

Genome Rearrangements : Examples & Applications

Guillaume Bourque
Genome Institute of Singapore

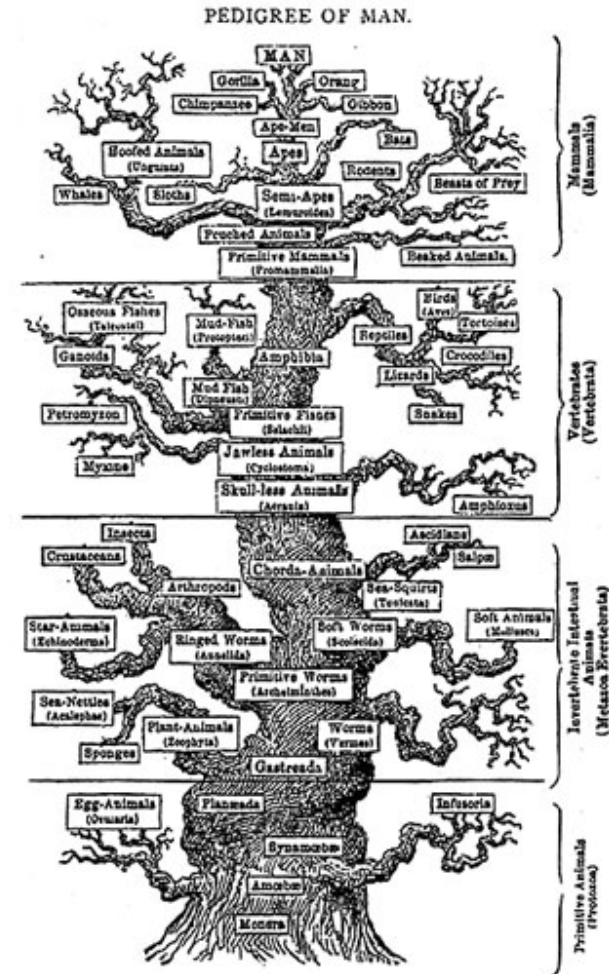
April 2009

Outline

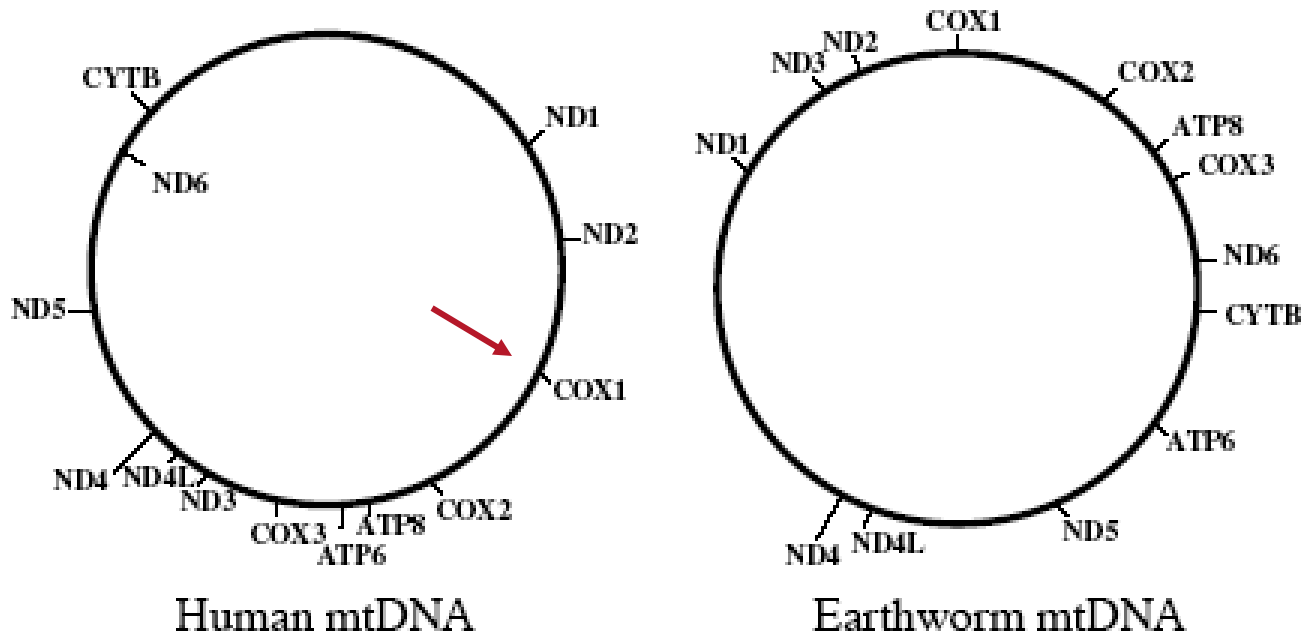
- Sequencing large genomes
- Motivation
- Definitions and pairwise algorithms for genome rearrangements
- Genome rearrangement phylogenies
- Applications and latest developments

Early phylogenetic trees

- Evolution and genetic relationships between species have been represented by **phylogenetic trees** for over a century.
- Haeckel's "Pedigree of Man" Diagram, 1866.



Two mtDNA

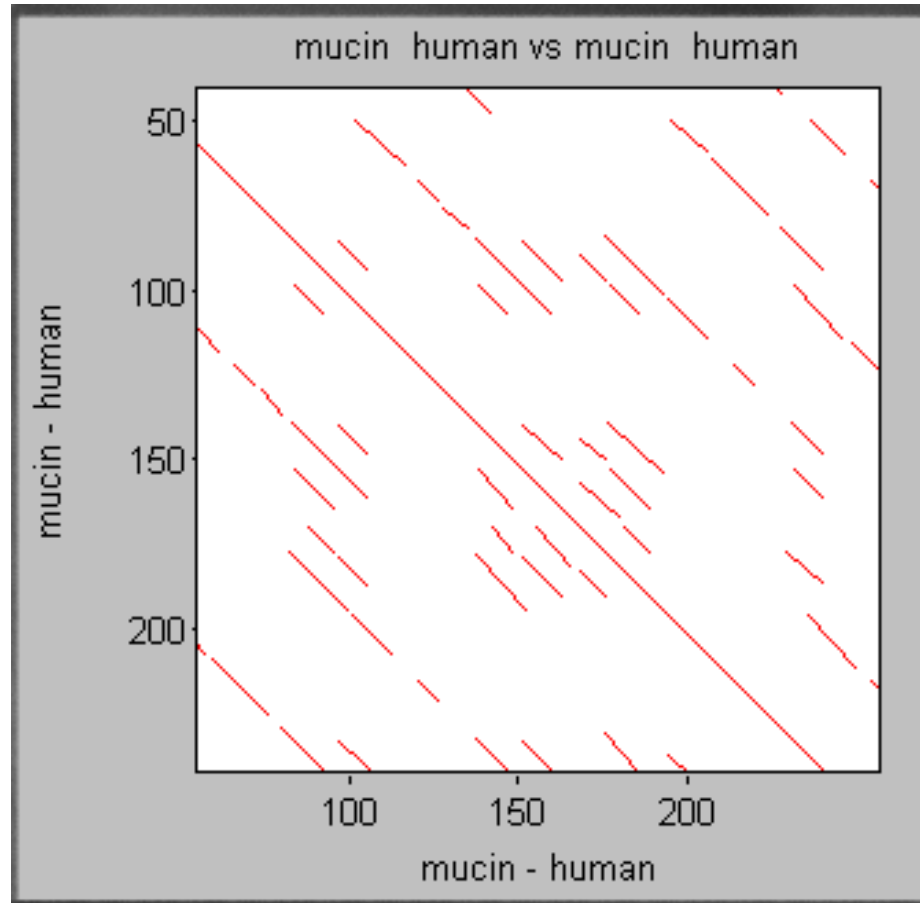


Human:	1	2	3	4	5	6	7	8	9	10	11	12	13
Earthworm:	1	2	3	5	-10	11	4	9	7	8	12	6	13

Dot Plots

- “Invented” in 1970 by Gibbs & McIntyre
- Good for quick graphical overview
- Simplest method for sequence comparison
- Inter-sequence comparison
- Intra-sequence comparison
 - Identifies internal repeats
 - Identifies domains or “modules”

