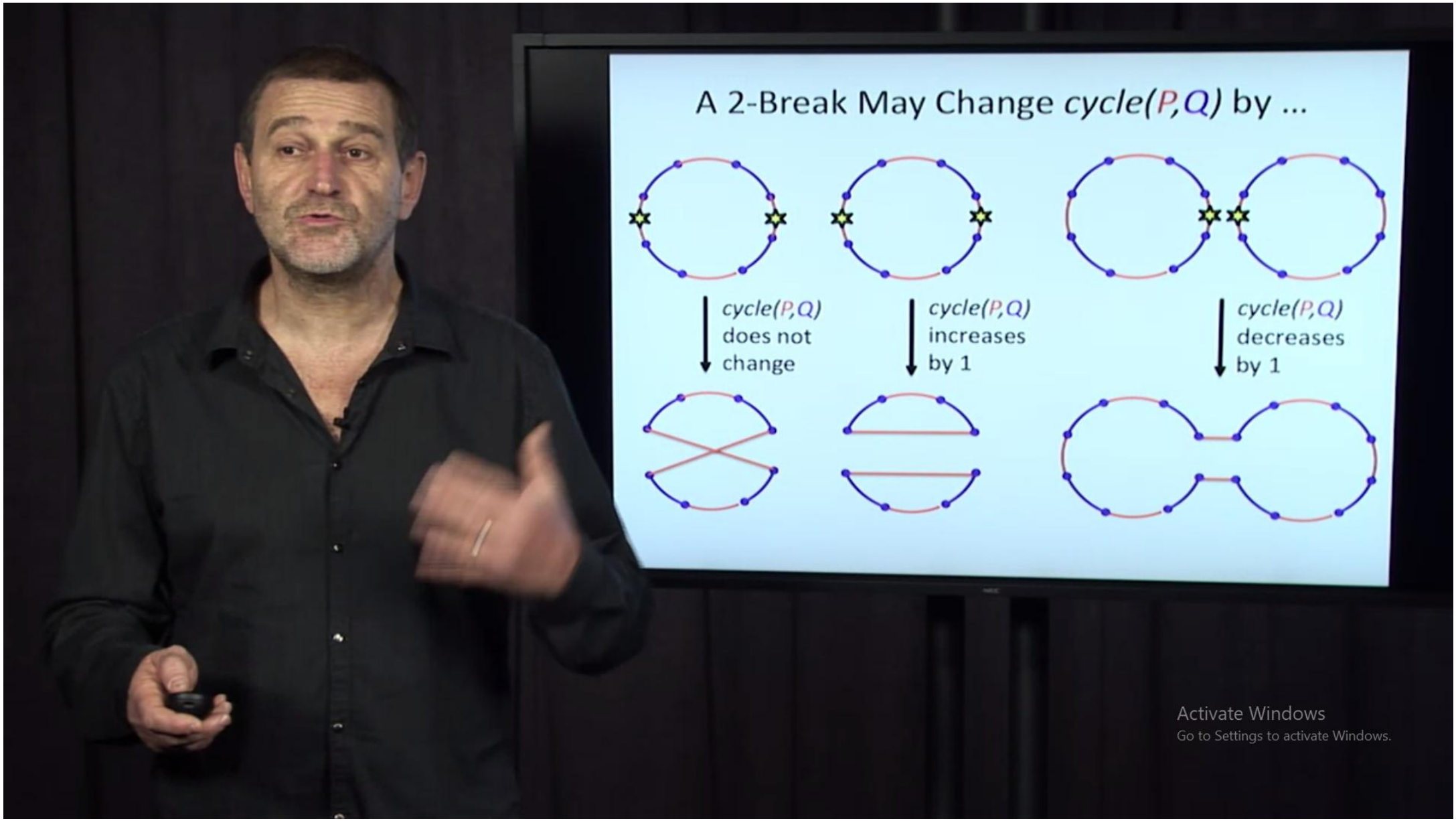


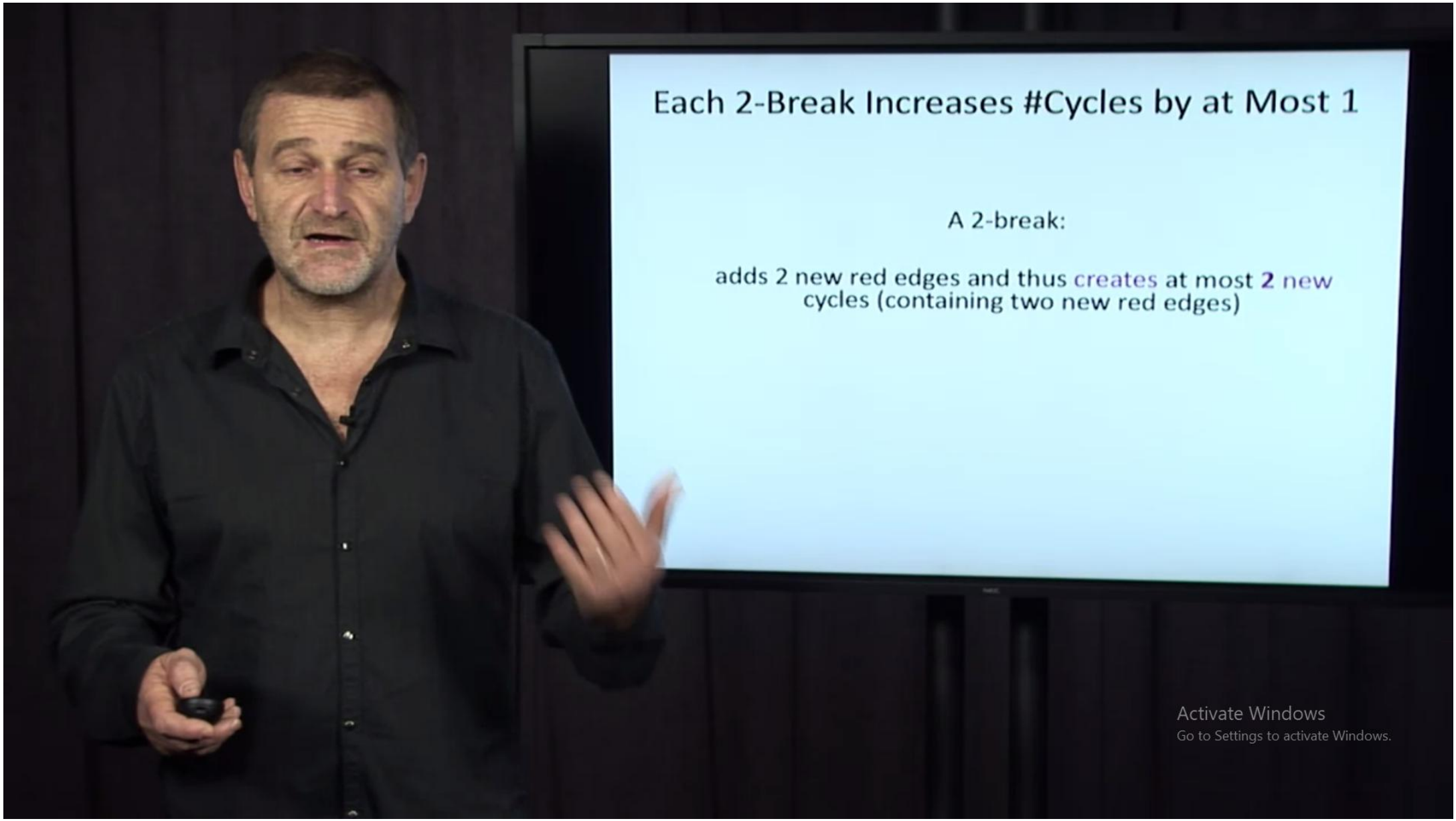
cycle(P,Q) does not change



A 2-Break May Change *cycle(P,Q)* by …



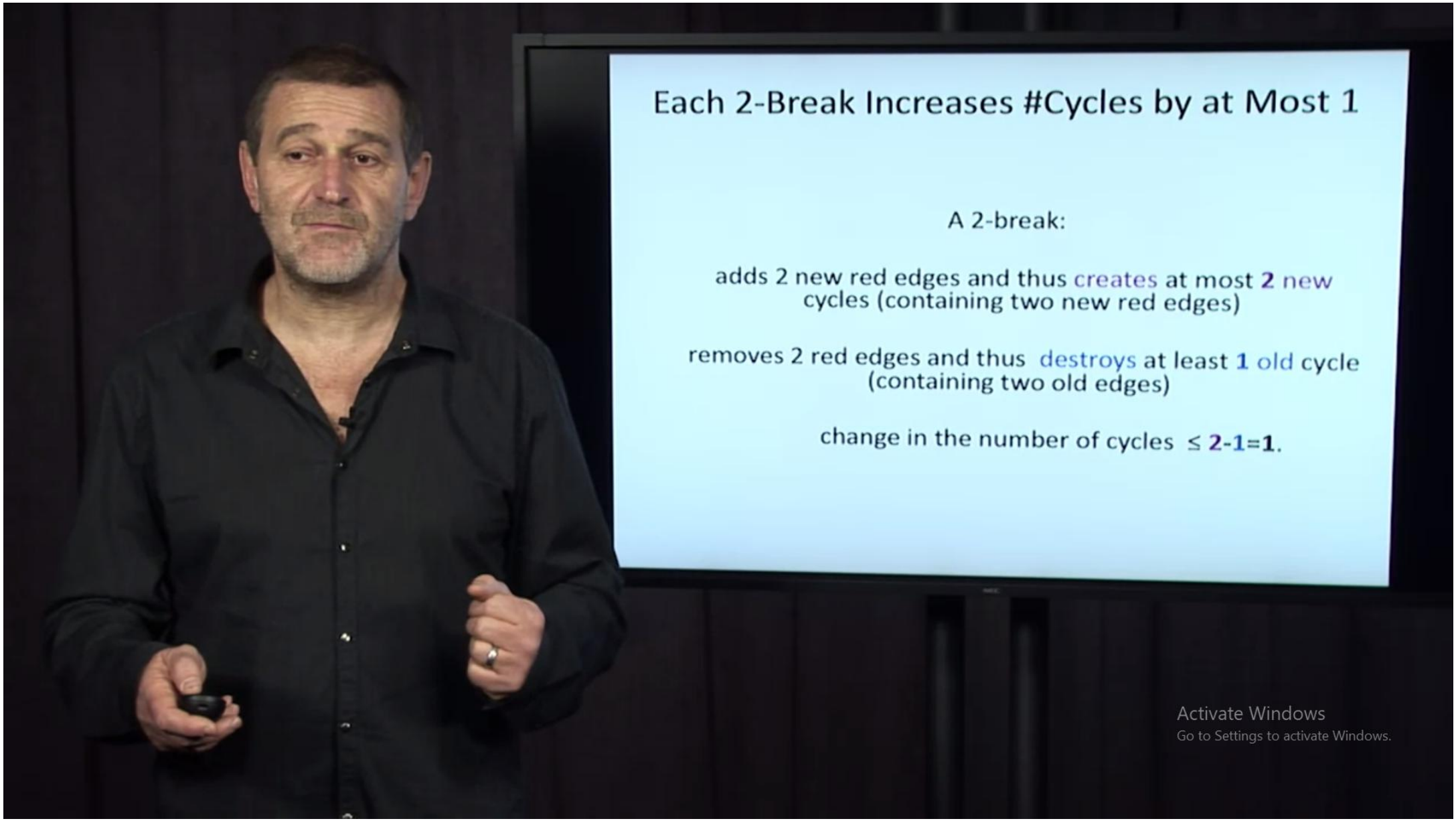



Each 2-Break Increases #Cycles by at Most 1
A 2-break:
adds 2 new red edges and thus creates at most 2 new cycles (containing two new red edges)
Activate Windows
Go to Settings to activate Windows.

