Combinatorial Problems and Algorithms in Comparative Genomics

Max Alekseyev

University of South Carolina, Columbia, SC, U.S.A.

2011

Лаборатория Алгоритмической Биологии

- Организована Павлом Певзнером в январе 2011 года на базе Академического университета РАН
- Финансируется “мегагрантом” Министерства образования и науки РФ
- Вебсайт: http://bioinf.aptu.ru

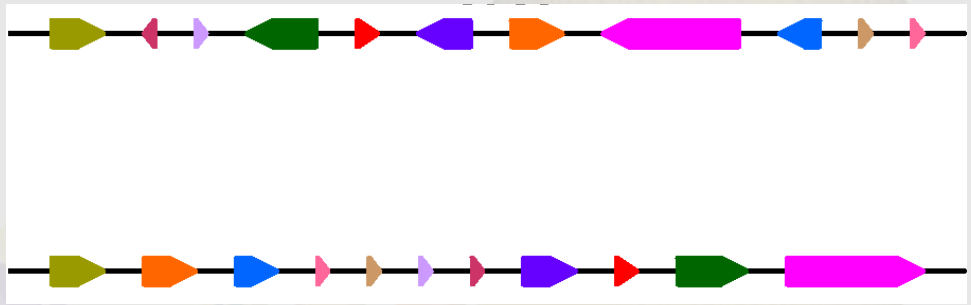
- Имеются исследовательские вакансии разных рангов (от старшекурсников до кандидатов наук).
  Требования к претендентам:
  - Наличие фундаментальной подготовки по математике и/или алгоритмике
  - Умение программировать на C++

Лаборатория Алгоритмической Биологии

- При участии Лаборатории в Академическом университете:
  - На кафедре *Математических и Информационных Технологий* открыт набор в магистратуру по *алгоритмической биоинформатике*
  - С осени 2011 года организуется аспирантура по направлению *биоинформатика*
- **7 мая с 11:00 до 12:30** в актовом зале Академического университета состоится лекция Павла Певзнера о *вычислительной протеомике*.

Genome Rearrangements

Mouse X chromosome

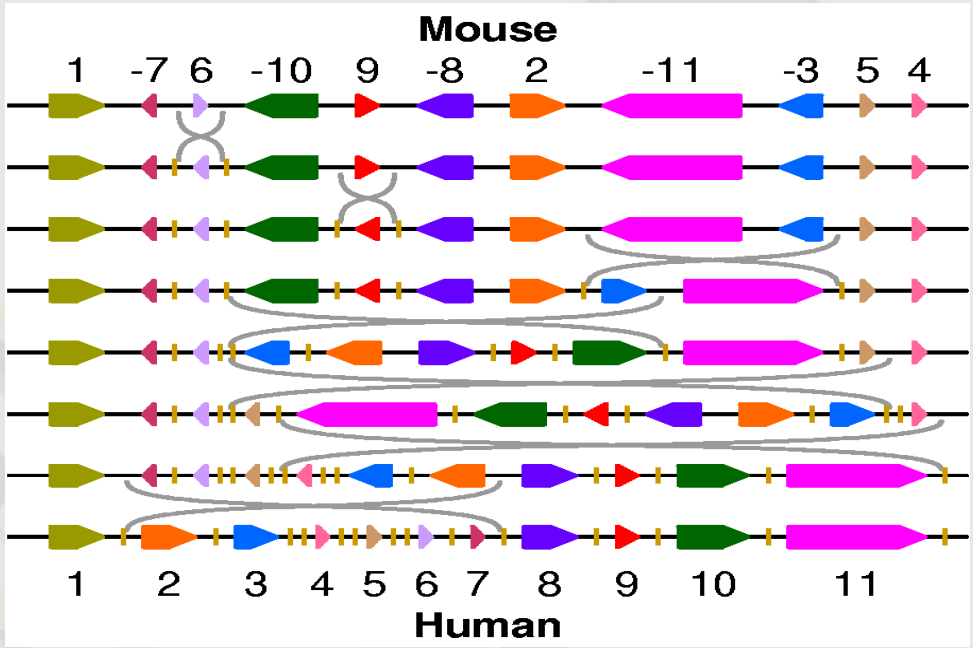


Unknown ancestor
~ 80 M years ago

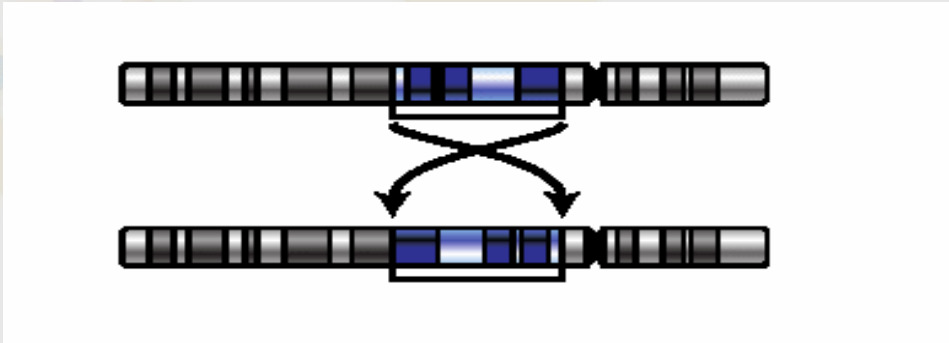
Human X chromosome

Genome Rearrangements: Evolutionary Scenarios

Unknown ancestor
~ 80 M years ago



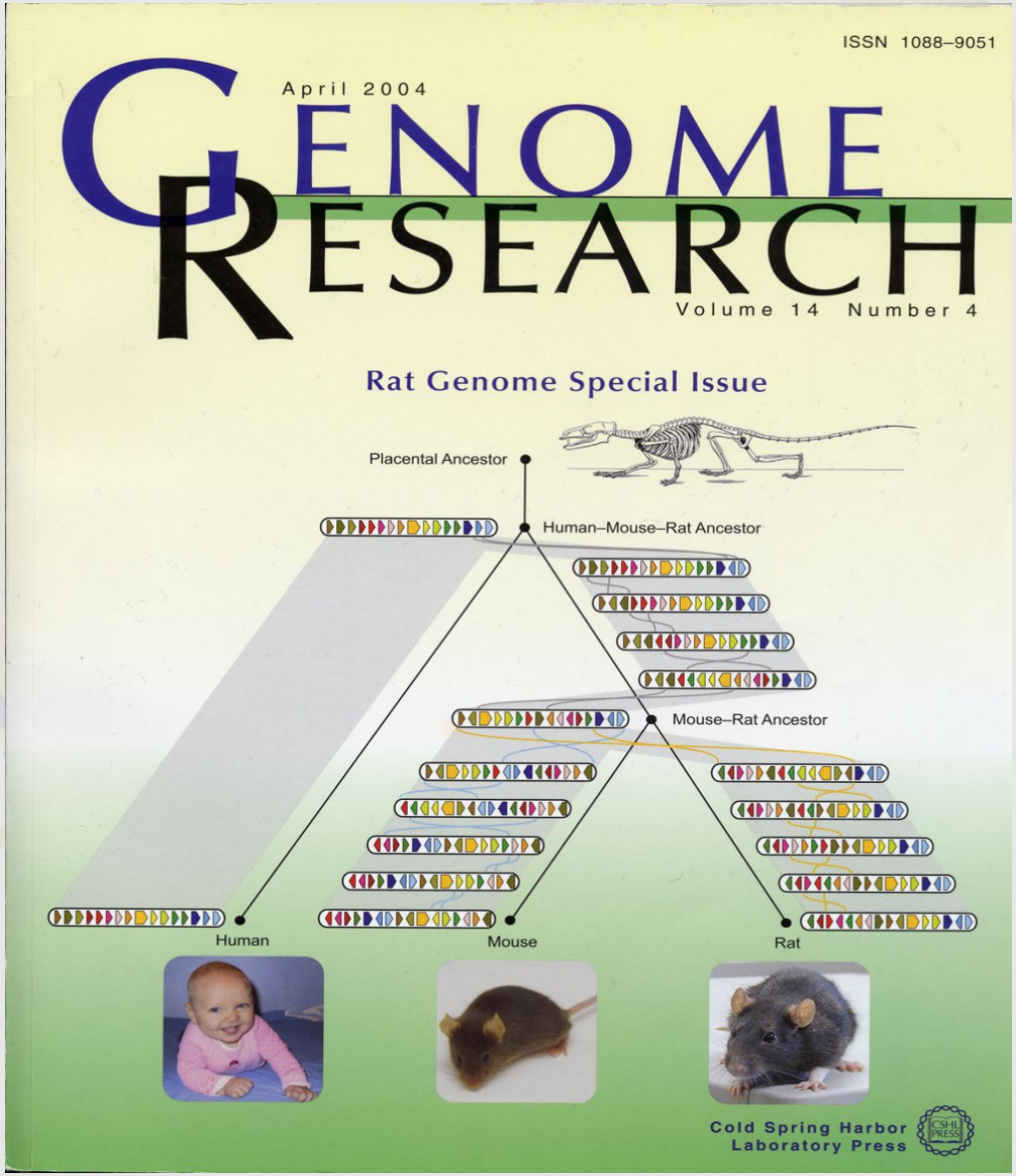
- ✔ What is the evolutionary scenario for transforming one genome into the other?
- ✔ What is the organization of the ancestral genome?
- ✔ Are there any rearrangement hotspots in mammalian genomes?



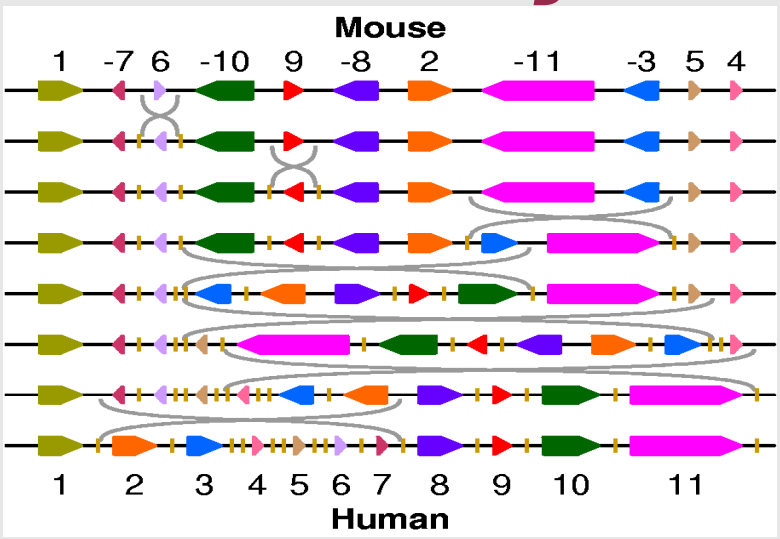
Reversal (inversion) flips a segment of a chromosome

Genome Rearrangements: Ancestral Reconstruction

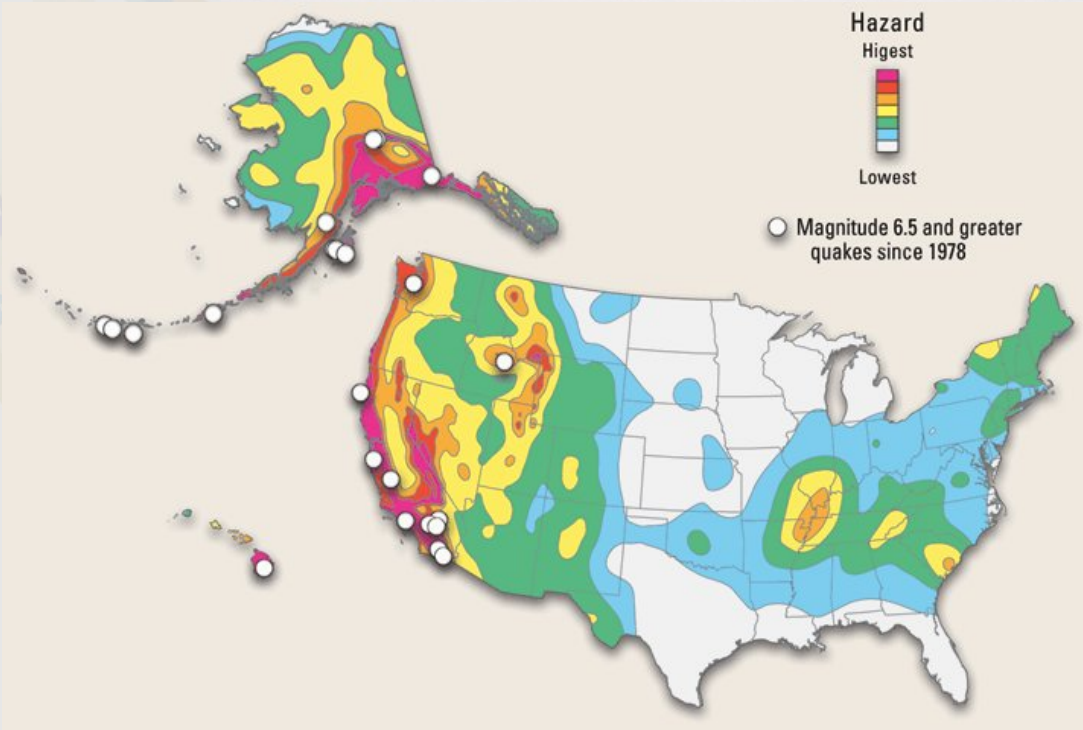
- ✔ What is the evolutionary scenario for transforming one genome into the other?
- ✔ What is the organization of the ancestral genome?
- ✔ Are there any rearrangement hotspots in mammalian genomes?



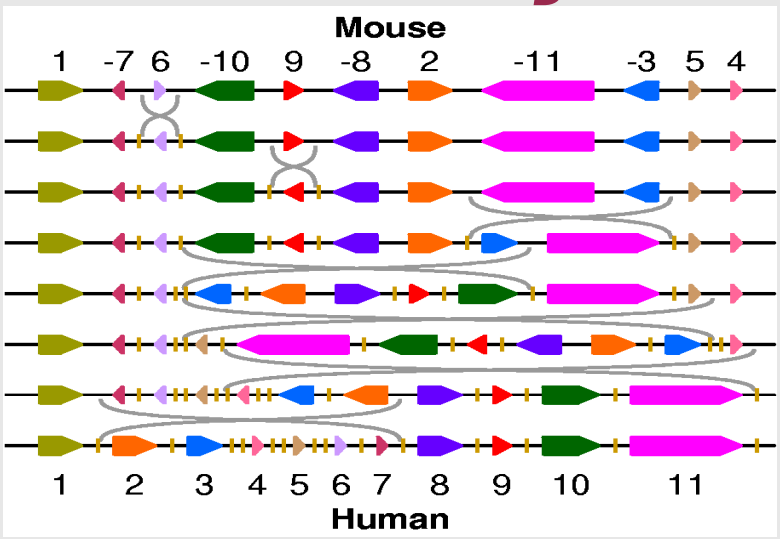
Genome Rearrangements: Evolutionary "Earthquakes"



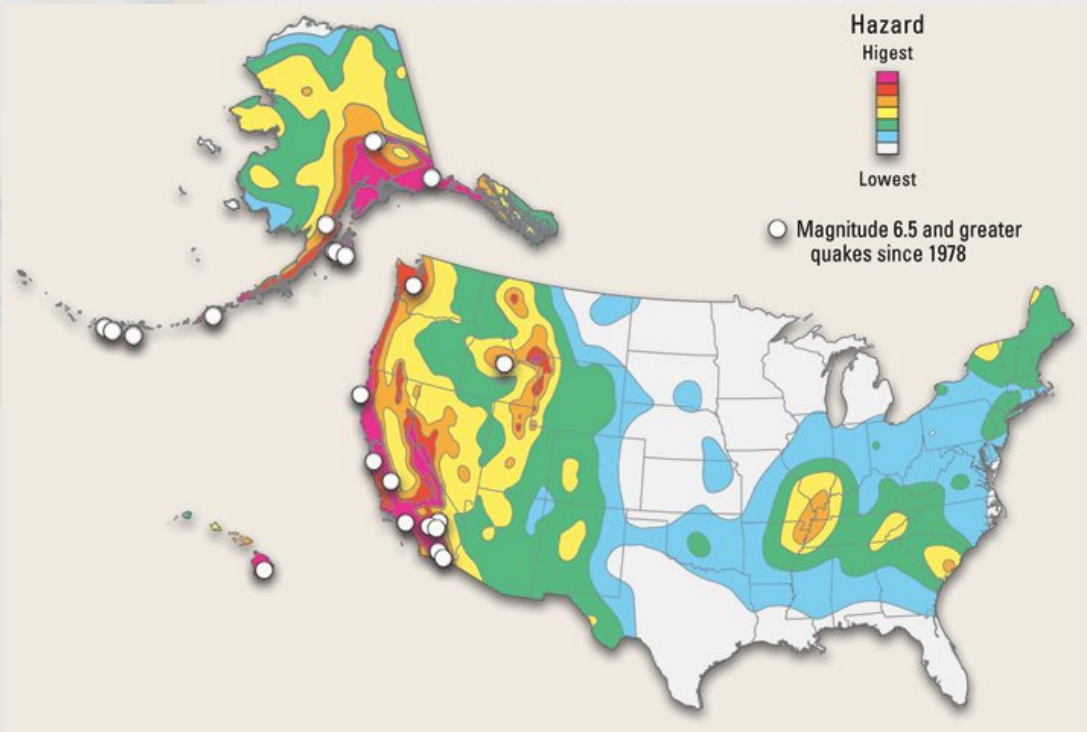
- ✔ What is the evolutionary scenario for transforming one genome into the other?
- ✔ What is the organization of the ancestral genome?
- ✔ Are there any rearrangement hotspots in mammalian genomes? (controversy in 2003-2008)



Genome Rearrangements: Evolutionary "Earthquakes"

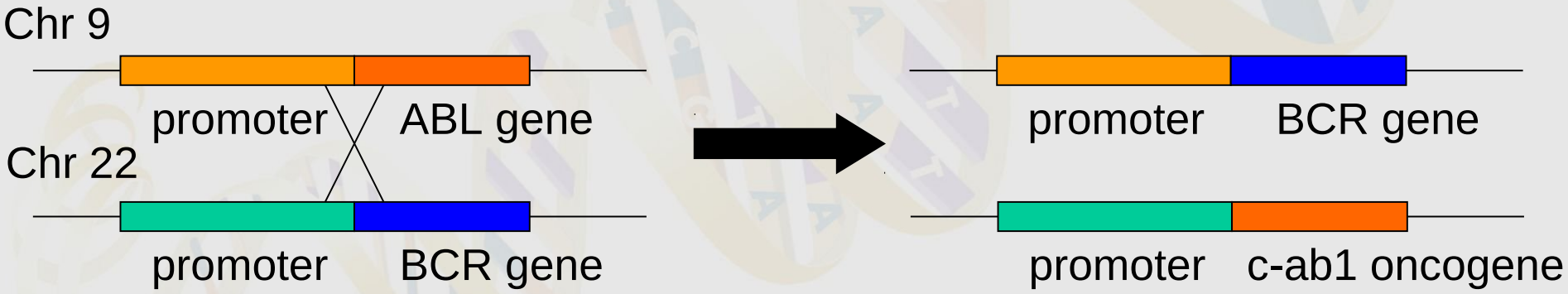


- ✔ What is the evolutionary scenario for transforming one genome into the other?
- ✔ What is the organization of the ancestral genome?
- ✔ **Where are the rearrangement hotspots in mammalian genomes?**



Rearrangement Hotspots in Tumor Genomes

- ✔ Rearrangements may disrupt genes and alter gene regulation.
- ✔ Example: rearrangement in leukemia yields "Philadelphia" chromosome:



