

Estimate of Genome-Wide Mutation Rate Difference between Human and Mouse Ancestral Repeats

HoJoon Lee, Sankar Subramanian, and Sudhir Kumar

Center for Evolutionary Functional Genomics,

The Biodesign Institute

Arizona State University, Tempe, AZ 85287-5301, USA.

CONTENTS

1. Introduction

2. Methods

3. Results

4. Discussion

5. Reference

CONTENTS

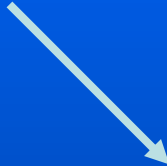
1. Introduction

2. Methods

3. Results

4. Discussion

5. Reference



Molecular Evolution

Neutral Substitution

Phylogenetic Analysis

Interspersed Repeats

History

Molecular evolution

“Nothing in biology makes sense except of in the light of evolution”

Dobzhansky, 1973

“Nothing in molecular biology makes sense except of in the light of molecular evolution”

HoJoon Lee, 2005

Molecular evolution

- **Molecular Evolution**

- : “At the molecular level, evolution is a process of mutation with selection.” (Pevsner 2003)

- **Mutation**

- : Mutation indicates any change in DNA

- The primary cause of evolution

- ex) **substitution**, insertions, deletions, and inversions

- **Fixation**

- : Mutations are established in the population

Molecular evolution

- **The three effects on fitness by mutations of DNA**

- 1. Negative (or purifying)**

Mutations may be harmful, reducing the chance of surviving or reproducing of progeny

- 2. Positive**

Mutations may enhance fitness by making organisms to adapt to changes in the environment

- 3. Neutral**

Mutations have no effect on fitness due to no phenotypic change

Neutral Substitution

- Under effectively neutral of selection (Kimura 1983),

$$\textit{Mutation rate} = \textit{Substitution rate}$$

$$\begin{aligned} \text{Rate of substitution} &= \text{Total mutation rate} \cdot \text{fixation probability} \\ s &= 2N \cdot u \cdot (1/2N) \\ &= u \end{aligned}$$

Where u = mutation rate, N = the number of diploid population

- **Mutation rate** is considered as the rate at which changes are integrated into a nucleotide sequence
- **Substitution (or fixation) rate** refers to the rate at which changes are established in the population

Neutral Substitution

The mutation rate can be estimated from the neutral substitution rate!!

Interspersed Repeats

- **Neutral Sites**

1. Fourfold degenerate sites (i.e. third position in codon)
2. Introns & Intergenic regions
3. Pseudo genes
4. Interspersed repeats

Interspersed Repeats

- 46% of the human genome and 37.5 % of the mouse genome are recognized as interspersed repeats

- 4 types of interspersed repeats

 - LINE, SINE, LTR, and DNA transposon

- Ancestral repeats

 - Interspersed repeats reside in both the human genome and the mouse genome

- Our list contains 135 ancestral repeat families that reported in Waterston *et al.* (2002)

Interspersed Repeats

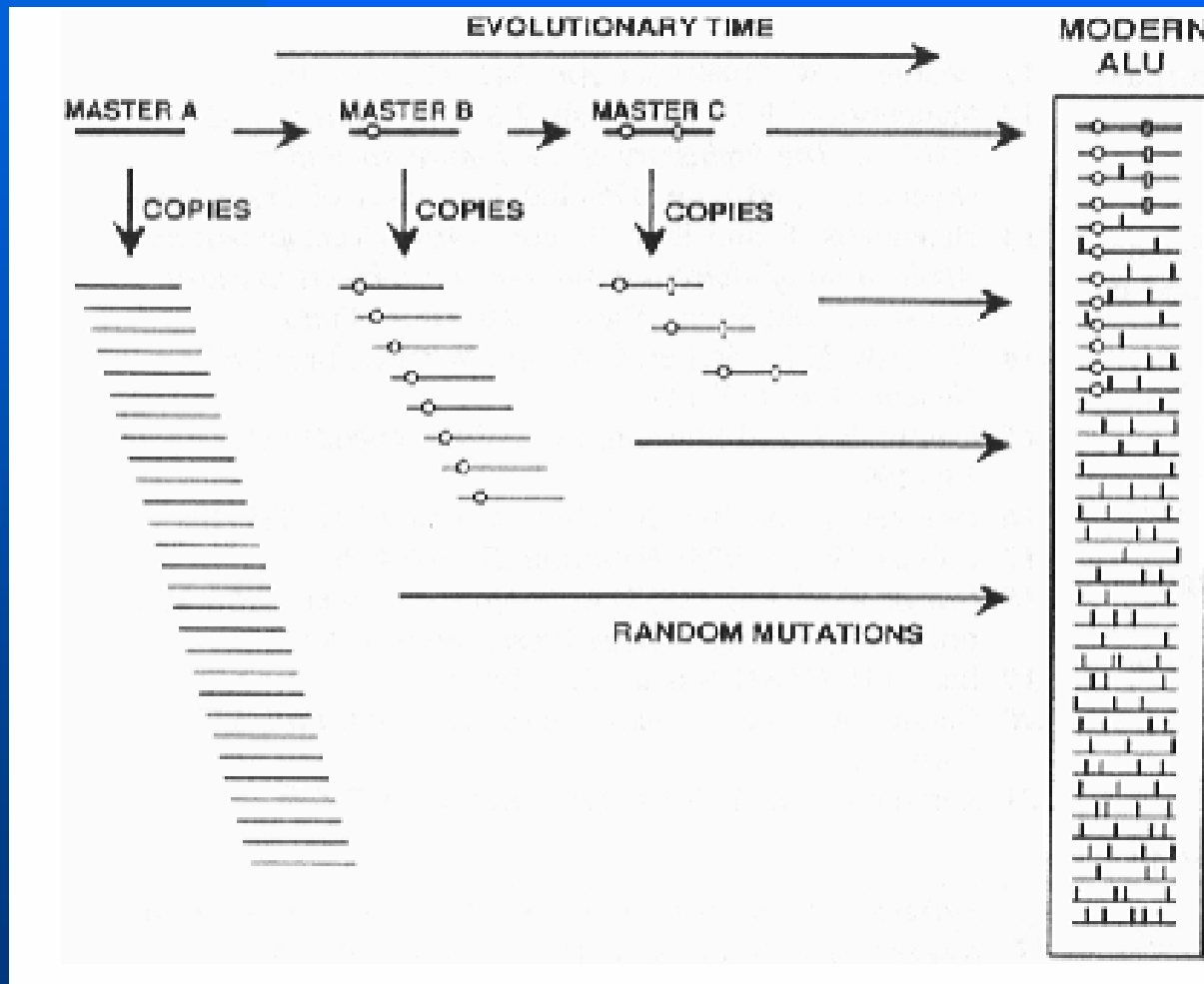
Classes of interspersed repeat in the human genome

			Length	Copy number	Fraction of genome
LINEs	Autonomous		6–8 kb	850,000	21%
	Non-autonomous		100–300 bp		
Retrovirus-like elements	Autonomous		6–11 kb	450,000	8%
	Non-autonomous		1.5–3 kb		
DNA transposon fossils	Autonomous		2–3 kb	300,000	3%
	Non-autonomous		80–3,000 bp		

From Waterston *et al.* (2002)