Each 2-Break Increases #Cycles by at Most 1

A 2-break:

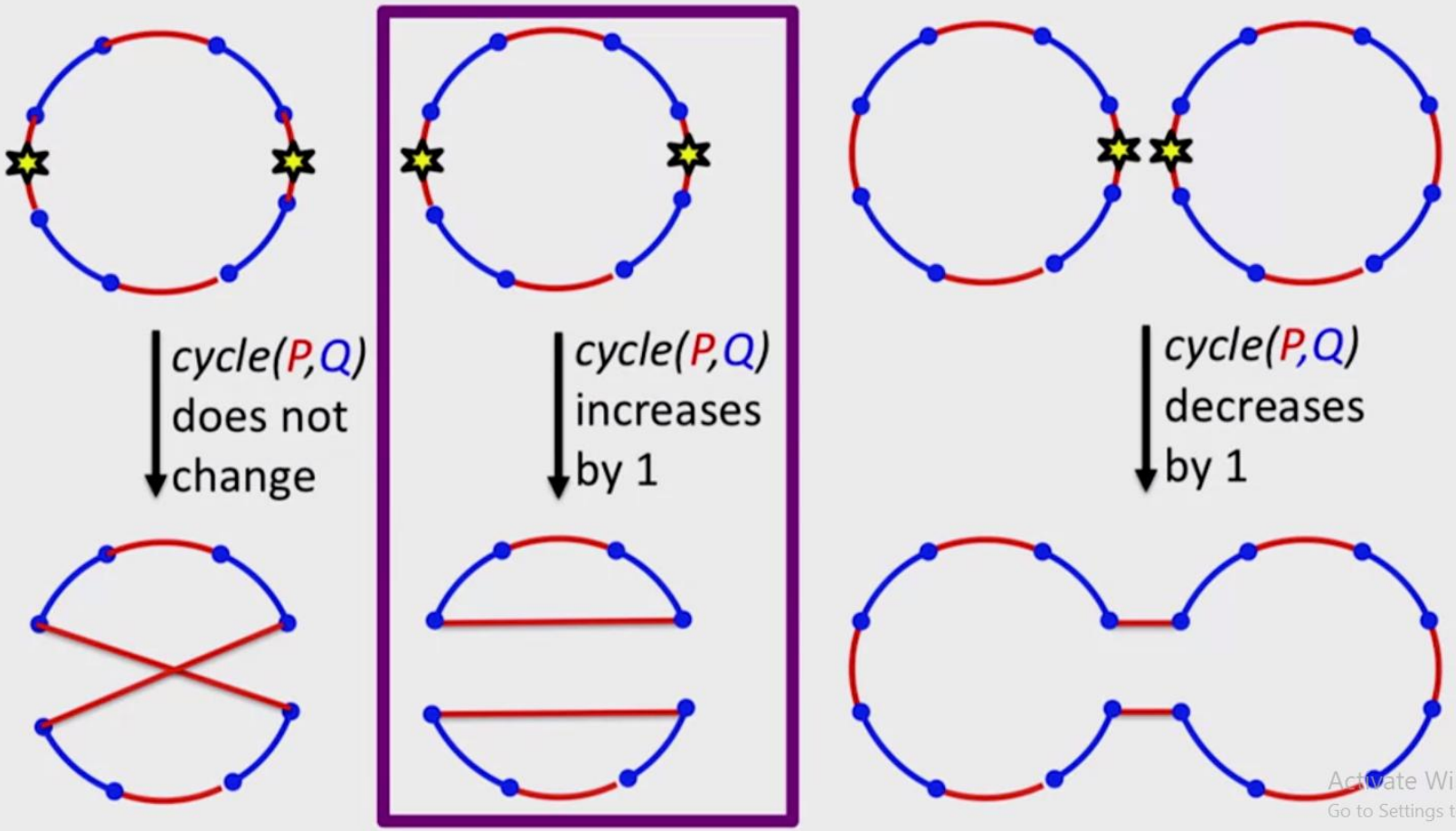
adds 2 new red edges and thus creates at most **2** new cycles (containing two new red edges)


Each 2-Break Increases #Cycles by at Most 1
A 2-break:
adds 2 new red edges and thus creates at most 2 new cycles (containing two new red edges)
removes 2 red edges and thus destroys at least 1 old cycle (containing two old edges)
change in the number of cycles ≤ 2-1=1.
Activate Windows
Go to Settings to activate Windows.

2-Break Distance Theorem

- A 2-break increases #cycles by at most 1.
- There exists a 2-break increasing #cycles by 1.

There Exists a 2-Break Increasing *cycle(P,Q)* by 1



2-Break Distance Theorem

- A 2-break increases #cycles by at most 1.
- There exists a 2-break increasing #cycles by 1.
- Every sorting by 2-breaks must increase #cycles by *blocks(P,Q) - cycle(P,Q)*

2-Break Distance Theorem

- A 2-break increases #cycles by at most 1.
- There exists a 2-break increasing #cycles by 1.
- Every sorting by 2-breaks must increase #cycles by *blocks(P,Q) - cycle(P,Q)*
- 2-break distance between genomes *P* and *Q*:

$$d(P,Q) = blocks(P,Q) - cycle(P,Q)$$

2-Break Distance between Human and Mouse Genomes

- *Human* and *Mouse* genomes can be decomposed into **280** synteny blocks (at least 0.5 million nucleotides in length)
- The breakpoint graph on these blocks has **35** cycles
- The 2-break distance between *Human* and *Mouse:*

$$d(H,M) = blocks(H,M) - cycle(H,M) = \mathbf{280-35=245}$$

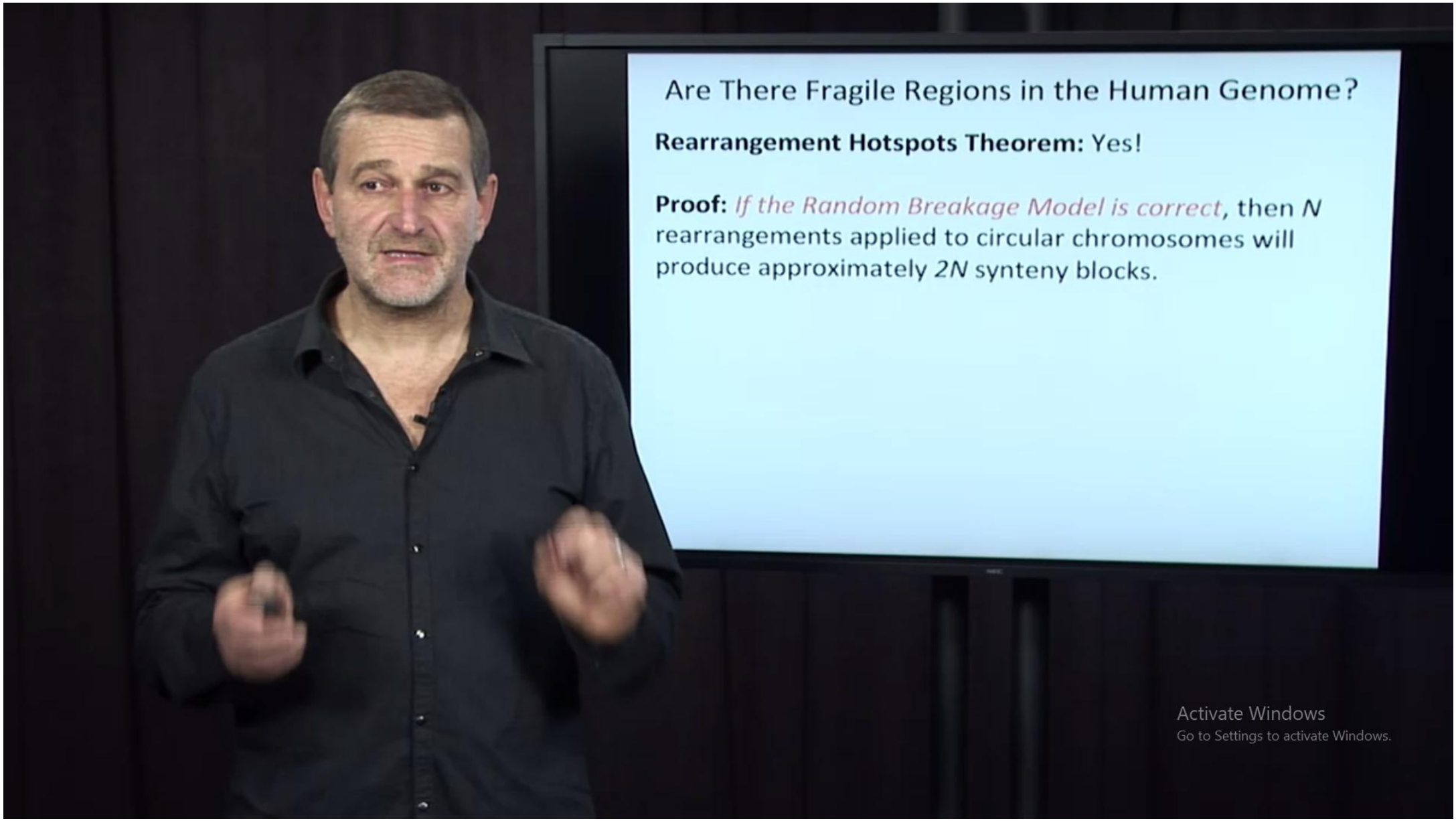
2-Break Distance between Human and Mouse Genomes

- *Human* and *Mouse* genomes can be decomposed into **280** synteny blocks (at least 0.5 million nucleotides in length)
- The breakpoint graph on these blocks has **35** cycles
- The 2-break distance between *Human* and *Mouse:*

$$d(H,M) = blocks(H,M) - cycle(H,M) = \mathbf{280-35=245}$$

- There are numerous **245**-step scenarios.
- The true scenario may have more than **245** steps.


Press Esc to exit full screen
Are There Fragile Regions in the Human Genome?
Activate Windows
Go to Settings to activate Windows.


Are There Fragile Regions in the Human Genome?
Rearrangement Hotspots Theorem: Yes!
Proof: If the Random Breakage Model is correct, then N rearrangements applied to circular chromosomes will produce approximately 2N synteny blocks.
Activate Windows
Go to Settings to activate Windows.

Are There Fragile Regions in the Human Genome?

Rearrangement Hotspots Theorem: Yes!

Proof: *If the Random Breakage Model is correct*, then *N* rearrangements applied to circular chromosomes will produce approximately *2N* synteny blocks.

Are There Fragile Regions in the Human Genome?

Rearrangement Hotspots Theorem: Yes!

Proof: *If the Random Breakage Model is correct*, then *N* rearrangements applied to circular chromosomes will produce approximately *2N* synteny blocks.

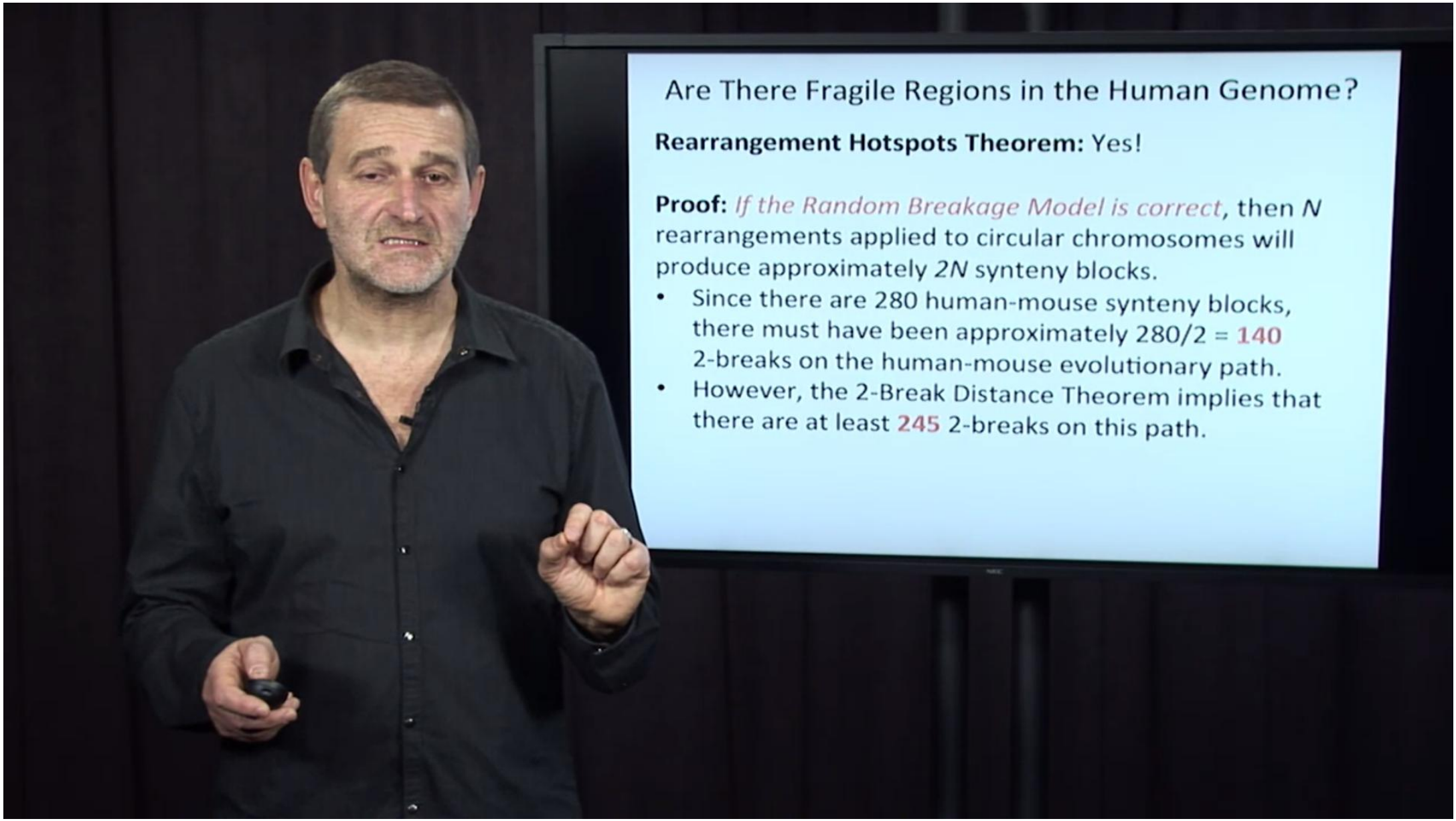
- Since there are 280 human-mouse synteny blocks, there must have been approximately 280/2 = **140** 2-breaks on the human-mouse evolutionary path.