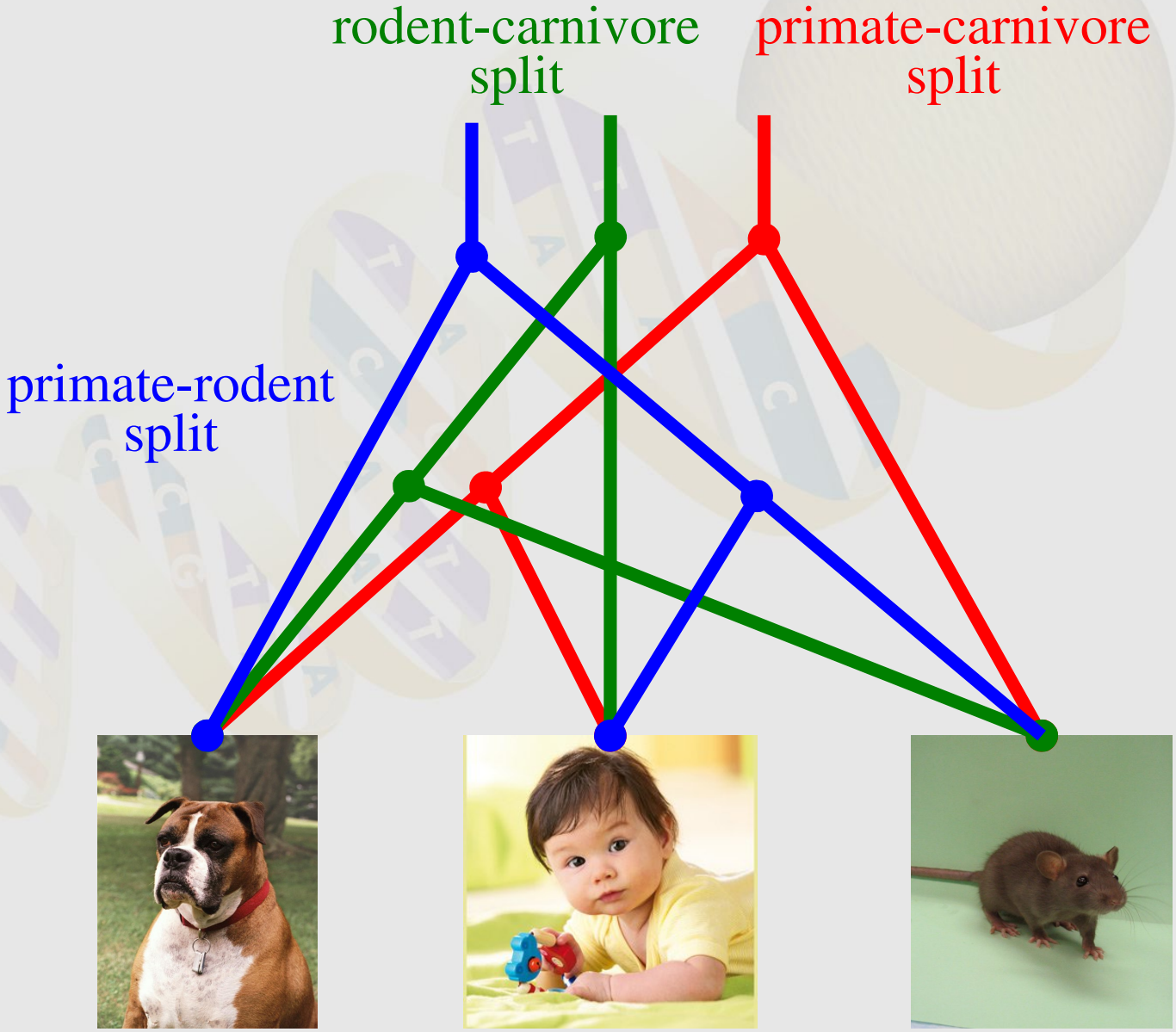
- ✔ Thousands of individual rearrangements hotspots known for different tumors.

Biological Problem:

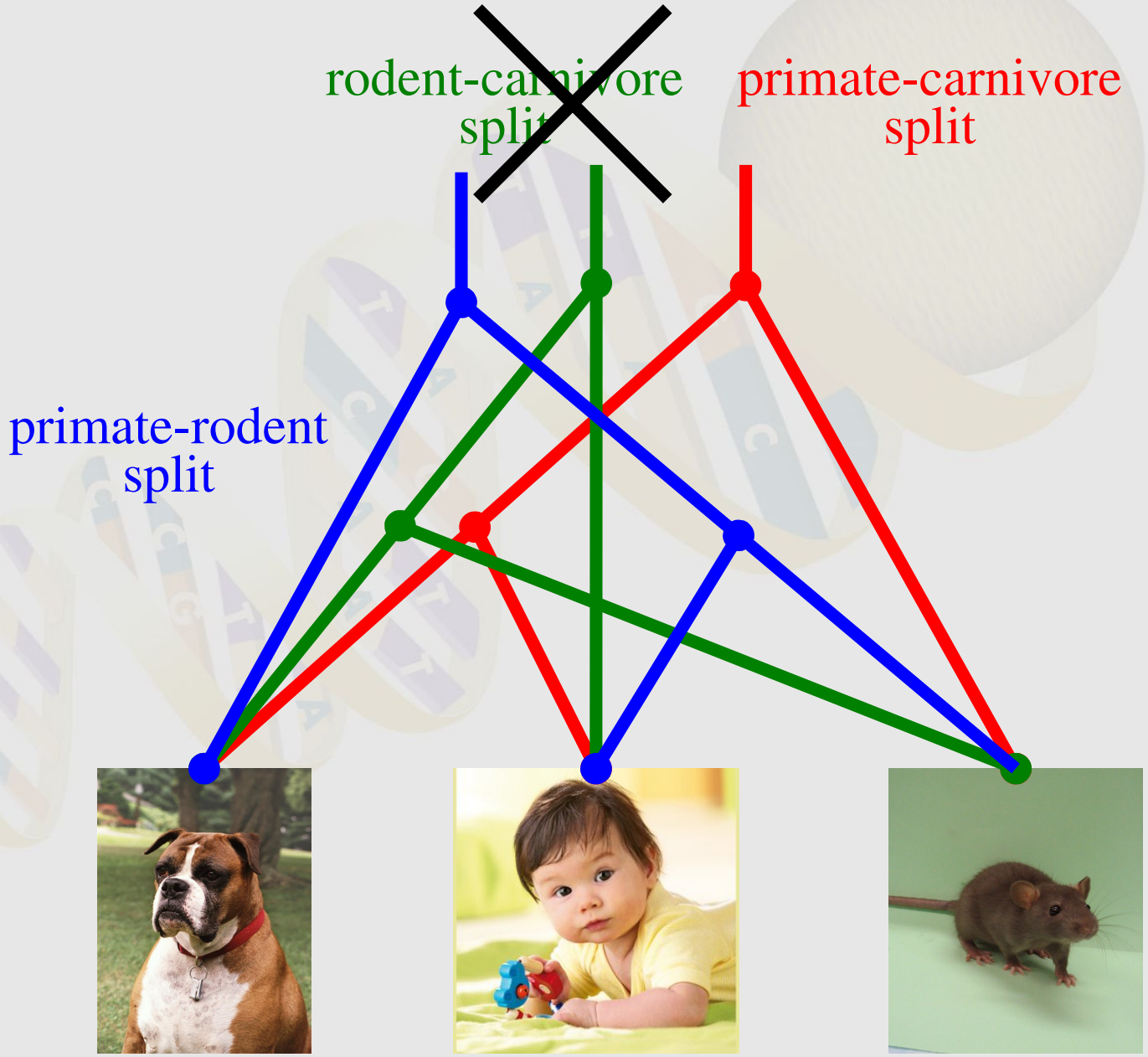
Who are evolutionary closer to humans: mice or dogs?

Who is “Closer” to Us: Mouse or Dog?

Primate – Rodent – Carnivore Split



Primate – Rodent – Carnivore Split



Primate–Rodent vs. Primate–Carnivore Split

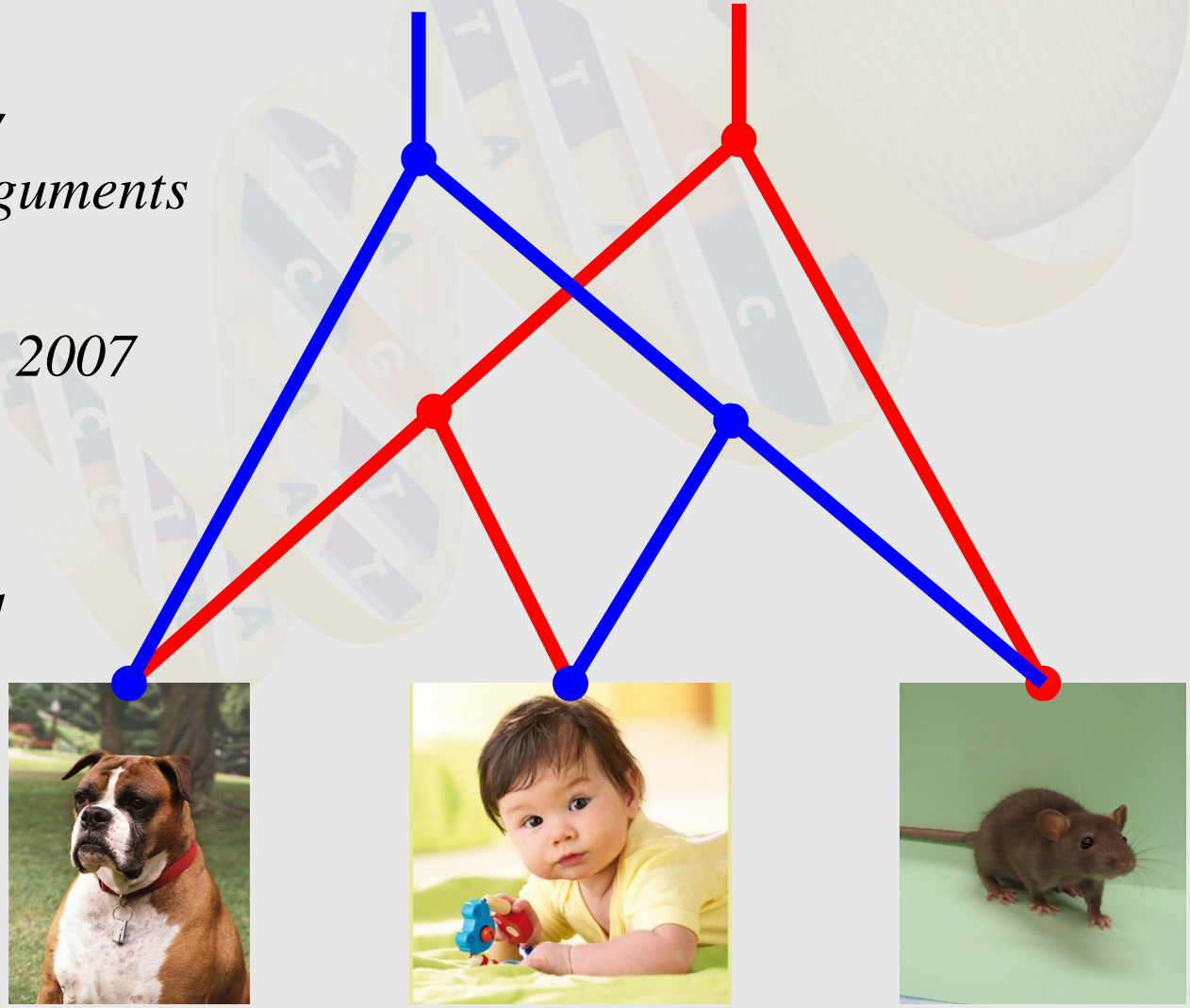
July 2007 and up
new papers supporting the primate-carnivore split

April 2007
Lunter et al., PLoS CB 2007 refuted Cannarozzi et al. arguments

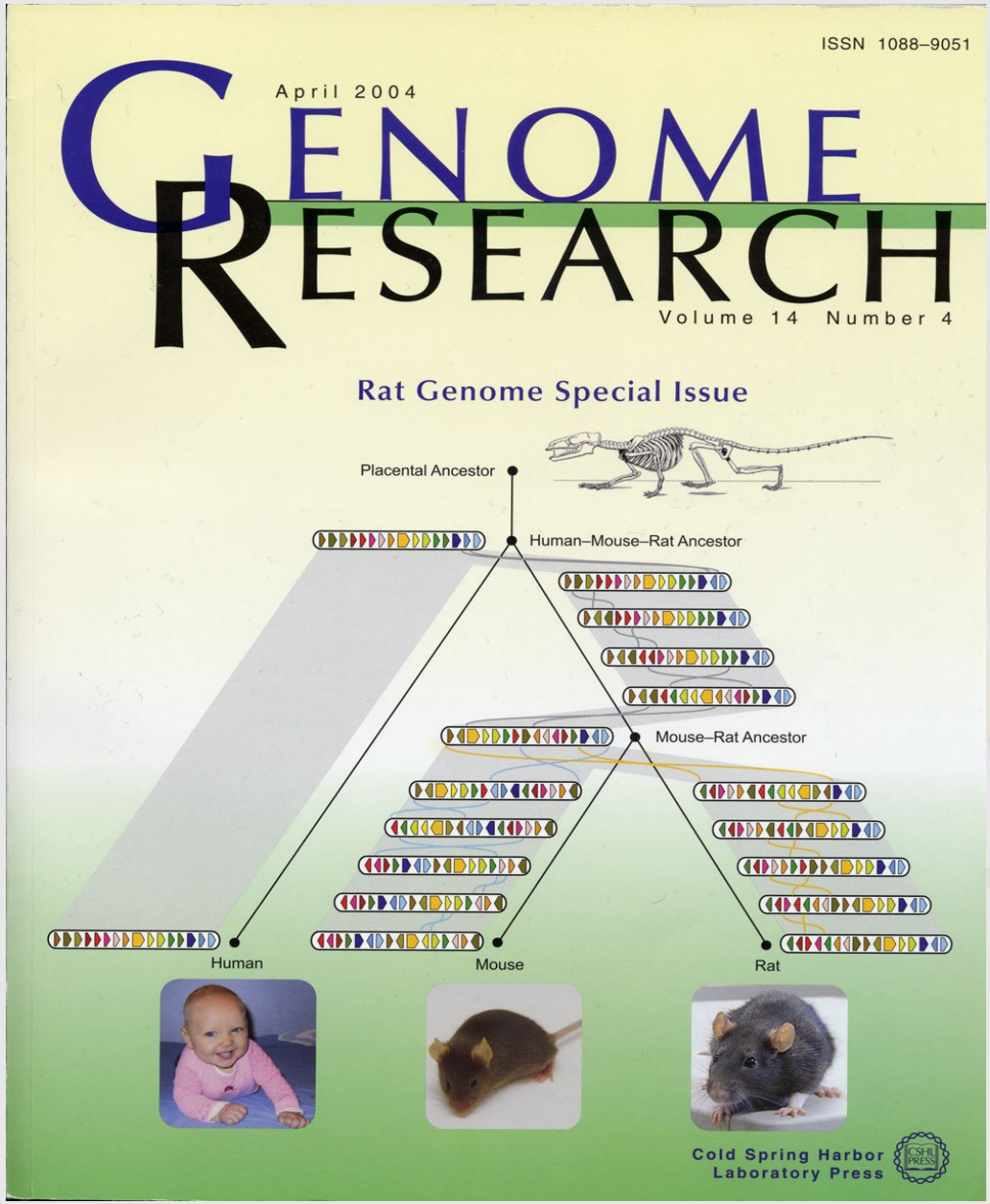
January 2007
Cannarozzi et. al., PLoS CB 2007 argued for the primate-carnivore split

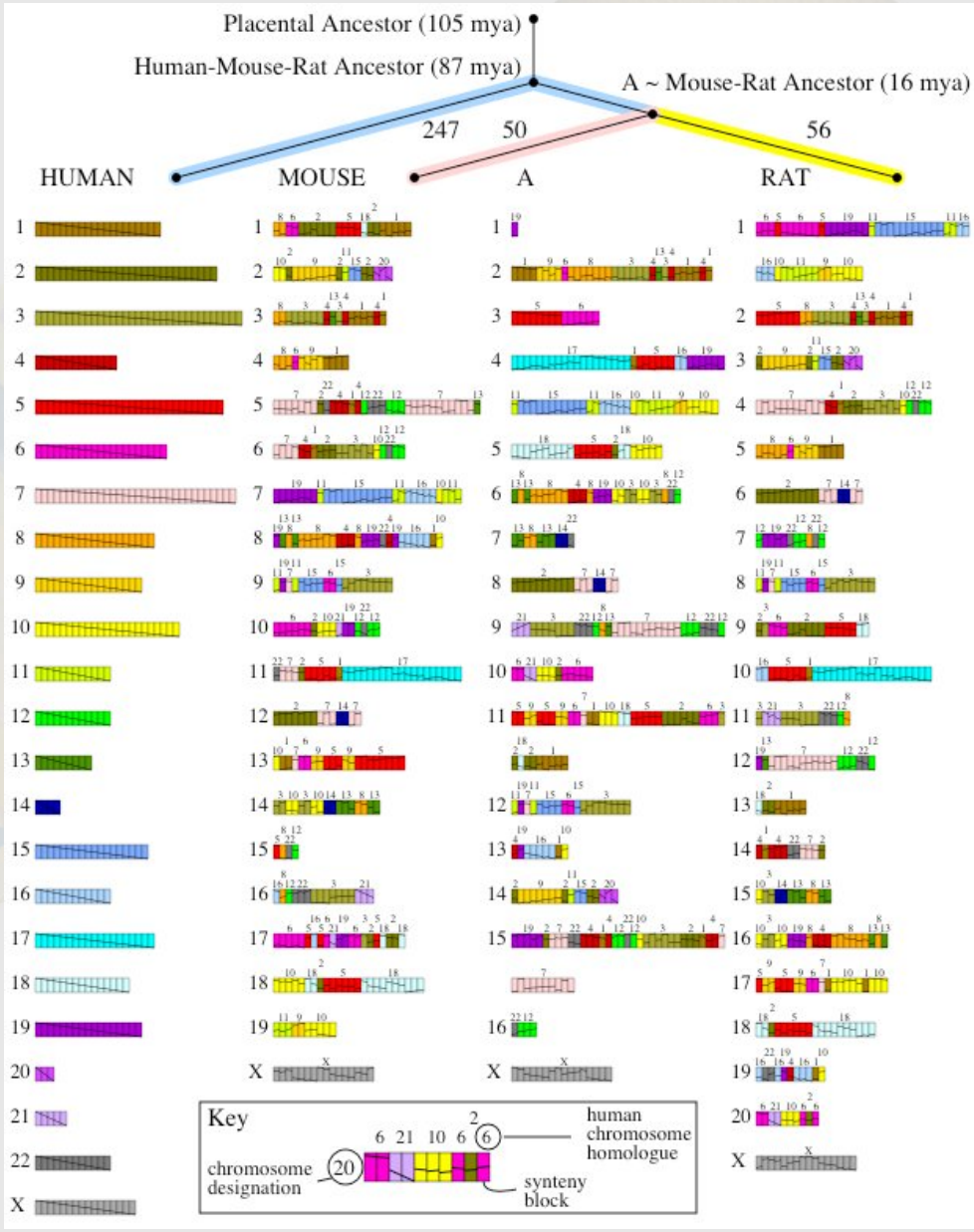
2001
Murphy et. al., Science 2001 set a new dominant view: the primate-rodent split