Interspersed Repeats

- Evolutionary History of Interspersed Repeats



Prescott *et al.* (1992)

Historical Controversies

Authors	Year	Data Set	Way to Estimate	Relative Rate*
Laird et. al and Kohne et. al	1969, 1970	Coding sequence of α , β chain of hemoglobin	DNA-DNA hybridization	$\gg 10.0$
Wu and Li	1985	Synonymous sites of 24 protein coding gene	Relative rate	~ 2.0
Kumar and Subramanian	2002	4-fold degenerate sites of 11 genes	Relative rate	1.2 - 1.7
Waterston et. al	2002	Ancestral repeats of human and mouse genomes	Relative divergence from consensus sequence	~ 2.0

* Relative rate = mouse / human

Historical Controversies

Laird *et al.* (1969)
Kohne *et al.* (1970)
Wu and Li (1984)
Waterson *et al.* (2002)
Hardison *et al.* (2003)

VS

Kumar and Subramanian (2002)

+

Lee *et al.* (2005)

INTRODUCTION

The purpose of our research is to study relative mutation rate between human and mouse genomes using ancestral repeats by phylogenetic analysis

CONTENTS

1. Introduction

2. Methods

3. Results

4. Discussion

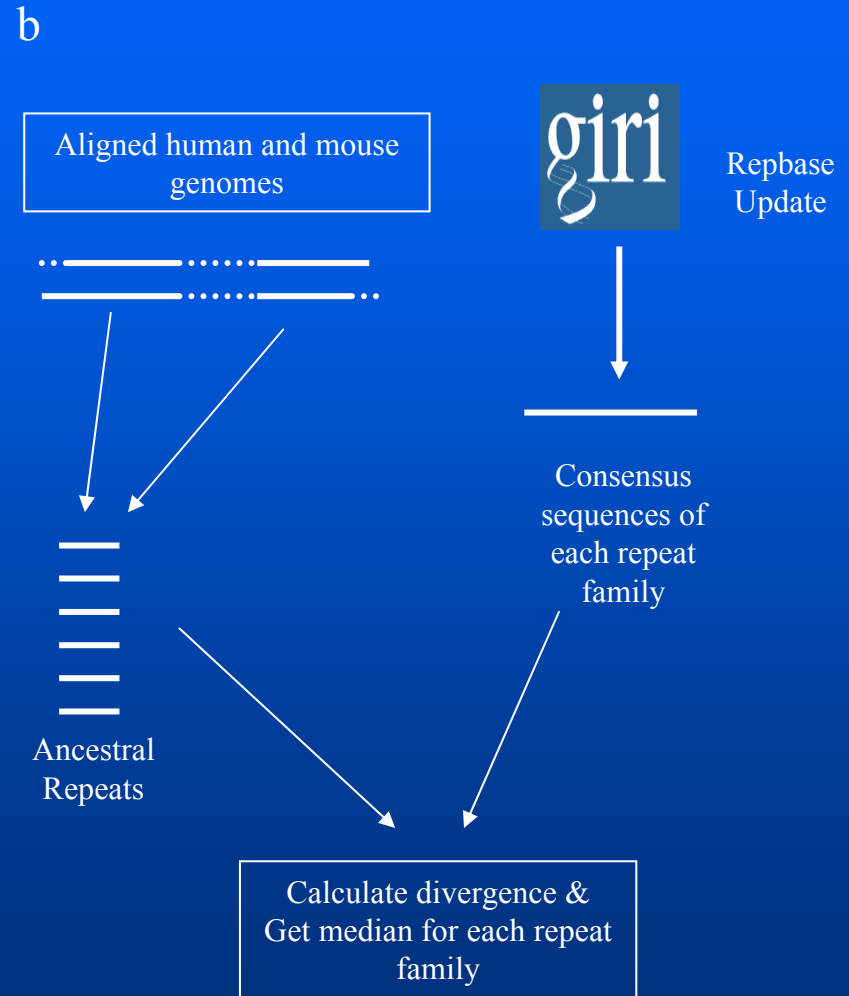
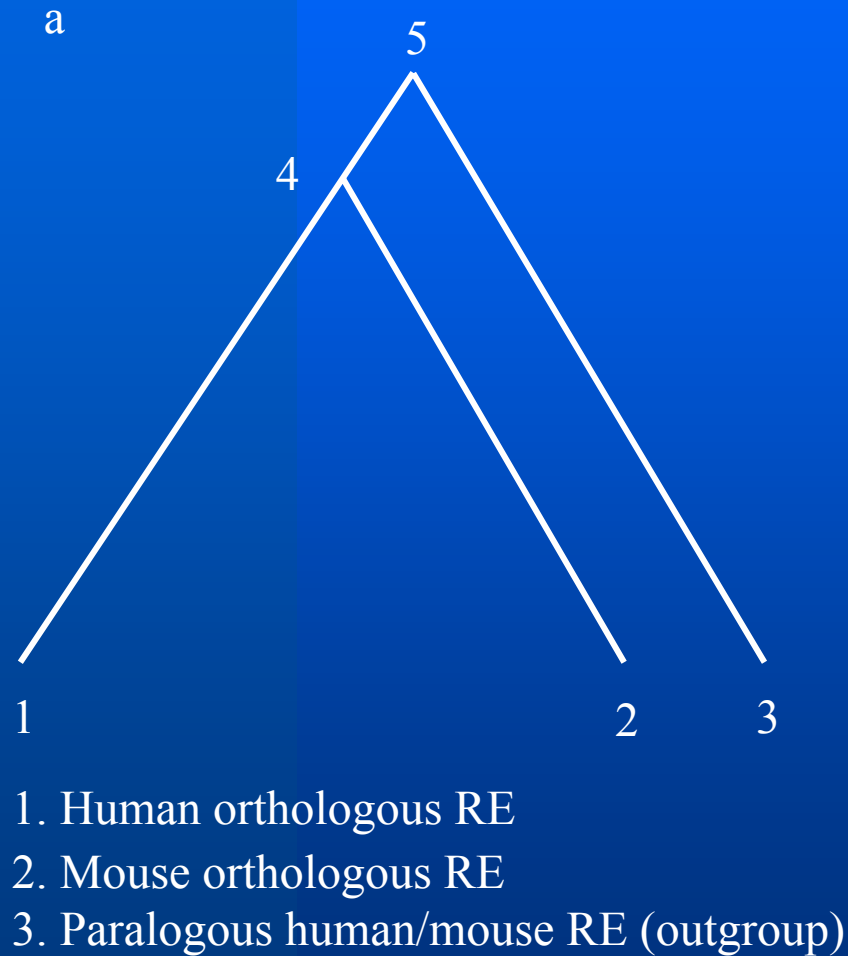
5. Reference

Historical Controversies

Authors	Year	Data Set	Way to Estimate	Relative Rate*
Laird et. al and Kohne et. al	1969, 1970	Coding sequence of α , β chain of hemoglobin	DNA-DNA hybridization	$\gg 10.0$
Wu and Li	1985	Synonymous sites of 24 protein coding genes	Relative rate	~ 2.0
Kumar and Subramanian	2002	4-fold degenerate sites of 11 genes	Relative rate	1.2 - 1.7
Waterston et. al	2002	Ancestral repeats of human and mouse genomes	Relative divergence from consensus sequence	~ 2.0