Are There Fragile Regions in the Human Genome?

Rearrangement Hotspots Theorem: Yes!

Proof: *If the Random Breakage Model is correct*, then *N* rearrangements applied to circular chromosomes will produce approximately *2N* synteny blocks.

- Since there are 280 human-mouse synteny blocks, there must have been approximately 280/2 = **140** 2-breaks on the human-mouse evolutionary path.
- However, the 2-Break Distance Theorem implies that there are at least **245** 2-breaks on this path.


Are There Fragile Regions in the Human Genome?
Rearrangement Hotspots Theorem: Yes!
Proof: If the Random Breakage Model is correct, then N rearrangements applied to circular chromosomes will produce approximately 2N synteny blocks.
• Since there are 280 human-mouse synteny blocks, there must have been approximately 280/2 = 140 2-breaks on the human-mouse evolutionary path.
• However, the 2-Break Distance Theorem implies that there are at least 245 2-breaks on this path.

A Contradiction!

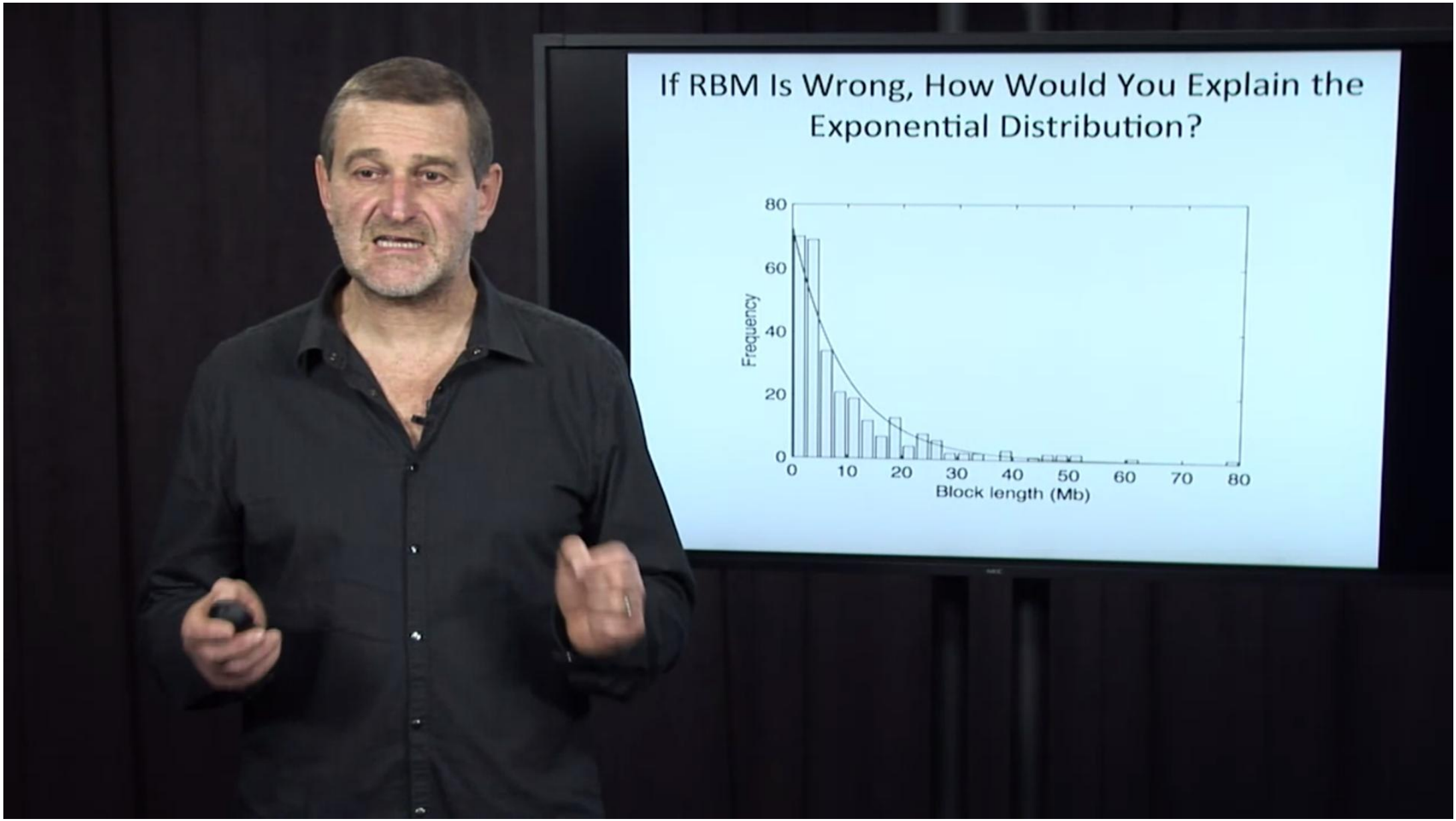
Since 245 is much larger than 140, we arrived at a contradiction implying that **one of our assumptions is incorrect! Which one?**

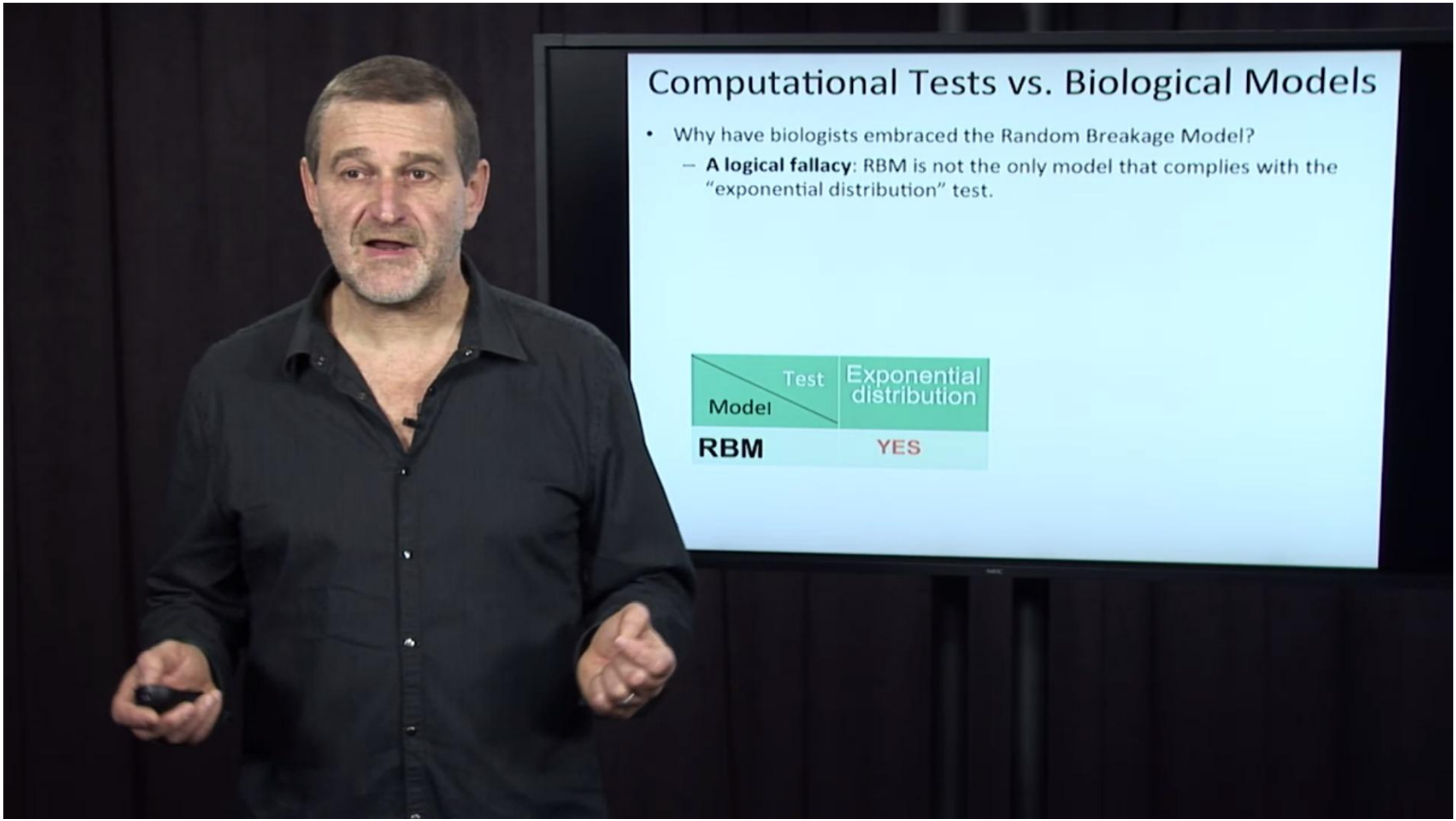
A Contradiction!

Since 245 is much larger than 140, we arrived at a contradiction implying that **one of our assumptions is incorrect! Which one?**

But the only assumption we made in this proof was:

"If the Random Breakage Model is correct..."


If RBM Is Wrong, How Would You Explain the Exponential Distribution?
Frequency
0
20
40
60
80
0
10
20
30
40
50
60
70
80
Block length (Mb)


Computational Tests vs. Biological Models
• Why have biologists embraced the Random Breakage Model?
– A logical fallacy: RBM is not the only model that complies with the "exponential distribution" test.
Test
Model
Exponential distribution
RBM
YES

Computational Tests vs. Biological Models

- Why have biologists embraced the Random Breakage Model?
  - A logical fallacy: RBM is not the only model that complies with the "exponential distribution" test.
- Why was RBM refuted?
  - It does not comply with the observed "breakpoint reuse."

Test / Model	Exponential distribution	Breakpoint reuse
RBM	**YES**	**NO**

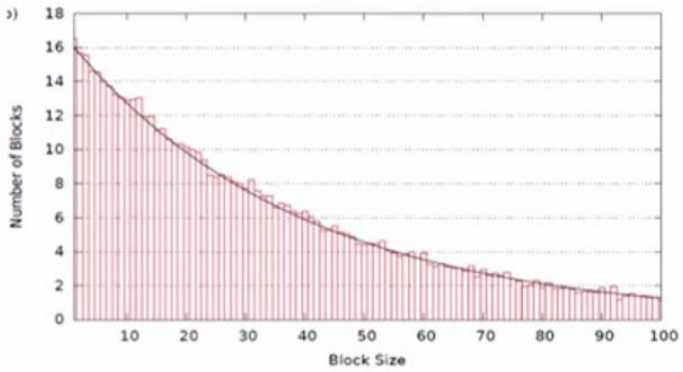
A New Model Needed

- Why have biologists embraced the Random Breakage Model?
  - A logical fallacy: RBM is not the only model that complies with the "exponential distribution" test.
- Why was RBM refuted?
  - It does not comply with the "breakpoint reuse" observed in genomes.
- **Is there a model that complies with both the "exponential distribution" and the "breakpoint reuse" tests?**

Model \ Test	Exponential distribution	Breakpoint reuse
RBM	**YES**	**NO**
???	**YES**	**YES**

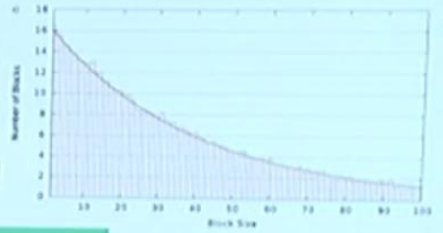
Does FBM Explain BOTH Exponential Distribution and Rearrangement Hotspots?

- A small number of short fragile regions explain rearrangement hotspots.
- If the fragile regions are somewhat randomly distributed throughout the genome, the synteny blocks follow the exponential distribution!