before 2001
most biologists believed in the primate-carnivore split

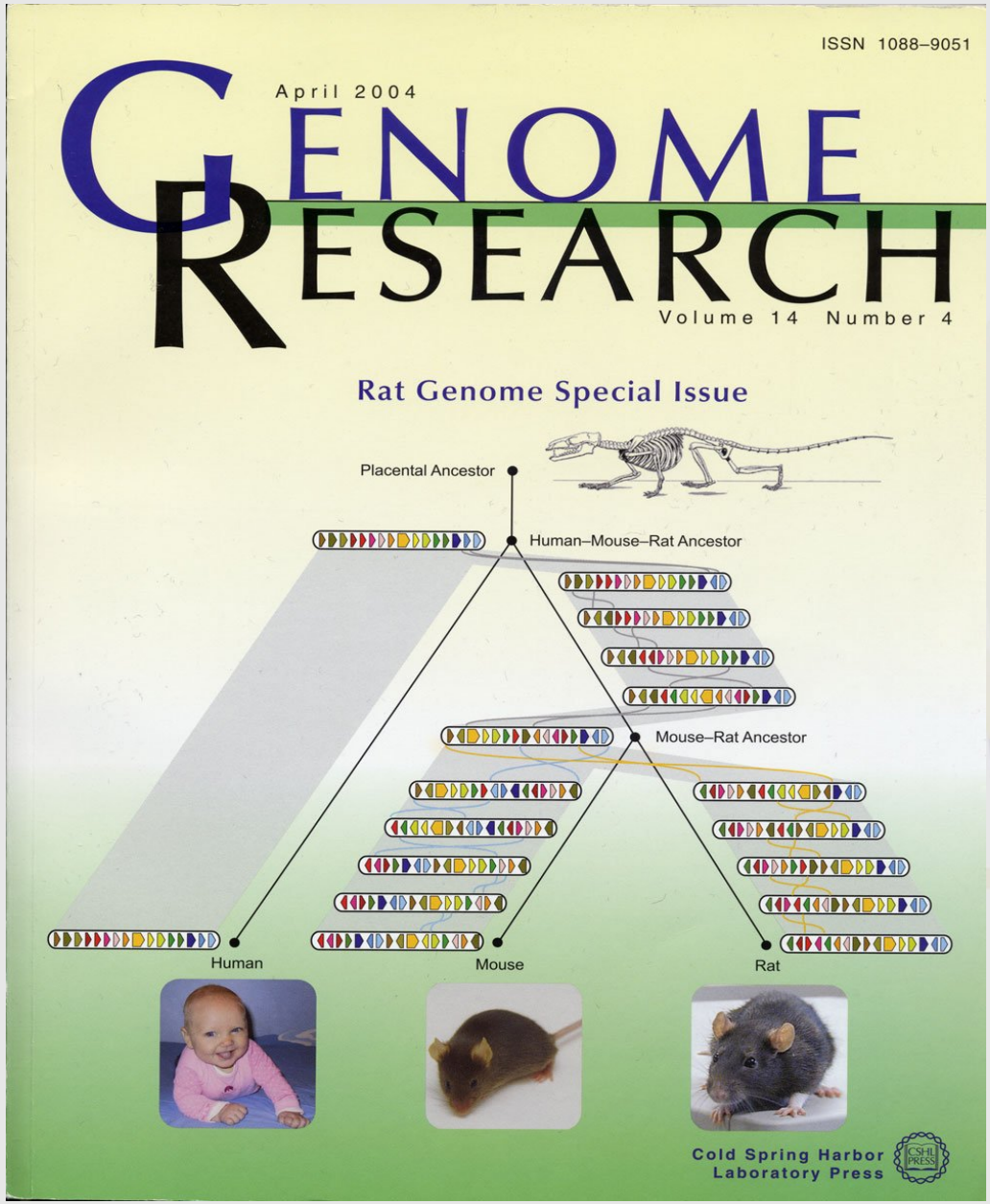


Reconstruction of Ancestral Genomes: Human / Mouse / Rat





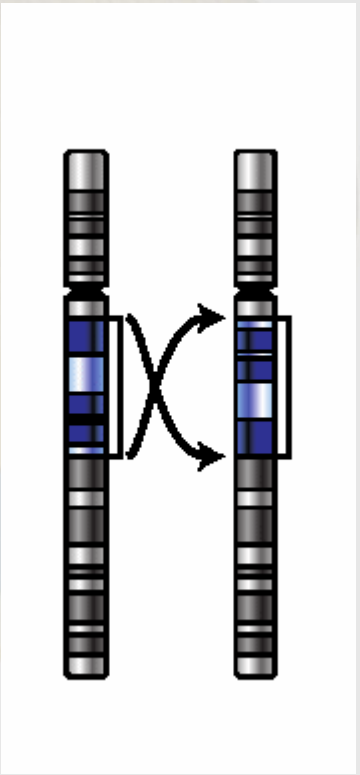
Reconstruction of MANY Ancestral Genomes: Can It Be Done?



Algorithmic Background:

Genome Rearrangements and Breakpoint Graphs

Unichromosomal Genomes: Reversal Distance



- A *reversal* flips a segment of a chromosome.
- For given genomes P and Q, the number of reversals in a shortest series, transforming one genome into the other, is called the **reversal distance** between P and Q.
- Hannenhalli and Pevzner (*FOCS 1995*) gave a polynomial-time algorithm for computing the reversal distance.

Prefix Reversals

- A *prefix reversal* flips a prefix a permutation.
- Pancake Flipping Problem: sort a given stack (permutation) of pancakes of different sizes with the minimum number of flips of any number of top pancakes.

Discrete Mathematics 27 (1979) 47–57.

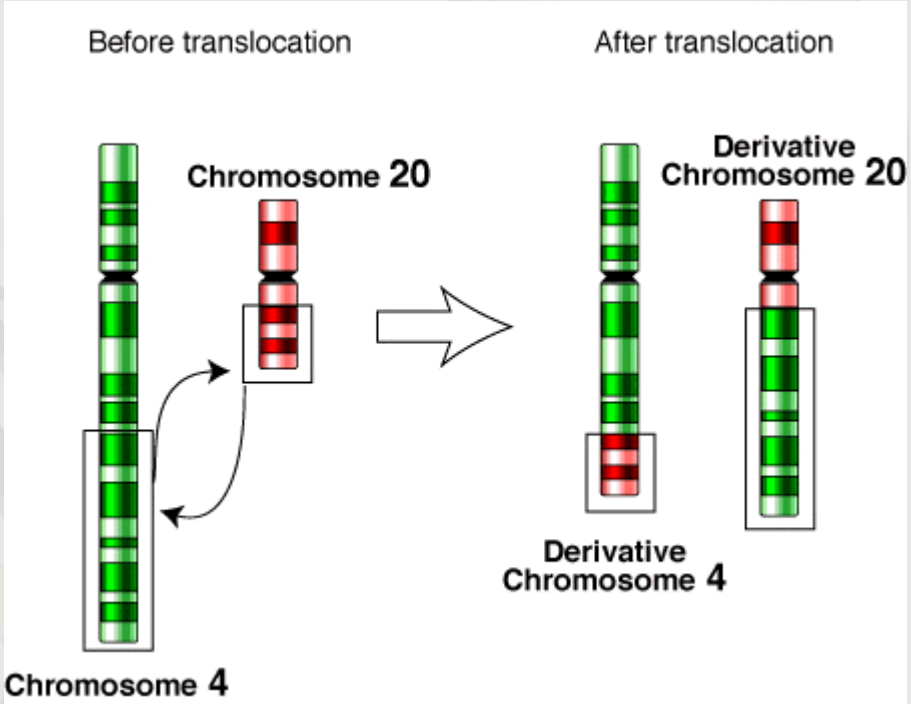
BOUNDS FOR SORTING BY PREFIX REVERSAL

William H. GATES
Microsoft, Albuquerque, New Mexico

Christos H. PAPADIMITRIOU*†
Department of Electrical Engineering, University of California, Berkeley, CA 94720, U.S.A.

Multichromosomal Genomes: Genomic Distance

- **Genomic Distance** between two genomes is the minimum number of *reversals*, *translocations*, *fusions*, and *fissions* required to transform one genome into the other.



- Hannenhalli and Pevzner (STOC 1995) extended their algorithm for computing the reversal distance to computing the genomic distance.

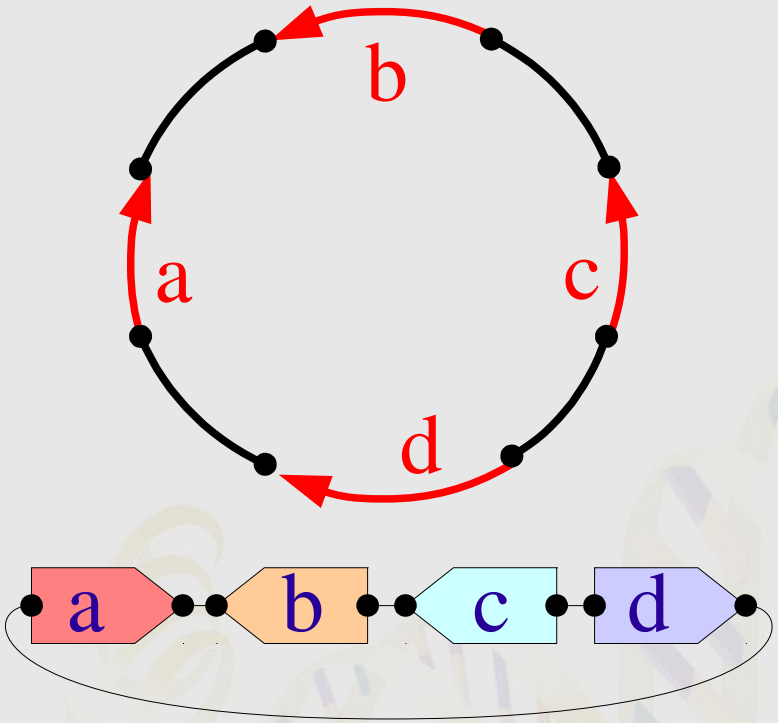
- These algorithms were followed by many improvements: *Kaplan et al. 1999, Bader et al. 2001, Tesler 2002, Ozery-Flato & Shamir 2003, Tannier & Sagot 2004, Bergeron 2001-07, etc.*

HP Theory Is Rather Complicated: Is There a Simpler Alternative?

- ✔ HP theory is a key tool in most genome rearrangement studies. However, it is rather complicated that makes it difficult to apply in complex setups.

- ✔ To study genome rearrangements in multiple genomes, we use ***2-break*** rearrangements, also known as DCJ (*Yancopoulus et al., Bioinformatics 2005*).

Simplifying HP Theory: Switch from Linear to Circular Chromosomes

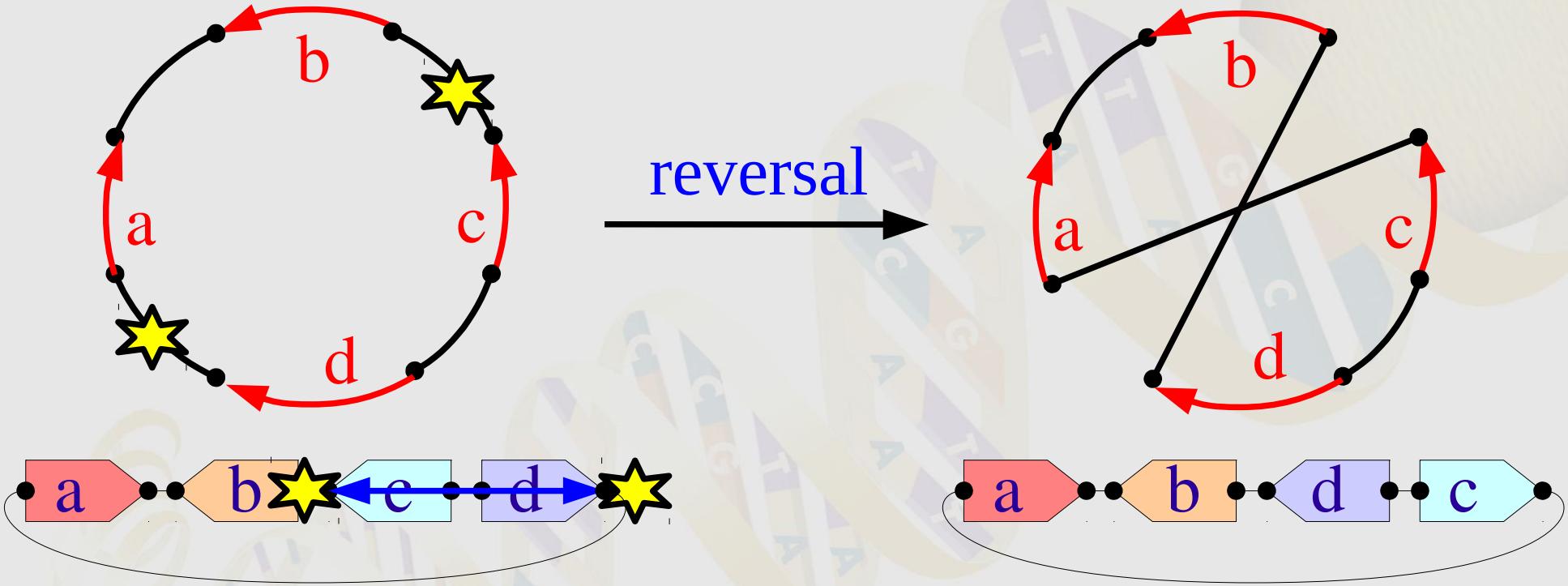


A chromosome can be represented as a *cycle* with *directed red* and ***undirected black*** edges, where:

red edges encode blocks and their directions;

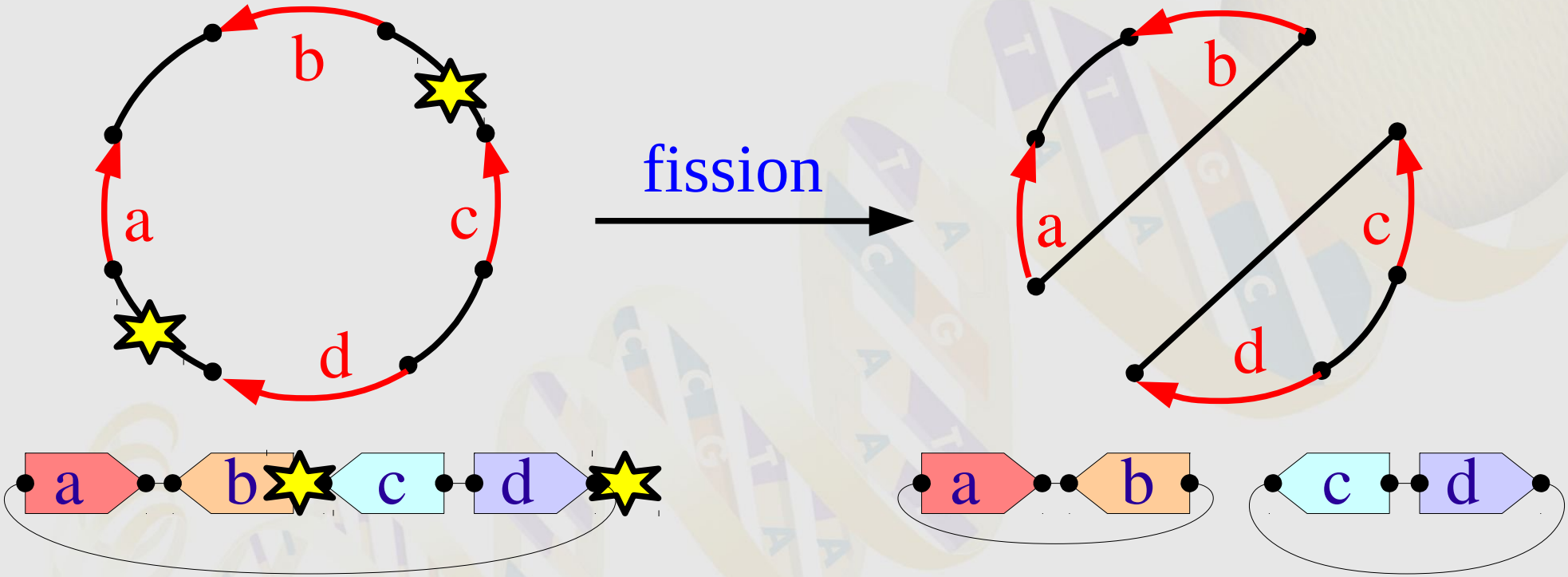
adjacent blocks are connected with black edges.

Reversals on Circular Chromosomes



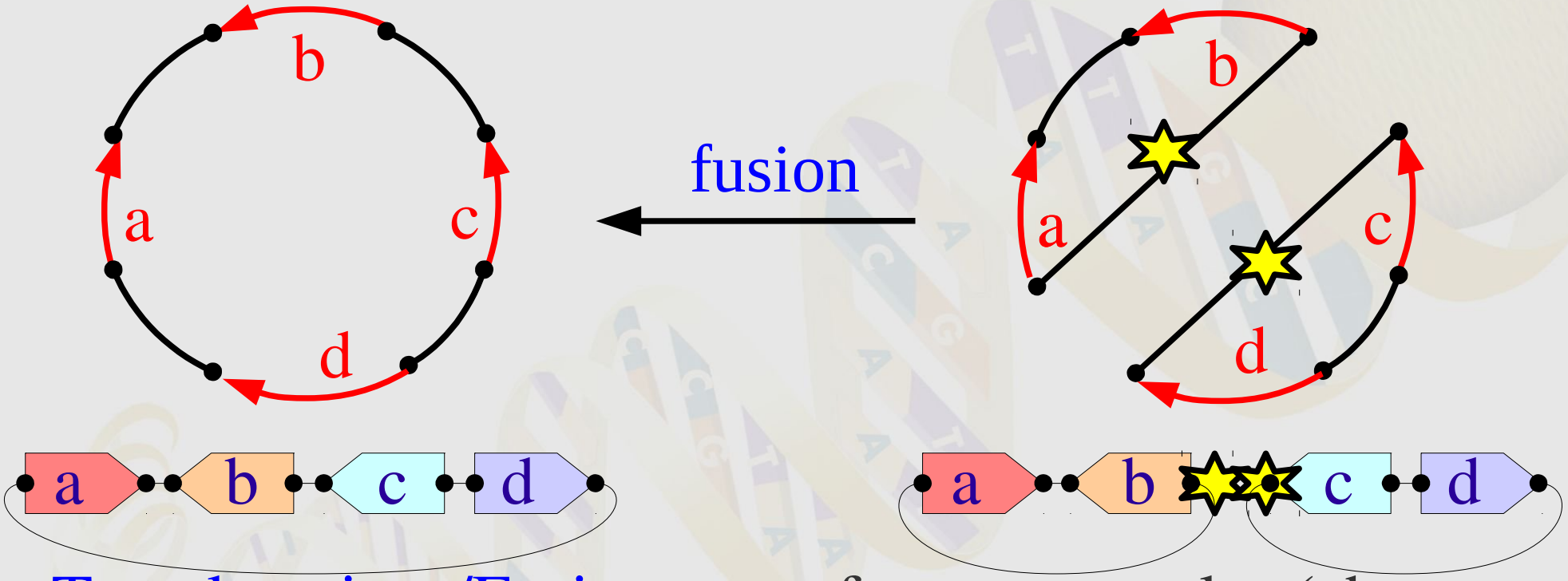
Reversals replace two black edges with two other black edges

Fissions



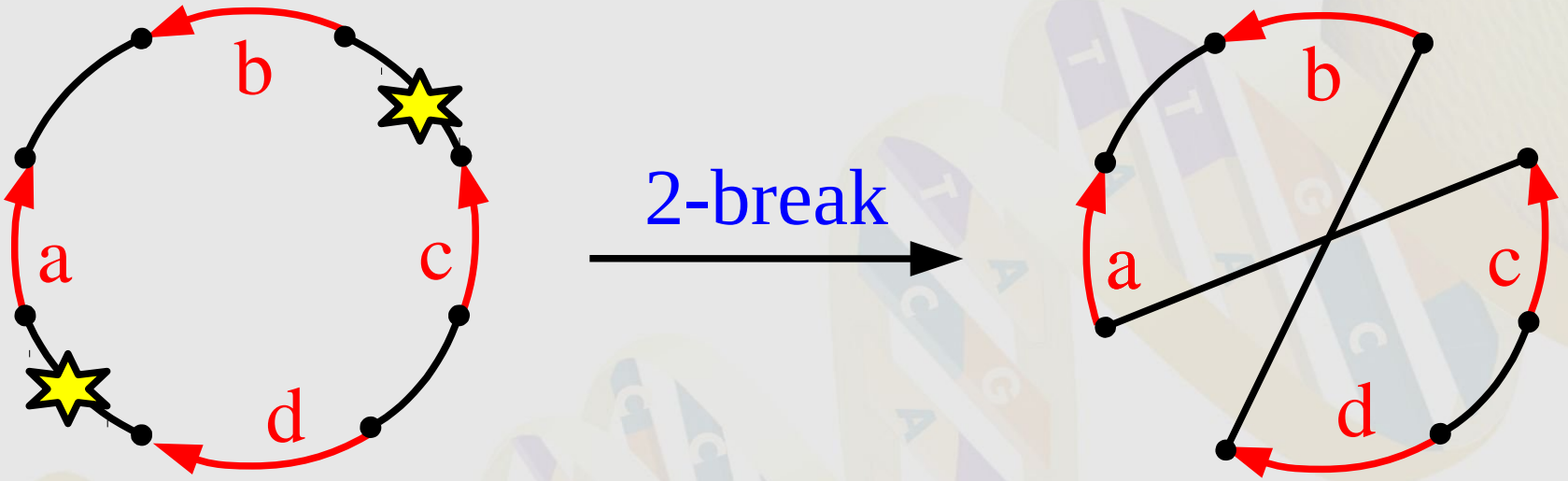
- ✔ Fissions split a single cycle (chromosome) into two.
- ✔ Fissions replace two black edges with two other black edges.

