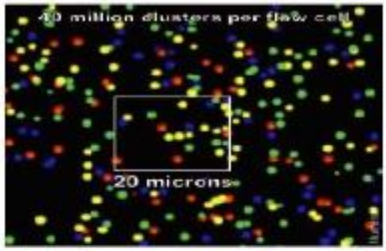
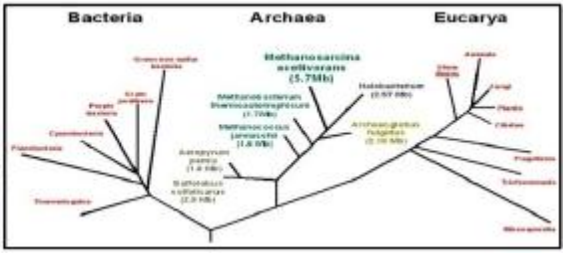
Bioinformatics: Introduction and Methods

Le Zhang

Computer Science Department, Southwest University

```
TAACCCTAACCCTAACCCTAACCCTAACCCTA
CCTAACCCTAACCCTAACCCTAACCCTAACCC
CCCTAACCCCTAACCCTAACCCTAACCCTAAC
AACCCTAACCCTAACCCTAACCCCTAACCCTA
ACCCTAACCCCAACCCCAACCCCAACCCCAAC
CTACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCCTAACCCTAA
```



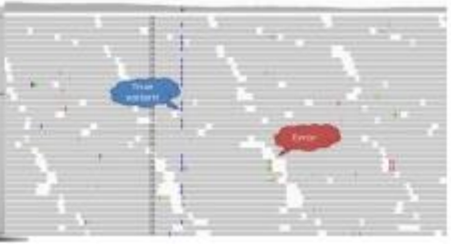


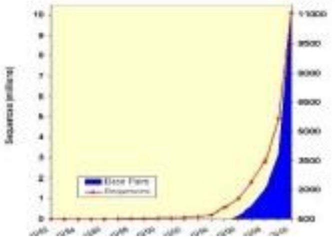
Case Study 2: From Dry to Wet, an Evolutionary Story

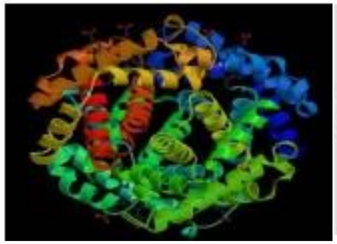
Le Zhang, Ph. D.