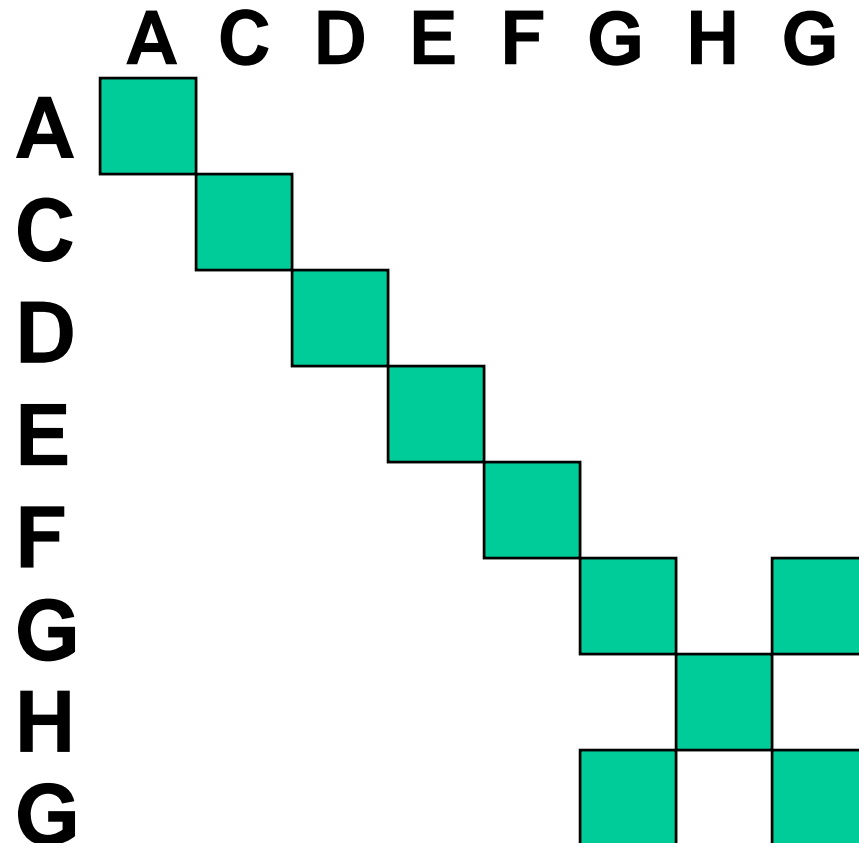
Dot Plots & Internal Repeats



Dot Plot Algorithm


- Take two sequences (A & B), write sequence A out as a row (length= m) and sequence B as a column (length= n)
- Create a table or “matrix” of “ m ” columns and “ n ” rows
- Compare each letter of sequence A with every letter in sequence B. If there’s a match mark it with a dot, if not, leave blank

Dot Plot Algorithm



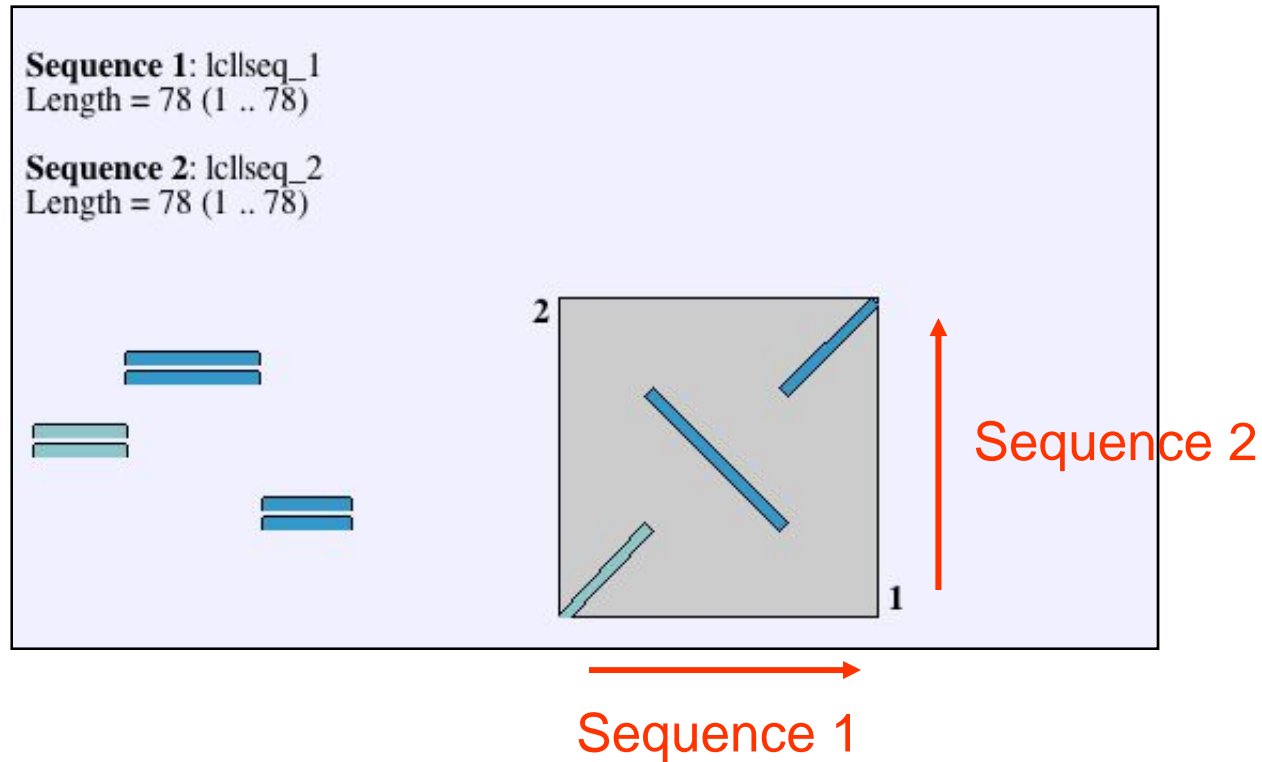
Blast 2 sequences :: Example

1: GGCACAAATCCAAATCCAAATCCGGGTTGGGGTTGGGGTTGGGGTTGCGACACATTTGGCCTGTCGTCGTCCGTCGTC
2: GGCACAAATCCAAATCCAAATCCAAATGTGTCGCAACCCCAACCCCAACCCCAACCCCTGGCCTGTCGTCGTCCGTCGTC



Need to reverse complement

Blast 2 sequences :: Output (1)



Blast 2 sequences :: Output (2)

```

      |||
      |||

Score = 64.1 bits (33), Expect = 5e-08
Identities = 33/33 (100%), Gaps = 0/33 (0%)
Strand=Plus/Minus

Query  24  GGGTTGGGGTTGGGGTTGGGGTTGCGACACATT  56
          |||
Sbjct  56  GGGTTGGGGTTGGGGTTGGGGTTGCGACACATT  24

      |||
      |||

Score = 44.9 bits (23), Expect = 0.028
Identities = 23/23 (100%), Gaps = 0/23 (0%)
Strand=Plus/Plus

Query  1  GGCACAAATCCAAATCCAAATCC  23
          |||
Sbjct  1  GGCACAAATCCAAATCCAAATCC  23

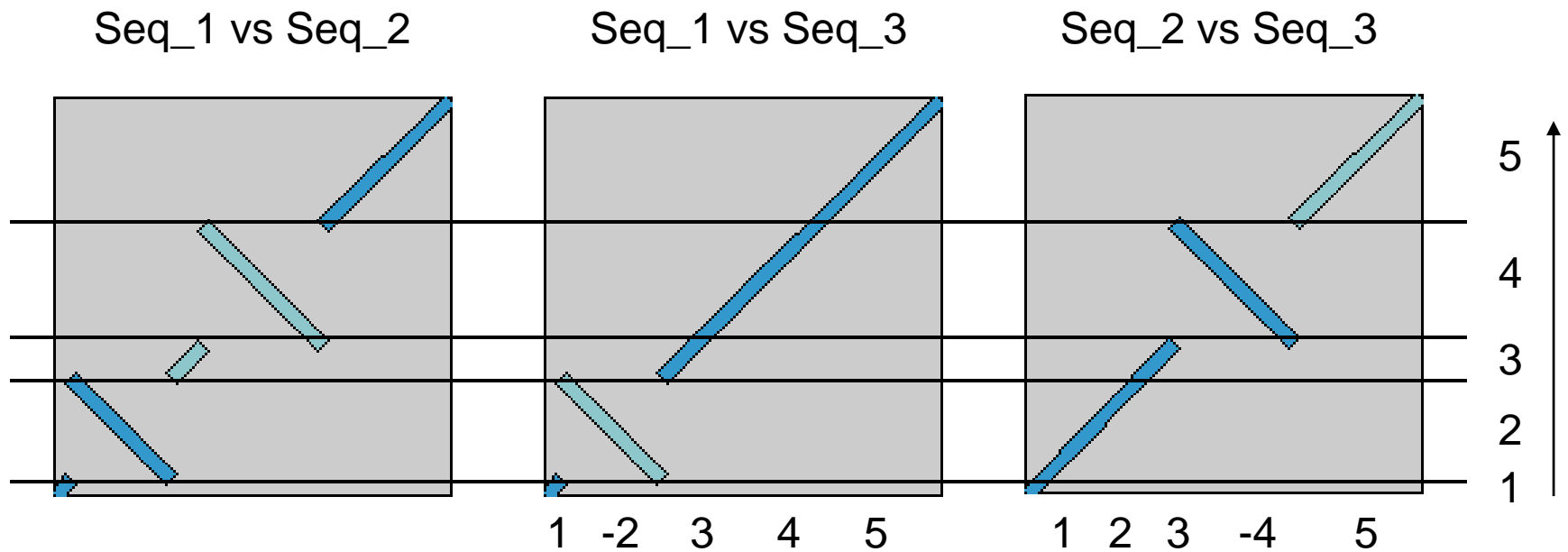
      |||
      |||

Score = 43.0 bits (22), Expect = 0.11
Identities = 22/22 (100%), Gaps = 0/22 (0%)
Strand=Plus/Plus

Query  57  TGGCCTGTCGTCGTCGTCGTC  78
          |||
Sbjct  57  TGGCCTGTCGTCGTCGTCGTC  78

```

What if you have 3 sequences...