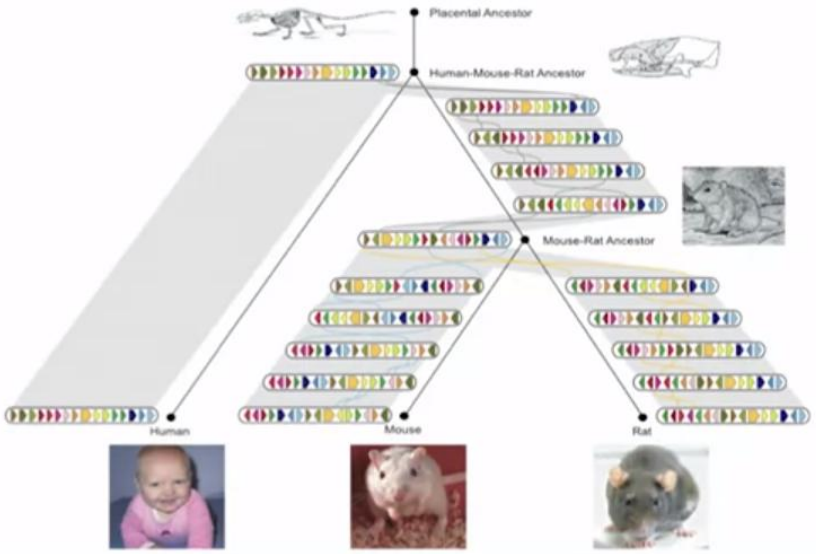
Does FBM Explain BOTH Exponential Distribution and Rearrangement Hotspots?

- A small number of short fragile regions explain rearrangement hotspots.
- If the fragile regions are somewhat randomly distributed throughout the genome, the synteny blocks follow the exponential distribution!



Model \ Test	Exponential distribution	Breakpoint reuse
RBM	YES	NO
FBM	YES	YES

Information About Multiple Genomes Enables a New Test



Multiple Breakpoint Reuse Test: analysis of breakpoints across multiple genomes

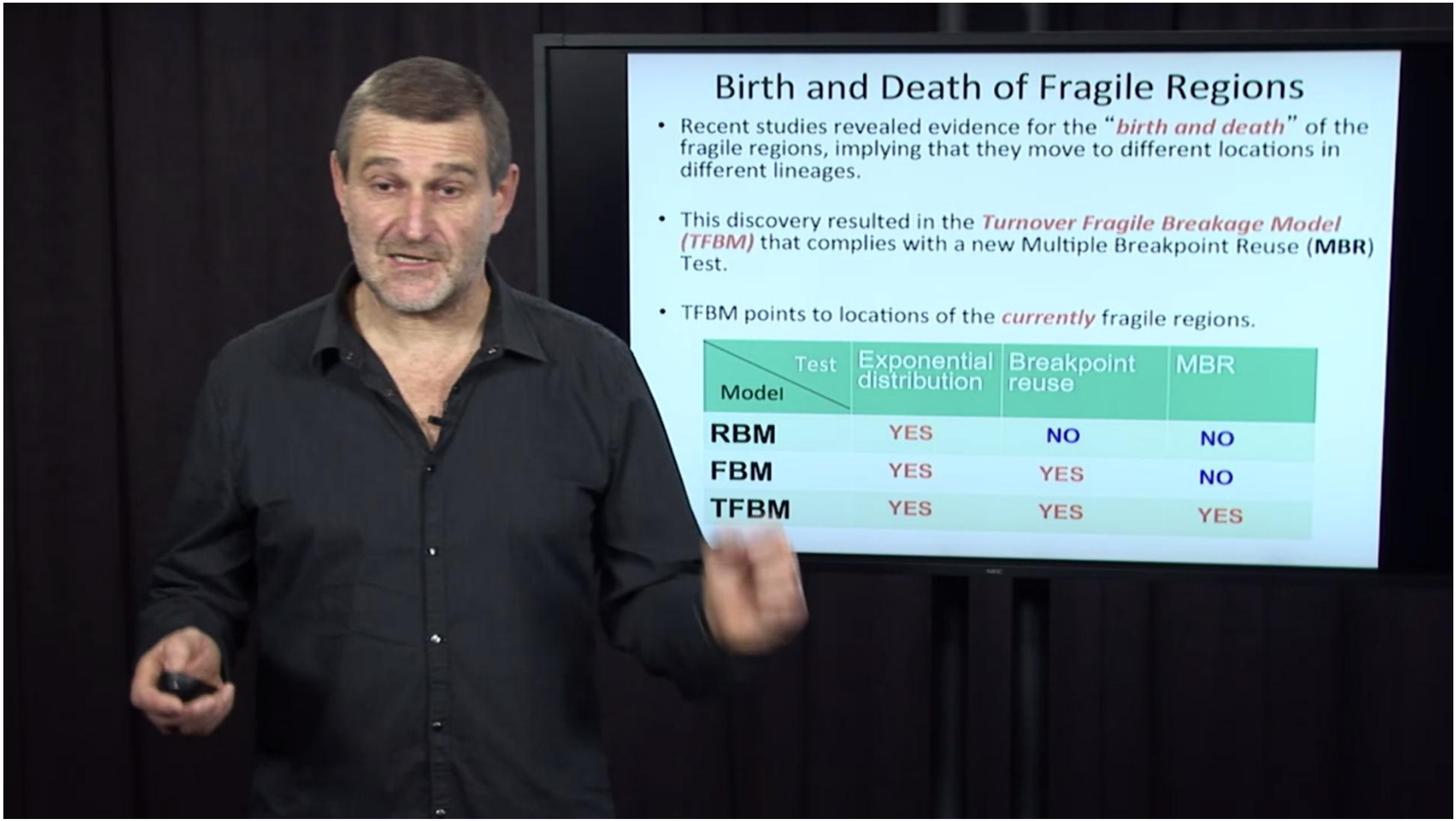
Model \ Test	Exponential distribution	Breakpoint reuse	A New Test?
RBM	**YES**	**NO**	**NO**
FBM	**YES**	**YES**	**NO**

Birth and Death of Fragile Regions

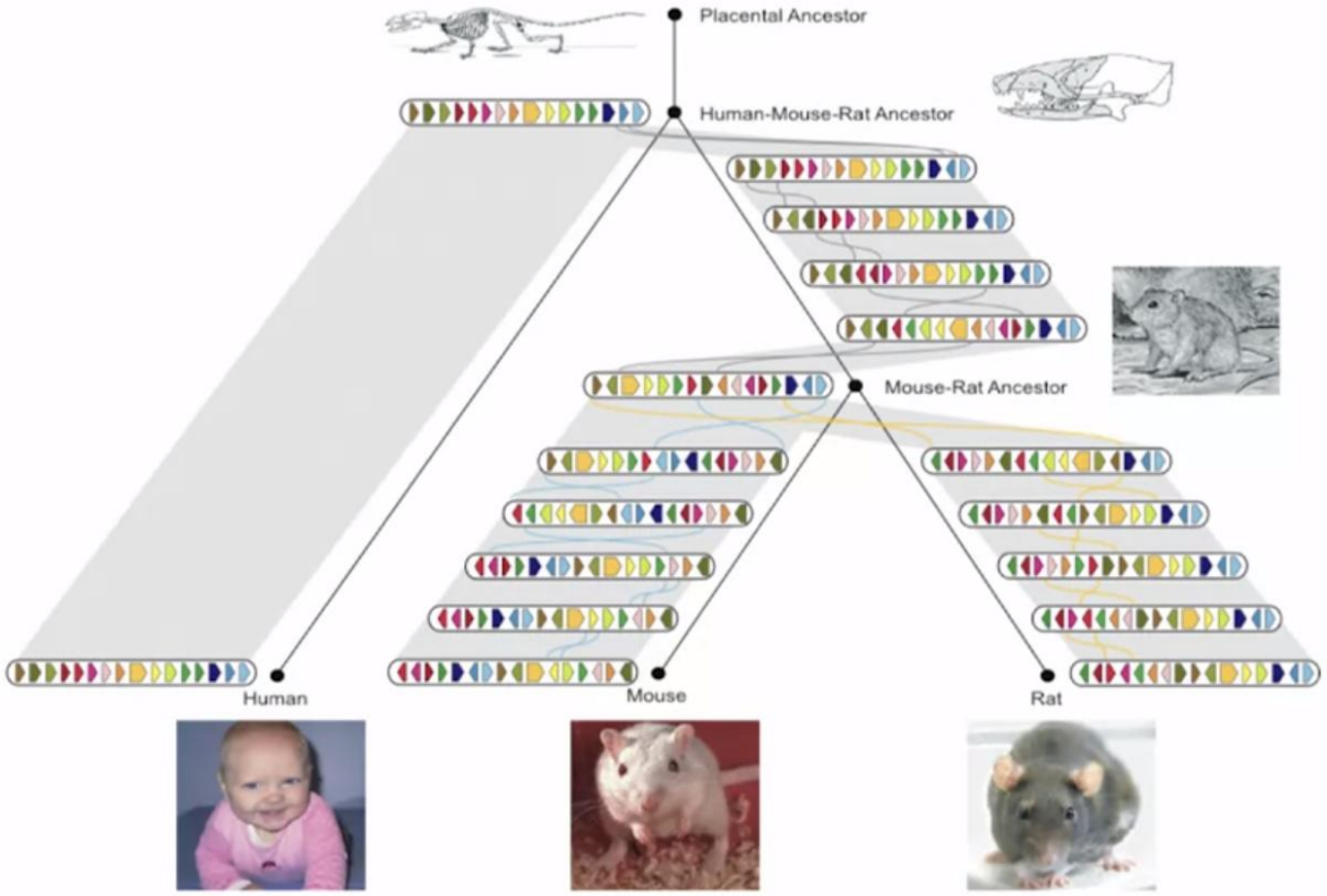
- Recent studies revealed evidence for the "***birth and death***" of the fragile regions, implying that they move to different locations in different lineages.

- This discovery resulted in the ***Turnover Fragile Breakage Model (TFBM)*** that complies with a new Multiple Breakpoint Reuse (**MBR**) Test.

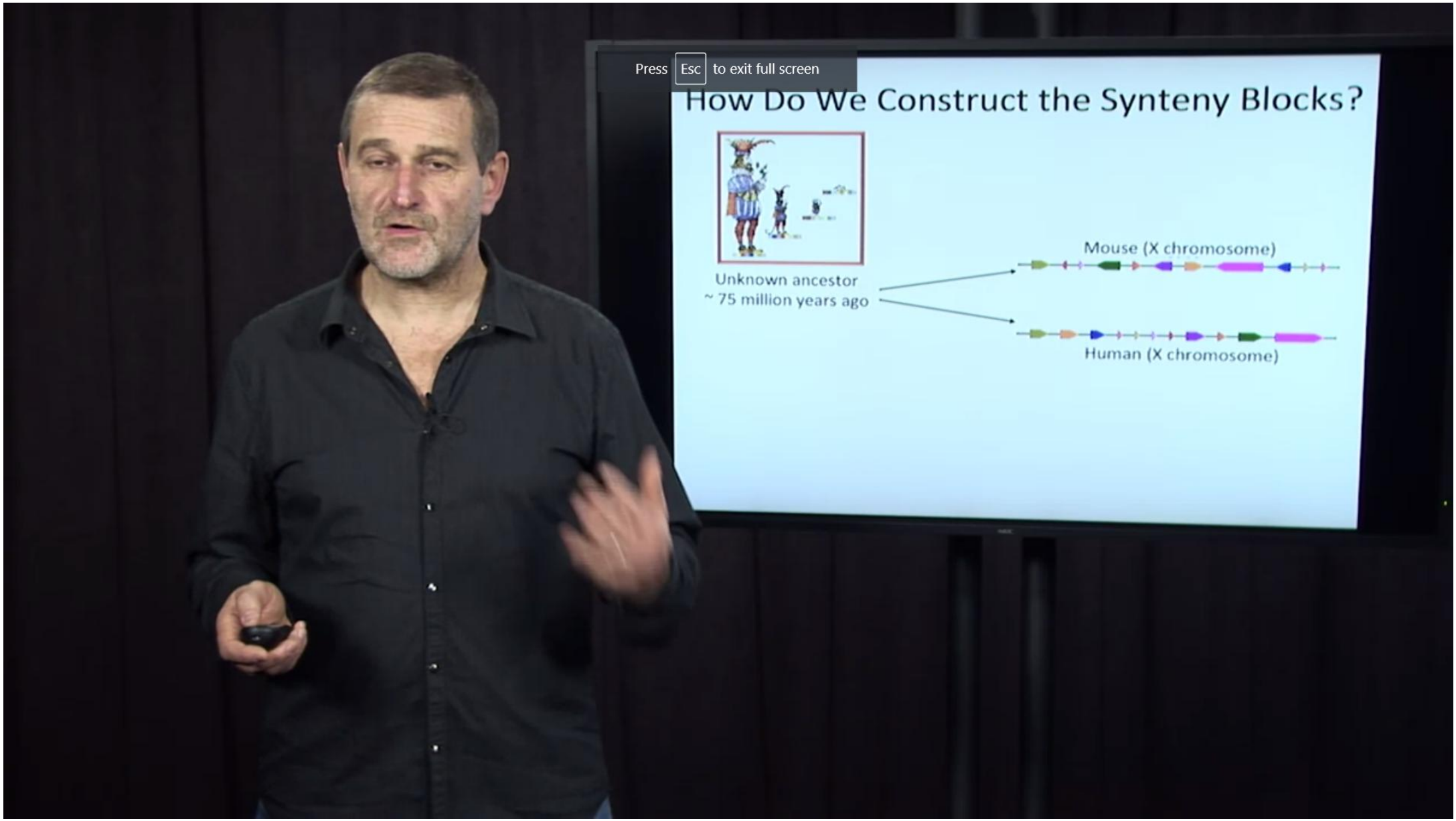
- TFBM points to locations of the ***currently*** fragile regions.