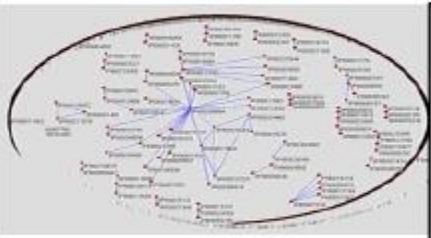
Computer Science Department

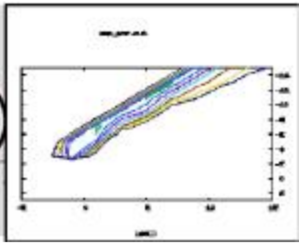
Southwest University



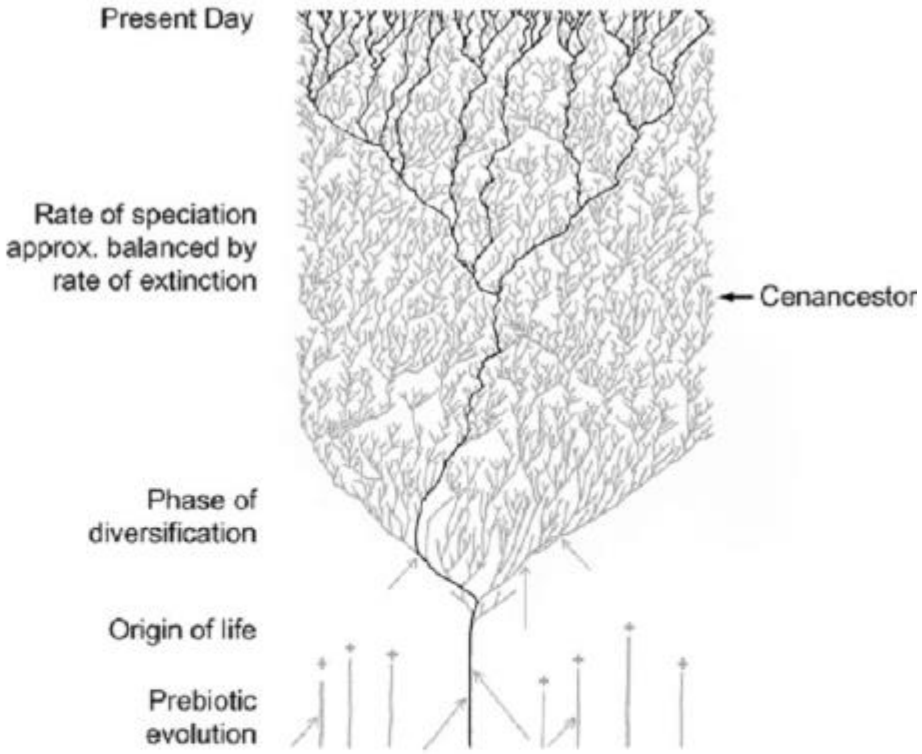




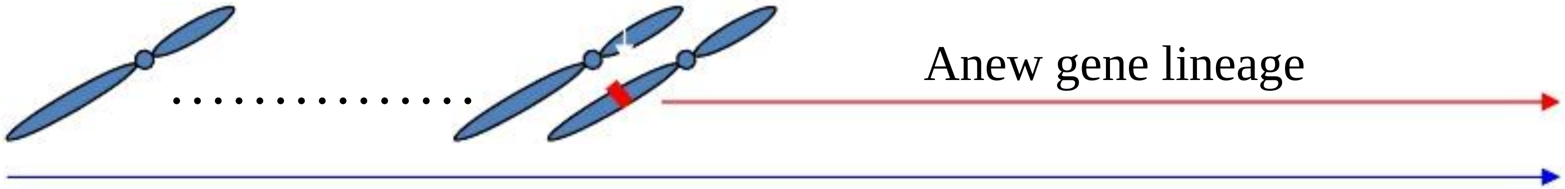




Newly emerging genes are key sources of evolutionary novelty



(Source: *Trends in Genetics* 20(4):182)