* Relative rate = mouse / human

Phylogenetic Analysis



Phylogenetic Analysis

1. Selection of homologous sequences
2. Multiple sequence alignment
3. Tree building
4. Measure of divergence
5. Tree evaluation

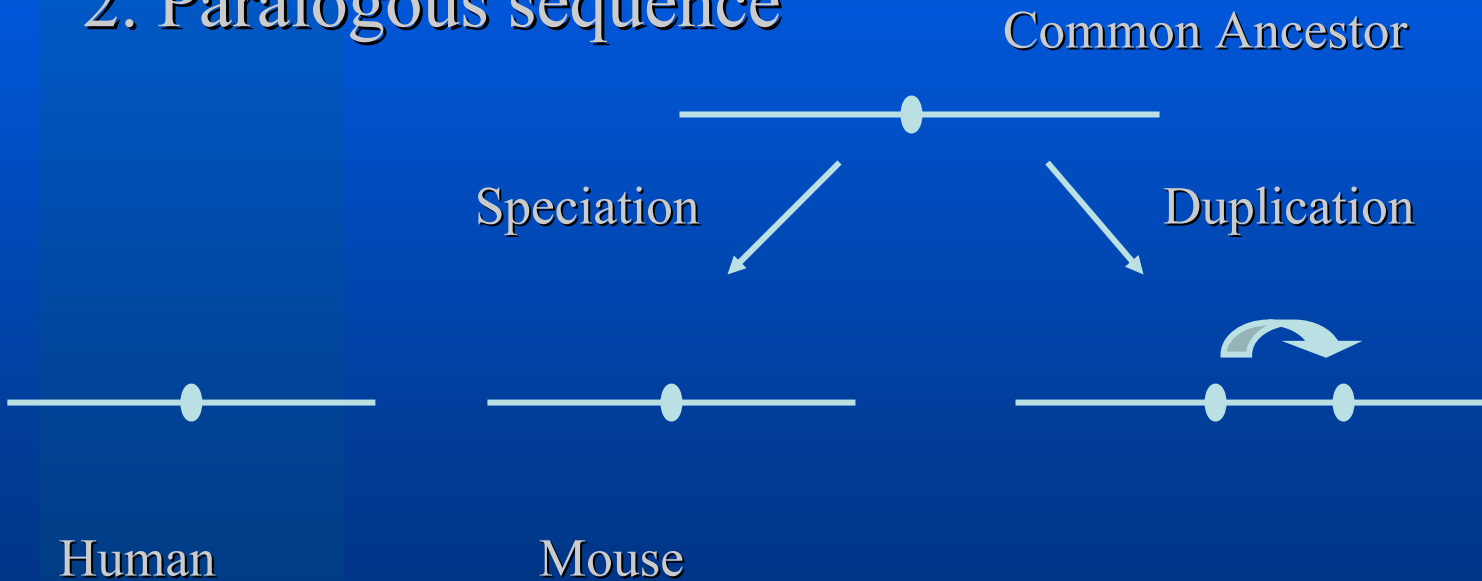
Phylogenetic Analysis

1. Selection of homologous sequences

- This step is prerequisite for inferring molecular evolution
- Homologous sequence
 - The sequences are descended from a common ancestor
 - Determination of homologous sequences by similarity

Phylogenetic Analysis

- Homologous sequence
 1. Orthologous sequence
 2. Paralogous sequence



Phylogenetic Analysis

2. Multiple sequence alignment

- Compare homologous sites in an evolutionary sense
- Homologous residue are aligned by sequence alignments
- Progressive sequence algorithm (Feng and Doolittle 1987)
- CLUSTALW (Thompson *et al.* 1994)

Phylogenetic Analysis

3. Tree building

- Distance method
- Maximum parsimony
- Maximum likelihood

Phylogenetic Analysis

4. Measure of divergence

- We can not simply count the number of mismatch (p-distance) between to estimate the divergence between sequences

ACAGTGCAGTG - AG	10 aligned sites
AT - GTGGA- TGAA -	8 identical sites
	2 mismatch sites

$$p \text{ distance} = 2 / 10 = 0.2$$

Phylogenetic Analysis

4. Measure of divergence

C	C→A	Single substitution
A→C→T→G	A	Sequential substitution
C→G	C→A	Coincidental substitution
T→A	T→A	Parallel substitution
A→C→T	A→T	Convergent substitution
C	C→T→C	Back substitution
Sequence 1	Sequence 2	

From Li 1997

- Model of DNA evolution

ex) Juke-Cantor, Kimura 2-parameters, and so on

Phylogenetic Analysis

5. Tree evaluation

- Two types of errors
 - : Topological and branch length
- Criteria of evaluation
 - : Consistency and robustness
- Bootstrap (Felsenstein 1985) is the most common procedure

Overview of our approach

Getting Interspersed Repeats



Identification of Orthologs



Selection Outgroup



Alignment sequences



Calculating Branch length

Overview of our approach

Getting Interspersed Repeats



Identification of Orthologs



Selection Outgroup



Alignment sequences

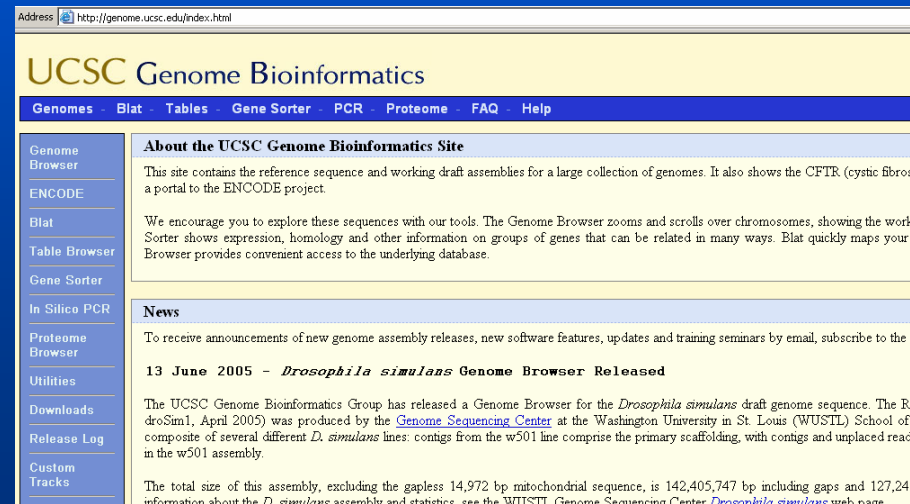


Calculating Branch length

1. Selection of homologous sequences
2. Multiple sequence alignment
3. Tree building
4. Measure of divergence
5. Tree evaluation

METHODS

- Data collection from UCSC genome browser
 1. Whole human (build35) and mouse (build33) genome
 2. ChromOut (hg17 and mm5)
 3. All Repeat elements from 135 ancestral repeat families
 4. Discard Repeat Elements (REs) less than 50 bp