Birth and Death of Fragile Regions
• Recent studies revealed evidence for the "birth and death" of the fragile regions, implying that they move to different locations in different lineages.
• This discovery resulted in the Turnover Fragile Breakage Model (TFBM) that complies with a new Multiple Breakpoint Reuse (MBR) Test.
• TFBM points to locations of the currently fragile regions.
Test / Model
Exponential distribution
Breakpoint reuse
MBR
RBM YES NO NO
FBM YES YES NO
TFBM YES YES YES

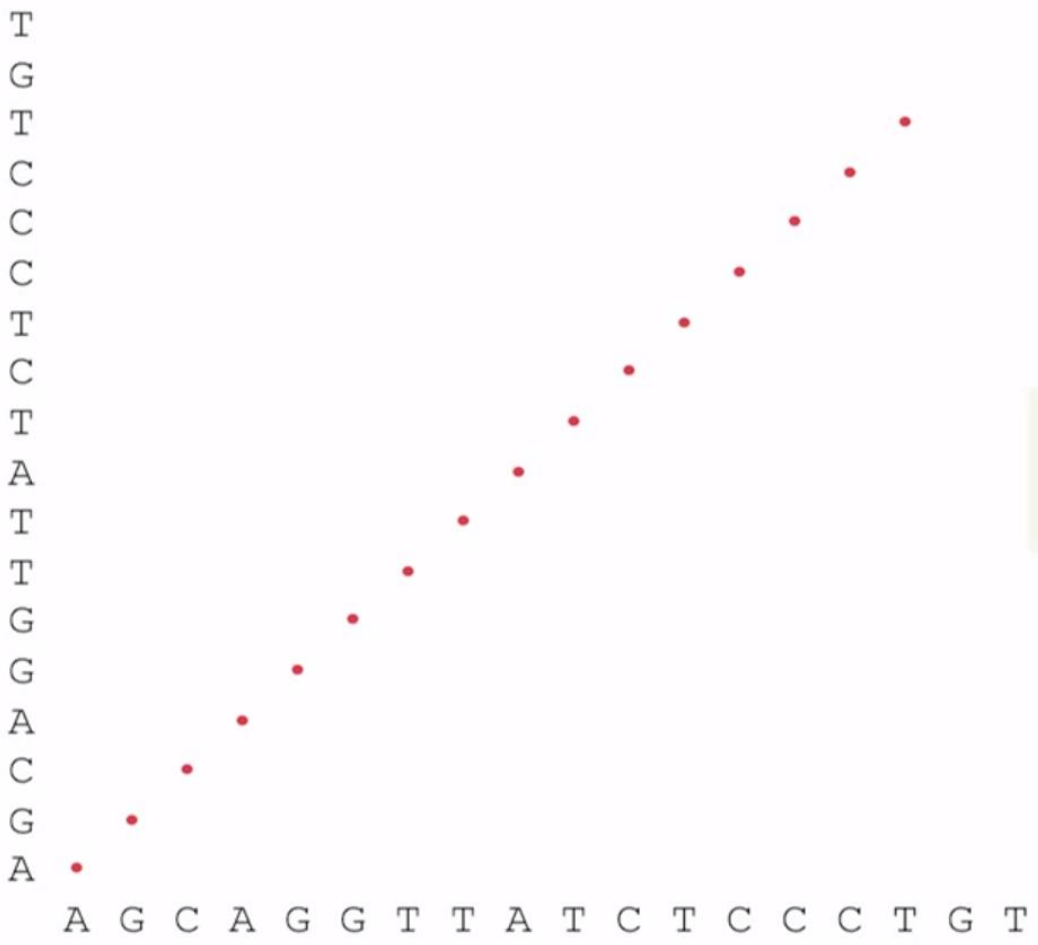
Where Are the Fragile Regions Located? What Causes Fragility?



Press Esc to exit full screen

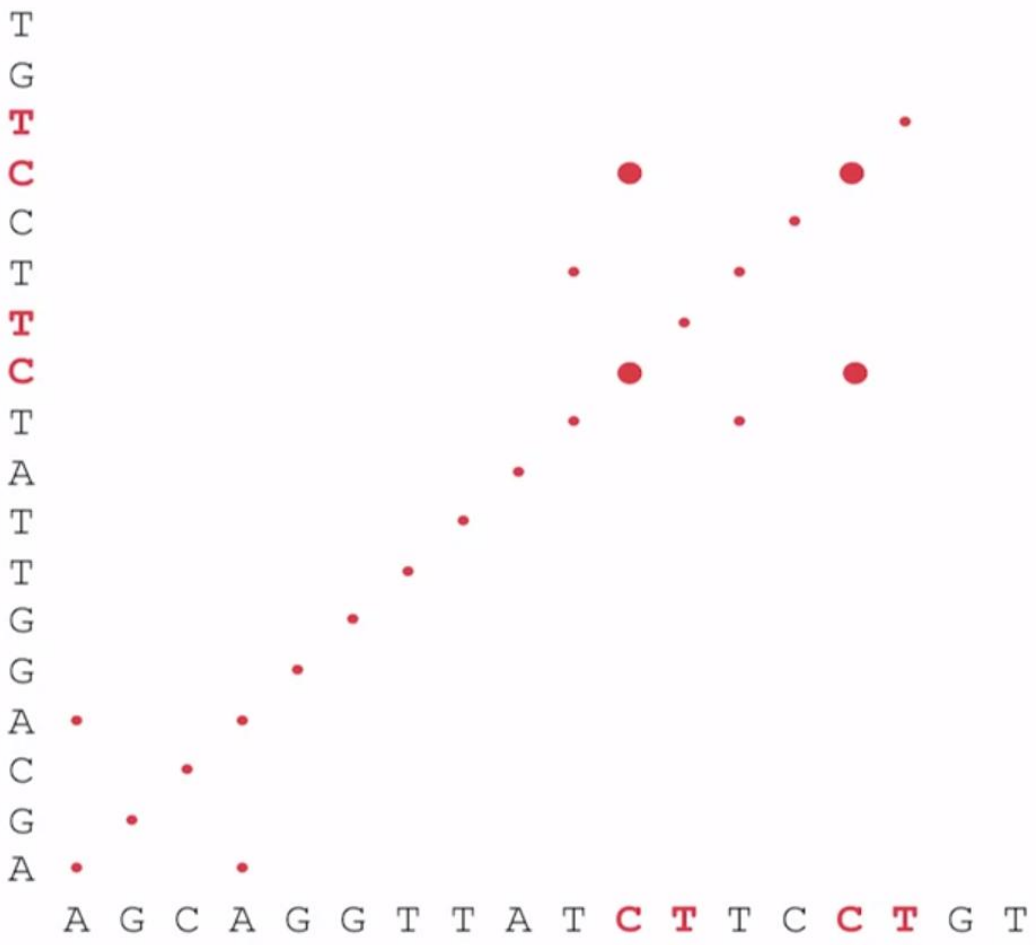


Comparing Genome with Itself (in 2-D)



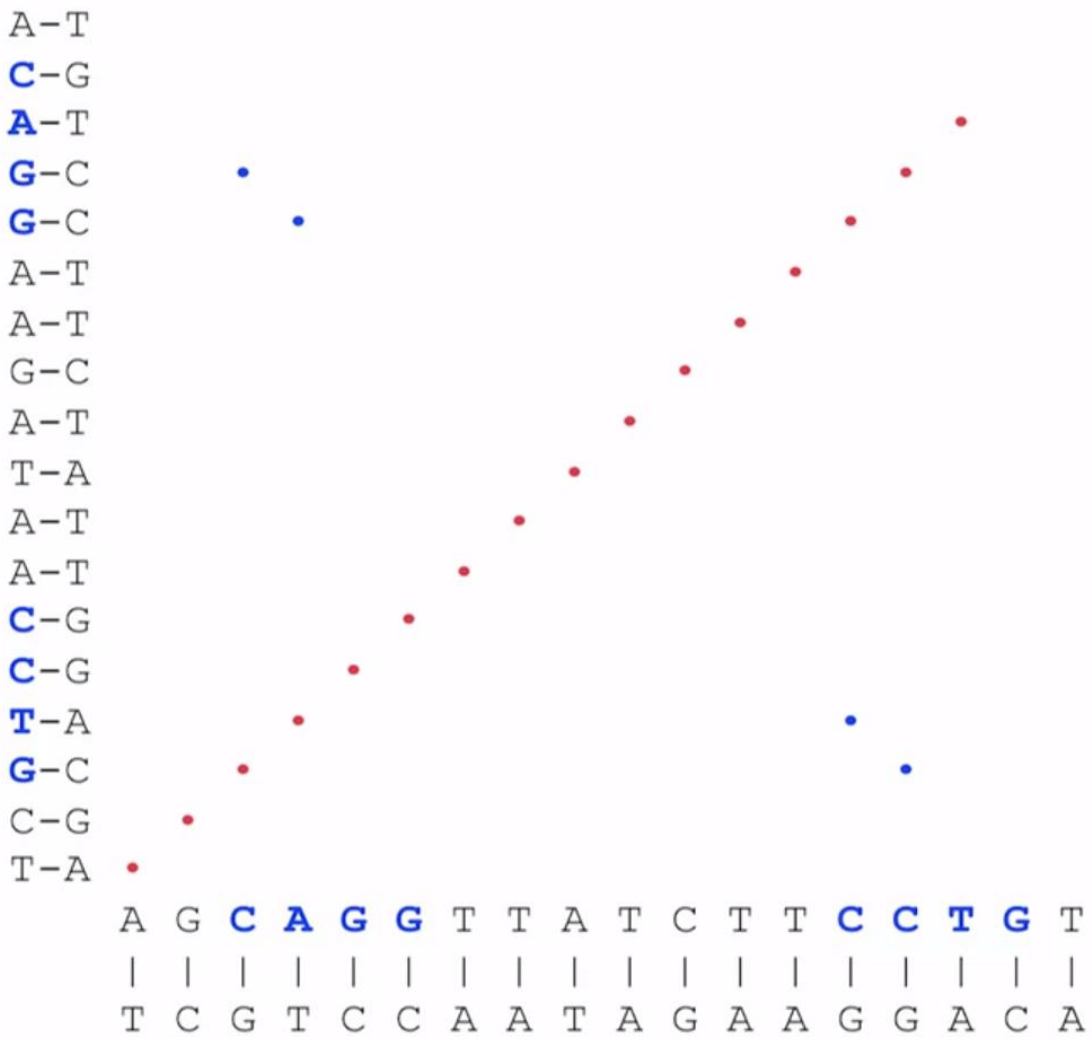
Identical *k*-mers
(*k*=3)