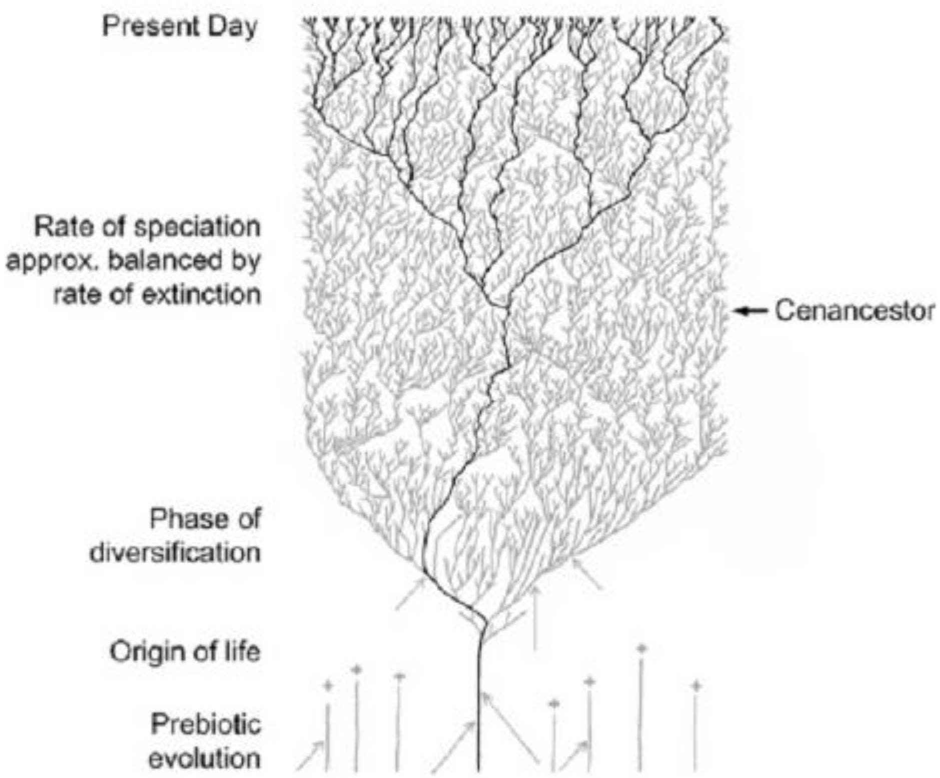


Anew gene was born

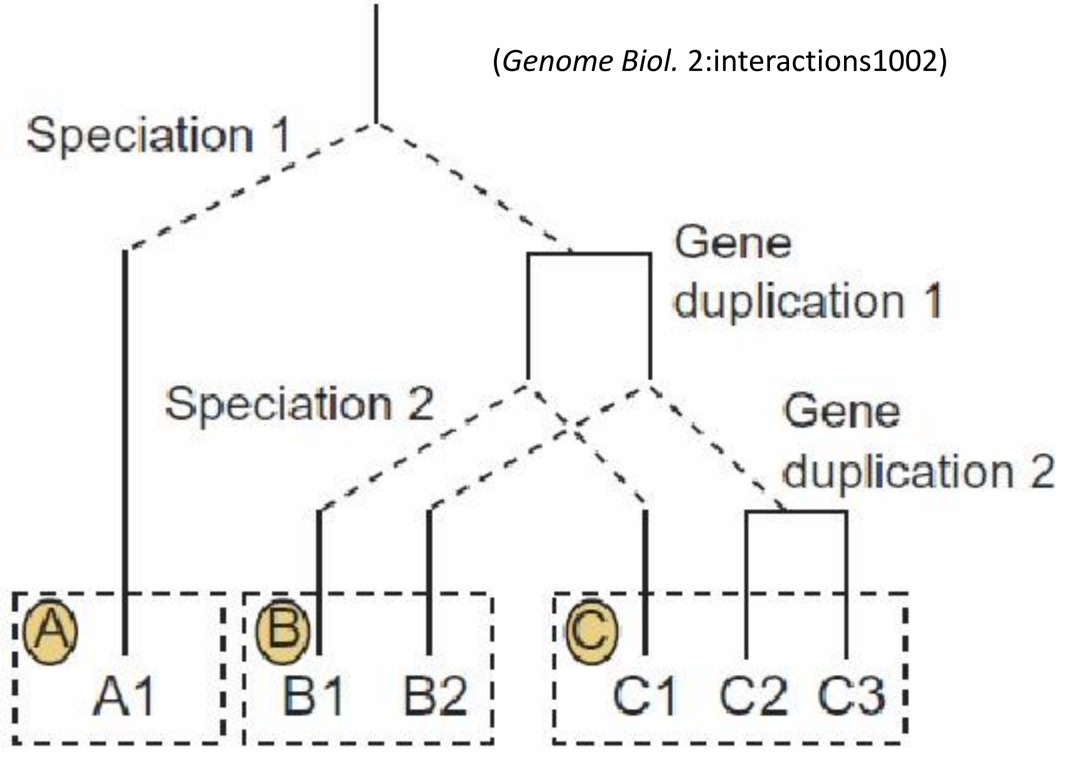
Anew gene lineage

Present

Million YearsAgo



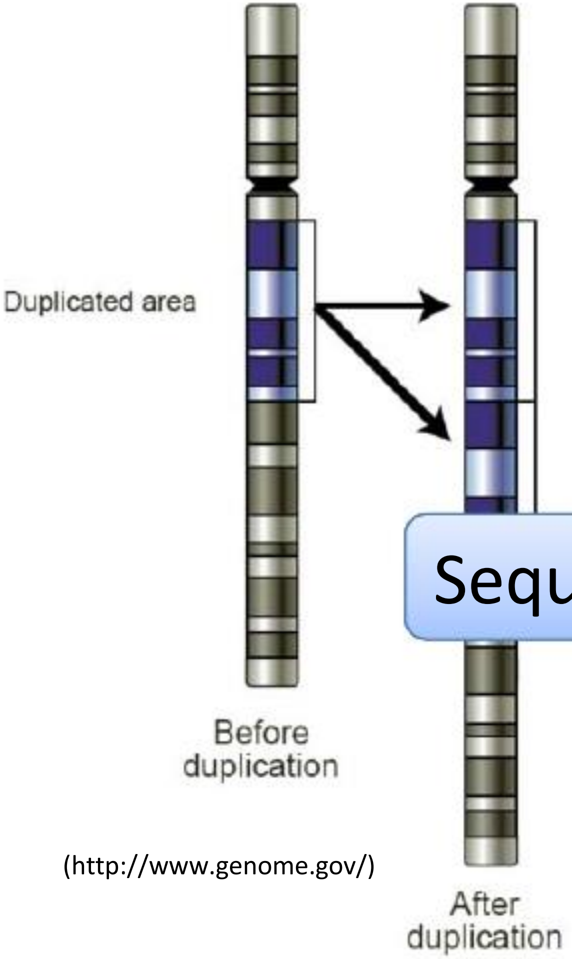
(Source: *Trends in Genetics* 20(4):182)



(*Genome Biol.* 2:interactions1002)

- Orthologs: homologuous genes result from **speciation event**
- Paralogs: homologuous genes result from **duplication event**

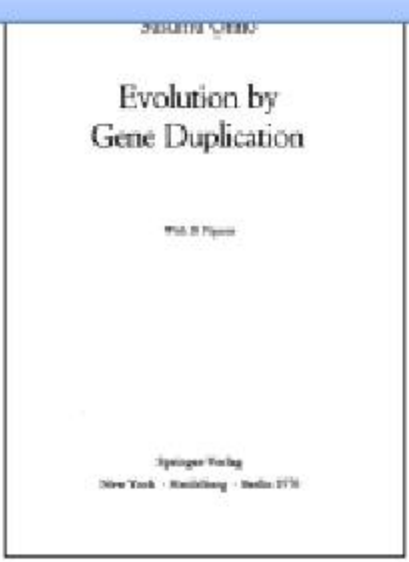
Sequence Divergence ~ Functional Novelty



(http://www.genome.gov/)