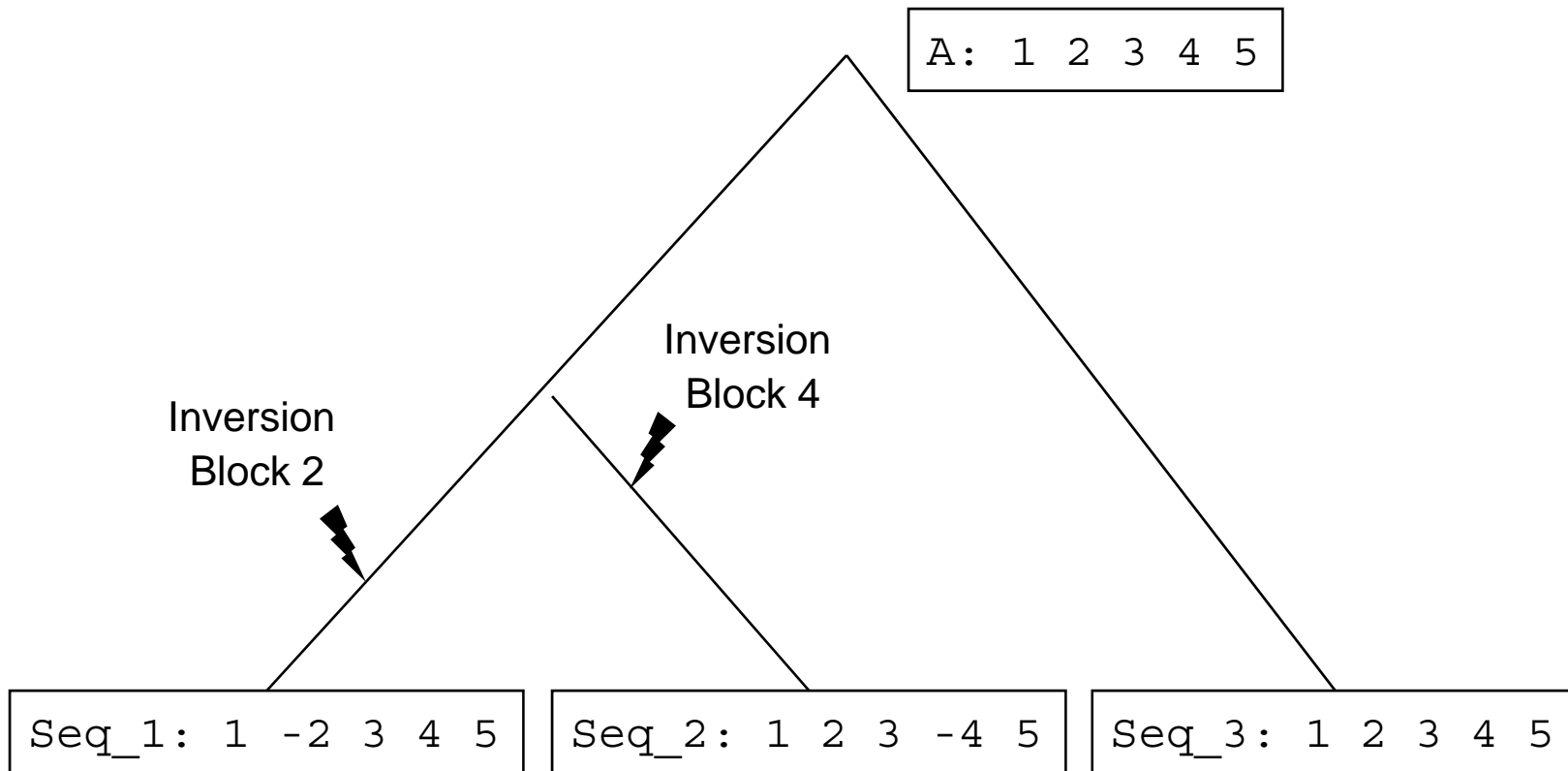


Seq_1	:	1	-2	3	4	5
Seq_2	:	1	2	3	-4	5
Seq_3	:	1	2	3	4	5

What is the tree?

Seq_1 : 1 -2 3 4 5
Seq_2 : 1 2 3 -4 5
Seq_3 : 1 2 3 4 5

Rearrangement Phylogeny



Practicals

Pancake Flipping Problem

- The chef is sloppy; he prepares an unordered stack of pancakes of different sizes
- The waiter wants to rearrange them (so that the smallest winds up on top, and so on, down to the largest at the bottom)
- He does it by flipping over several from the top, repeating this as many times as necessary



Christos Papadimitriou and Bill Gates flip pancakes

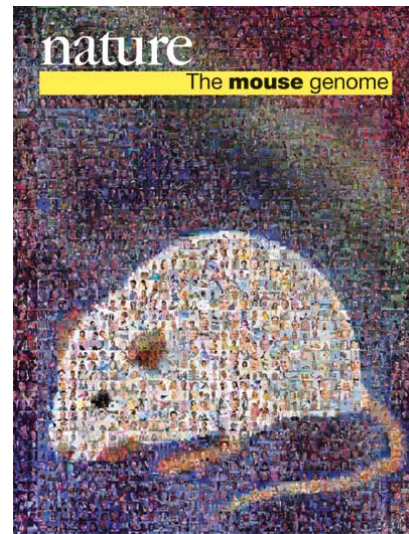
William Gates and Christos Papadimitriou showed in the mid-1970s that this problem can be solved by at most $\frac{5}{3}(n + 1)$ *prefix reversals*

Big genomes

Human, 2001



Mouse, 2002



Rat, 2004



Chicken, 2004



Each of these genomes has over 1 billion base pairs.

Recent Projects

- Human-Mouse-Rat
- Human-Mouse-Rat-Chicken
- 8 Mammalian genomes

Two Step Process

The analysis of genome rearrangements in large genomes (e.g. mammalian genomes) is typically divided into two components:

- 1) Identify orthologous blocks (either based on sequence similarity or orthologous genes).
- 2) Find rearrangement scenario that best explains the observed block arrangements (i.e. the most parsimonious).

Genes or Sequence

- Reasons to use sequence data instead of genes:
 - Avoids annotation problems even works in non-coding regions.
 - Less sensitive to large gene families.
 - Preserves information on micro-rearrangements (rearrangements within conserved blocks).
 - ...
- Reasons to use genes instead of sequence data:
 - Focuses the analysis on important regions of the genome.
 - Thresholds are length independent.
 - Less sensitive to random repeat sequences

Human-Mouse-Rat

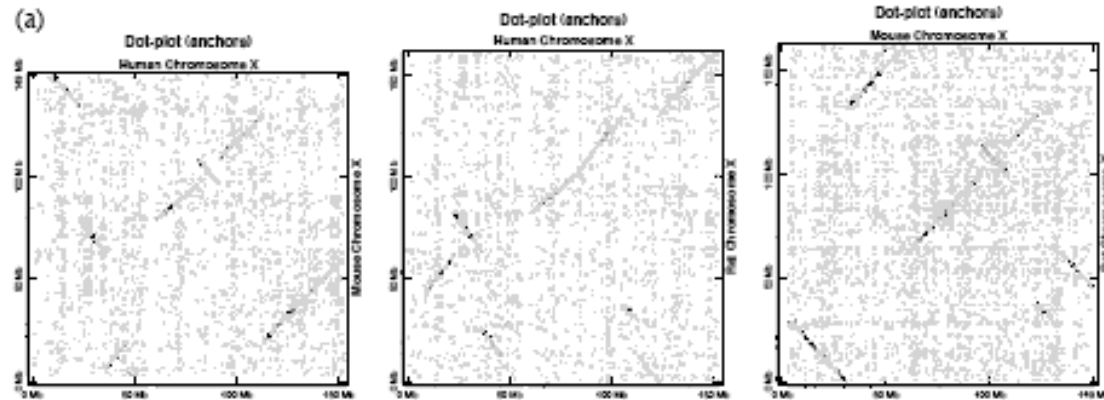
- Rat Genome Consortium (Baylor College, Celera, UCSC, Penn State, Berkeley, Stanford, UCSD, etc.)
- Detailed sequence data with millions of short local alignments.
- Reconstruction of the putative genomic architecture of the ancestral murid rodent genome.
- Detailed history of X chromosome.

Finding Conserved Segments

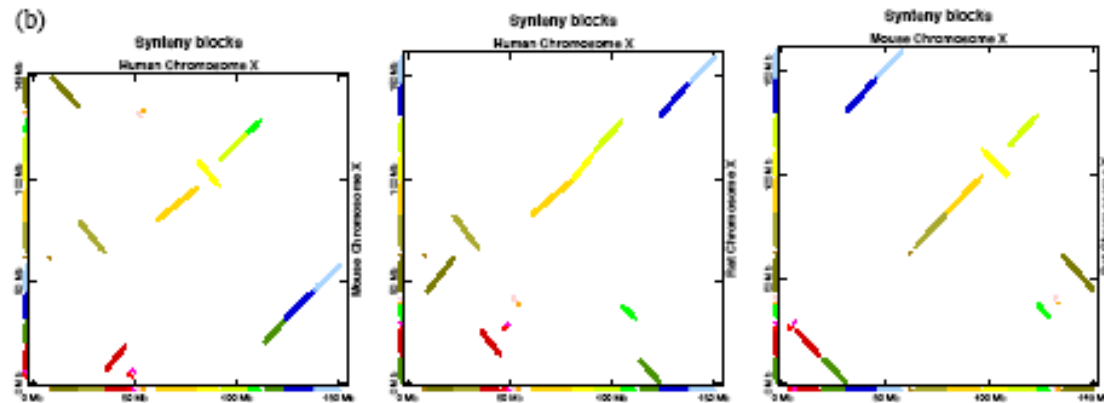
- Mask repeats (using RepeatMasker)
- Find good local alignments (using PatternHunter).
For Human/Mouse/Rat, PatternHunter found about 2.1, 2.2 and 14.6 million alignments for HM, HR and MR respectively (ranging from 30 to 24k bp).
- Remove remaining repeats found in alignments to create 291,000 unique three way anchors.
- Link pairs of anchors when the Manhattan distance is small and keep connected component whose span is over a certain threshold (e.g. at 300 kb we found 391 blocks).
- Done with GRIMM-Synteny by Glenn Tesler.