From Identical 3-mers to Identical 2-mers



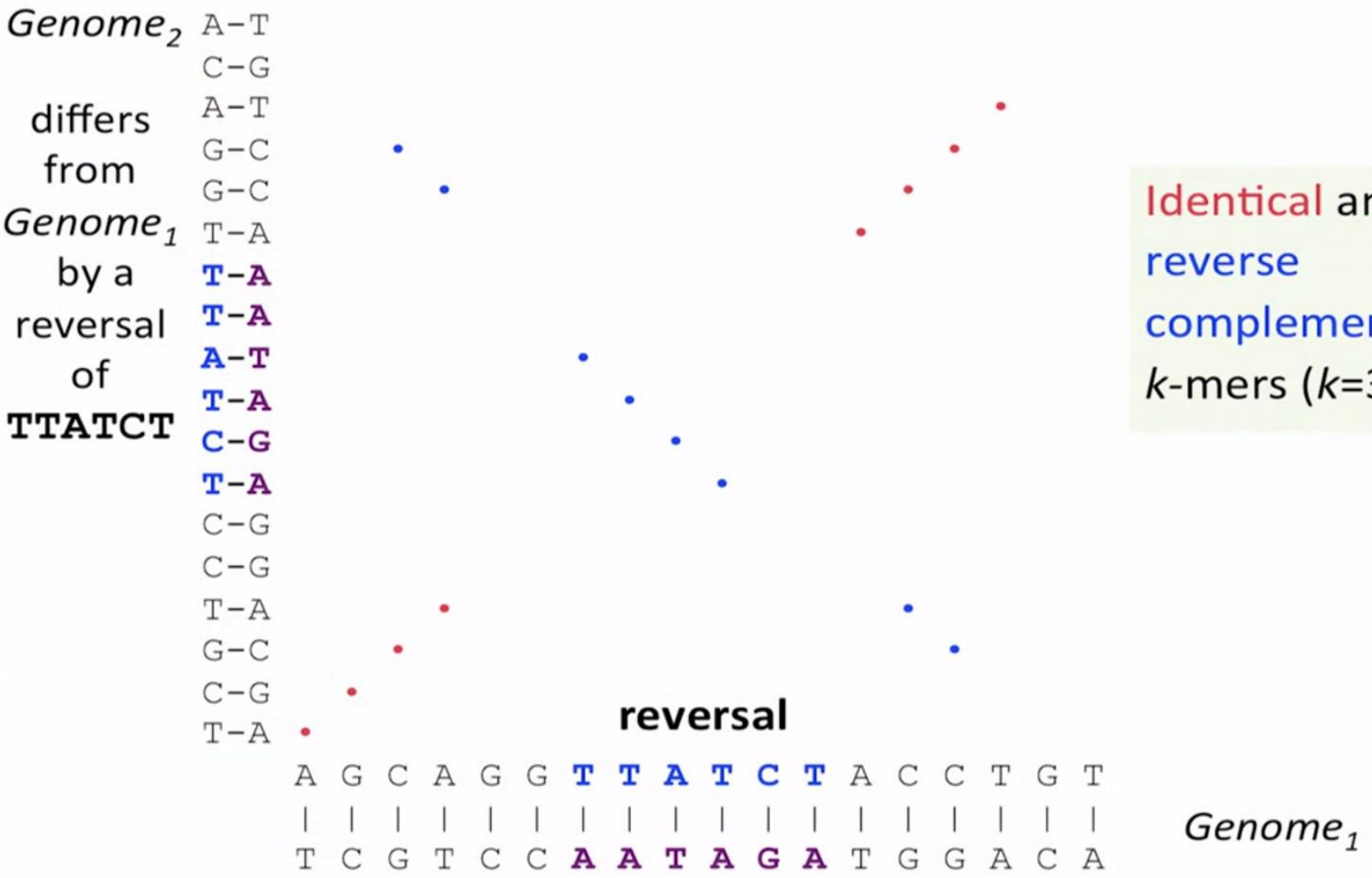

We Forgot about the Complementary Strand
A-T
C-G
A-T
G-C
G-C
A-T
A-T
G-C
A-T
T-A
A-T
A-T
C-G
C-G
T-A
G-C
C-G
T-A
A G C A G G T T A T C T T C C T G T
| | | | | | | | | | | | | | | | |
T C G T C C A A T A G A A G G A C A
Identical and reverse complementary k-mers (k=3).

We Forgot about the Complementary Strand



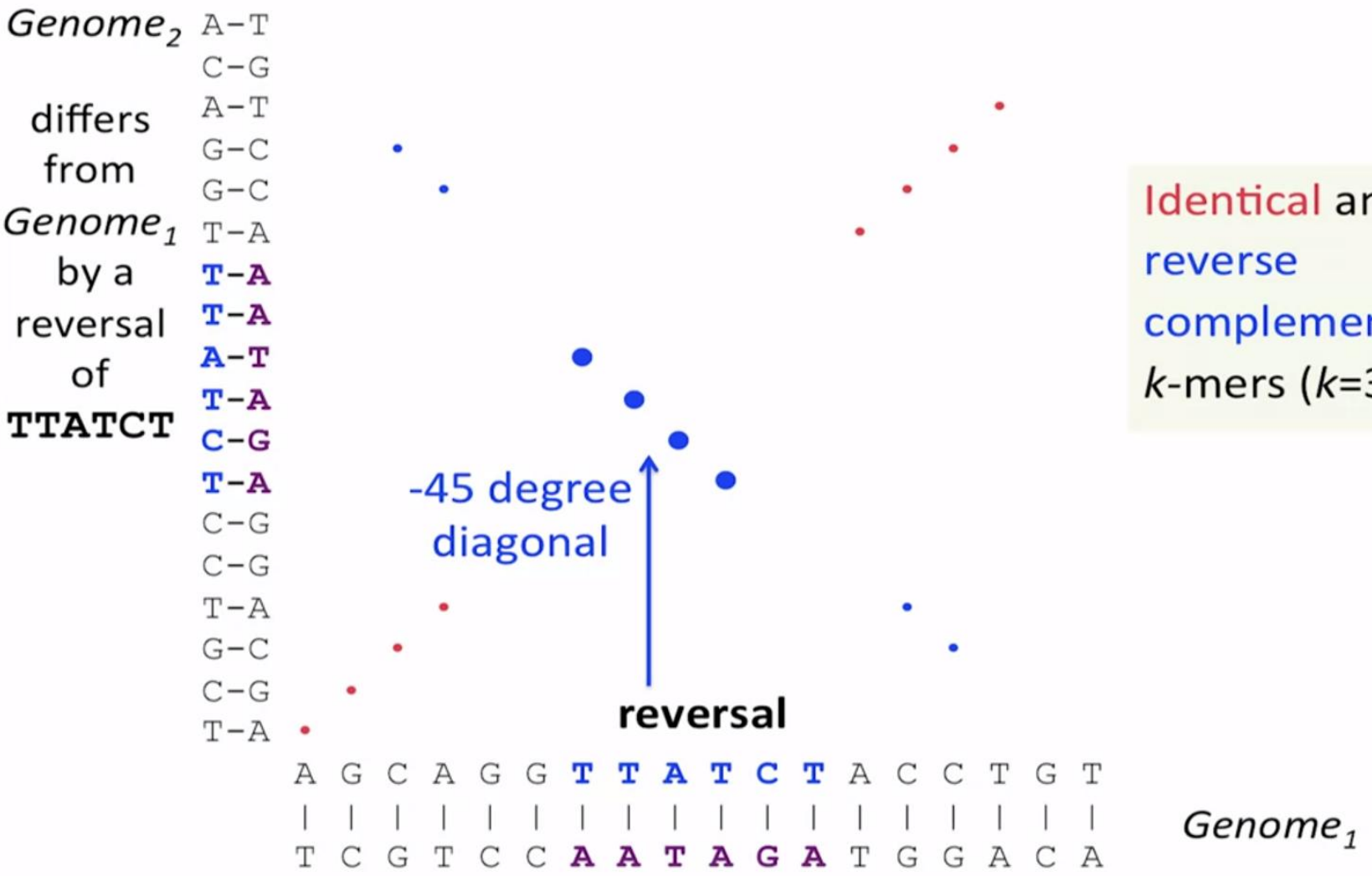
Identical and reverse complementary *k*-mers (*k*=3).

Comparing Different Genomes



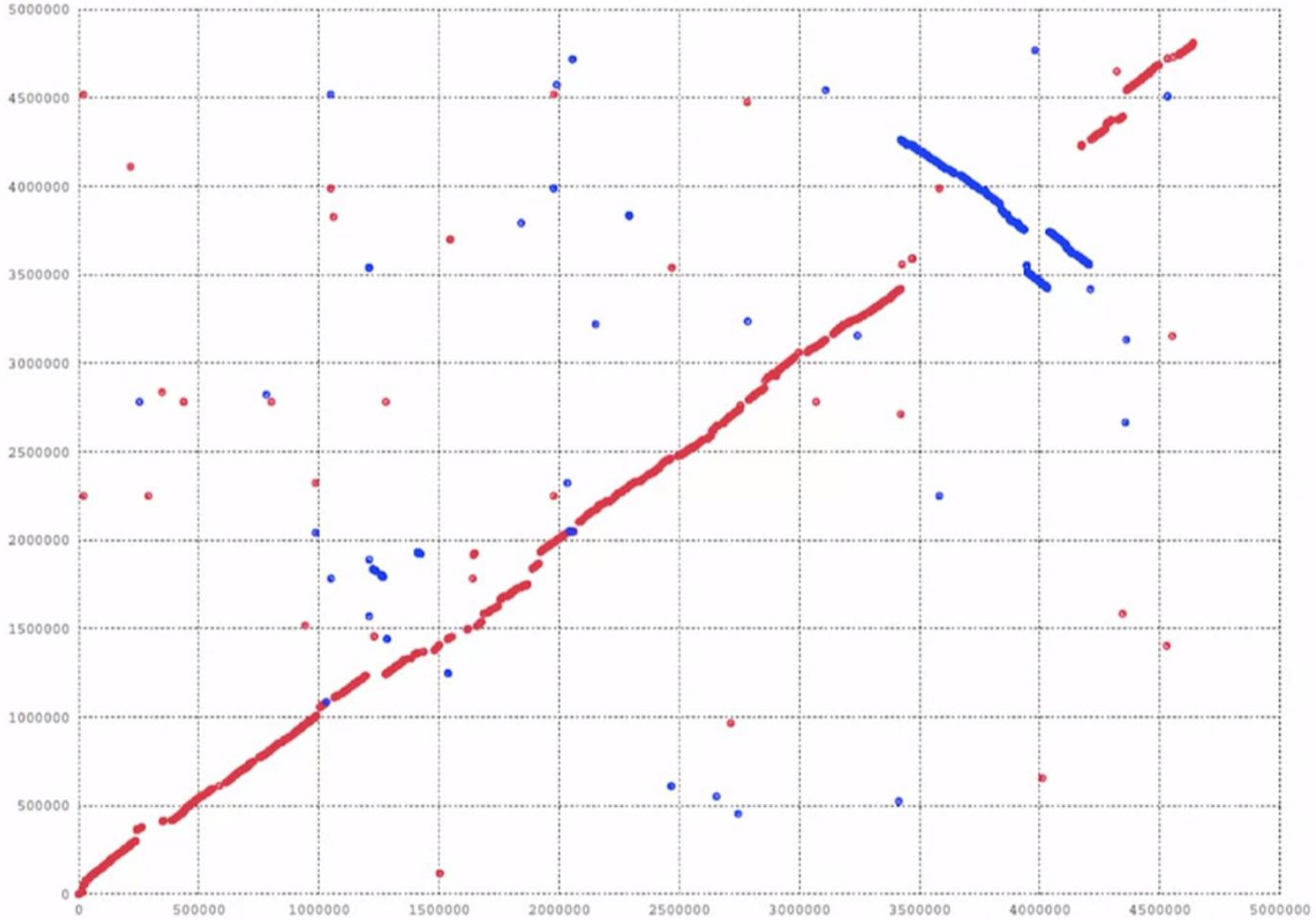
Identical and reverse complementary k-mers (k=3).