Susumu Ohno

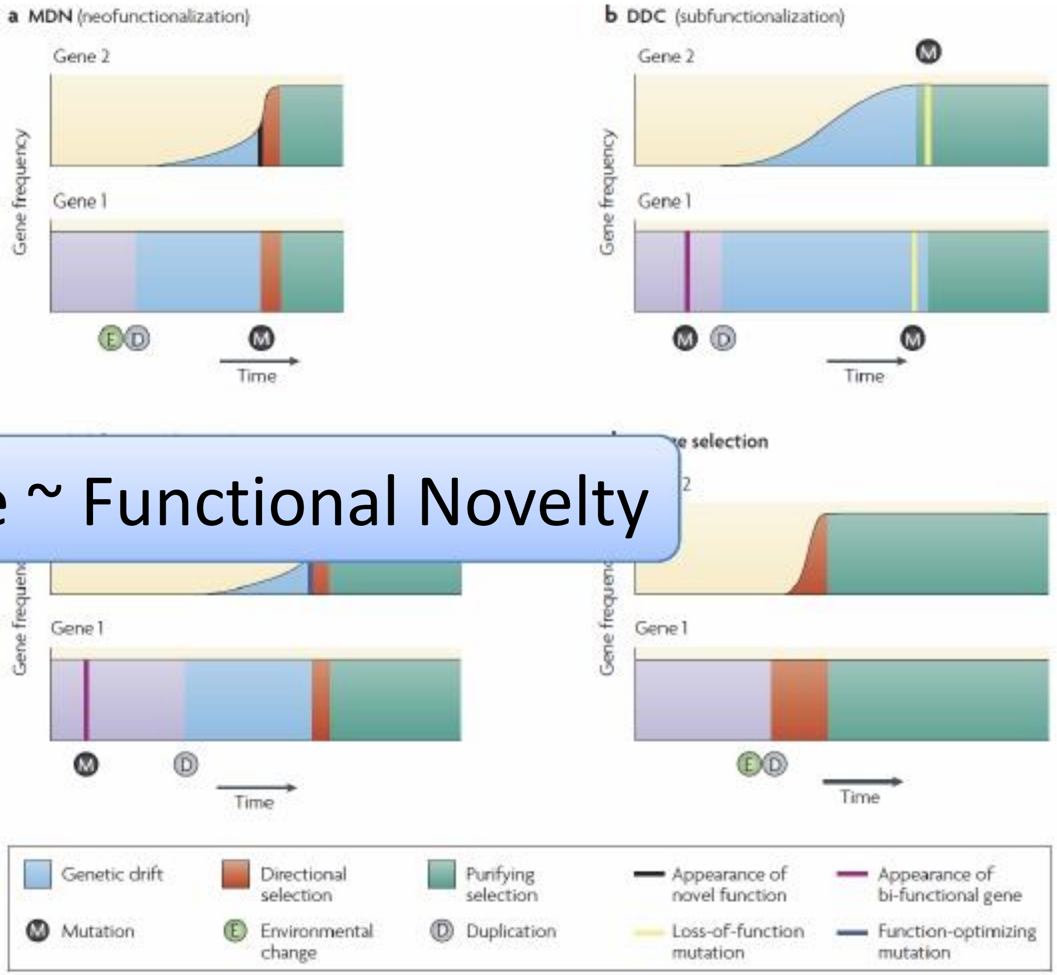
Susumu Ohno
February 1, 1928–January 13, 2000


Susumu Ohno

Evolution by Gene Duplication

With 8 Figures

Springer-Verlag
New York · Heidelberg · Berlin 1970


a MDN (neofunctionalization)
b DDC (subfunctionalization)
e selection
Gene 2
Gene 1
Gene frequency
Time
M
E
D
Genetic drift
Directional selection
Purifying selection
Appearance of novel function
Appearance of bi-functional gene
Mutation
Environmental change
Duplication
Loss-of-function mutation
Function-optimizing mutation

(Conant *et al*. 2008)