Chromosome X

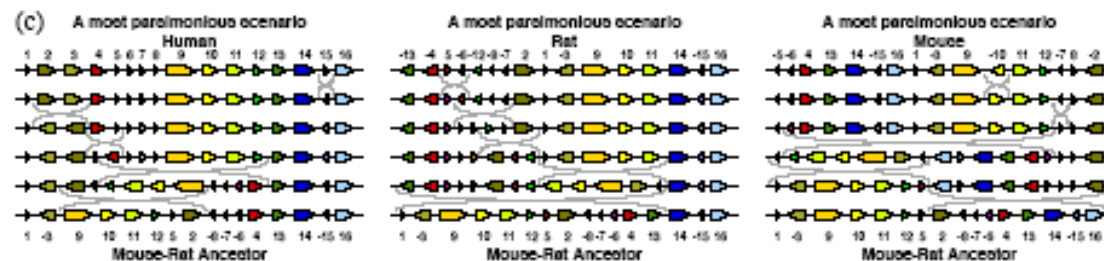
two way similarities
(PatternHunter)



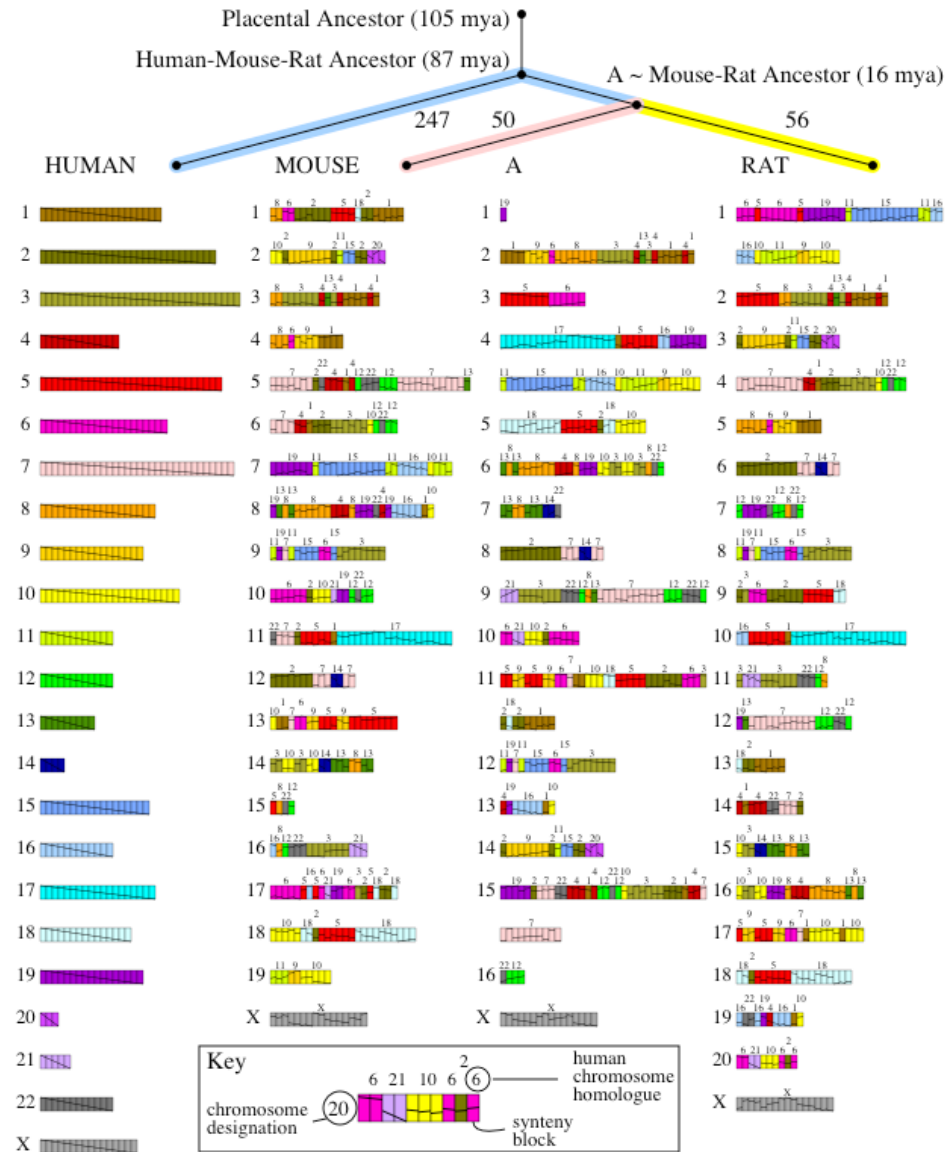
synteny blocks
(GRIMM-Synteny)



rearrangement
scenario
(MGR)



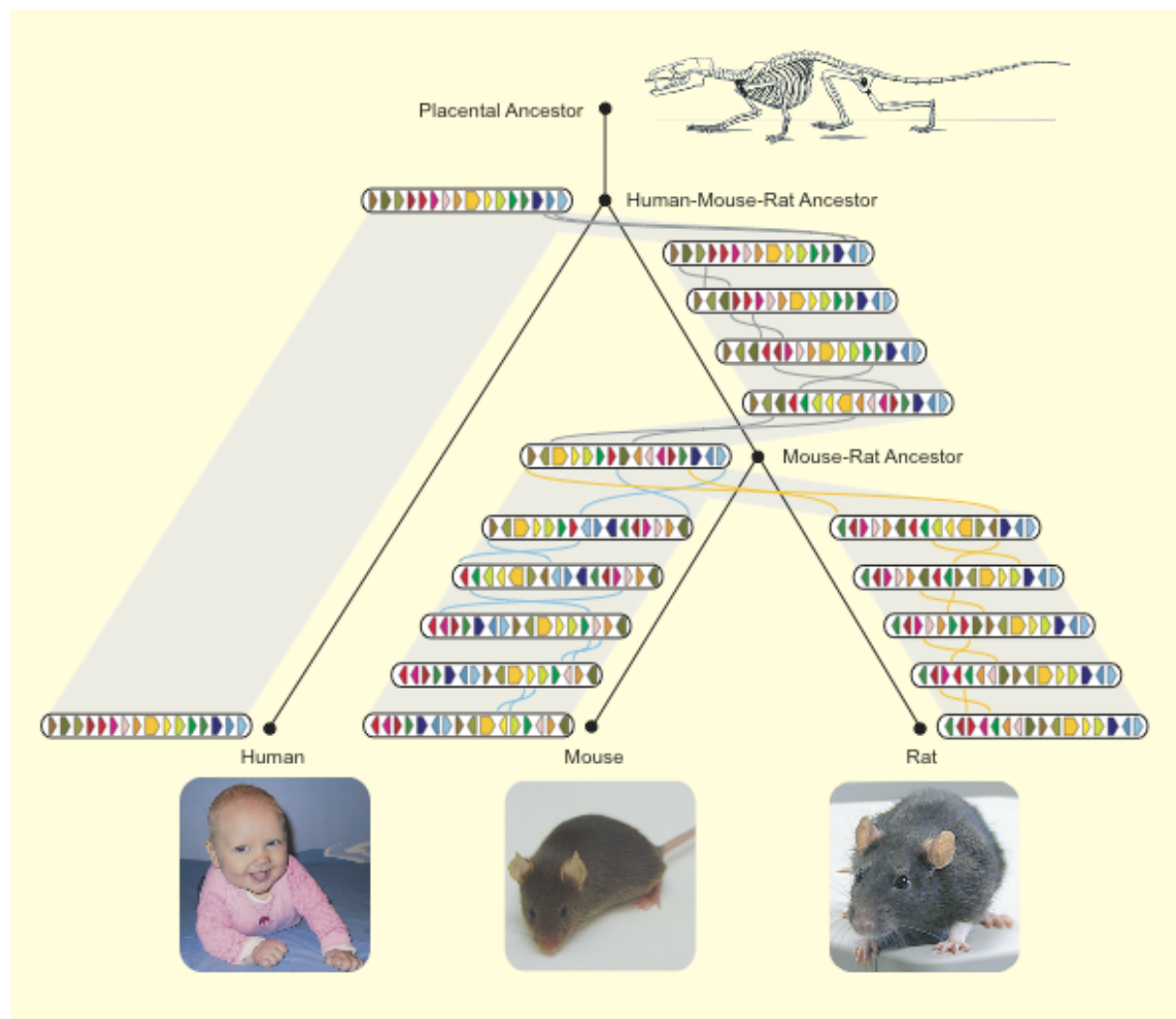
Human-Mouse-Rat Consortium



Chromosome X Scenario

- X chromosome consists of 16 human-mouse-rat orthologous segments of at least 300 kb in size
- In the most parsimonious scenario these were created by 15 inversions in the descent from the primate-rodent ancestor.
- This scenario is guaranteed to be minimum and moreover it can be shown that it is unique.
- This scenario was “roughly” threshold independent.
- Using outgroup data from cat, cow and dog, we found that all or most of these events occurred in the rodent lineage:
 - 0 (or 1) in the human lineage,
 - 5 (or 4) prior to the divergence of rat and mouse
 - 5 in the rat lineage
 - 5 in the mouse lineage