METHODS

Getting IRs



Identification of Orthologs



Selection Outgroup



Alignment sequences



Calculating Branch length

Reciprocal Best BLAST Hits

Query Set

1_MIR3_H
2_MIR3_H
3_MIR3_H
4_MIR3_H
5_MIR3_H
⋮
N_MIR3_H

Database Set

1_MIR3_H	1_MIR3_M
2_MIR3_H	2_MIR3_M
3_MIR3_H	3_MIR3_M
4_MIR3_H	4_MIR3_M
5_MIR3_H	5_MIR3_M
⋮	⋮
N_MIR3_H	N ₁ _MIR3_M



724_MIR3_H
62999_MIR3_H
1121_MIR3_M
16750_MIR3_H
9295_MIR3_M
65370_MIR3_H
29225_MIR3_H
24798_MIR3_H

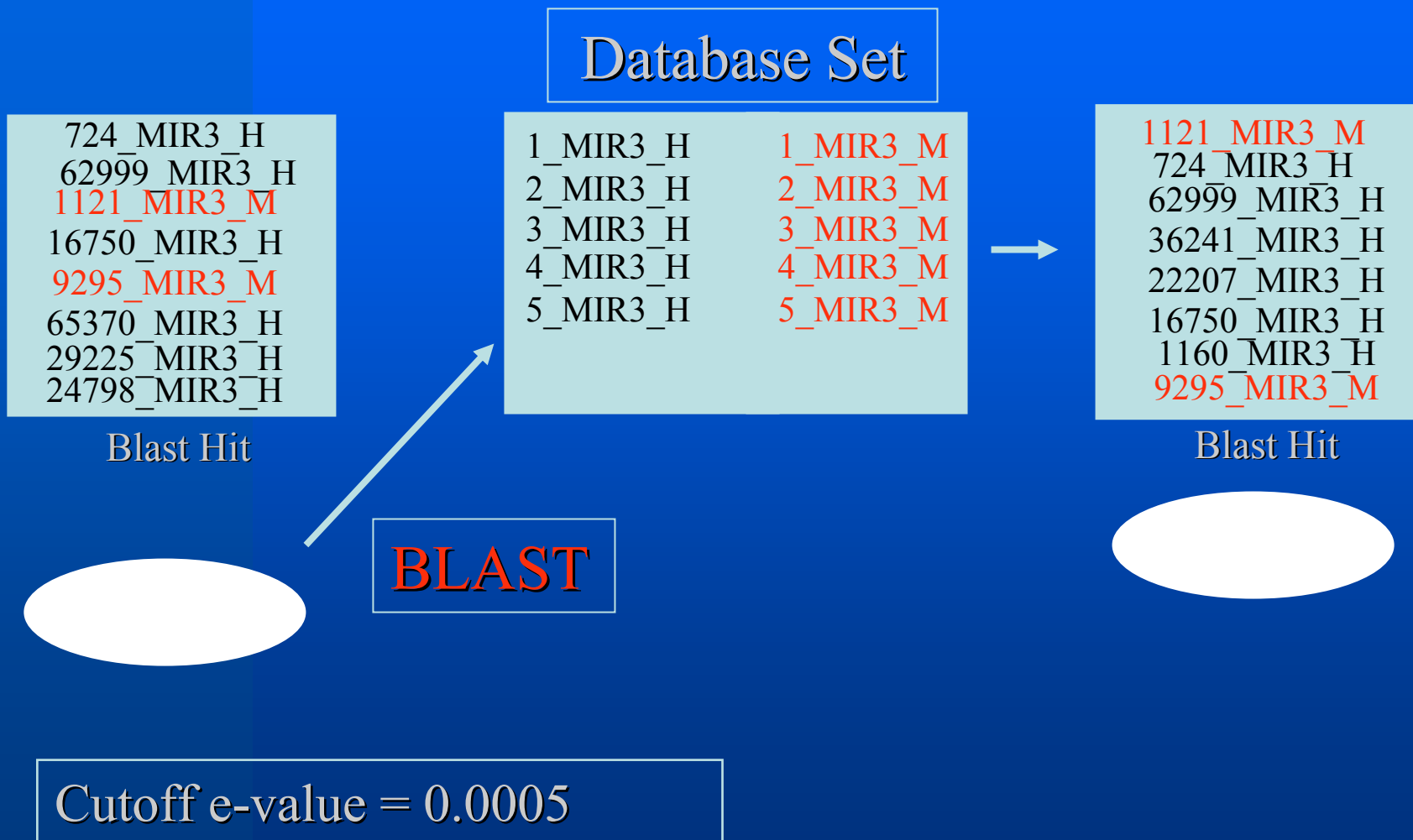
Blast Hit

BLAST

724_MIR3_H

Cutoff e-value = 0.0005

Reciprocal Best BLAST Hits



METHODS

724_MIR3_H
1121_MIR3_M

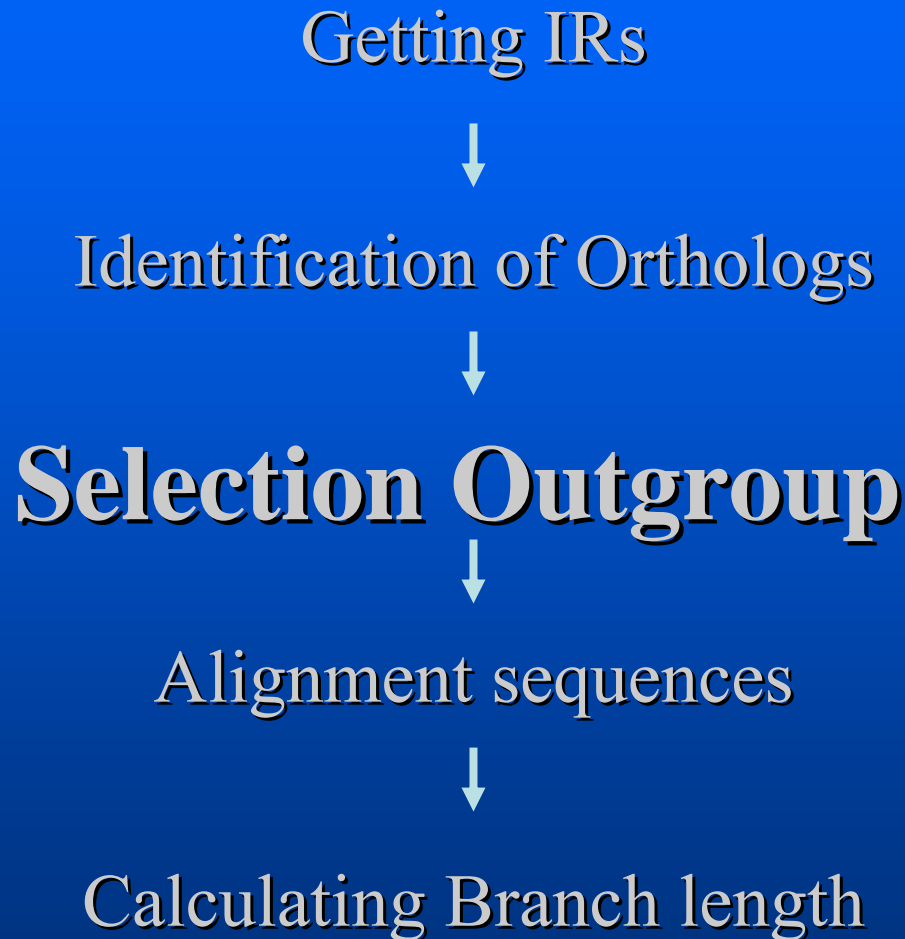
First iteration

1121_MIR3_M
724_MIR3_H

Second iteration

Orthologous repeats

METHODS



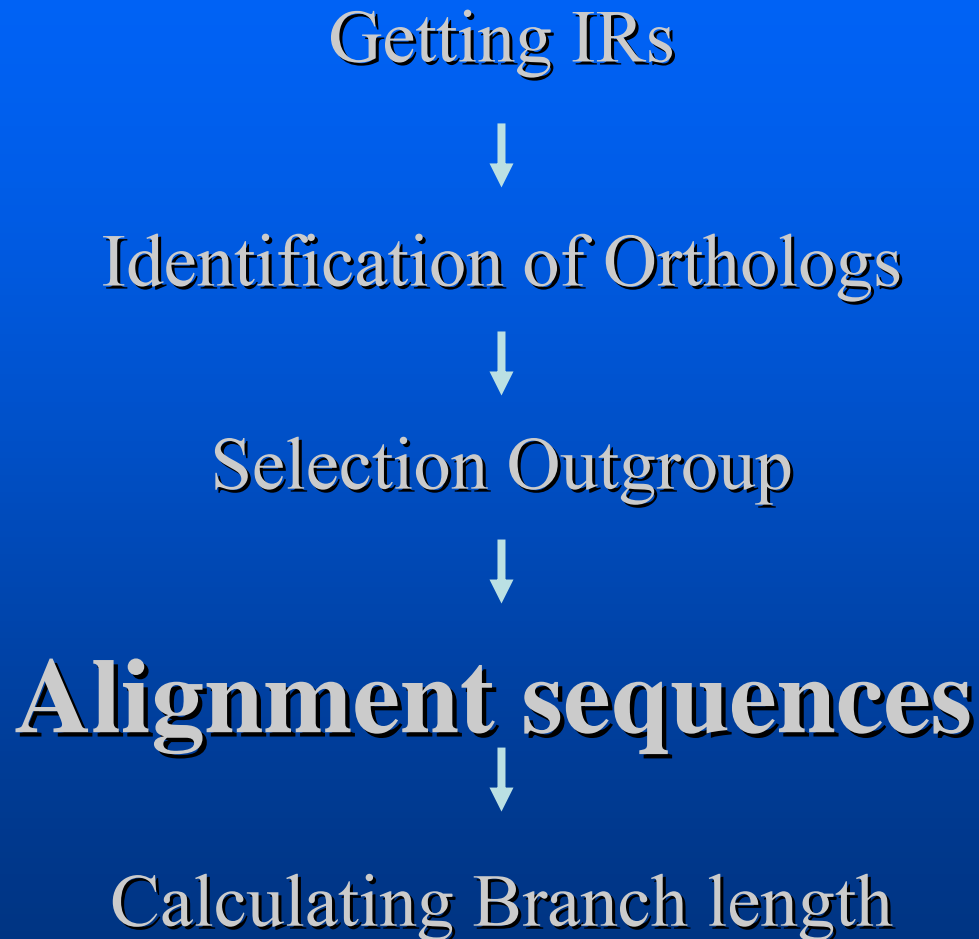
METHODS

724_MIR3_H
62999_MIR3_H
1121_MIR3_M
16750_MIR3_H
9295_MIR3_M
65370_MIR3_H
29225_MIR3_H
24798_MIR3_H

1121_MIR3_M
724_MIR3_H
62999_MIR3_H
36241_MIR3_H
22207_MIR3_H
16750_MIR3_H
1160_MIR3_H
9295_MIR3_M

Consistent outgroup if these two sequences identical
Inconsistent outgroup if not identical

METHODS



Sequence alignment

- **Align three sequences by “Profile alignment”**

Make sure that orthologous pair aligned first

Done by CLUSTALW

METHODS

Getting IRs



Identification of Orthologs



Selection Outgroup



Alignment sequences



Calculating Branch length

Branch length

- **Calculating branch length by Maximum Likelihood**
 - **General time reversible (GTR; Lanave 1984), HKY (1985), and GTR + gamma**
 - **implanted in PAUP***
 - **average over all branch length in a repeat family**