Translocations / Fusions



- ✔ Translocations/Fusions transform two cycles (chromosomes) into a single one.
- ✔ They also replace two black edges with two other black edges.

2-Breaks



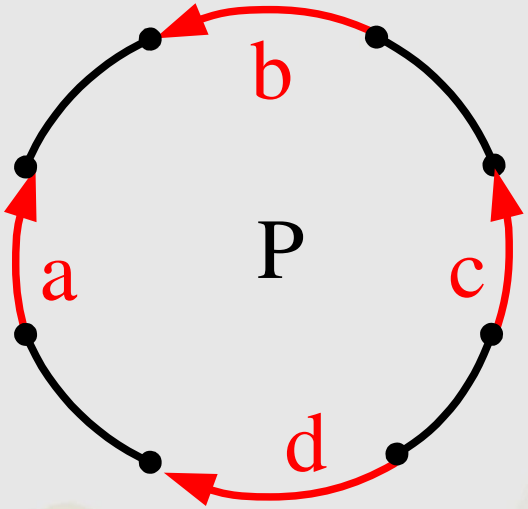
- ✔ 2-Break replaces *any pair* of black edges with another pair forming matching on the same 4 vertices.
- ✔ Reversals, translocations, fusions, and fissions represent all possible types of 2-breaks.

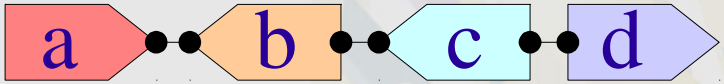
2-Break Distance

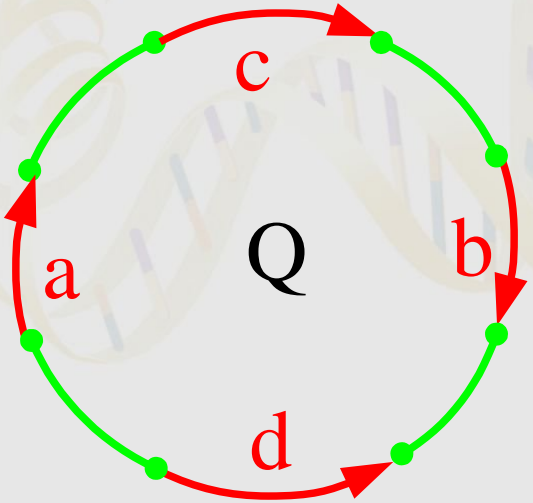
- The **2-Break distance** ***dist(P,Q)*** between genomes ***P*** and ***Q*** is the minimum number of 2-breaks required to transform ***P*** into ***Q***.

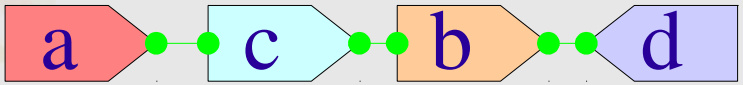
- In contrast to the genomic distance, the 2-break distance is easy to compute.

Two Genomes as Black-Red and Green-Red Cycles

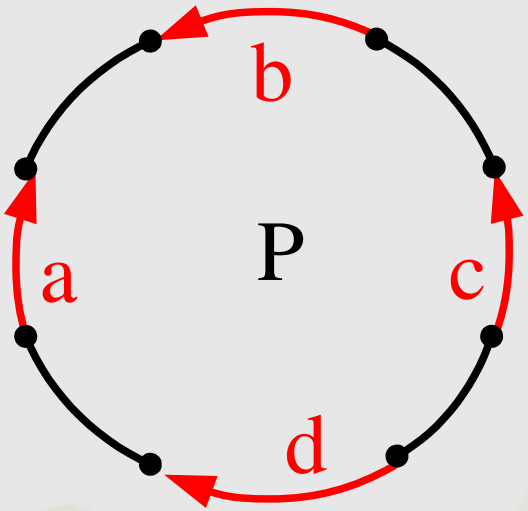


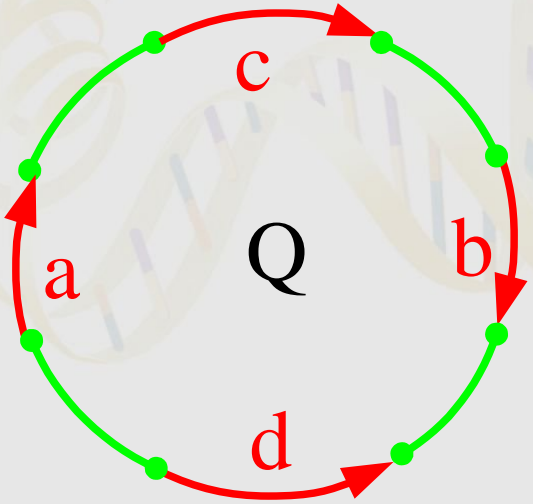


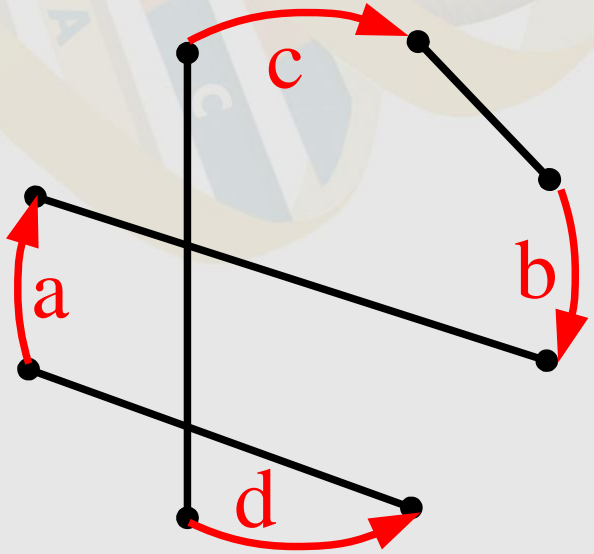




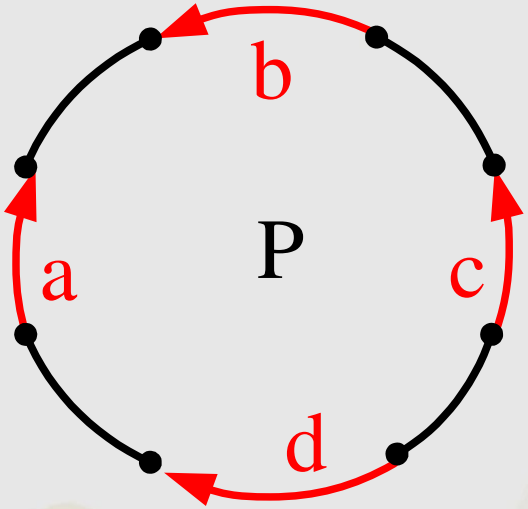
Rearranging P in the Q order

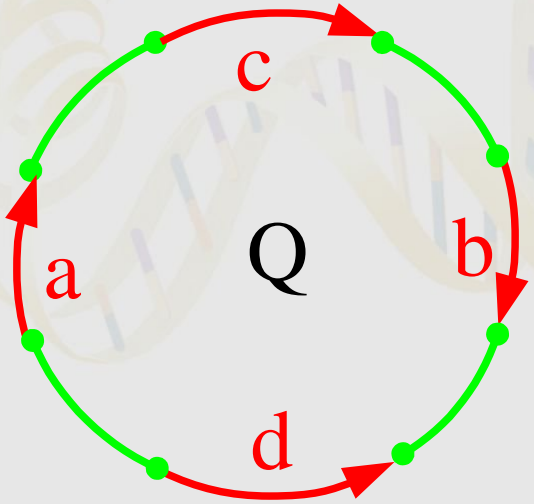




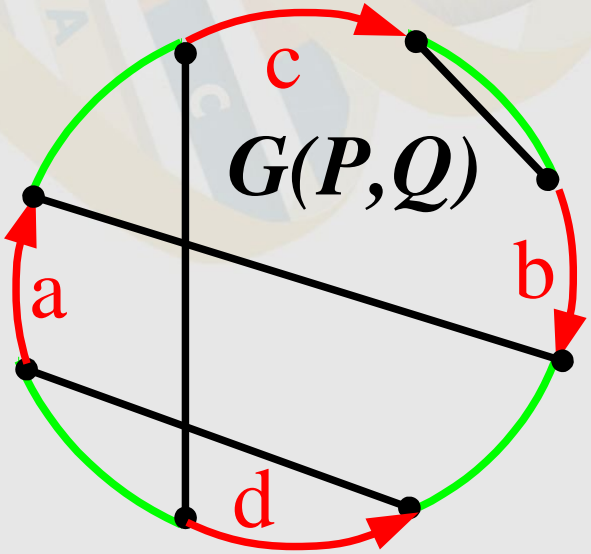


Breakpoint Graph = Superposition of Genome Graphs: Gluing Red Edges with the Same Labels



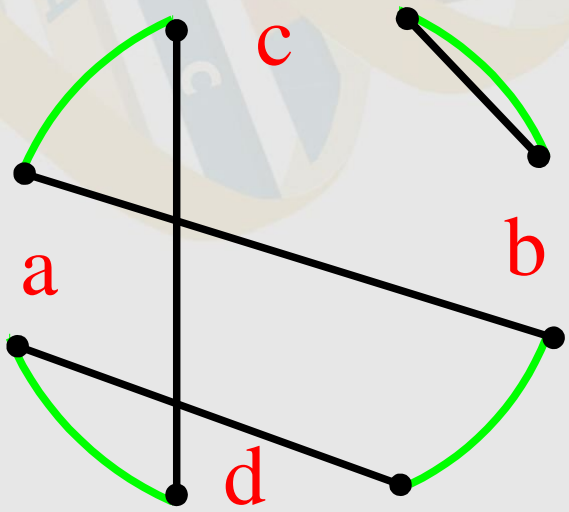


Breakpoint Graph
(Bafna & Pevzner, FOCS 1994)



Black-Green Cycles

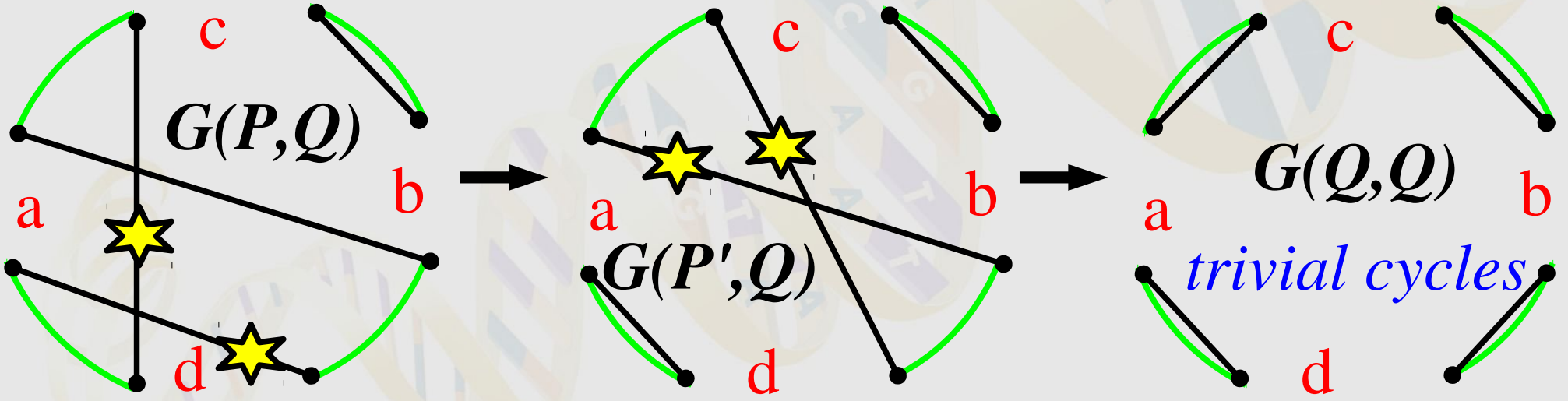
- Black and green edges represent perfect matchings in the breakpoint graph. Therefore, together these edges form a collection of black-green alternating cycles (where the color of edges alternate).



- The number of black-green cycles ***cycles(P,Q)*** in the breakpoint graph ***G(P,Q)*** plays a central role in computing the 2-break distance between ***P*** and ***Q***.

Rearrangements Change Cycles

Transforming genome ***P*** into genome ***Q*** by 2-breaks corresponds to transforming the breakpoint graph ***G(P,Q)*** into the breakpoint graph ***G(Q,Q)***.



cycles(P,Q) = 2

cycles(P',Q) = 3

cycles(Q,Q) = 4

= blocks(P,Q)

Transforming P into Q by 2-breaks

$$P = P_0 \xrightarrow{\text{2-breaks}} P_1 \rightarrow \ldots \rightarrow P_d = Q$$

$$G(P,Q) \rightarrow G(P_1,Q) \rightarrow \ldots \rightarrow G(Q,Q)$$

cycles(P,Q) cycles $\rightarrow$ ... $\rightarrow$ ***blocks(P,Q)*** cycles

# of black-green cycles increased by
blocks(P,Q) - cycles(P,Q)

How much each 2-break can contribute to this increase?

2-Break Distance

- Any 2-Break increases the number of cycles by at most one (***Δcycles ≤ 1***)
- Any non-trivial cycle can be split into two cycles with a 2-break (***Δcycles = 1***)
- Every sorting by 2-break must increase the number of cycles by ***blocks(P,Q) - cycles(P,Q)***
- The **2-Break Distance** between genomes P and Q:

$$dist(P,Q) = blocks(P,Q) - cycles(P,Q)$$

(cp. *Yancopoulos et al., 2005, Bergeron et al., 2006*)

Multi-Break Rearrangements

- The standard rearrangement operations (*reversals, translocations, fusions, and fissions*) make *2 breakages* in a genome and glue the resulting pieces in a new order.

- *k*-Break rearrangement operation makes *k breakages* in a genome and glues the resulting pieces in a new order.

- Rearrangements are rare evolutionary events and biologists believe that *k*-break rearrangements are unlikely for $k>3$ and relatively rare for $k=3$ (at least in the mammalian evolution).

- Also, in radiation biology, chromosome aberrations for $k>2$ (indicative of chromosome damage rather than evolutionary viable variations) may be more common, e.g., complex rearrangements in irradiated human lymphocytes (*Sachs et al., 2004; Levy et al., 2004*).

3-Break Distance: Focus on Odd Cycles

- A cycle is called *odd* if it contains an odd number of black edges.
- The *3-Break Distance* between genomes P and Q is:

$$d_3(P,Q) = (\#blocks - cycles^{odd}(P,Q)) / 2$$

Multi-Break Rearrangements

- We proposed exact formulas for the *k-break distance* between multi-chromosomal circular genomes as well as a linear-time algorithm for computing it. (**MA & PP, *Theor. Comput. Sci. 2008***)
- The exact formulas for $d_k(P,Q)$ becomes complex as k grows, e.g.:

Corollary 2. *The 4-break distance between a black matching P and a gray matching Q is*

$$d_4(P, Q) = \left\lceil \frac{|P| - c_1(P, Q) - \lfloor c_2(P, Q)/2 \rfloor}{3} \right\rceil$$

where $c_i(P, Q)$ is the number of black-gray cycles containing i modulo 3 black edges.

Corollary 3. *The 5-break distance between a black matching P and a gray matching Q is*

$$d_5(P, Q) = \left\lceil \frac{|P| - c_1(P, Q) - \min\{c_2(P, Q), c_3(P, Q)\} - \lfloor \max\{0, c_3(P, Q) - c_2(P, Q)\}/3 \rfloor}{4} \right\rceil$$

where $c_i(P, Q)$ is the number of cycles containing i modulo 4 black edges.

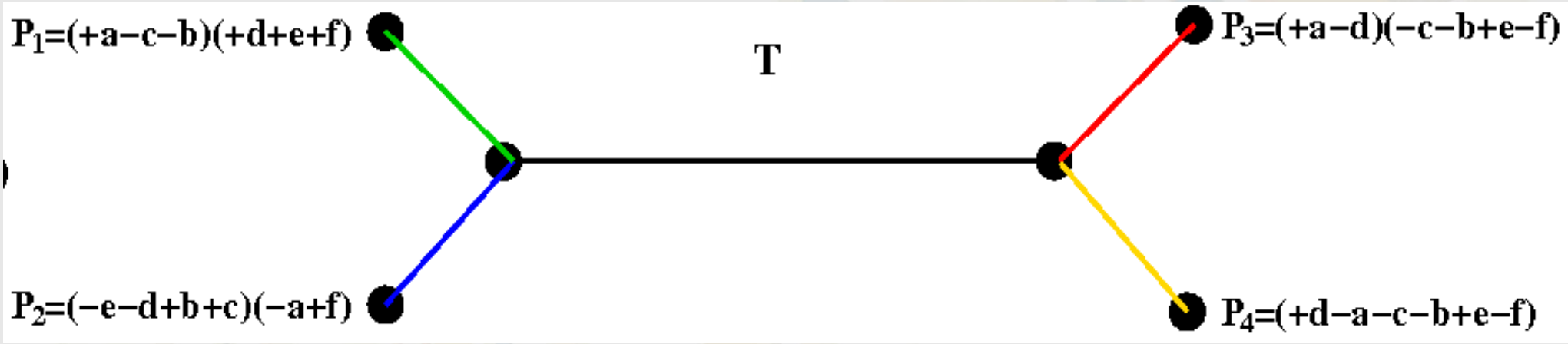
- The formula for $d_{20}(P,Q)$ is estimated to contain over 1,500 terms.

Algorithmic Problem:

Reconstruction of Ancestral Genomes

Ancestral Genomes Reconstruction in a Nutshell

- Given a set of genomes, reconstruct genomes of their common ancestors.



- The evolutionary tree of these genomes may be known or unknown.

Existing Tools for Ancestral Genomes Reconstruction

- ✔ **GRAPPA**: *J. Tang, B. Moret et al. (2001)*
- ✔ **MGR**: *G. Bourque and P. Pevzner (2002)*
- ✔ **InferCARs**: *J. Ma, D. Haussler et al. (2006)*
- ✔ **EMRAE**: *H. Zhao and G. Bourque (2007)*
- ✔ **MGRA**: *M. Alekseyev and P. Pevzner (2009)*

Ancestral Genomes Reconstruction Problem (with a known phylogeny)

- **Input:** a set of ***k*** genomes and a phylogenetic tree ***T***
- **Output:** genomes at the internal nodes of the tree ***T***
- **Objective**: minimize the total sum of the genomic distances along the branches of ***T***

- NP-complete in the "simplest" case of ***k=3***.
- ***What makes it hard?***

Breakpoints Are "Footprints" of Re-arrangements on the "Ground" of Genomes

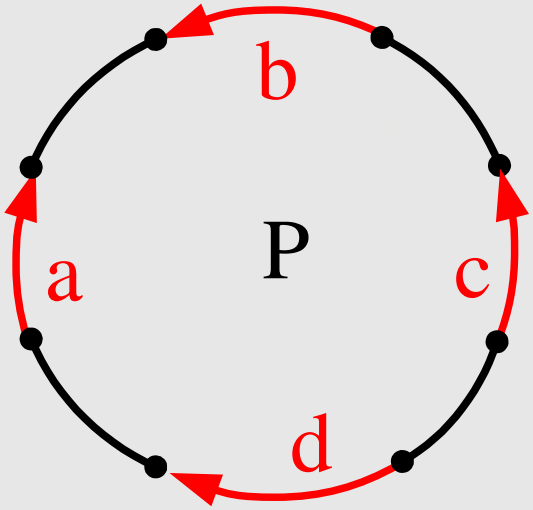
- ✔ NP-complete in the "simplest" case of ***k=3***.
- ✔ ***What makes it hard? BREAKPOINTS RE-USES (resulting in messy "footprints")!***

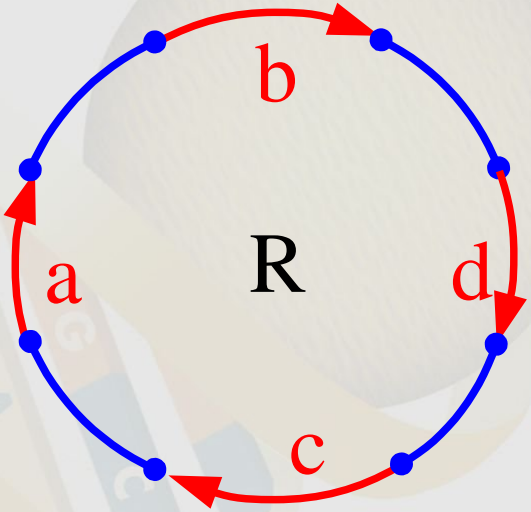
  Ancestral Genome Reconstructions of MANY Genomes (i.e., for **large *k***) may be easier to solve.