History of Chromosome X



cover of *Genome Research* April
2004

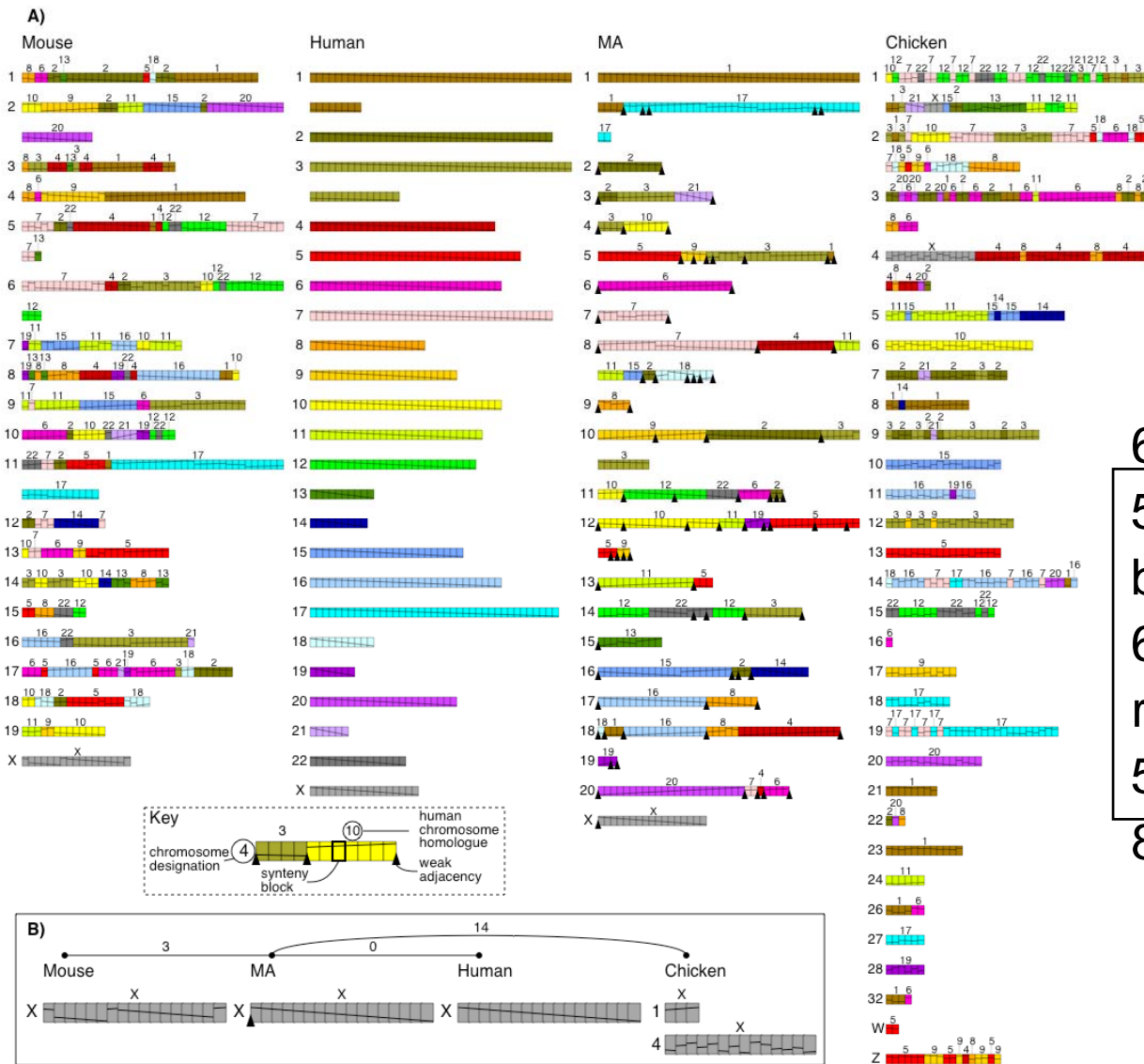
Chicken Consortium

- Analysis of genome rearrangements in Human, Mouse, Rat and Chicken.
- Used two approaches concurrently to identify orthologous blocks:
 - With four-way orthologous genes
 - With four-way unique sequence alignments
- Employed chicken as an outgroup to:
 - Reconstruct the putative architecture of two ancestors (mouse-rat and human-mouse-rat ancestors).
 - Estimate the rates of rearrangements in the different lineages.
- Analyze micro-rearrangements within orthologous blocks.

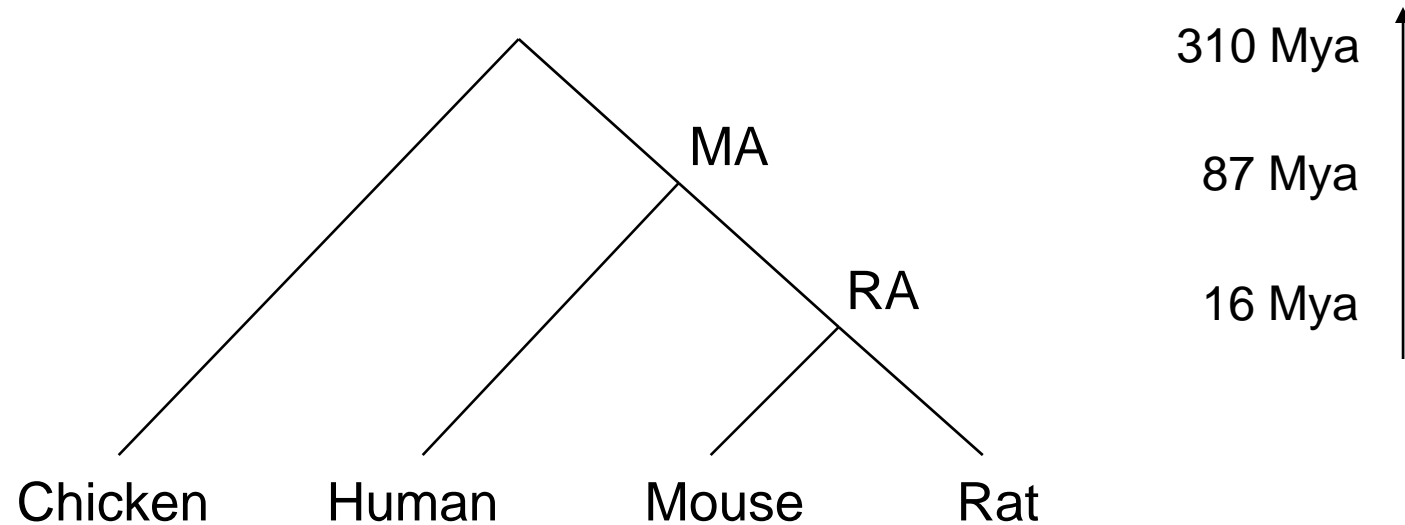
Many Solutions

- Again, because the initial pairwise distances are substantial, it is possible to find alternative ancestors with the same total number of rearrangements.
- Starting from the initial solution recovered by MGR, we explored neighboring alternative solutions using:
 - Depth-first-search approach
 - Breadth-first-search approach
- We keep the first 6000 alternative ancestors (3000 from each) and partitioned the ancestral adjacencies into two categories:
 - Strong adjacencies (adjacencies preserved in all alternative ancestors)
 - Weak adjacencies (... not preserved ...)