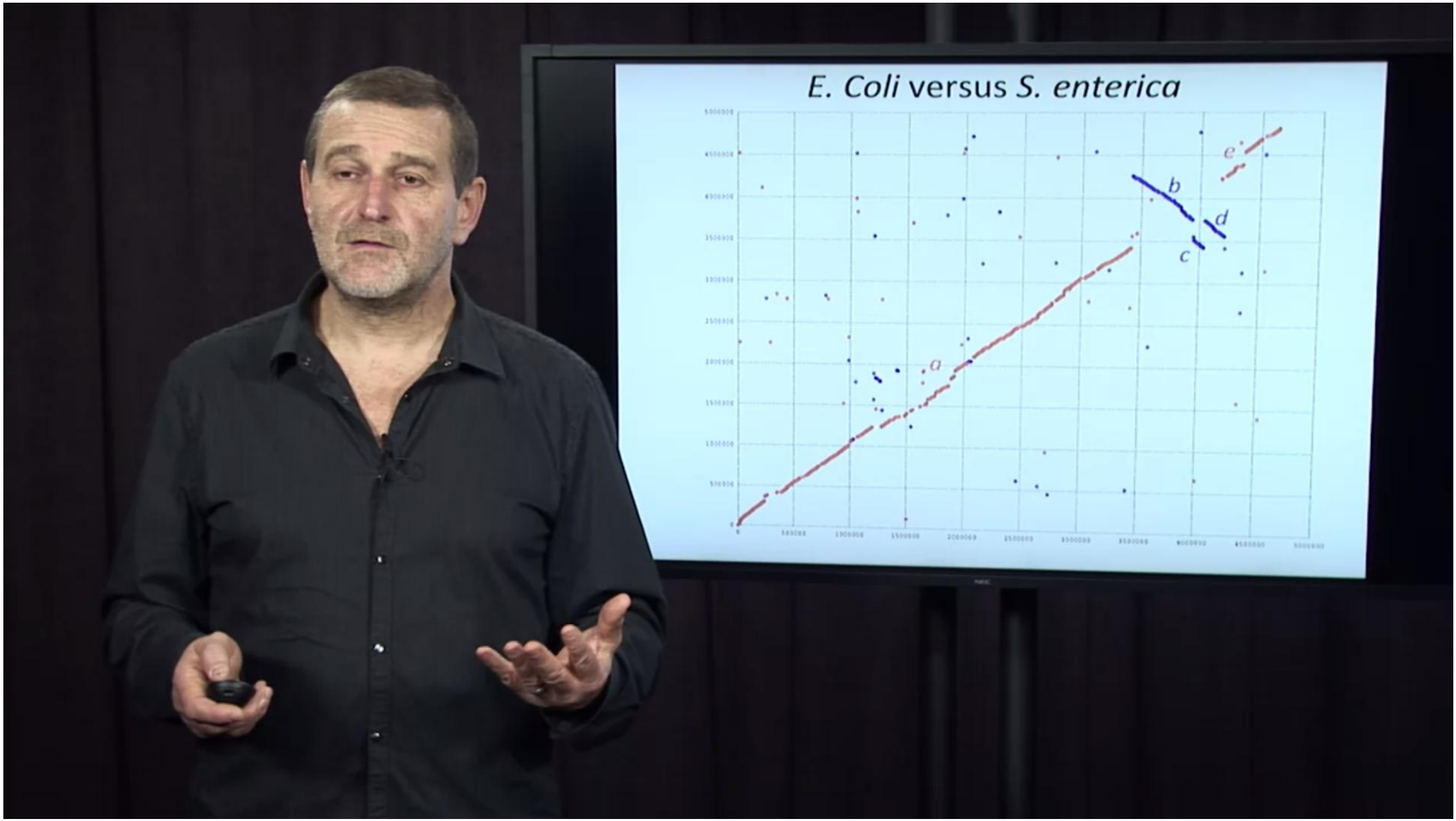
Comparing Different Genomes



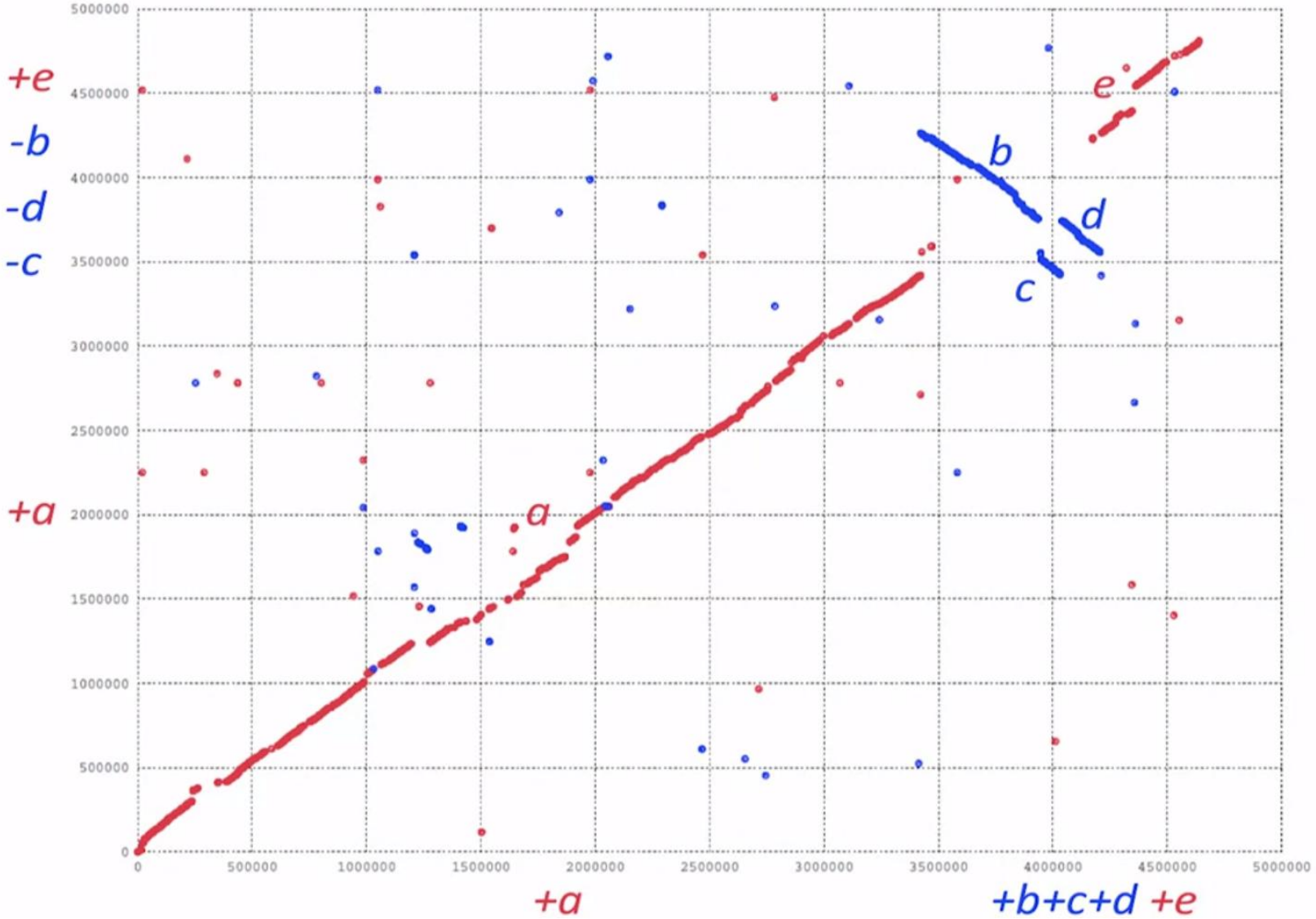
Identical and reverse complementary k-mers (k=3).

E. Coli versus *S. enterica*




E. Coli versus S. enterica
a
b
c
d
e

E. Coli versus *S. enterica*



Synteny Blocks as Diagonals in Genomic Dot Plot

Finding Synteny Blocks Problem. Find diagonals in the genomic dot-plot.

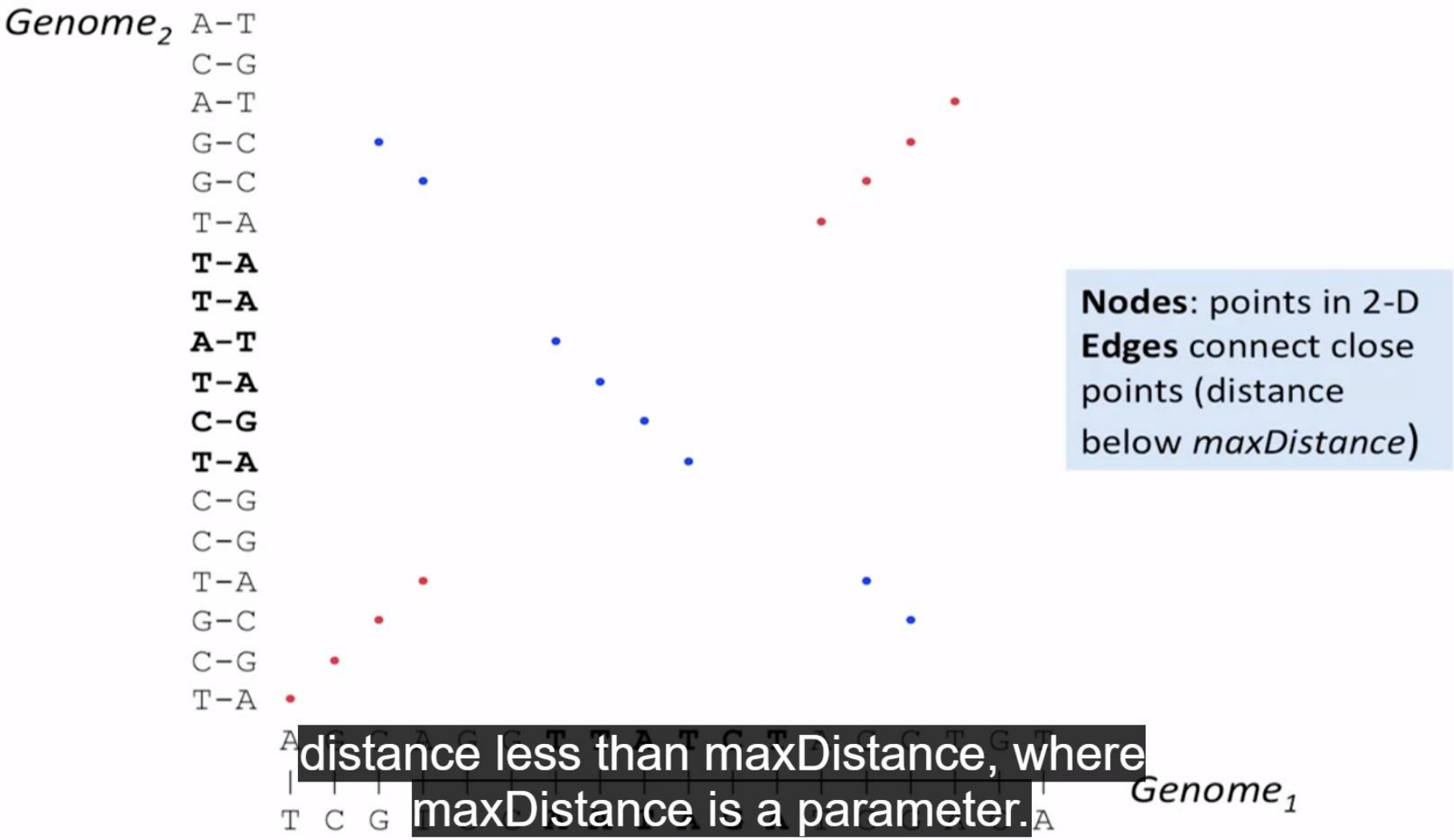
- **Input.** A set of points *DotPlot* in 2-D.
- **Output.** A set of diagonals in *DotPlot* representing synteny blocks.

Synteny Blocks as Diagonals in Genomic Dot Plot

Finding Synteny Blocks Problem. Find diagonals in the genomic dot-plot.

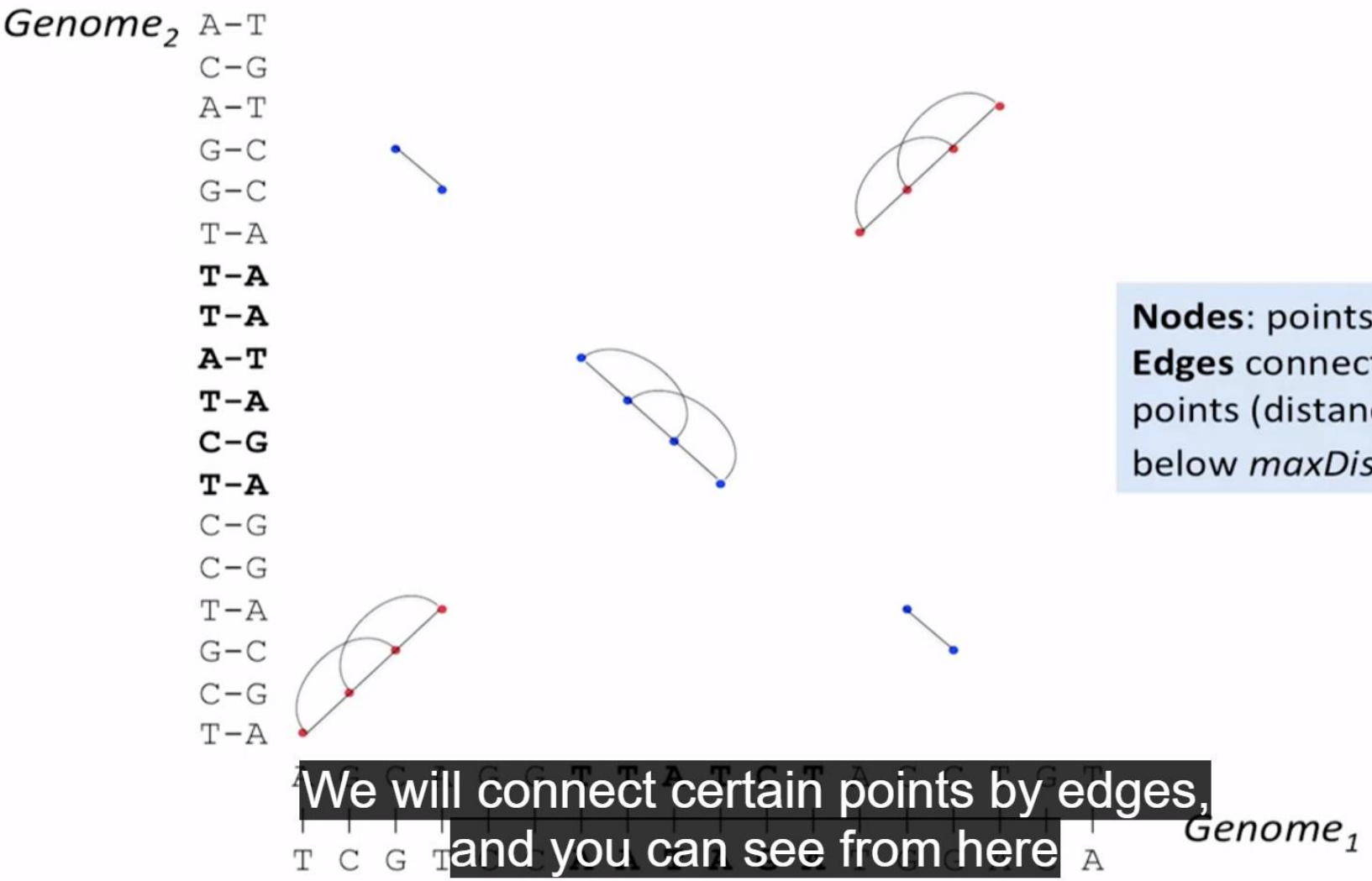
- **Input.** A set of points *DotPlot* in 2-D.
- **Output.** A set of diagonals in *DotPlot* representing synteny blocks.