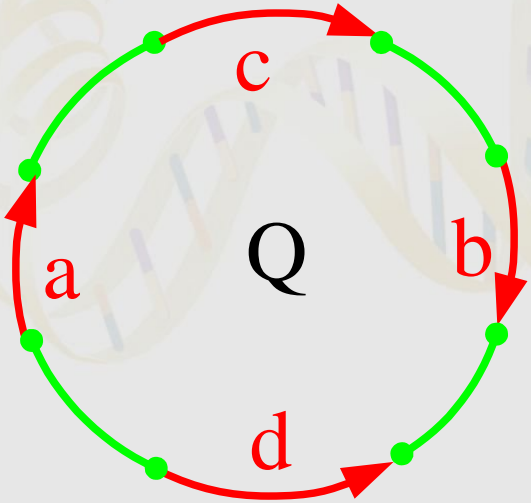
Solution:

Multiple Breakpoint Graphs and MGRA Algorithm

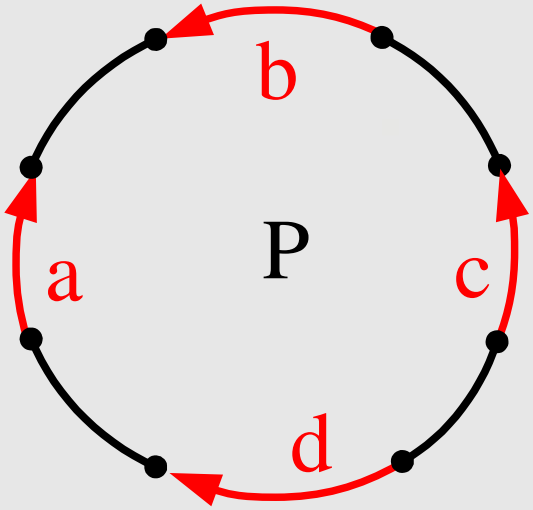
How to Construct Breakpoint Graph for Multiple Genomes?

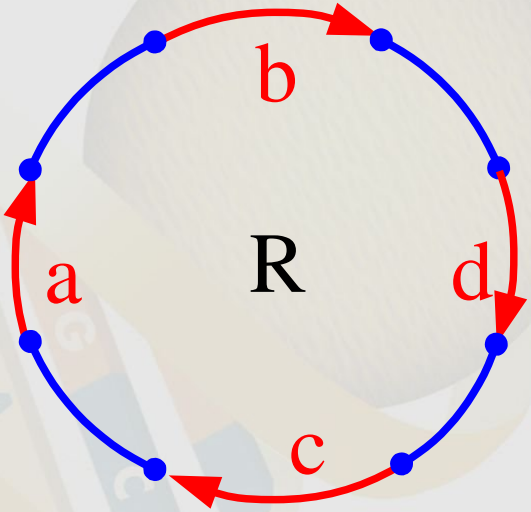


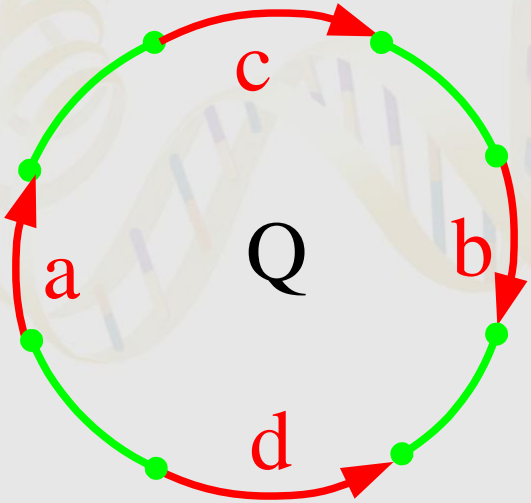


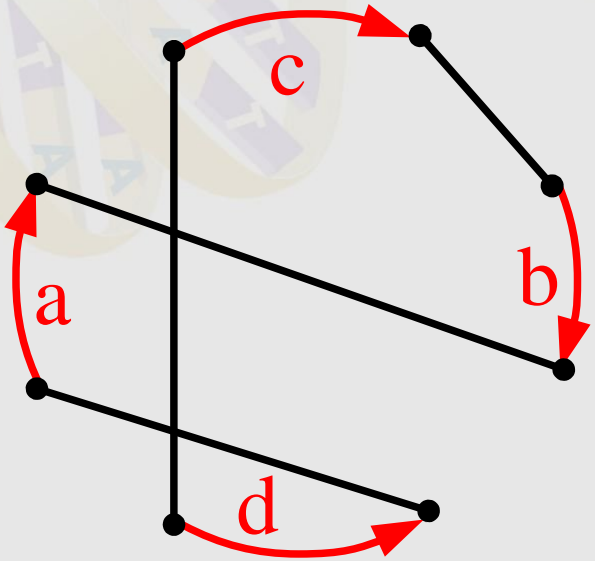


Constructing Multiple Breakpoint Graph: rearranging P in the Q order

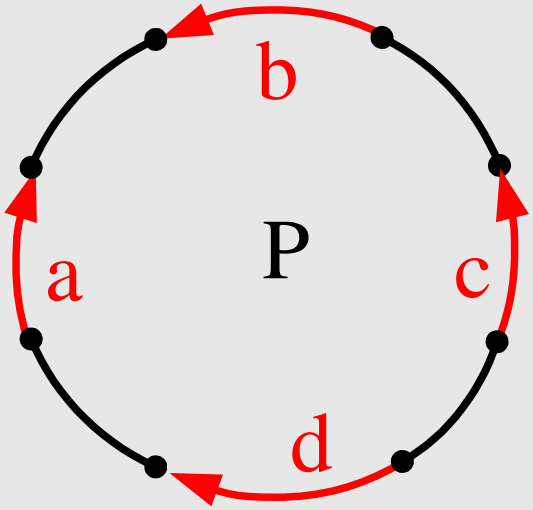


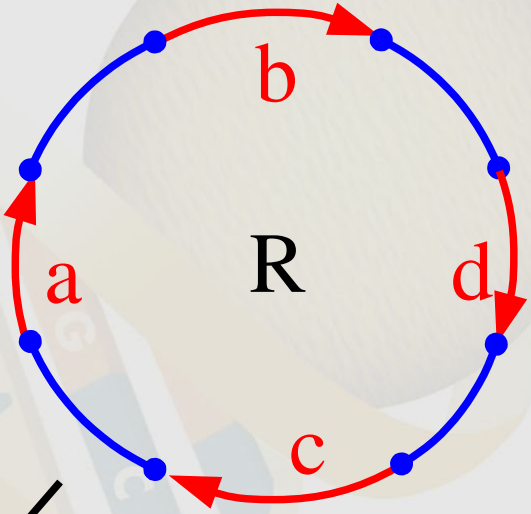


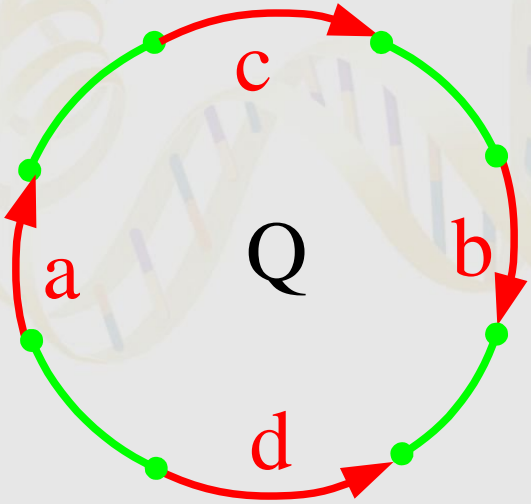


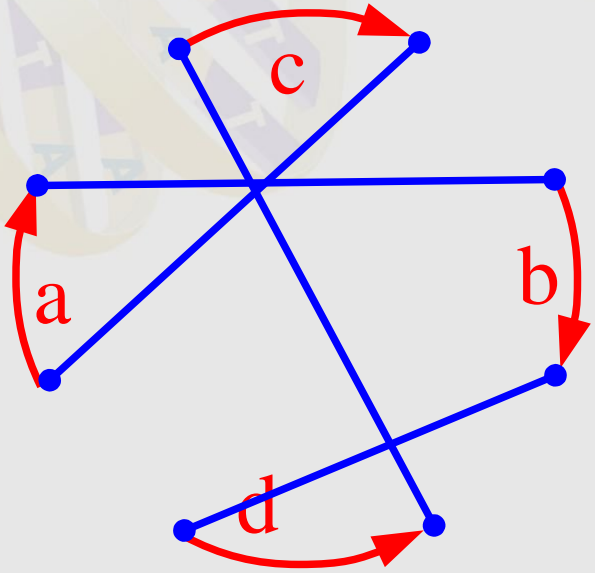


Constructing Multiple Breakpoint Graph: rearranging R in the Q order

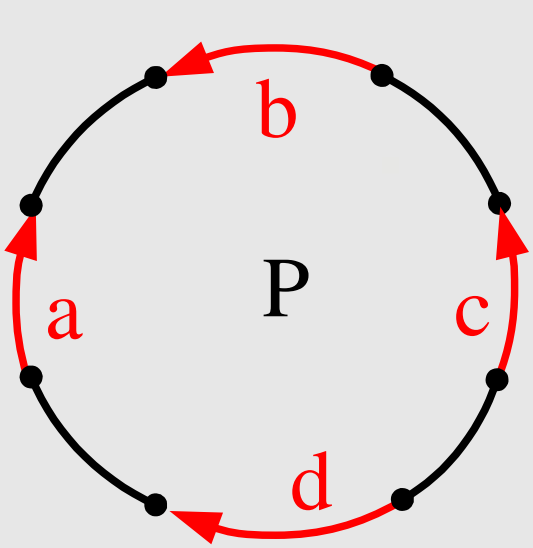


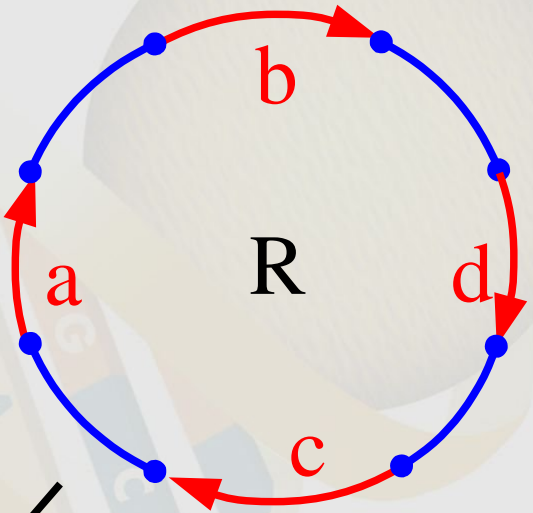


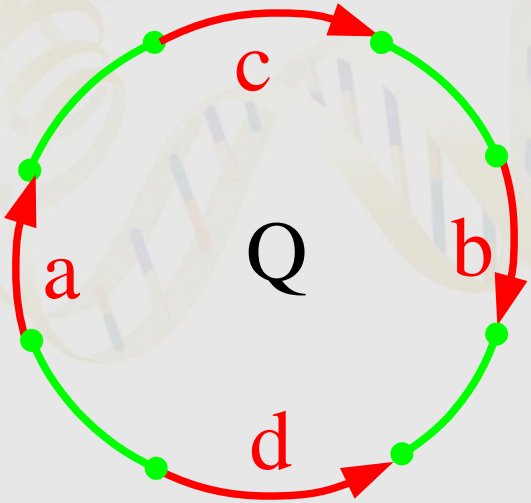


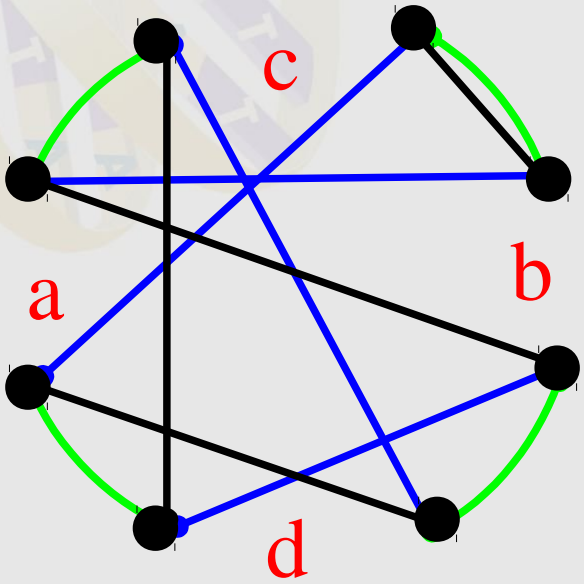


Multiple Breakpoint Graph: Still Gluing Red Edges with the Same Labels





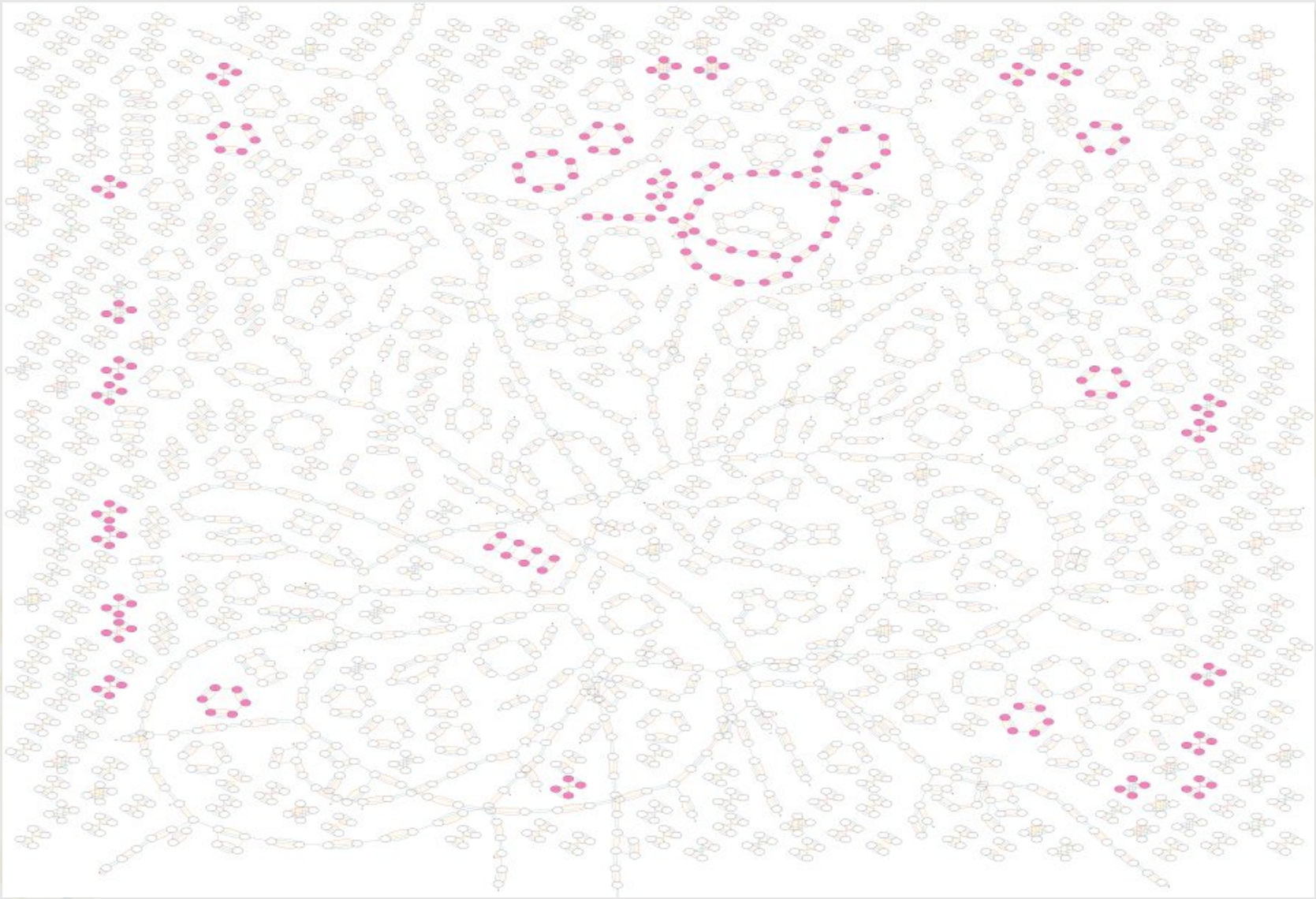




Multiple Breakpoint Graph

G(P,Q,R)

Multiple Breakpoint Graph of 6 Genomes



Multiple Breakpoint Graph ***G(M,R,D,Q,H,C)*** of the ***M****ouse,* ***R****at,* ***D****og,* *maca****Q****ue,* ***H****uman, and* ***C****himpanzee* genomes.

k=2 Genomes: Two Ways of Sorting by 2-Breaks

Transforming P into Q with *"black" 2-breaks*:

$$P = P_0 \rightarrow P_1 \rightarrow \ldots \rightarrow P_{d-1} \rightarrow P_d = Q$$

$$G(P,Q) \rightarrow G(P_1,Q) \rightarrow \ldots \rightarrow G(P_d,Q) = G(Q,Q)$$

Transforming Q into P with *"green" 2-breaks*:

$$Q = Q_0 \rightarrow Q_1 \rightarrow \ldots \rightarrow Q_d = P$$

$$G(P,Q) \rightarrow G(P,Q_1) \rightarrow \ldots \rightarrow G(P,Q_d) = G(P,P)$$

Let's combine these two ways...

Sorting By 2-Breaks: Meet In The Middle

- Let X be *any genome* on a path from P to Q:

$$P = P_0 \rightarrow P_1 \rightarrow \ldots \rightarrow P_m = X = Q_{m-d} \leftarrow \ldots \leftarrow Q_1 \leftarrow Q_0 = Q$$

- 2-Breaks at the left and right hand sides of X *are independent*!

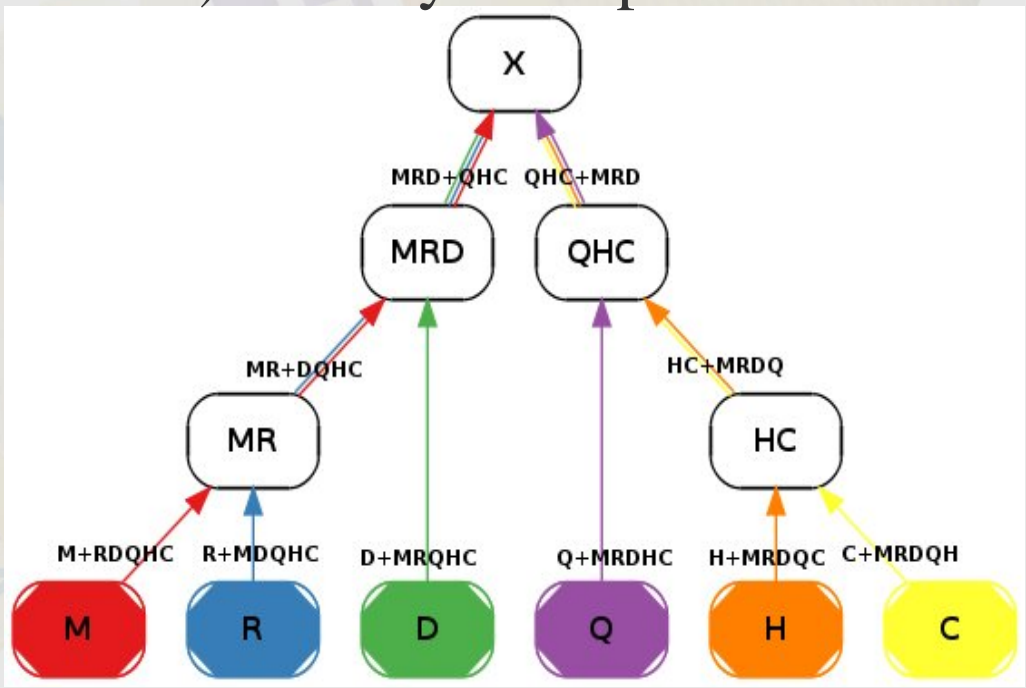
- Sorting By 2-Breaks Problem is equivalent to finding a *shortest* transformation of $G(P,Q)$ into a set of trivial cycles $G(X,X)$ (an identity breakpoint graph of *a priori unknown* genome X):

$$G(P_0,Q_0) \rightarrow G(P_1,Q_0) \rightarrow G(P_1,Q_1) \rightarrow G(P_1,Q_2) \rightarrow \ldots \rightarrow G(X,X)$$

- The **"black"** and **"green"** 2-breaks may arbitrarily alternate.