Human-Rodent ancestor

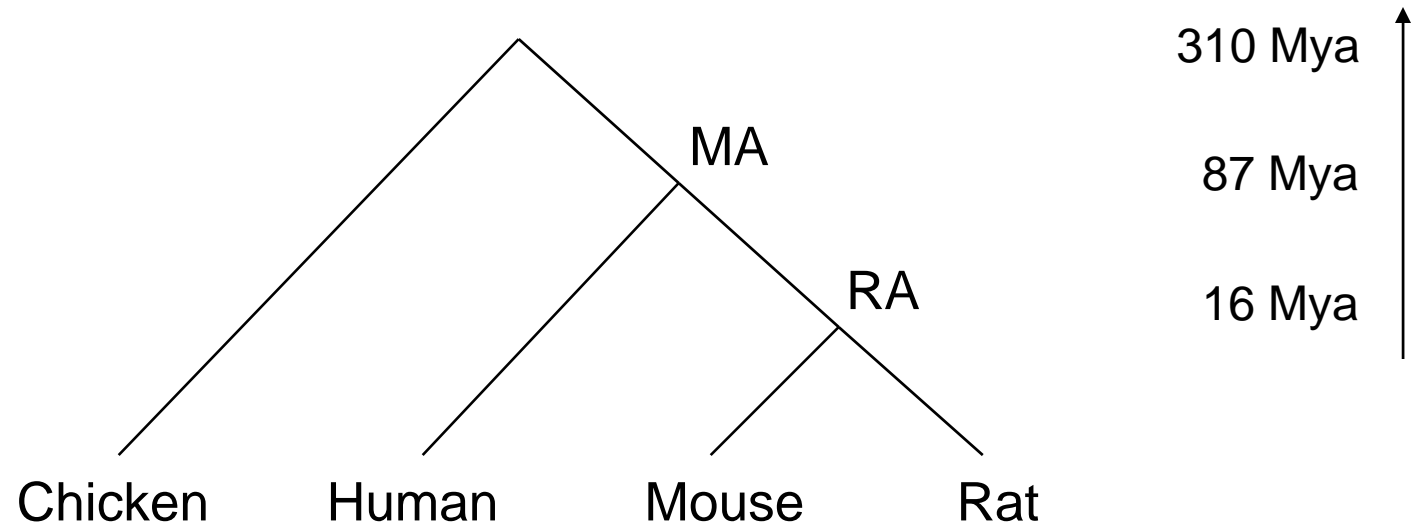


Rates of Rearrangements



	All rearrangements							Inter-chromosomal						
	Gene-based (geneX)				Sequence-based			Gene-based (geneX)				Sequence-based		
Run	4	7	10	20	100K	200K	300K	4	7	10	20	100K	200K	300K
<i>MA-Chicken</i>	4.2	5.3	4.4	4.7	3.9	3.7	4.9	2.5	3.4	2.7	3.4	2.3	2.1	3.7
<i>RA-Mouse</i>	0.4	0.5	0.5	0.5	0.5	0.5	0.7	0.6	0.8	0.7	0.8	0.5	0.5	0.7
<i>RA-Rat</i>	0.5	0.6	0.5	0.6	0.5	0.5	0.5	0.4	0.5	0.5	0.6	0.4	0.3	0.5
<i>MA-RA</i>	1.4	1.7	1.4	1.4	1.5	1.5	2.1	2.1	2.4	2.0	1.9	1.6	1.6	2.3
<i>MA-Human</i>	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Rates of Rearrangements

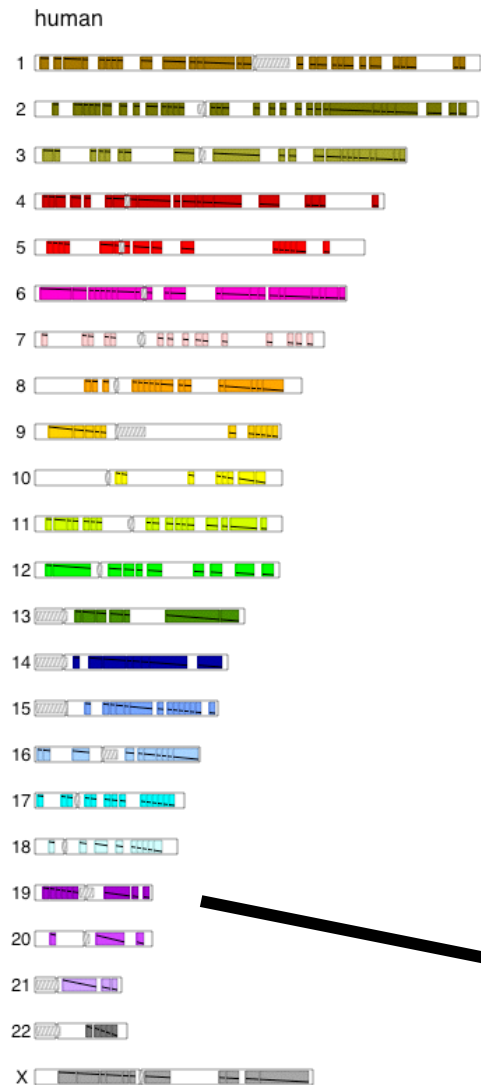


- High ratio of intra-chromosomal vs. inter-chromosomal rearrangements on the chicken edge.
- Similar number of rearrangements on the MA-chicken and on the MA-Mouse edge.
- Rate of inter-chromosomal rearrangements: 0.19 per Mya on MA-chicken, 0.34 on MA-human and 1.1 on MA-mouse.

8 Mammalian Genomes

- Integrate data from 3 sequenced mammals (human-mouse-rat) and 5 high density RH map of other mammals (cat-dog-pig-cow-horse).
- Use a slightly different method to recover the conserved segments since we are dealing with ordered data with no base-pair position but similar ideas of proximity and minimum cluster size.
- Look at the predicted ancestral chromosomal associations and compare with other lower resolution studies.
- Test breakpoint reuse hypothesis.

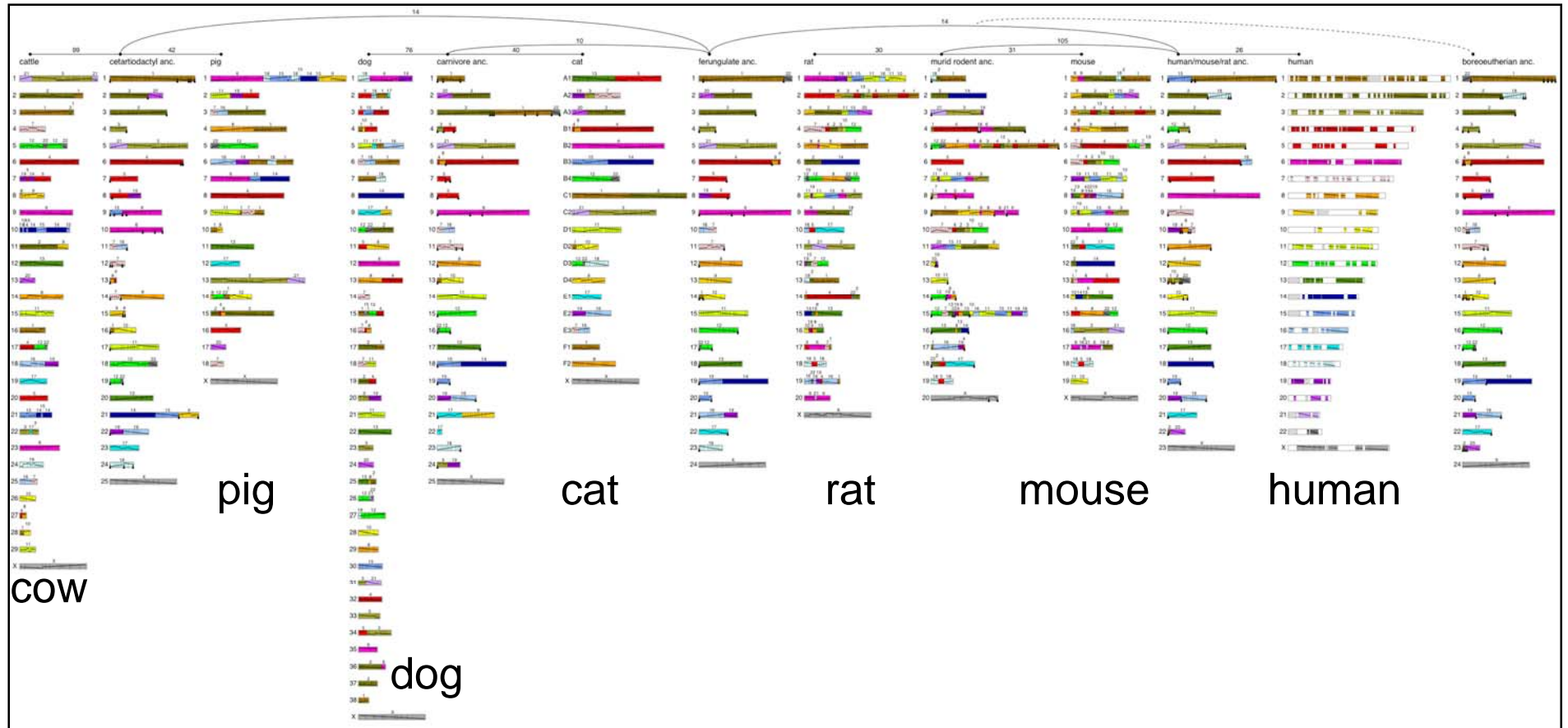
Coverage of blocks in Human



- 307 conserved blocks
- Blocks are color coded based on human chromosome homologue
- Blocks are traversed by a diagonal line to indicate relative order (in the other genomes)