Connecting Closely Located Points in Genomic Dot-Plot



distance less than maxDistance, where maxDistance is a parameter.

Connecting Closely Located Points in Genomic Dot-Plot



Nodes: points in 2-D
Edges connect close points (distance below *maxDistance*)

We will connect certain points by edges, and you can see from here

Synteny Block Generation Algorithm

Synteny(*DotPlot,maxDistance,minSize*)

maxDistance: gap size

minSize: minimum synteny block size

- Form a graph whose node set is the set of points in *DotPlot*
- Connect two nodes by an edge if the 2-D distance between them is < *maxDistance.* The connected components in the resulting graph define synteny blocks
- Delete small synteny blocks (length < *minSize*)

Let's form a graph whose node set is the set of points in DotPlot, and

Two Synteny Block Generation Algorithms: Which One is Better?

Synteny(*DotPlot,maxDistance,minSize*)

- Form a graph whose node set is the set of points in *DotPlot*
- Connect two nodes by an edge if the 2-D distance between them is < *maxDistance.* The connected components in the resulting graph define synteny blocks
- Delete small synteny blocks (length < *minSize*)

Amalgamate(*DotPlot,maxDistance,minSize*)

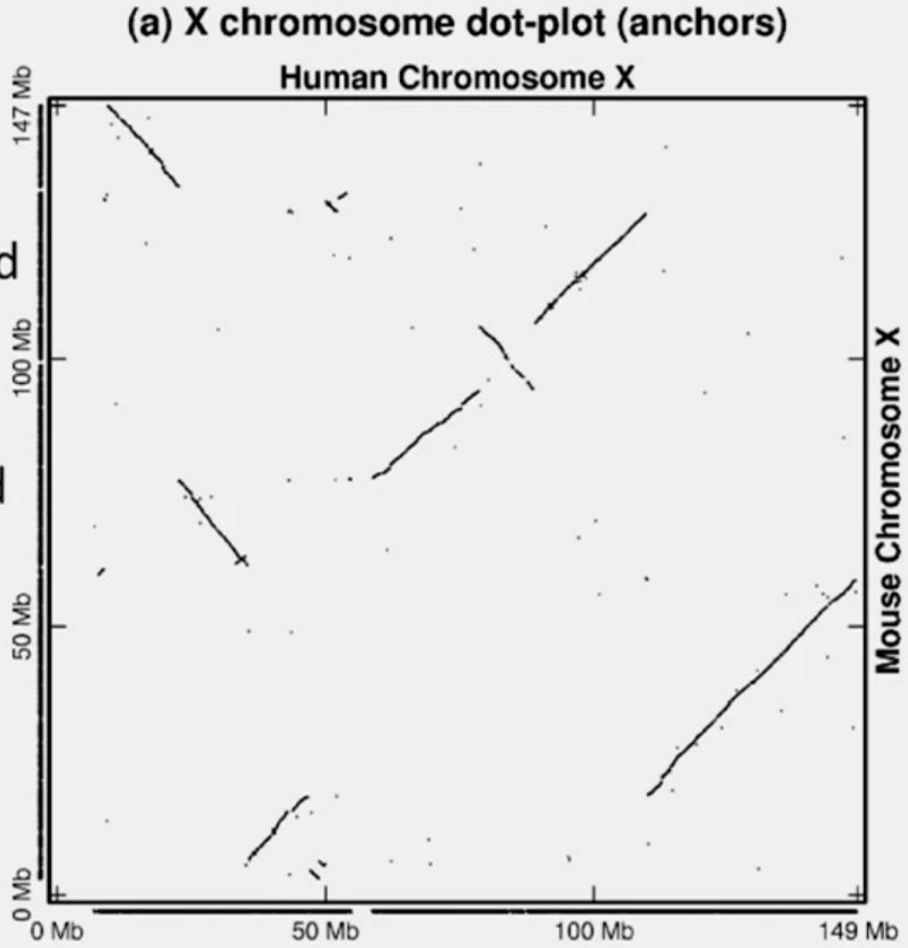
- Define each point in *DotPlot* as a separate block and iteratively amalgamate the resulting blocks
- Amalgamate two blocks if they contain two points that are separated by < *maxDistance* in another genome.
- Delete small synteny blocks (length < *minSize*)

to solve, because I have never defined an

Genomic Dot Plot (Human vs. Mouse X Chromosome)

≈25,000 anchors (regions of shared similarity).

Anchors enlarged for visibility.



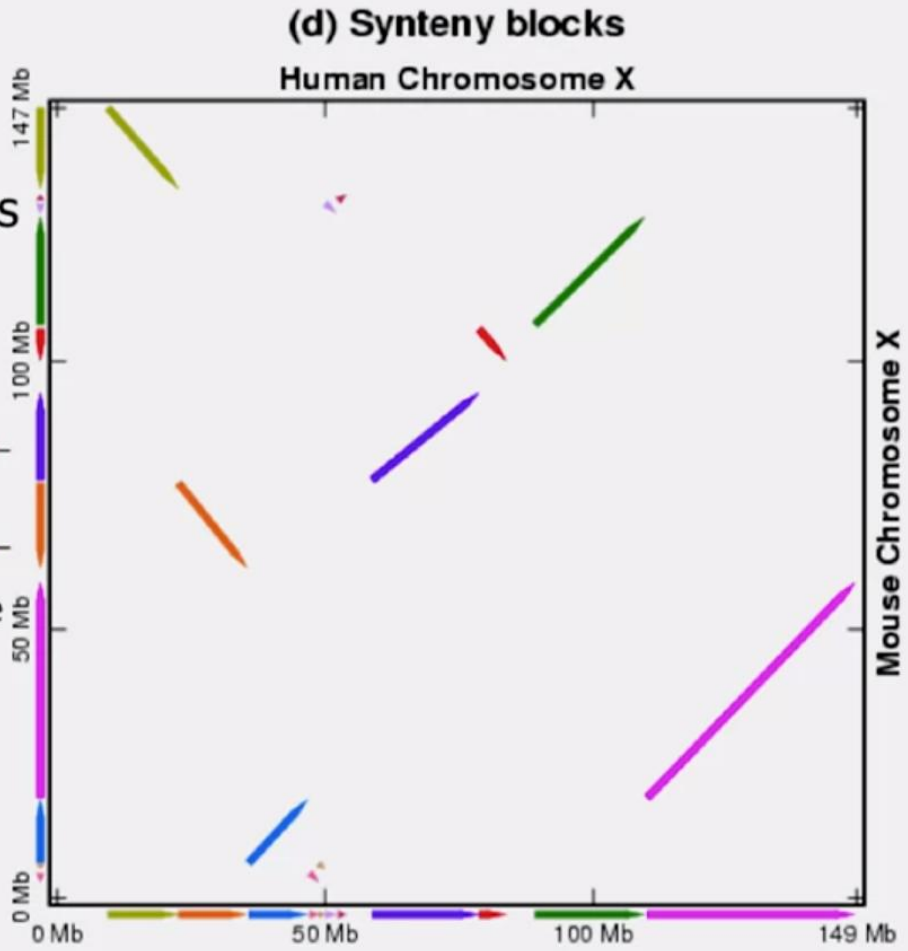

Synteny Blocks in 2-D
11 synteny blocks
(d) Synteny blocks
Human Chromosome X
Mouse Chromosome X
147 Mb
100 Mb
50 Mb
0 Mb
0 Mb
50 Mb
100 Mb
149 Mb

Synteny Blocks in 2-D

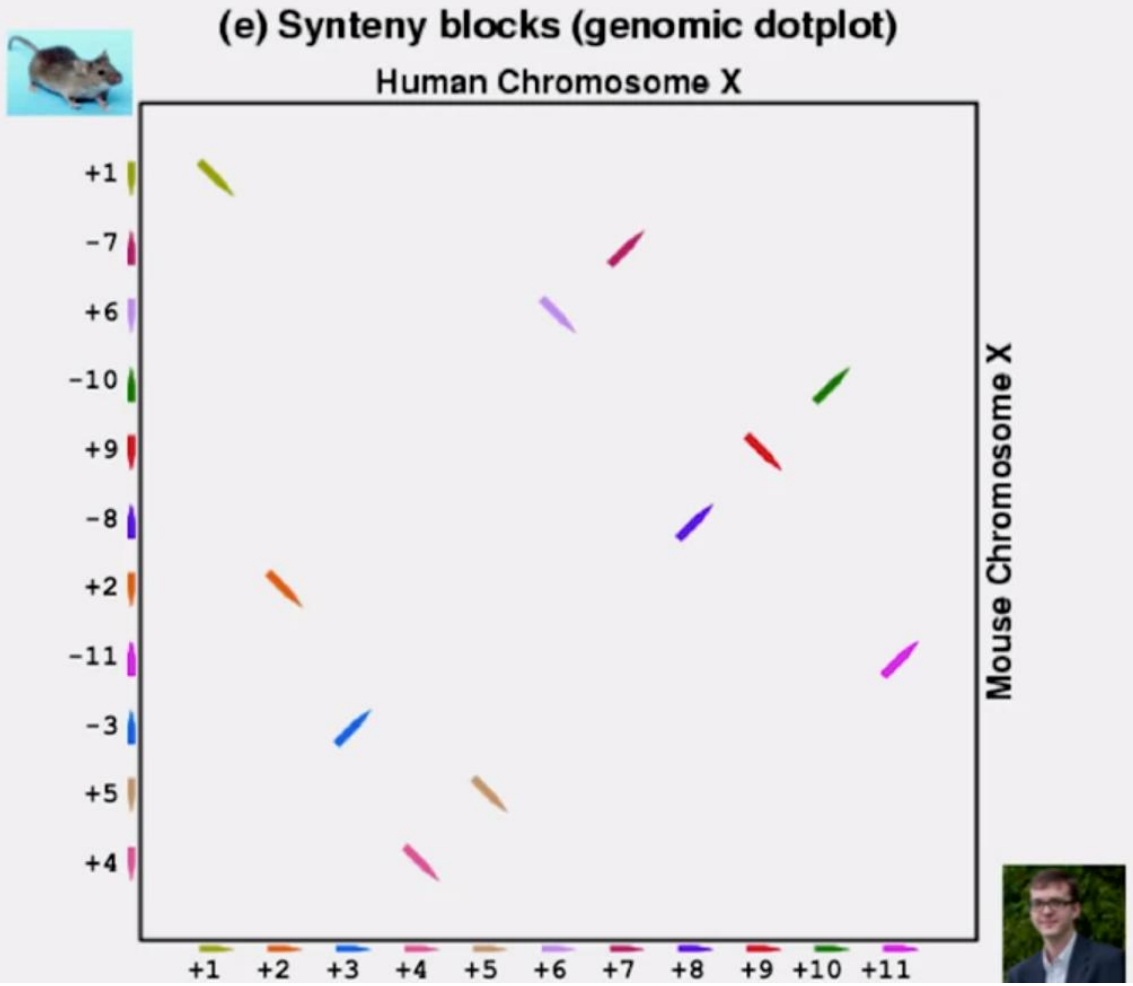
11 synteny blocks

Mouse X chromosome

Human X chromosome



Ignoring the Length of the Synteny Blocks



Constructing Synteny Blocks and Reconstructing Rearrangement History of All Mammals