MGRA: Transformation into an Identity Breakpoint Graph

- We find a transformation of the multiple breakpoint graph $G(P_1,P_2,...,P_k)$ with *reliable rearrangements* (recognized from their "footprints") into *some (a priori unknown!)* identity multiple breakpoint graph $G(X,X,...,X)$:
  $G(P_1,P_2,...,P_k) \rightarrow ... \rightarrow G(X,X,...,X)$



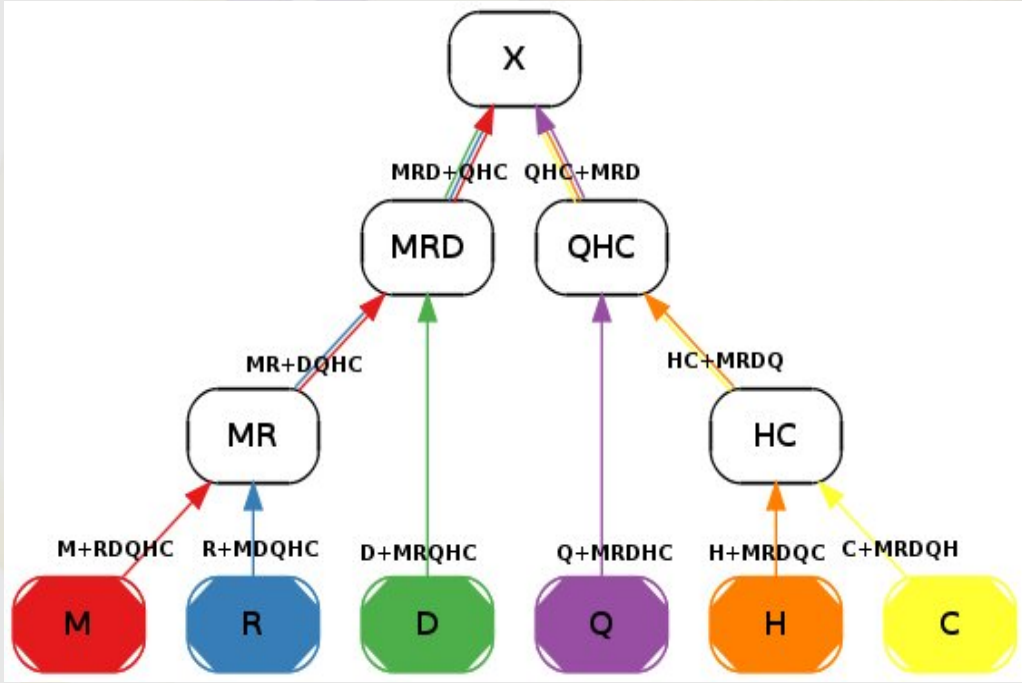
- Each rearrangement is consistent with the given tree T and thus is assigned to some branch of T.

- Rearrangements are applied in *arbitrary order* that ideally (if no extensive breakpoint re-uses) does not affect the result. Previously applied rearrangements may reveal "footprints" of new ones.

Tree-Consistent Rearrangements

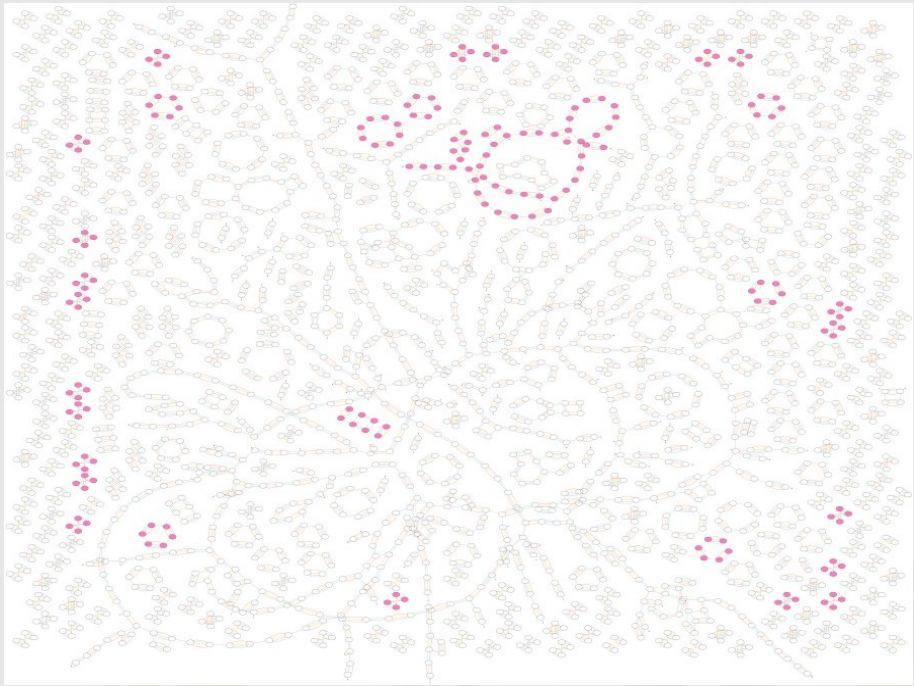
- Each branch of the given tree ***T*** defines *two complementary groups of genomes*, to each of which the same 2-breaks may be applied simultaneously.

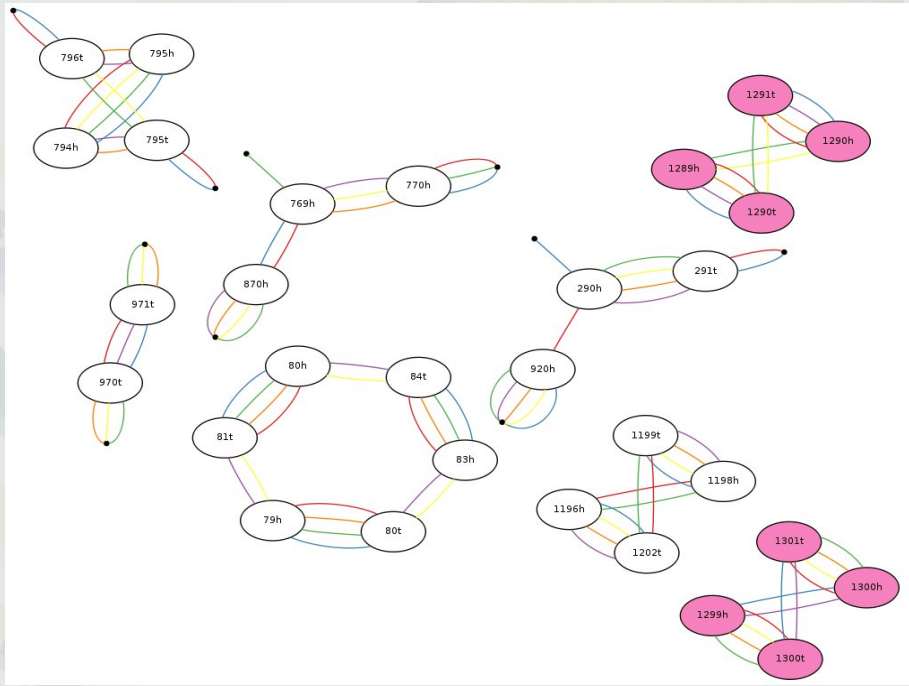
- For example, the branch labeled ***MR+DQHC*** defines groups ***{M, R}*** (*Mouse* and *Rat*) and ***{D,Q,H,C}*** (*Dog, macaQue, Human, Chimpanzee*).
  But there are no groups like ***{M,C}*** or ***{R,D,H}***.



- So, we can apply the same rearrangement to ***M*** and ***R*** simultaneously, viewing it as happening in their common ancestor (denoted ***MR***) along the ***MR+DQHC*** branch.

When All Reliable 2-Breaks Are Identified and “Undone”

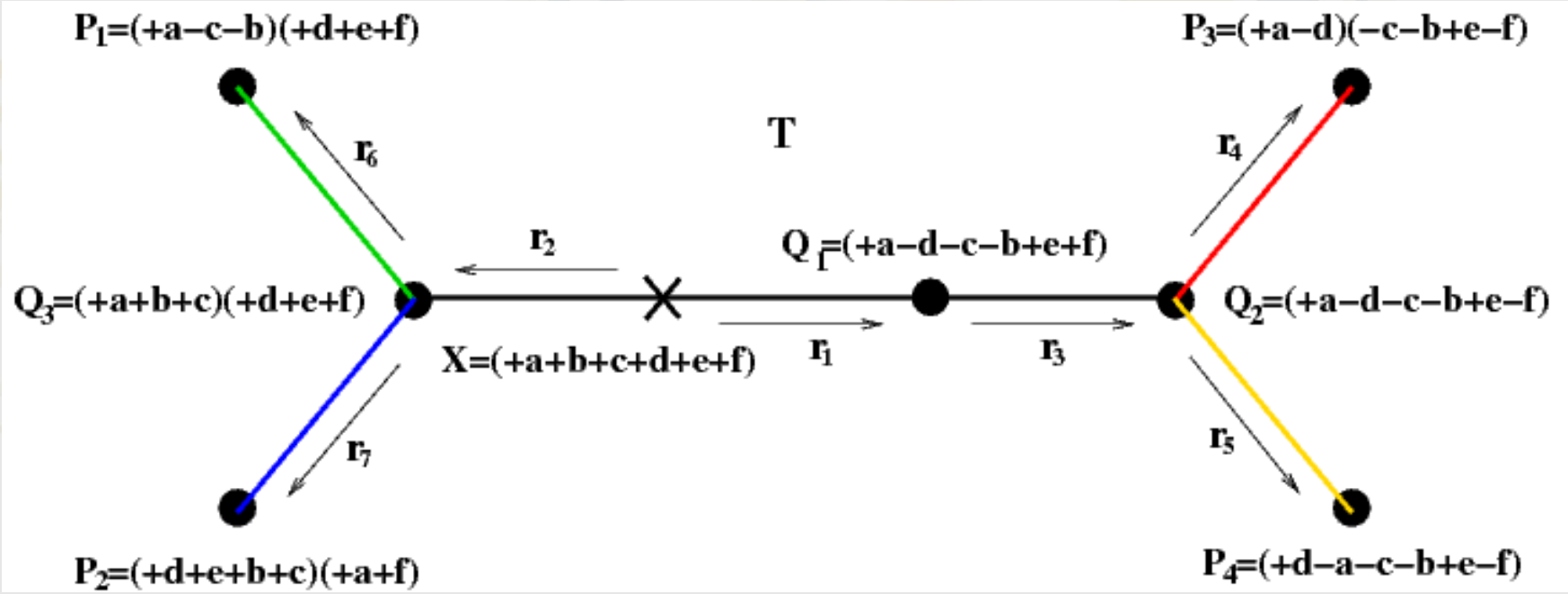




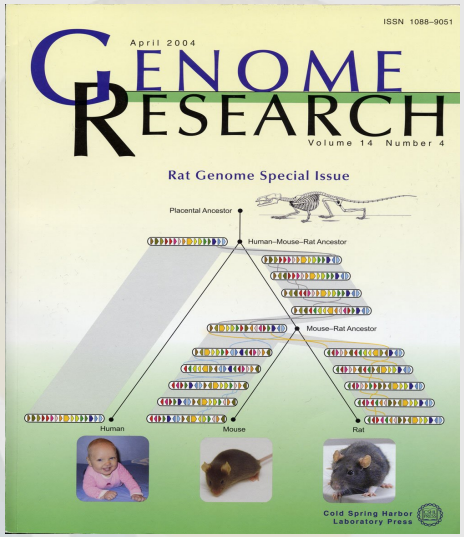
- ✔ The multiple breakpoint graph is reduced dramatically!
- ✔ The remaining (non-trivial) components can be processed manually in the case-by-case fashion.

MGRA: Reconstruction of the Ancestral Genomes

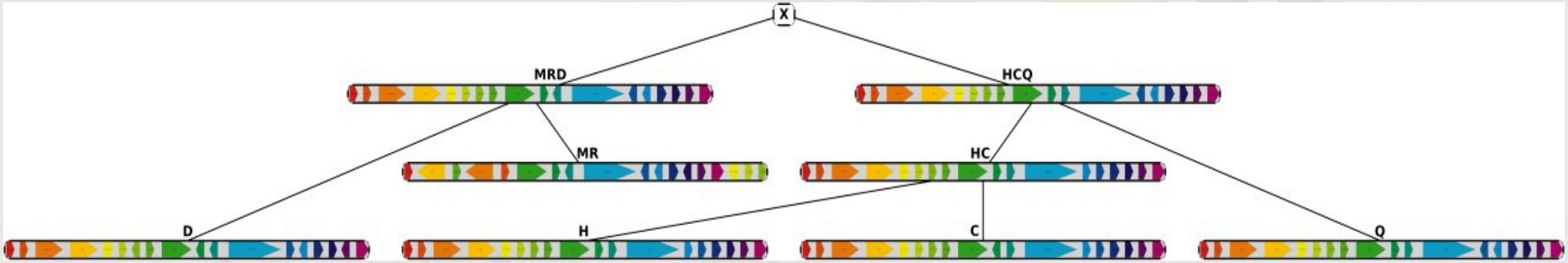
- The resulting identity breakpoint graph $G(X,X,...,X)$ defines its underlying genome X.
- The *reverse transformation* is applied to the genome X to transform it into each of the original genomes $P_1, P_2, ..., P_k$.
- This transformation traverses all internal nodes of T and thus defines the ancestral genome at every node.



Reconstructed X Chromosomes

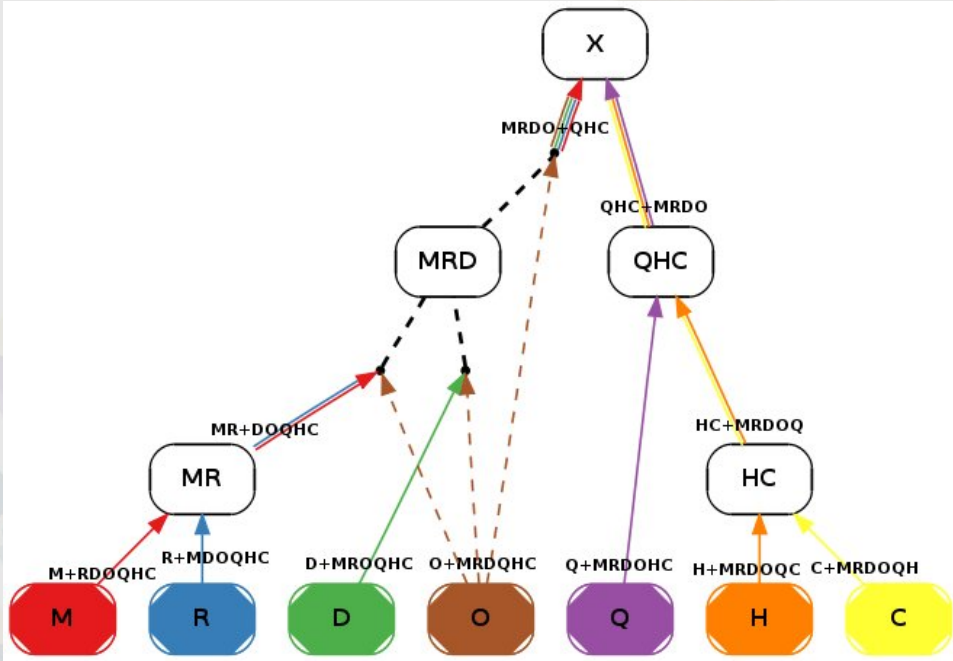


- The *Mouse, Rat, Dog, macaQue, Human, Chimpanzee* genomes and their reconstructed ancestors:



If The Evolutionary Tree Is Not Known

- For the set of 7 mammalian genomes: *Mouse, Rat, Dog, macaQue, Human, Chimpanzee*, and *Opossum*, the evolutionary tree ***T*** is not known.



- Depending on the primate – rodent – carnivore split, *three topologies are possible* (only two of them are viable).
- However, these three topologies share many common branches in ***T*** (***confident branches***). We can restrict the transformation only to such branches in order to simplify the breakpoint graph, not breaking an evidence for either of the topologies.

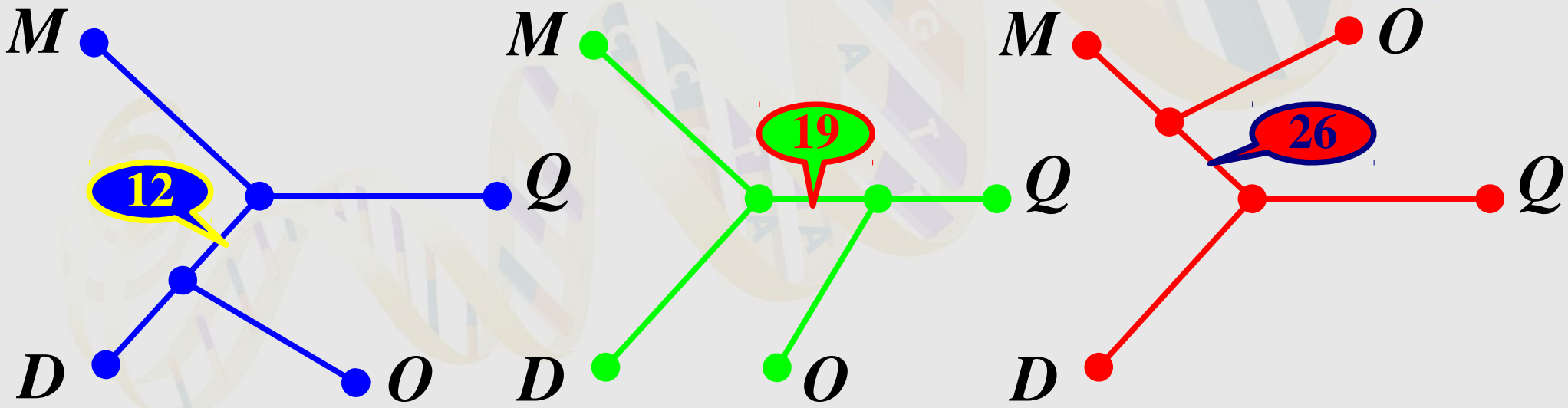
Resolving The Primate-Rodent-Carnivore Split Controversy

- We reduced the multiple breakpoint graph ***G(M,D,Q,O)*** (of representatives of each family and an outgroup) with reliable 2-breaks on the confident groups of genomes.

- ***What would be an evidence for one topology over the others?***

Rearrangement Evidence For The Primate-Carnivore Split

- ✔ Each of the three topologies has an unique branch in the tree. A *single rearrangement* assigned to such a branch would correspond to least *two rearrangements* if this branch is absent.



- ✔ We observed the prevalence of rearrangements' "foot-prints" specific to the primate – carnivore split.

Biological Problem:

Why and Where Genome Re-arrangements Happen?

Chromosome Breakage Models

- *Chromosome Breakage Models* specify how chromosomes are broken by rearrangements.

- While the exact mechanism of rearrangements is not known, such models try to explain as many as possible statistical characteristics observed in real genomes.

- The more characteristics are captured by a model, the better is this model.

- The choice of a model is particularly important in simulations that aim creation of *simulated genomes* whose characteristics should match those of real genomes.

Testing Models

- Given a characteristic observed in real genomes and a chromosome breakage model, we can test whether the model explains this characteristic.
- **Test:** Simulate genomes using the model and check if the simulated genomes possess the required characteristic.
- As soon as new characteristic in real genomes is discovered, the existing models can be tested against it.
- If they fail, this calls for a new model that would explain all previously known characteristics as well as the new one.

Susumu Ohno: Rearrangements occur randomly

Ohno, 1970, 1973

- **Random Breakage Hypothesis:** Genomic architectures are shaped by rearrangements that occur randomly.