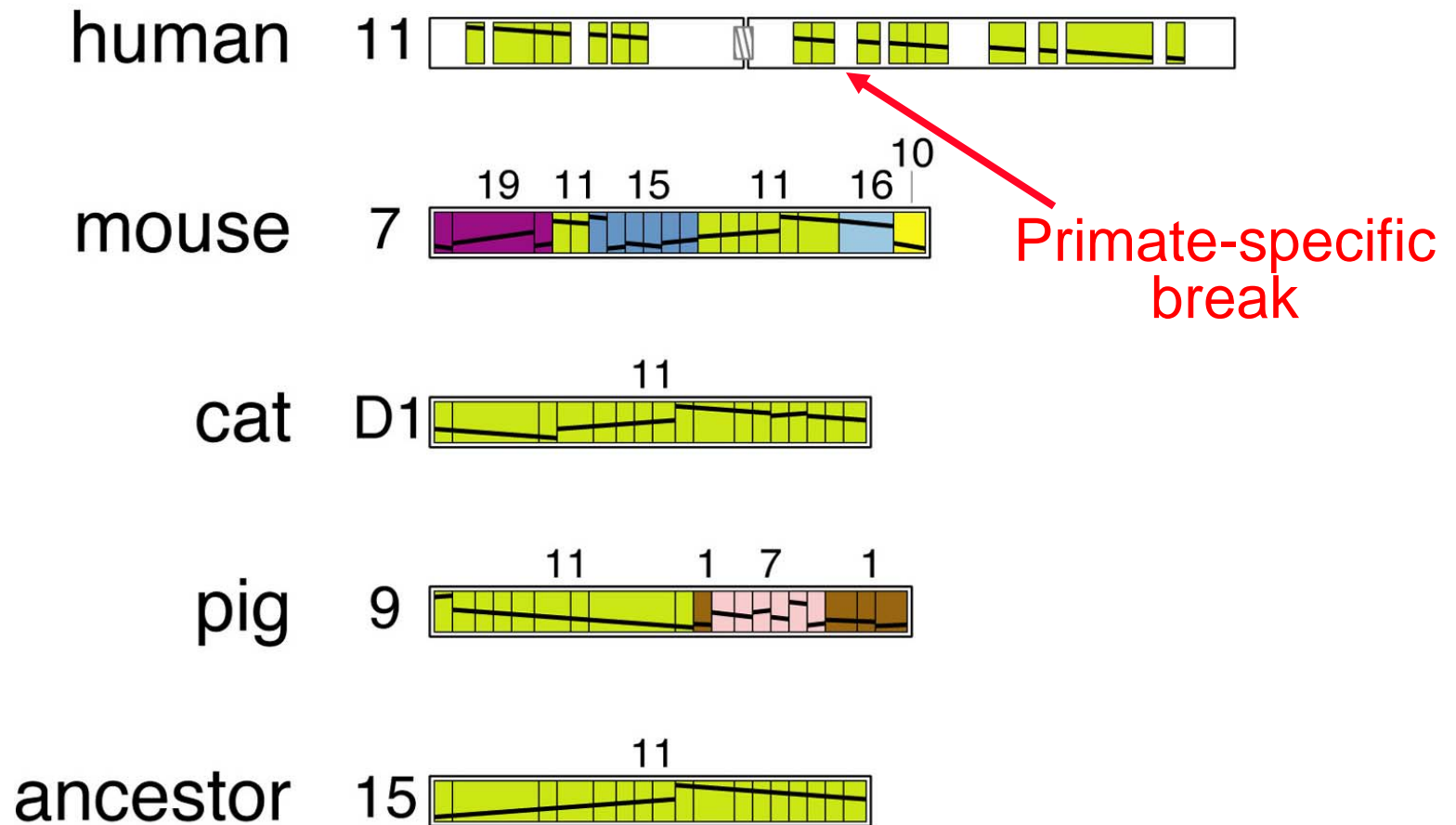


Evolutionary scenario

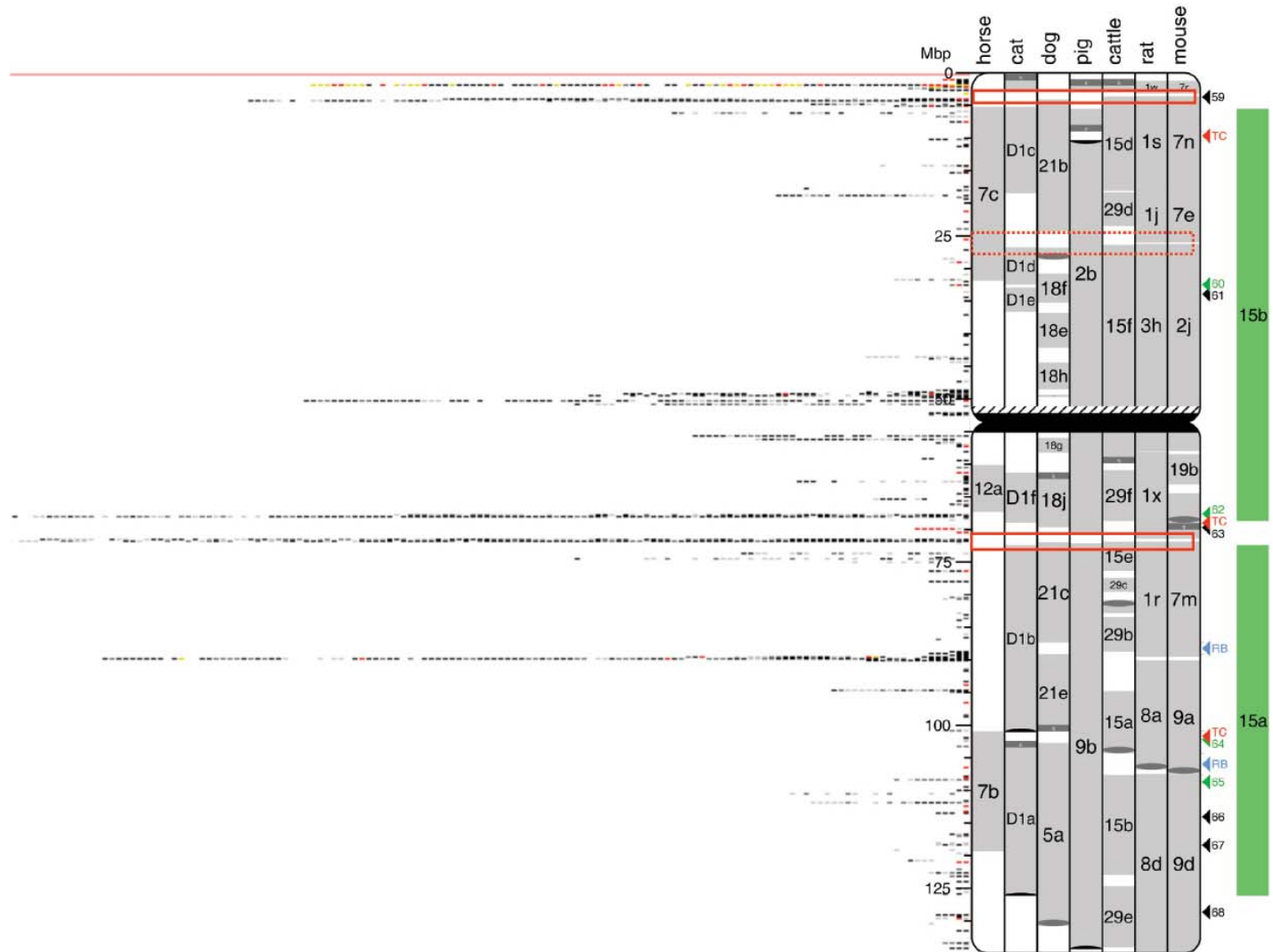


Murphy et al, Science, 2005

Human Chromosome 11



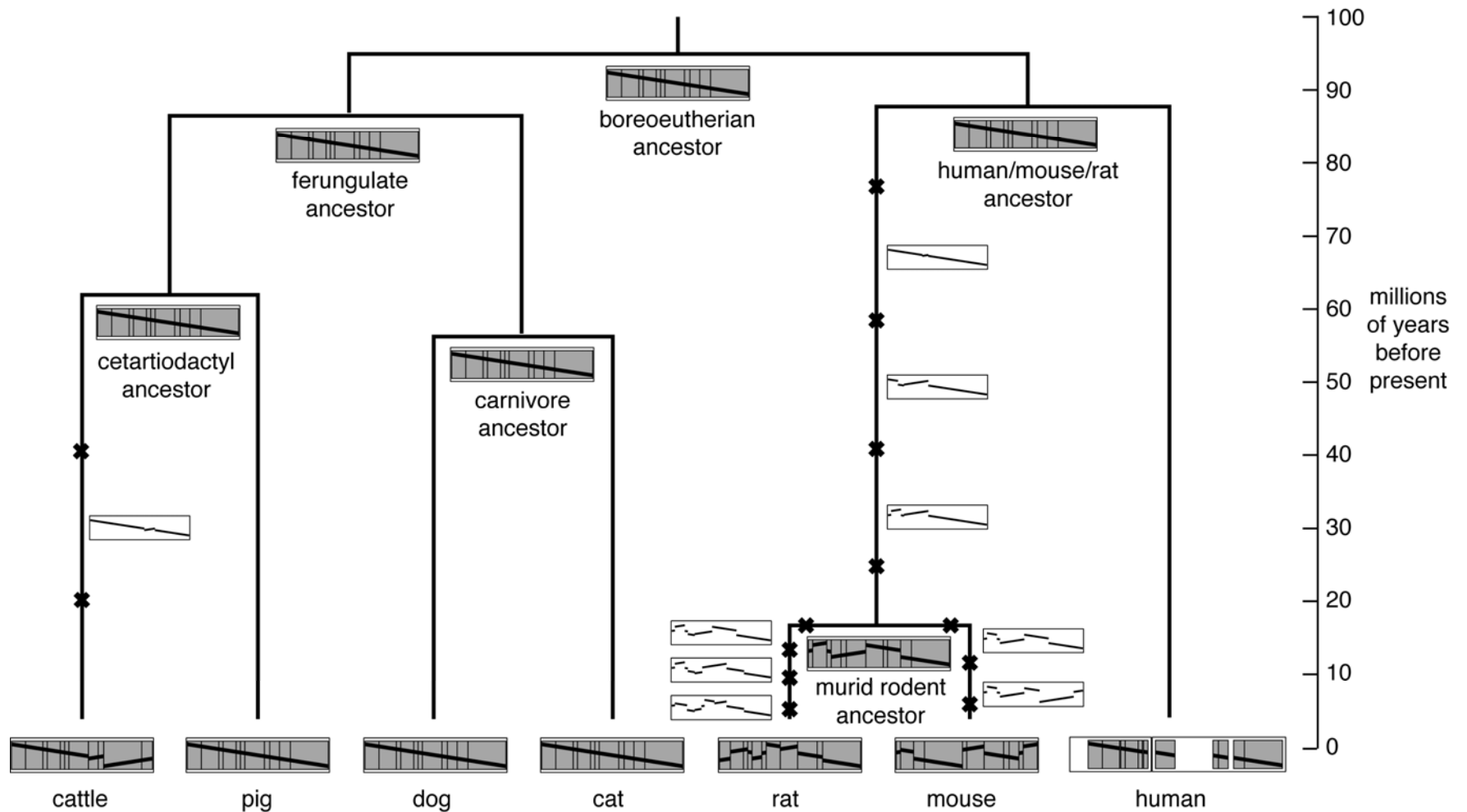
Human Chromosome 11



Overview of the Results

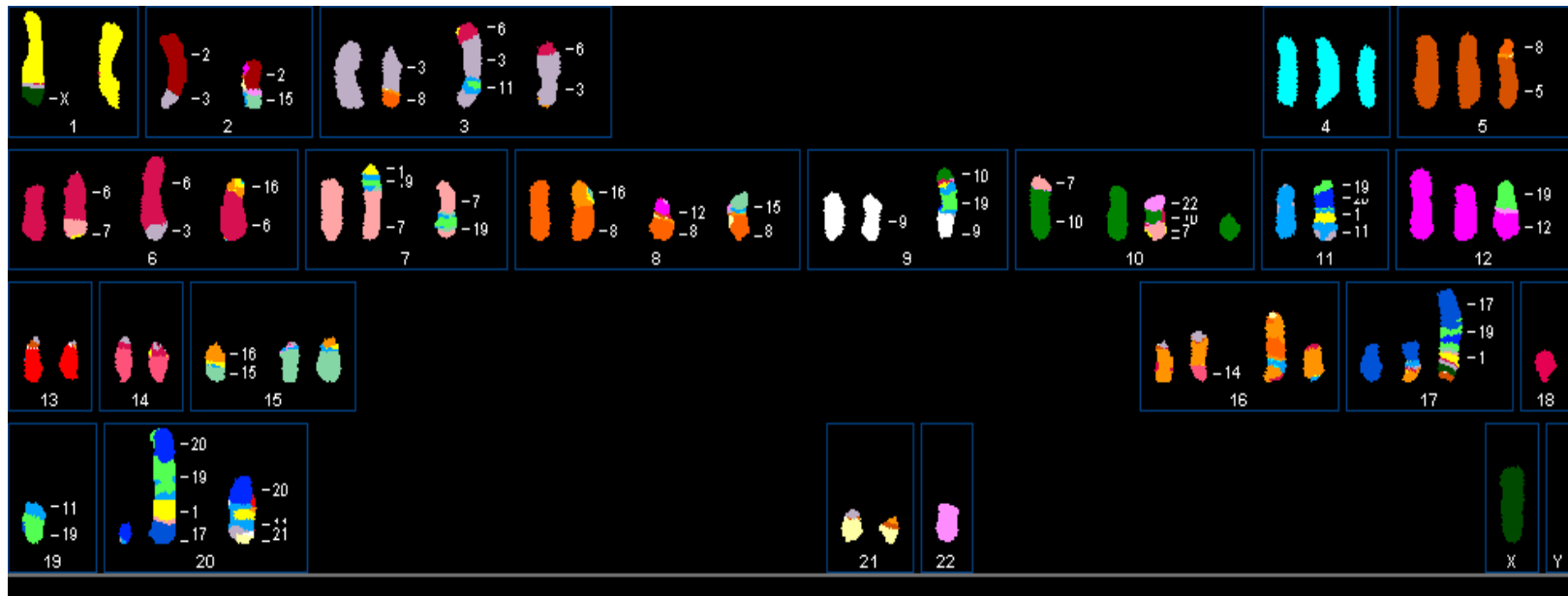
- Nearly 20% of chromosome breakpoint regions were reused.
- Gene-density is higher in evolutionary breakpoint regions.
- Segmental duplications populate the majority of primate-specific breakpoints.
- There is correlation between human cancer breakpoints and evolutionary breakpoints.

X chromosome evolution



Cancer Rearrangements...

Highly rearranged cancer genome



Provided by Nalla Palanisamy, GIS

- Karyotype of an MCF7 cell shows a highly rearranged cancer genome.
- DNA fragments from this genome won't always be directly mappable to the normal human genome.

Evolutionary breakpoints and cancer

- Initially we found a correlation between evolutionary breaks and recurrent cancer breakpoints.
- What about breaks found in many different types of cancer?

<i>BCL2</i> ●▼	<i>ELL</i> ^b ◆●▲	<i>HOXD11</i> ^b ◆	<i>MN1</i> ◆○▲△	<i>PRCC</i> □
<i>BCL3</i> ³ ▼	<i>EP300</i> ³ ◆	<i>HOXD13</i> ^b ◆	<i>MSF</i> ◆	<i>PRDM16</i> ^b ◆▲
<i>BCL6</i> ⁶ ●▼	<i>EPS15</i> ◆	<i>HPR</i> □	<i>MSI2</i> ○	<i>PRKAR1A</i> □
<i>BCL7A</i> ▼	<i>ERG</i> ^b ◆●△	<i>HSRNAFEV</i> ³ △	<i>MTCP1</i> ◇	<i>PRRX1</i> ^b ◆
<i>BCL8</i> ▼	<i>ERVWE1</i> ▼	<i>IGH@</i> ●▼■◇	<i>MUC1</i> ▼	<i>PSIP1</i> ^b ◆
<i>BCL9</i> ●▼	<i>ETS1</i> ^b ▼	<i>IGK@</i> ▼■	<i>MYB</i> ^b □	<i>PVT1</i> ●▼
<i>BCL10</i> ▼	<i>ETV1</i> ^b △	<i>IGL@</i> ●▼■	<i>MYC</i> ^b ●▼■◇	<i>RABEP1</i> ▲
<i>BCL11A</i> ^b ▼	<i>ETV4</i> ^b △	<i>IL2</i> ◇	<i>MYH11</i> ◆○▲	<i>RAD51L1</i> △