CONTENTS

1. Introduction

2. Methods

3. Results

4. Discussion

5. Reference

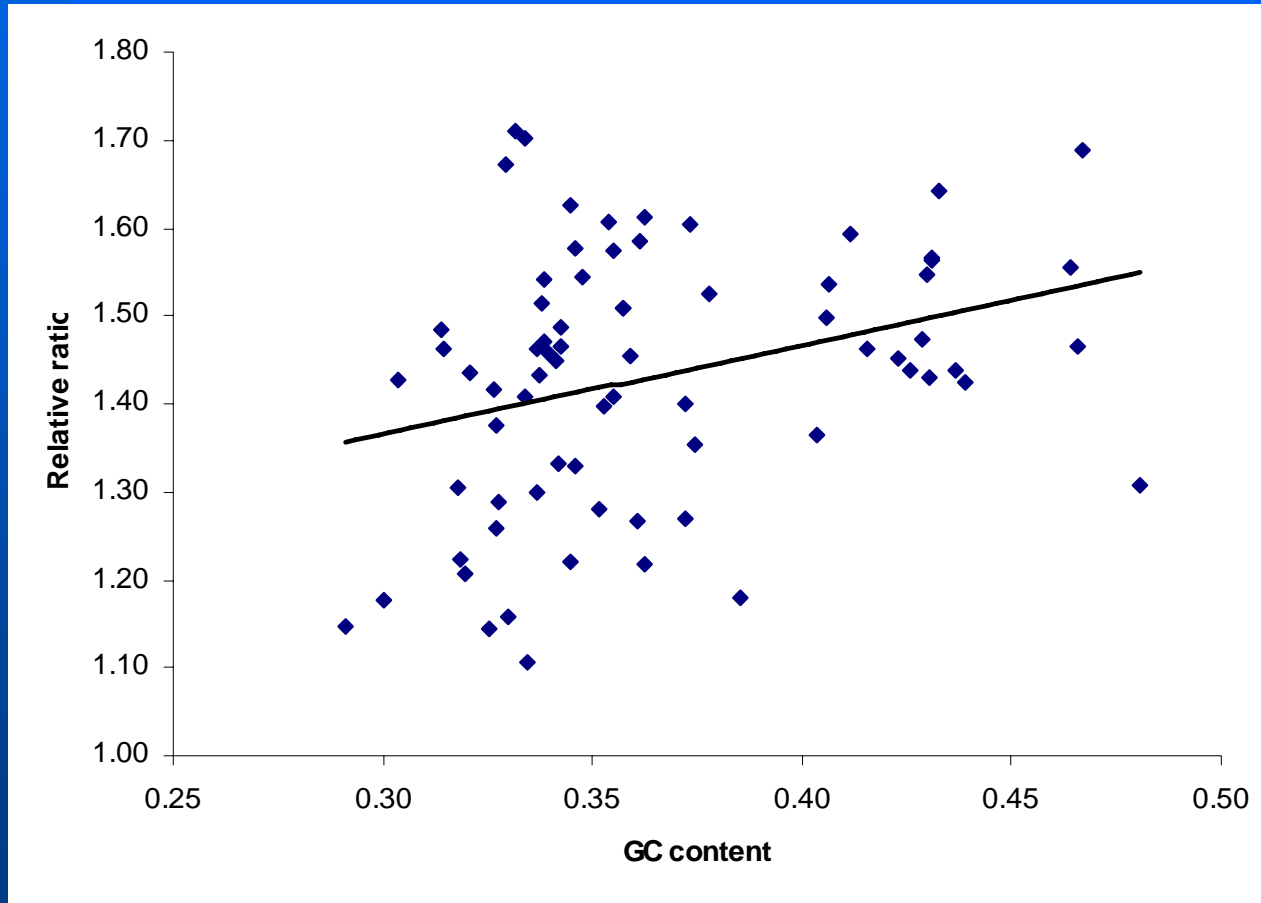
30 largest repeat families

Interspersed Repeat		No of repeats		NO of orthologs		Divergence1		Relative Rate*
Family	Class	Human	Mouse	Total	Consistent Outgroup	Human	Mouse	
L2	LINE	385405	54736	17393	492(461,31)	0.293	0.271	0.925
MIR	SINE	194742	44069	15323	394(345,49)	0.279	0.306	1.100
MIR3	SINE	69054	9531	4054	167(146,21)	0.266	0.255	0.961
L1ME	LINE	32950	9214	2851	82(67,15)	0.326	0.301	0.923
L3	LINE	44837	9549	1300	71(62,9)	0.178	0.173	0.972
L1M4	LINE	47531	19998	5081	66(58,8)	0.276	0.311	1.125
MER5B	DNA	22704	4785	1881	60(57,3)	0.281	0.335	1.191
MER5A	DNA	33903	9163	2710	48(43,5)	0.303	0.301	0.991
SUM		1300457	266282	76723	1756(1590,166)	0.303	0.308	1.018

72 young repeat families

Repeat family	No. of Orthologs	Divergence		JC		Ratio (mouse/human)
		Human	Mouse	Human	Mouse	
Charlie9	71	0.22	0.23	0.25	0.28	1.11
L1MA6	485	0.16 (0.16)	0.25 (0.28)	0.18	0.31	1.72
L1MA7	415	0.16 (0.16)	0.25 (0.28)	0.19	0.31	1.63