Random Breakage Model (RBM)

- The random breakage hypothesis was embraced by biologists and has become *de facto* theory of chromosome evolution.

- Nadeau & Taylor, *Proc. Nat'l Acad. Sciences 1984*
  - First convincing arguments in favor of the **Random Breakage Model (RBM)**
  - RBM implies that there is no rearrangement hotspots
  - RBM was re-iterated in hundreds of papers

Fragile Breakage Model (FBM)

- *Pevzner & Tesler, PNAS 2003*

  - argued that every evolutionary scenario for transforming *Mouse* into *Human* genome must result in a large number of *breakpoint re-uses*, a contradiction to the RBM.
  - proposed the **Fragile Breakage Model (FBM)** that postulates existence of *rearrangement hotspots* and *vast breakpoint re-use*
  - FBM implies that the human genome is a mosaic of *solid* and *fragile* regions

Rebuttal of the Rebuttal

- ***Sankoff & Trinh, J. Comput. Biol. 2004***, presented arguments against the Fragile Breakage Model:

  "… we have shown that breakpoint re-use of the same magnitude as found in Pevzner and Tesler, 2003 may very well be artifacts in a context where NO re-use actually occurred."

Rebuttal of the Rebuttal of the Rebuttal

- ***Sankoff & Trinh, J. Comput. Biol. 2004***, presented arguments against the Fragile Breakage Model: *"… we have shown that breakpoint re-use of the same magnitude as found in Pevzner and Tesler, 2003 may very well be artifacts in a context where NO re-use actually occurred."*
- ***Peng et al., PLoS Comput. Biol. 2006***, found an error in the Sankoff–Trinh arguments.
- ***Sankoff, PLoS Comput. Biol. 2006***, *acknowledged the error: "Not only did we foist a hastily conceived and incorrectly executed simulation on an overworked RECOMB conference program committee, but worse — nostra maxima culpa — we obliged a team of high-powered researchers to clean up after us!"*

All Recent Studies Support FBM

A 1463 Gene Cattle–Human Comparative Map With Anchor Points Defined by Human Genome Sequence Coordinates

Annelie Everts-van der Wind,[1] Srinivas R. Kata,[3] Mark R. Band,[2] Mark Rebeiz,[1] Denis M. Larkin,[1] Robin E. Everts,[1] Cheryl A. Green,[1] Lei Liu,[2] Shreedhar Natarajan,[2] Tom Goldammer,[3] Jun Heon Lee,[1] Stephanie McKay,[3] James E. Womack,[3] and Harris A. Lewin[1,4]

Hotspots of mammalian chromosomal evolution

Jeffrey A Bailey*, Robert Baertsch†, W James Kent†, David Haussler‡ and Evan E Eichler*

Human, Mouse, and Rat Genome Large-Scale Rearrangements: Stability Versus Speciation

Shaying Zhao,[1,3] Jyoti Shetty,[1] Lihua Hou,[1] Arthur Delcher,[1] Baoli Zhu,[2] Kazutoyo Osoegawa,[2] Pieter de Jong,[2] William C. Nierman,[1] Robert L. Strausberg,[1] and Claire M. Fraser[1]

Dynamics of Mammalian Chromosome Evolution Inferred from Multispecies Comparative Maps

William J. Murphy,[1,3*†] Denis M. Larkin,[5*] Annelie Everts-van der Wind,[5*] Guillaume Bourque,[8] Glenn Tesler,[9] Loretta Auvil,[6] Jonathan E. Beever,[5] Bhanu P. Chowdhary,[1] Francis Galibert,[11] Lisa Gatzke,[6] Christophe Hitte,[11] Stacey N. Meyers,[5] Denis Milan,[12] Elaine A. Ostrander,[13] Greg Pape,[6] Heidi G. Parker,[13] Terje Raudsepp,[1] Margarita B. Rogatcheva,[5] Lawrence B. Schook,[5,7] Loren C. Skow,[1] Michael Welge,[6] James E. Womack,[2] Stephen J. O'Brien,[4] Pavel A. Pevzner,[10] Harris A. Lewin[5,7†]

Recurring genomic breaks in independent lineages support genomic fragility

Hanno Hinsch[1] and Sridhar Hannenhalli*[1,2]

Is mammalian chromosomal evolution driven by regions of genome fragility?

Aurora Ruiz-Herrera*, Jose Castresana† and Terence J Robinson*

Analysis of segmental duplications reveals a distinct pattern of continuation-of-synteny between human and mouse genomes

Michael R. Mehan · Maricel Almonte · Erin Slaten · Nelson B. Freimer · P. Nagesh Rao · Roel A. Ophoff

Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates

Hiroshi Kikuta,[1] Mary Laplante,[1] Pavla Navratilova,[1] Anna Z. Komisarczuk,[1] Pär G. Engström,[2,3] David Fredman,[2] Altuna Akalin,[2] Mario Caccamo,[4] Ian Sealy,[4] Kerstin Howe,[4] Julien Ghislain,[5] Guillaume Pezeron,[5] Philippe Mourrain,[4] Staale Ellingsen,[1,10] Andrew C. Oates,[6] Christine Thisse,[7] Bernard Thisse,[7] Isabelle Foucher,[8] Birgit Adolf,[9] Andrea Geling,[9,11] Boris Lenhard,[1,2,12] and Thomas S. Becker[1,13]

7E olfactory receptor gene clusters and evolutionary chromosome rearrangements

Y. Yue, T. Haaf

Kikuta et al., Genome Res. 2007: *"... the Nadeau and Taylor hypothesis is not possible for the explanation of synteny in rat."*

… With One Influential Exception

Reconstructing contiguous regions of an ancestral genome

Jian Ma,[1,5,6] Louxin Zhang,[2] Bernard B. Suh,[3] Brian J. Raney,[3] Richard C. Burhans,[1] W. James Kent,[3] Mathieu Blanchette,[4] David Haussler,[3] and Webb Miller[1]

Ma et al., Genome Res. 2006:

"Simulations … suggest that this frequency of breakpoint re-use is approximately what one would expect if breakage was equally likely for every genomic position … a careful analysis [of the RBM vs. FBM controversy] is beyond the scope of this study."

Our Contribution

- ✓ We reconcile the evidence for limited breakpoint reuse in **Ma et al., 2006** with the Fragile Breakage Model and reveal a ***rampant*** but ***elusive*** breakpoint reuse.

- ✓ We provide evidence for the "***birth and death***" of the fragile regions, implying that they move to different locations in different lineages, explaining why **Ma et al., 2006**, found limited breakpoint reuse between different branches of the evolutionary tree.

- ✓ We introduce the ***Turnover Fragile Breakage Model (TFBM)*** that accounts for the *"birth and death"* of the fragile regions and sheds light on a possible relationship between rearrangements and ***Matching Segmental Duplications***.

- ✓ TFBM points to locations of the ***currently*** fragile regions in the human genome.

Tests vs. Models

- ✓ Why biologists believe in RBM? Because RBM implies the exponential distribution of the sizes of the synteny blocks observed in real genomes.
- ✓ A flaw in this logic: RBM is not the only model that complies with the "exponential distribution" test.
- ✓ Why Pevzner and Tesler refuted RBM? Because RBM does not comply with the "breakpoint reuse" test: RBM implies low reuse but real genomes reveal high reuse.
- ✓ FBM complies with both the "exponential distribution" and "breakpoint reuse" tests.
- ✓ But is there a test that both RBM and FBM fail?

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

Tests vs. Models

- ✓ Why biologists believe in RBM? Because RBM implies the exponential distribution of the sizes of the synteny blocks observed in real genomes.
- ✓ A flaw in this logic: RBM is not the only model that complies with the "exponential distribution" test.
- ✓ Why Pevzner and Tesler refuted RBM? Because RBM does not comply with the "breakpoint reuse" test: RBM implies low reuse but real genomes reveal high reuse.
- ✓ FBM complies with both the "exponential distribution" and "breakpoint reuse" tests.
- ✓ RBM and FBM fail the ***Multispecies Breakpoint Reuse (MBR)*** test.

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

Tests vs. Models

- ✓ Why biologists believe in RBM? Because RBM implies the exponential distribution of the sizes of the synteny blocks observed in real genomes.
- ✓ A flaw in this logic: RBM is not the only model that complies with the "exponential distribution" test.
- ✓ Why Pevzner and Tesler refuted RBM? Because RBM does not comply with the "breakpoint reuse" test: RBM implies low reuse but real genomes reveal high reuse.
- ✓ FBM complies with both the "exponential distribution" and "breakpoint reuse" tests.
- ✓ ***TFBM passes all three tests.***

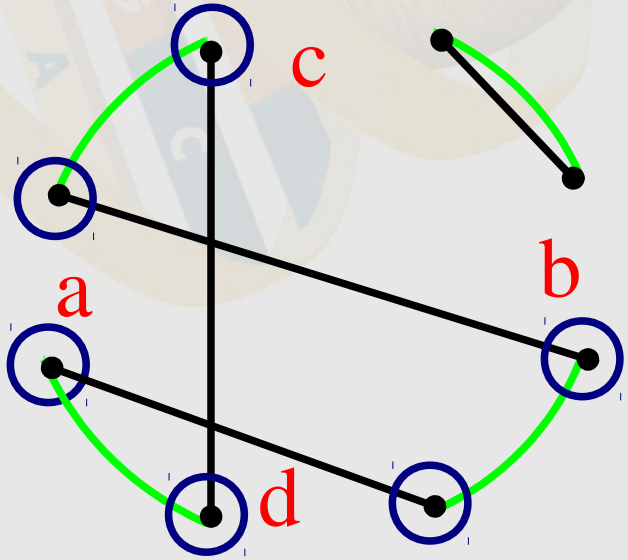
Model \ Test	Exponential distribution	Breakpoint reuse	MBR
RBM	**YES**	**NO**	**NO**
FBM	**YES**	**YES**	**NO**
TFBM	**YES**	**YES**	**YES**

Algorithmic Problem:

Breakpoint Re-use Analysis

Breakpoints Are Vertices in Non-trivial Cycles

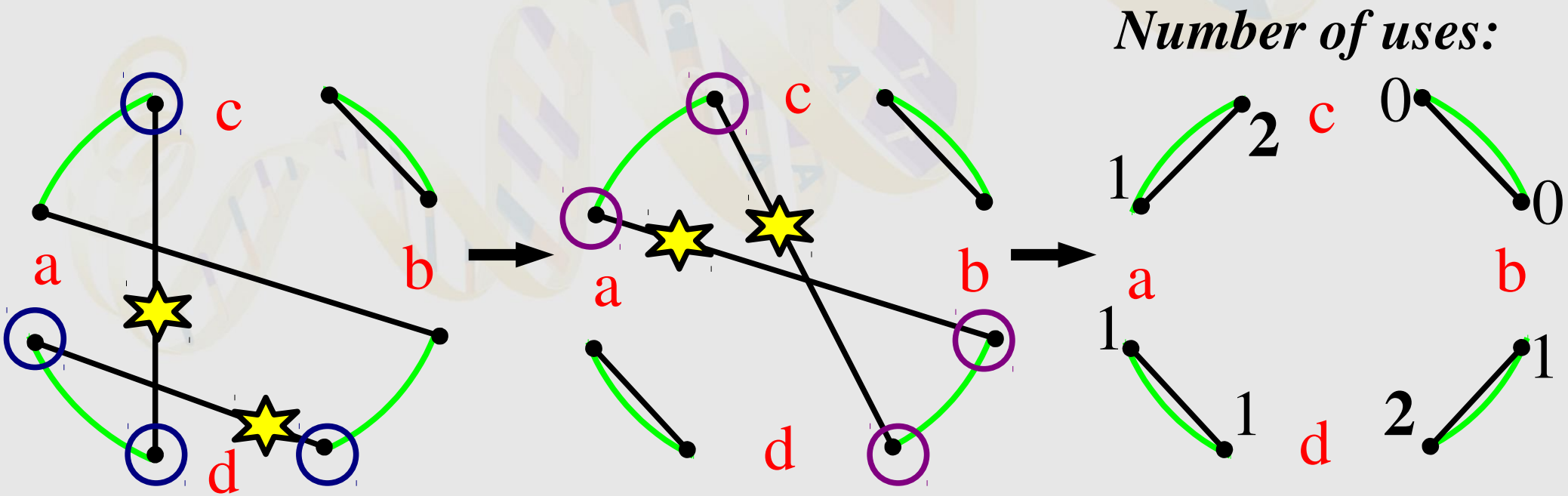
- Breakpoints correspond to regions in the genome that were broken by some re-arrangement(s).
- In the breakpoint graph, breakpoints correspond to vertices having two neighbors (while vertices with just one neighbor represent common adjacencies between synteny blocks).
- All ***vertices in non-trivial cycles*** in the breakpoint graph represent breakpoints.



Breakpoint Uses and Reuses

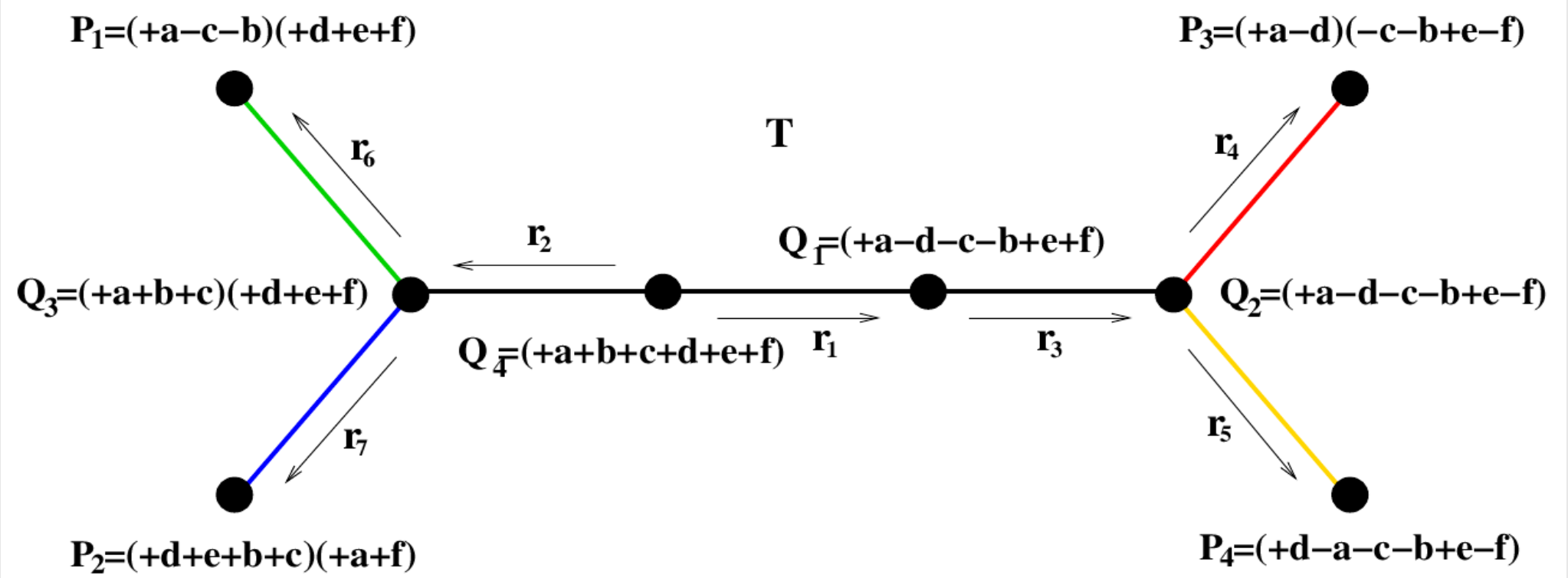
- Each 2-break ***uses*** four vertices (the endpoints of the affected edges).

- A vertex (breakpoint) is ***reused*** if it is used by at least two different 2-breaks (i.e., the number of uses > 1).



Intra- and Inter- Reuses

- For an evolutionary tree with known rearrangement scenarios, a breakpoint is ***intra-reused*** on some branch if it is used by at least two different 2-breaks along this branch.
- Similarly, a breakpoint is ***inter-reused*** across two branches if it is used on both these branches.



Rearrangement Scenarios Remain Ambiguous

- In mammalian evolution we know only genomes of existing species but do not know the ancestral genomes.
- While ancestral genomes can be reliably reconstructed, the exact rearrangement scenarios between them remain ambiguous.
- **Can we compute the number of breakpoint intra- and inter- reuses without knowing rearrangement scenarios?**

Number of Intra-Reuses (Lower Bound)

For a rearrangement scenario between genomes P and Q:

- ✔ The number of 2-breaks is at least $dist(P,Q)$
- ✔ Each 2-break uses 4 breakpoints
- ✔ The number of breakpoints is $2 \cdot blocks(P,Q)$
- ✔ Hence the total number of intra-reuses is:

$$\geq 4 \cdot dist(P,Q) - 2 \cdot blocks(P,Q)$$

Number of Inter-Reuses (Lower Bound)

For two branches (P,Q) and (P',Q') in the tree:

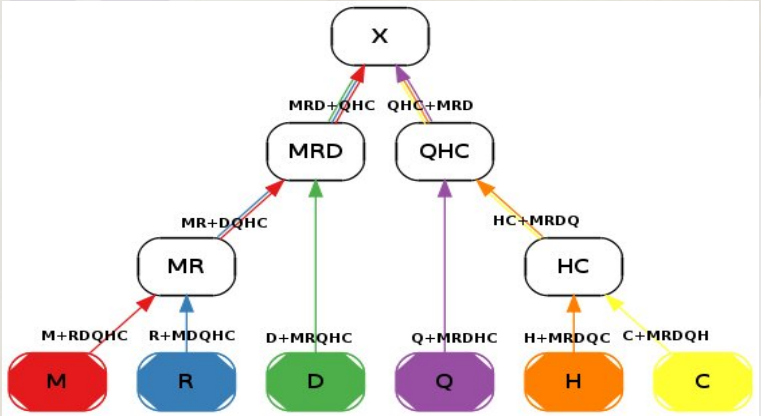
- Set V of the vertices in non-trivial cycles in $G(P,Q)$ represents the breakpoints between genomes P and Q
- Set V' of the vertices in non-trivial cycles in $G(P',Q')$ represents the breakpoints between genomes P' and Q'
- Hence, the number of inter-reuses is

  $\geq$ size of the intersection of V and V'

Surprising Irregularities in Breakpoint Reuse Across Various Pairs of Branches

- Statistics of breakpoint intra- and inter-reuses between the branches of the tree of six mammalian genomes:

	M+	R+	D+	Q+	H+	MR+	QH+
M+	84	68	20	4	5	58	15
R+		96	22	3	6	60	17
D+			174	17	19	98	64
Q+				12	10	25	18
H+					22	23	18
MR+						292	80
QH+							70



- Colors represent the "distance" between a pair of branches:
  red = adjacent branches;
  green = branches separated by one other branch;
  yellow = branches separated by two other branches.

- ***What is surprising about this Table?***

Solution:

Turnover Fragile Breakage Model and Multispecie Breakpoint Reuse Test

Turnover Fragile Breakage Model (TFBM)

- ✔ The Ma et al. observation and the statistics of inter-reuses indicates:

  Breakpoint inter-reuses mostly happen across adjacent branches of the evolutionary tree.

- ✔ Turnover Fragile Breakage Model (TFBM):

  Fragile regions are subject to a "birth and death" process and thus have limited lifespan.