<i>BCL11B</i> ^b ●	<i>ETV6</i> ^b ◆●○▲▼◇△□	<i>IL21R</i> ▼	<i>MYST3</i> ^b ◆▲	<i>RANBP2</i> △
<i>BCR</i> ◆●○▲▼■◇	<i>EVI1</i> ^b ◆○▲	<i>IL3</i> ●	<i>MYST4</i> ^b ◆▲	<i>RANBP17</i> ●
<i>BIRC3</i> ▼■	<i>EWSR1</i> ●△□	<i>IRF1</i> ^b ●	<i>NCKP1SD</i> ◆	<i>RAP1GDS1</i> ●
<i>BRCA2</i> ▲	<i>FCGR2B</i> ▼	<i>IRF4</i> ^b ■	<i>NCOA2</i> ◆	<i>RARA</i> ^b ◆
<i>BRD4</i> □	<i>FGFR1</i> ^a ◆○	<i>JAK2</i> ^a ●○	<i>NCOA4</i> □	<i>RBM15</i> ◆
<i>BTG1</i> ▼	<i>FGFR10P</i> ◆○	<i>JAZF1</i> ^b △	<i>NFKB2</i> ^b ▼■◇	<i>REL</i> ^b ●▼
<i>BVR1</i> ▼	<i>FGFR3</i> ^a ■◇	<i>JJAZ1</i> ^b △	<i>NONO</i> □	<i>RET</i> ^a □
<i>C12orf9</i> △	<i>FHIT</i> □	<i>KIAA1618</i> ◇	<i>NOTCH1</i> ^b ●	<i>RFG9</i> □
<i>CARS</i> △	<i>FIP1L1</i> ○	<i>KTN1</i> □	<i>NPM1</i> ◆▼◇	<i>ROS1</i> ^a △
<i>CBFA2T1</i> ^b ◆○▲	<i>FLI1</i> ^b △□	<i>LAF4</i> ^b ●	<i>NR4A3</i> ^b △	<i>RPL22P1</i> ◆○
<i>CBFA2T3</i> ^b ◆▲	<i>FN1</i> ◆	<i>LAMA4</i> △	<i>NRG1</i> □	<i>RPN1</i> ◆▲

Mitelman, Johansson & Mertens, Nature Genetics, 2004

Conclusions

- Genome rearrangements is relatively new field with many very interesting mathematical and computational problems.
- Promising new developments:
 - Maximum likelihood approaches for the analysis of rearrangements.
 - Phylogenies under conservations criterias.
 - More realistic approaches that combine reversals, translocations, duplications, etc.
 - ...
- Interesting biological problems:
 - Genome rearrangements in cancer.
 - Analysis of evolutionary breakpoints.
 - ...

Acknowledgments

- Pavel Pevzner and Glenn Tesler (UCSD)
- Bill Murphy (Texas A&M), Denis Larkin and Harris Lewin (U of Illinois)
- At GIS:
 - Ed Liu, Vinsensius Vega, Ken Sung, Kuo Ping Chiu, Hong Sain Ooi,
...
- At NUS:
 - Louxin Zhang (Math)
 - Hon Wai Leong (CS)