L1MA8	427	0.16 (0.15)	0.25 (0.27)	0.18	0.31	1.72
L1MA9	728	0.18 (0.18)	0.26 (0.28)	0.21	0.32	1.52
L1MA10	335	0.19 (0.19)	0.27 (0.29)	0.22	0.33	1.50
L1MB1	461	0.18 (0.18)	0.27 (0.29)	0.21	0.33	1.57
L1MB2	444	0.18 (0.18)	0.26 (0.28)	0.20	0.32	1.60
L1MC1	454	0.18 (0.17)	0.27 (0.28)	0.21	0.33	1.57
MLT1A	783	0.22 (0.21)	0.29 (0.31)	0.26	0.37	1.42
MLT1A0	1086	0.20 (0.19)	0.28 (0.3)	0.23	0.36	1.57
MLT1B	1275	0.19 (0.18)	0.28 (0.28)	0.22	0.34	1.55
MLT1C	1424	0.21 (0.21)	0.29 (0.30)	0.25	0.36	1.44
MER20	1372	0.20 (0.19)	0.27 (0.29)	0.23	0.33	1.43

RESULTS



$$R^2 = 0.1$$

CONTENTS

1. Introduction

2. Methods

3. Results

4. Discussion →

5. Reference

- Appropriate Outgroup
- Advantage of our approach
- Significance of our study

DISCUSSION

- **Appropriate Outgroup**

Consistent outgroup

Inconsistent outgroup

Genomics Average

Consistent

Inconsistent

Human outgroup

1.020

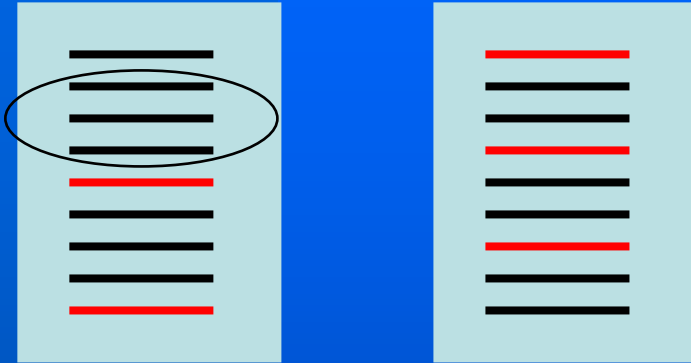
1.252

Mouse outgroup

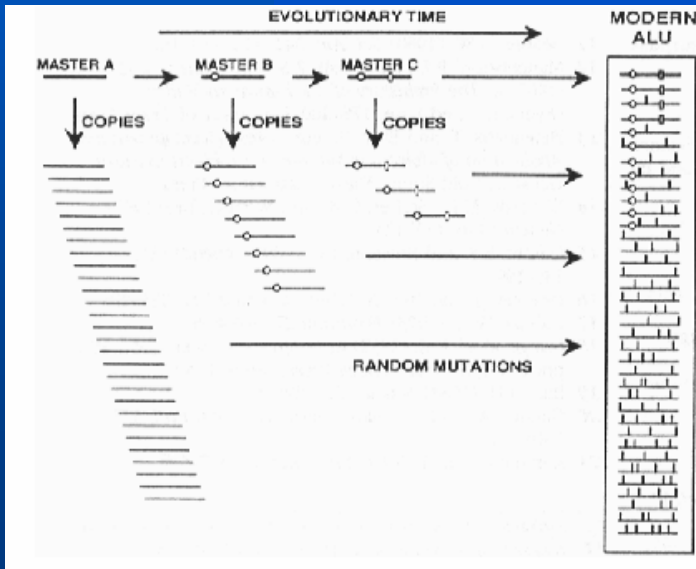
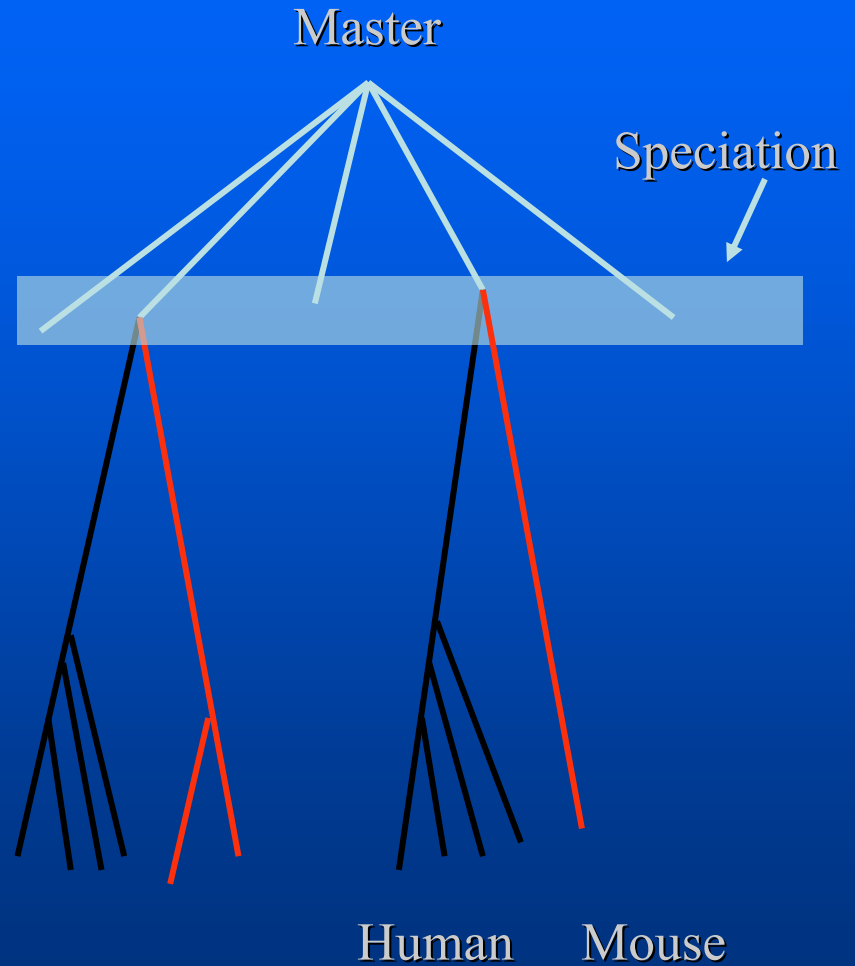
0.998