Simplest TFBM: Fixed Turnover Rate for Fragile Regions

- ***TFBM(m,n,x):***
  - genomes have ***m*** fragile regions
  - ***n*** (out of ***m***) fragile regions are *active*
  - each 2-break is applied to ***2*** (out of ***n***) randomly chosen active fragile regions
  - after each 2-break, ***x*** active fragile regions (out of ***n***) "die" and ***x*** new active fragile regions (out of ***m-n***) are "born"

- FBM is a particular case of TFBM with ***x=0***
- RBM is a particular case of TFBM with ***x=0*** and ***n=m***

Recognizing the "Birth and Death"

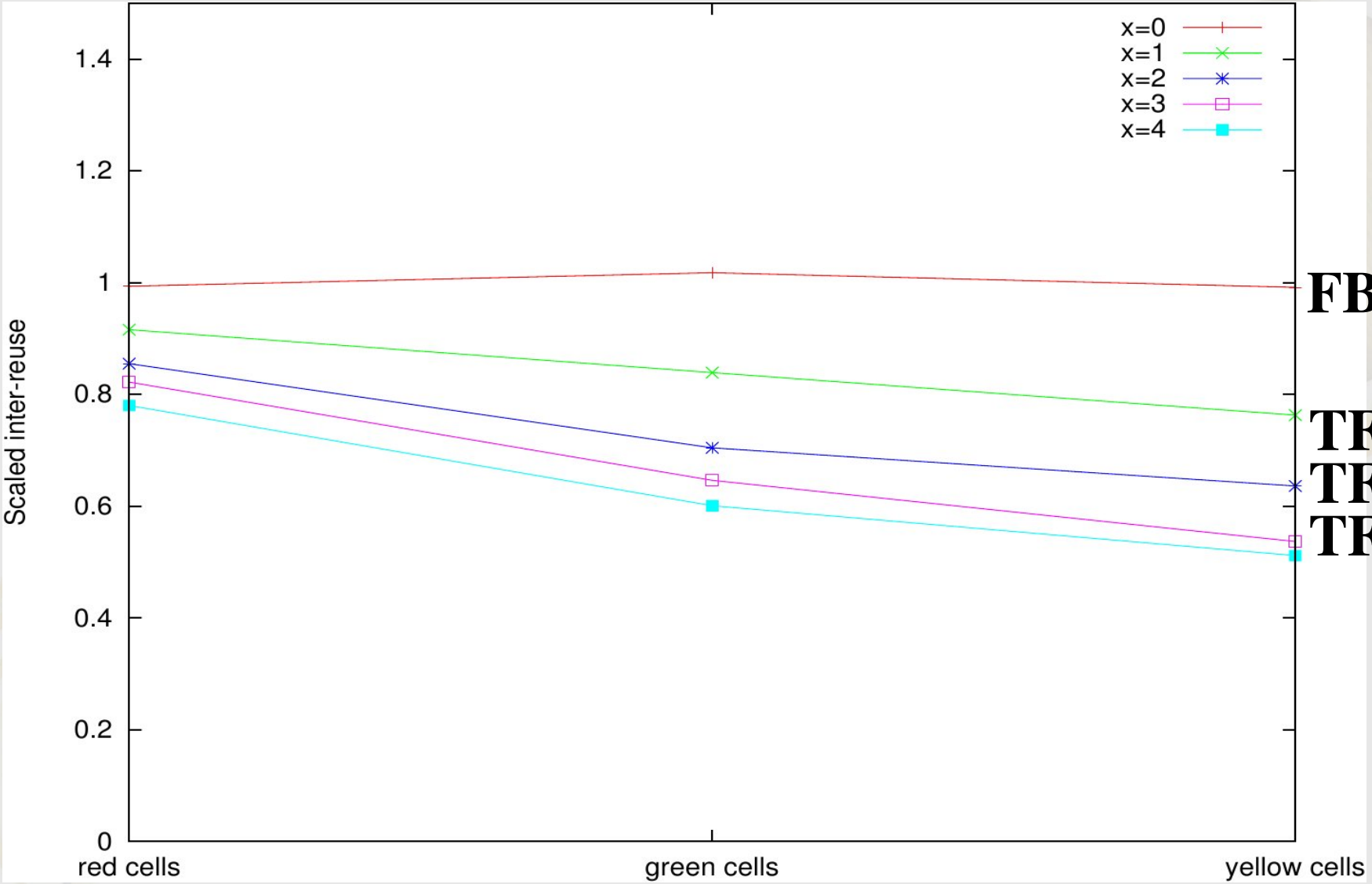
- Given an evolutionary tree with known rearrangement scenarios, how one would determine whether they followed TFBM with $x = 0$ (that is, FBM/RBM) or $x > 0$?

- Comparing breakpoint inter-reuse across different pairs of branches would help, but it also depends on the branch lengths that may differ significantly across the tree.

Scaled Breakpoint Reuse

- The number of breakpoint intra- and inter- reuses depends on the length of branches. To eliminate this dependency, we define the **scaled intra- and inter- reuse**:
  - We defined and expressed analytically:
    $E(t)$ = the expected number of intra-reuses along a branch of length **t**;
    $E(t_1,t_2)$ = the expected number of inter-reuses across branches of length **t_1** and **t_2**.
  - Scaled intra- and inter-reuse is the number of reuses divided by **$E(t)$** or **$E(t_1,t_2)$** respectively.

Scaled Inter-Reuse in Colored Cells (Simulated Genomes with Variable Branch Length)



Simulations for the case when ***n=900*** out of ***m=2000*** fragile regions are active and various turnover rate ***x=0..4.***

Measuring Reuse in the Whole Evolutionary Tree

- TFBM suggests that on average the number breakpoint reuses $\mathbf{br(r_1,r_2)}$ for 2-breaks $\mathbf{r_1}$ and $\mathbf{r_2}$ depends on the distance (in the evolutionary tree) between them. The larger is the distance, the smaller is $\mathbf{br(r_1,r_2)}$.

- Our goal is to define a ***single measure for the whole tree*** that would "describe" this trend and allow one to test whether the rearrangement process follow the TFBM with $x>0$.

Multispecies Breakpoint Reuse

- The **multispecies breakpoint reuse** is a function **R(L)** expressing averaged breakpoint reuse between pairs of rearrangements separated by **L** other rearrangements in the given tree.

- It can be explicitly defined as:

$$R(L) = \Sigma\, br(r_1, r_2) \,/\, \Sigma\, 1$$

where both sums are taken over all pairs of rearrangements r_1 and r_2 at distance **L** in the tree.

Multispecies Breakpoint Reuse Test

- For RBM/FBM, $R(L)$ is a constant.
- For TFBM with $x > 0$, $R(L)$ is a decreasing function.
- **MBR Test**: compute $R(L)$, and check if it is decreasing.

  (A stronger variant: determine x and check if $x>0$.)

Multispecies Breakpoint Reuse in TFBM (theoretic curve)

- For TFBM with parameters m, n, x, we derive an analytic formula:

$$R(L) = 8(m-n)/(mn) * (1 - xm/(n(m-n)))^L + 8/m$$

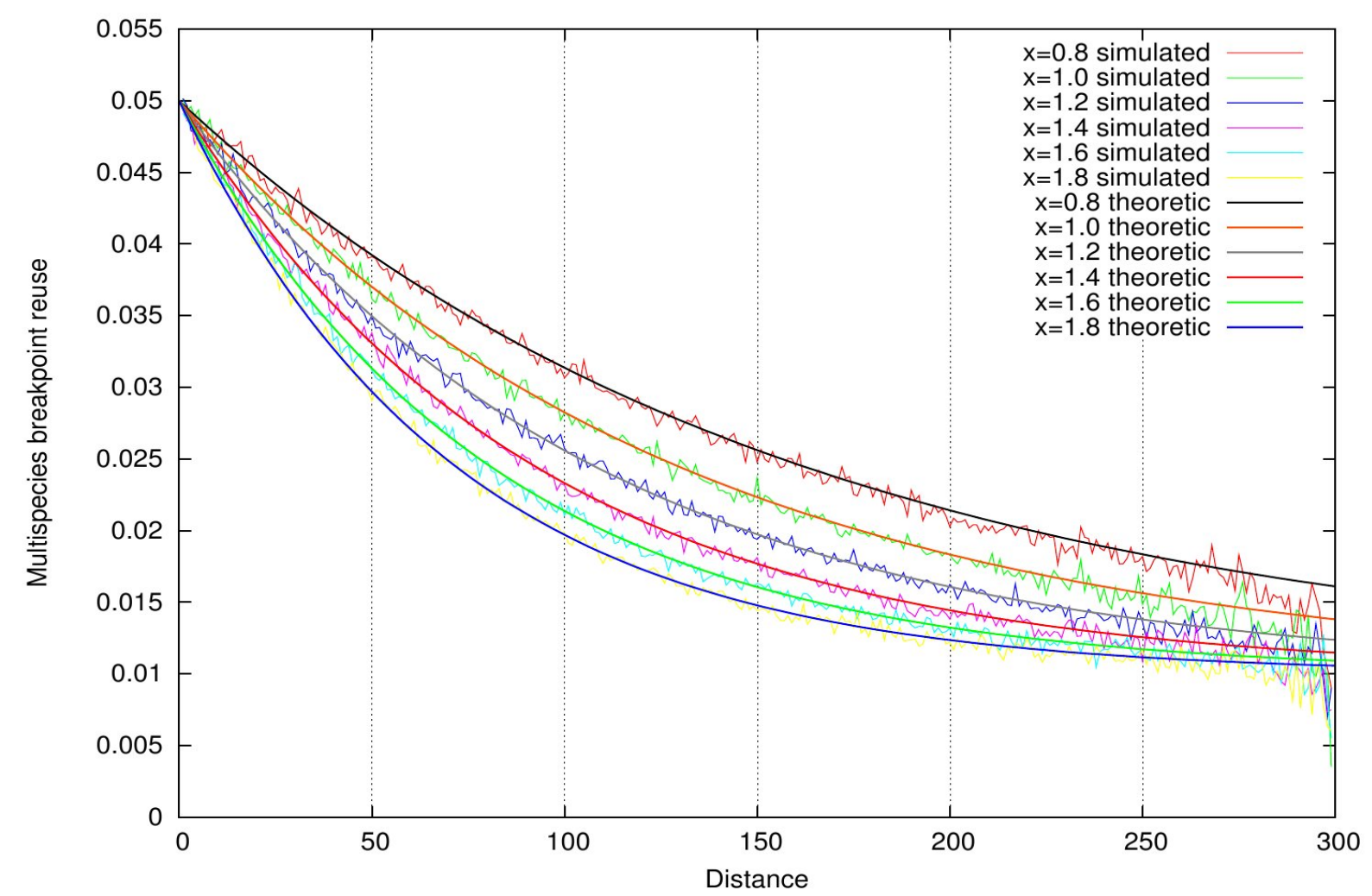
- For small L, $R(L)$ is approximated by a straight line:

$$8/n - 8x/n^2 \, L$$

which does not depend on m.

- Given $R(L)$, the parameters n and x can be determined from the value and slope of $R(L)$ at $L=0$.

Multispecies Breakpoint Reuse in TFBM (theoretic vs. empiric curve)



Simulations for the case when $n=160$ out of $m=800$ fragile regions are active and various turnover rate x

From Simulated to Real Genomes: Complications

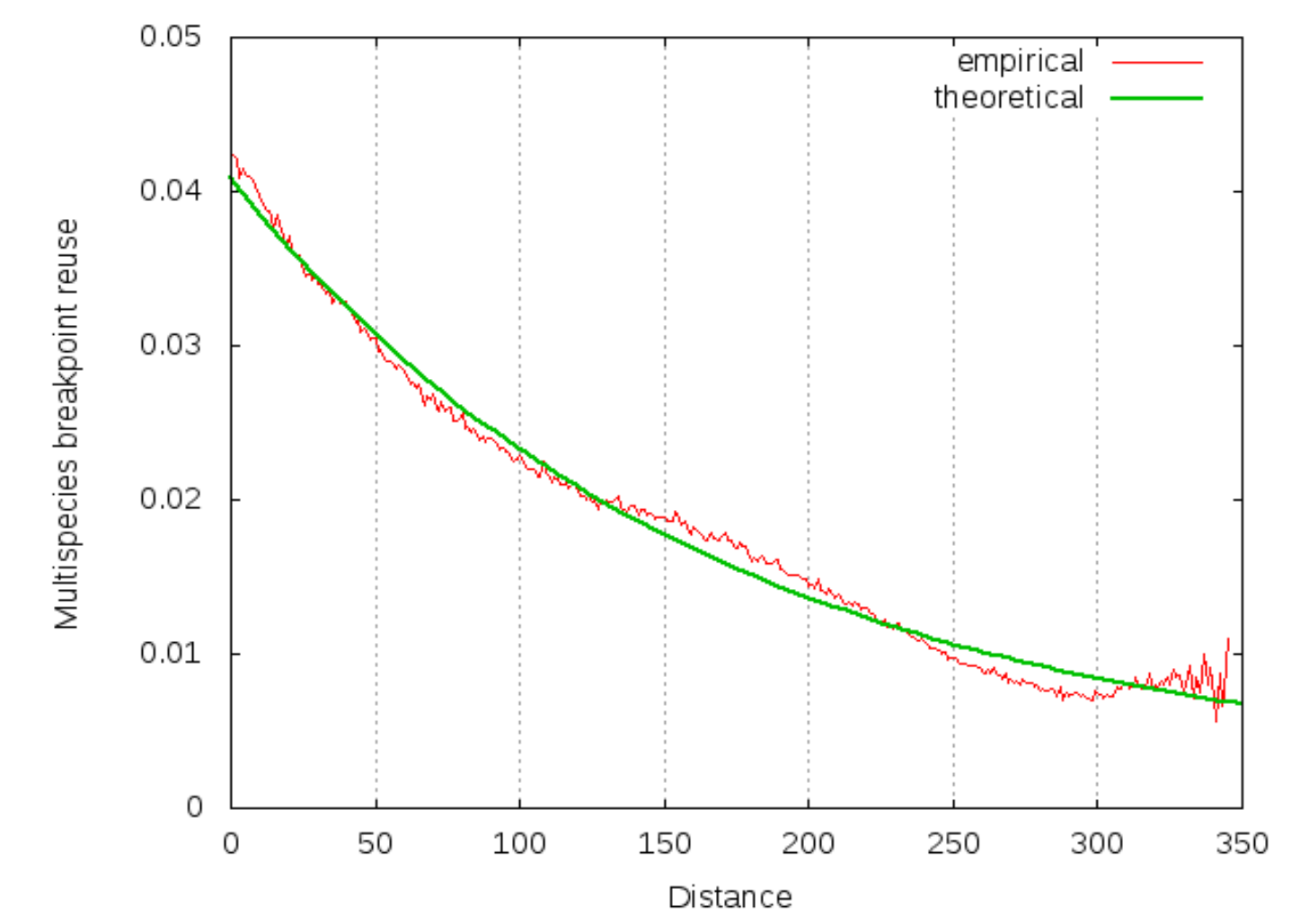
- It is easy to compute $R(L)$ for simulated genomes, whose rearrangement history is defined by simulations.

- For real genomes, while we can reliably reconstruct the ancestral genomes, the exact evolutionary scenarios between them remain ambiguous.

From Simulated to Real Genomes: Complications

- It is easy to compute $R(L)$ for simulated genomes, whose rearrangement history is defined by simulations.
- For real genomes, while we can reliably reconstruct the ancestral genomes, the exact evolutionary scenarios between them remain ambiguous.
- We can **sample random scenarios** instead.

Multispecies Reuse between Mammalian Genomes



✔ Best fit: $m \approx 4017$ $n \approx 196$ $x \approx 1.12$

Implications:

How will the Human Genome Evolve in the Next Million Years?

Prediction Power of TFBM

- ✔ Can we determine currently active regions in the human genome **H** from comparison with other mammalian genomes?

- ✔ RBM provides no clue
- ✔ FBM suggests to consider the breakpoints between **H** and *any* other genome
- ✔ TFBM suggests to consider the *closest* genome such as the macaque-human ancestor **QH**.
  Breakpoints in **G(QH,H)** are likely to be reused in the future rearrangements of **H**.

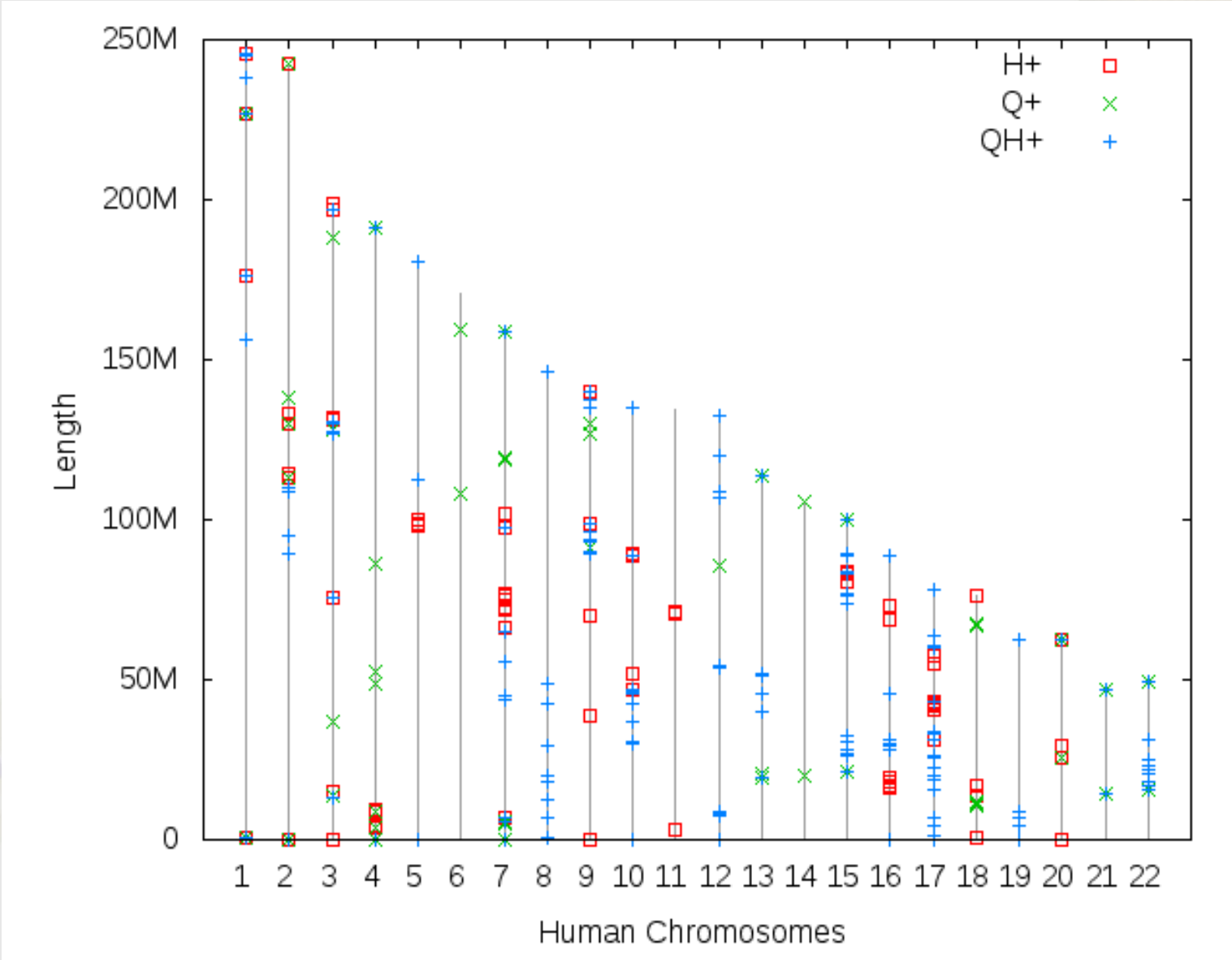
Validation of Predictions for the Macaque-Human Ancestor (QH)

Prediction of fragile regions on **(QH,H)** based on the mouse, rat, and dog genomes:

- ✔ Using mouse genome **M** as a proxy:
  accuracy 34 / 552 ≈ 6%
- ✔ Using mouse-rat-dog ancestor genome **MRD**:
  accuracy 18 / 162 ≈ 11%
- ✔ Using macaque genome **Q**:
  accuracy 10 / 68 ≈ 16%

(using synteny blocks larger than 500K)

Putative Active Fragile Regions in the Human Genome



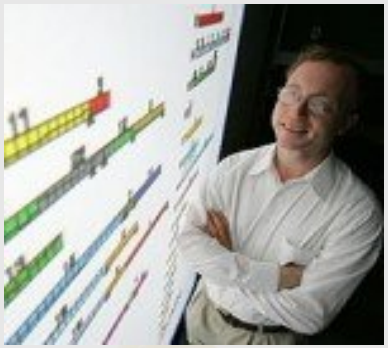
Unsolved Mystery: What Causes Fragility?

- ✔ *Zhao and Bourque, Genome Res. 2009,* suggested that fragility is promoted by *Matching Segmental Duplications*, a pair of long similar regions located within breakpoint regions flanking a rearrangement.

- ✔ TFBM is consistent with this hypothesis since the similarity between MSDs deteriorates with time, implying that MSDs are also subject to a "birth and death" process.

Acknowledgments

Pavel Pevzner, UC San Diego



Glenn Tesler, UC San Diego

Jian Ma, University of Illinois at Urbana-Champaign