0.797

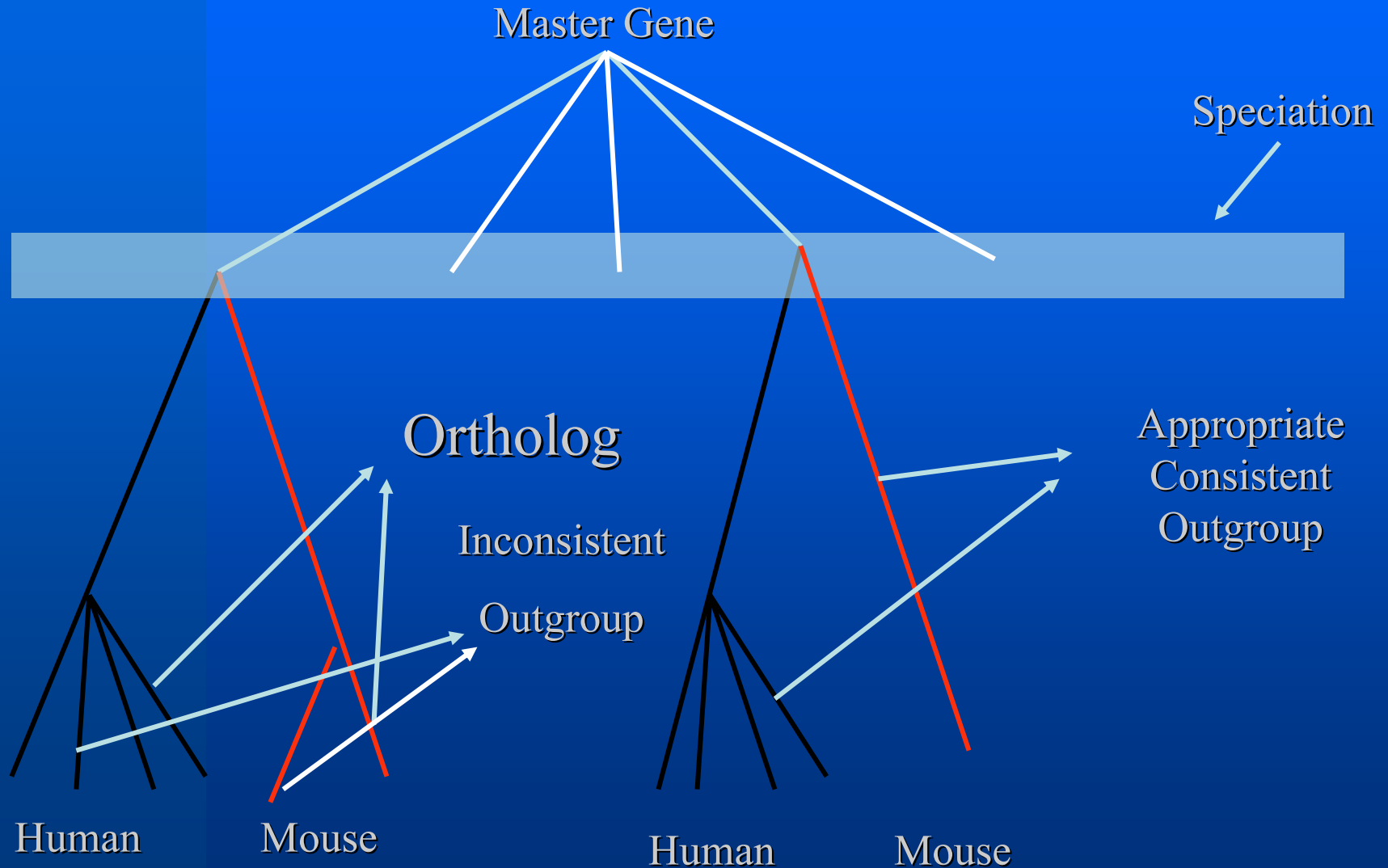
DISCUSSION



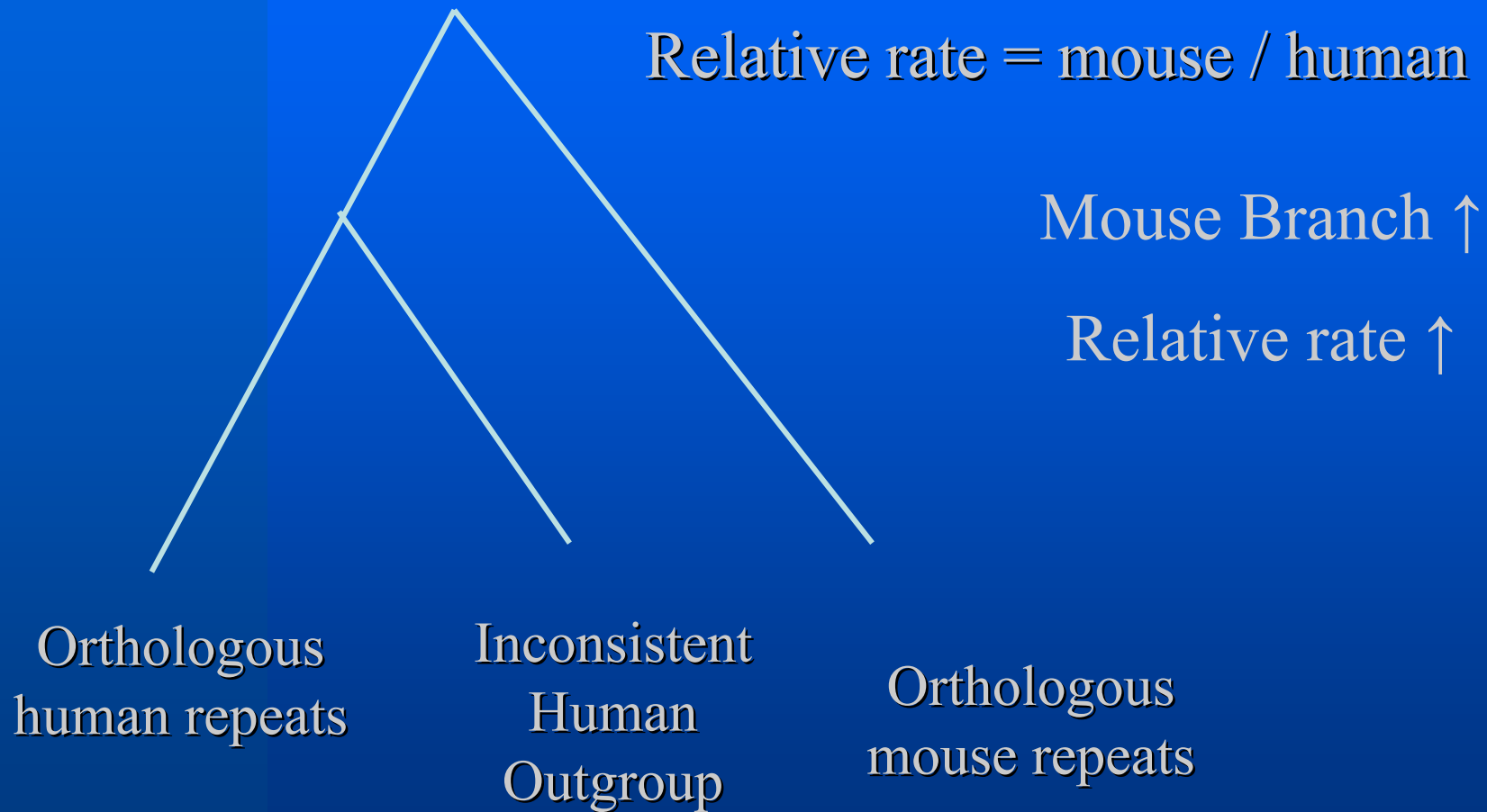
Blast hit



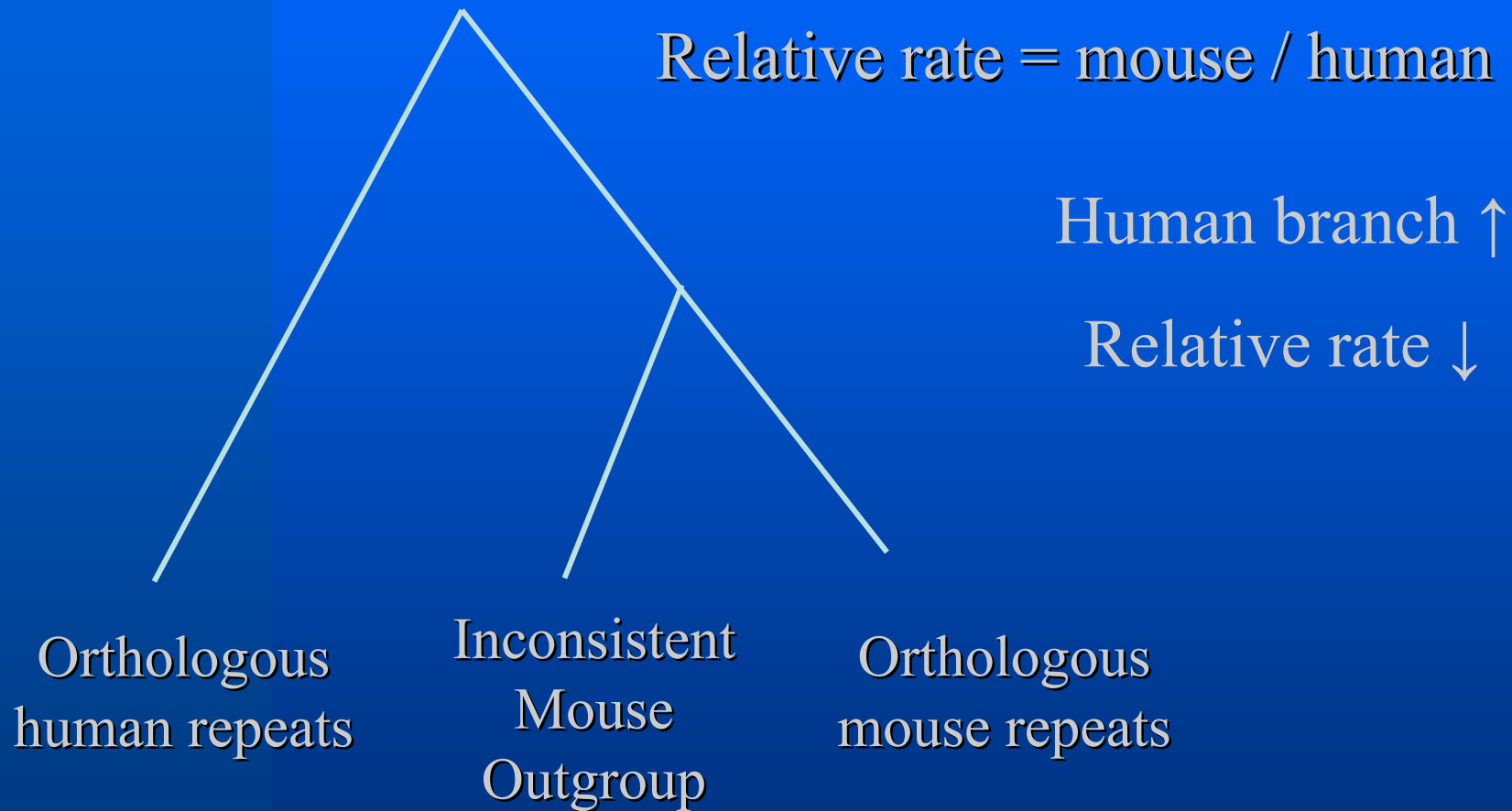
DISCUSSION



Inconsistent human outgroup



Inconsistent mouse outgroup



DISCUSSION

Advantage of our approach

- Large amount of data & week selection on four-fold degenerate sites.
- The assumption of Waterston et al., orthology-by-synteny may be negatively impacted by the fast rate of rearrangement (Bourque et al. 2004; Kumar et al. 2001; Waterston et al. 2002).
- GTR is a more sophisticated model to estimate the number of substitution

DISCUSSION

- Significance of our study

The knowledge of difference in mutation
between human and mouse



Comparison between human and mouse at
DNA level



The study of human genomes can be
enhanced by experiments on mouse

CONCLUSIONS

Our approach may provide the opportunity to get more reliable estimation of evolutionary rate in human and mouse

Our results indicate that genome-wide mutation rates differ slightly between human and mouse, and this difference is less than previously reported.

CONCLUSIONS

- **Future plan**

- Apply same methods to four-fold degenerate sites and compare results
- In relative rate test, use Marsupials as a reference
- Check the variation across the human and mouse genomes

ACKNOWLEDGEMENTS

I would like to thank Dr. Sudhir Kumar for his guidance and extensive support of this project and also Dr. Jeffrey Touchman and Dr. Andrzej Czygrinow for useful comments. I also appreciate to other members of the lab for their help, specially, Dr. Sankar Subramanian.

Thanks to EFG for internship funds and to NIH for research grants (to Sudhir Kumar).

THANK YOU !!

Comment??