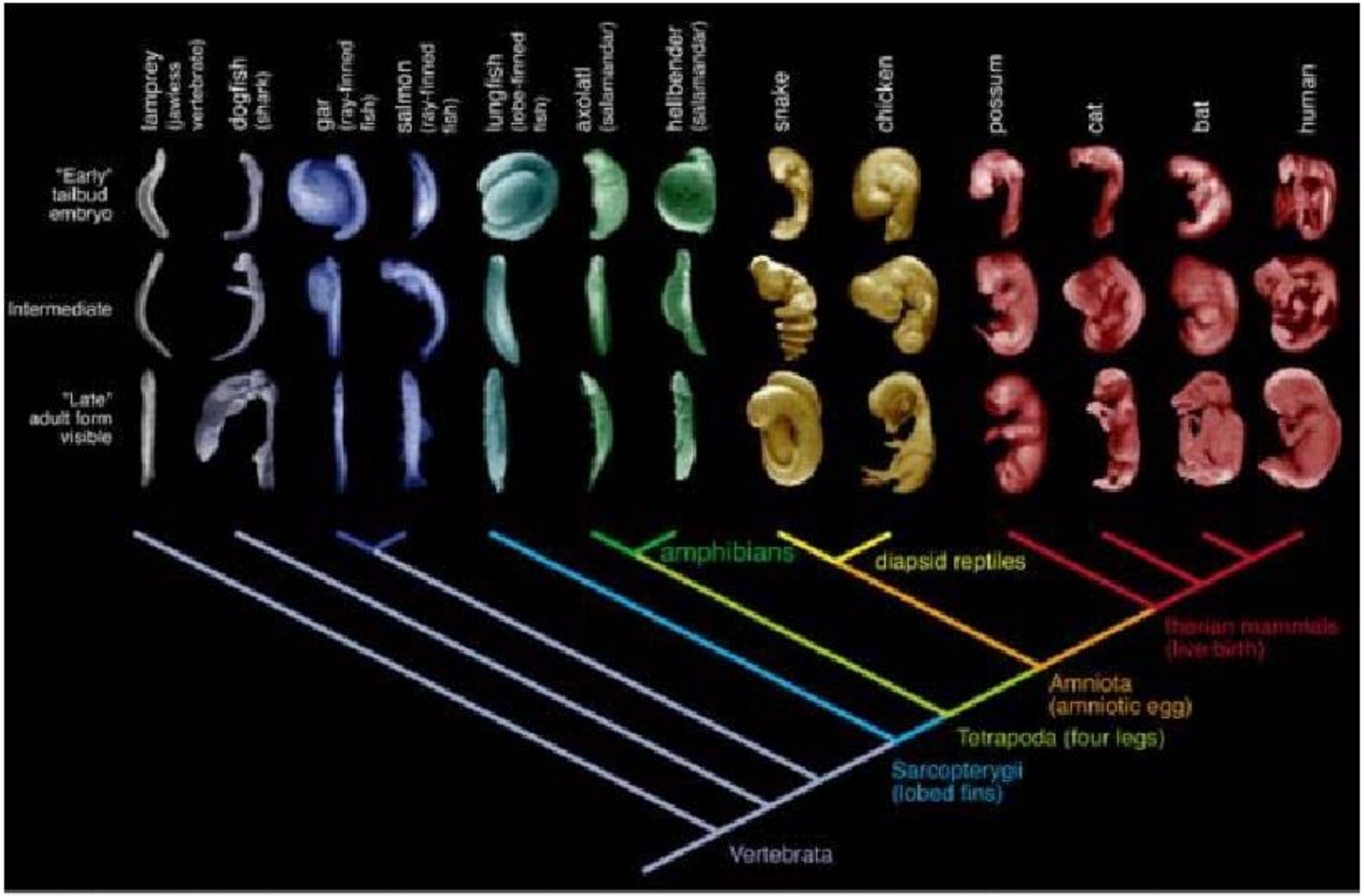
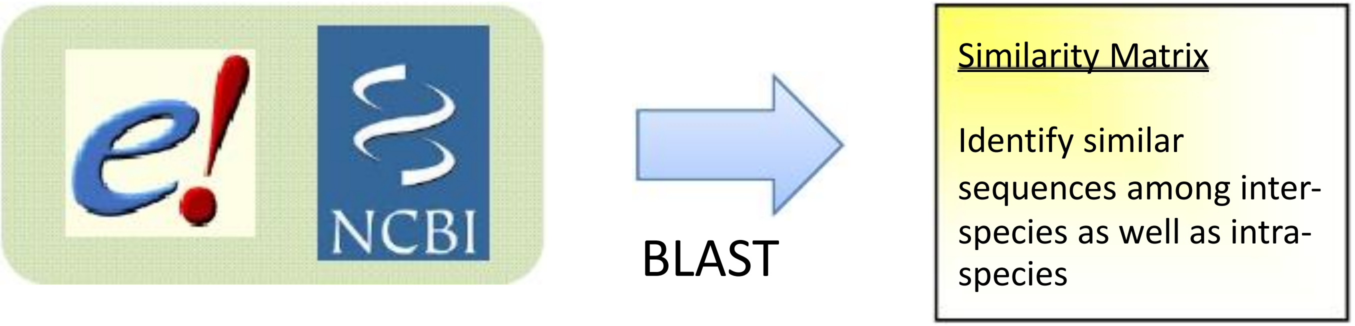


Figure 8. Developmental sequences of various vertebrates shown in phylogenetic context. Note the shared similarities of some closely related taxa, particularly the amniotes (modified from Richardson et al. 1998.) (Figure Source: ncseprojects.org/image/icons-evolution-figure-8)

Computationally screening for function divergent genes involved in early development regulation



Sequences from 14 species

BLAST

Similarity Matrix

Identify similar sequences among inter-species as well as intra-species

Insert an in-video survey here.

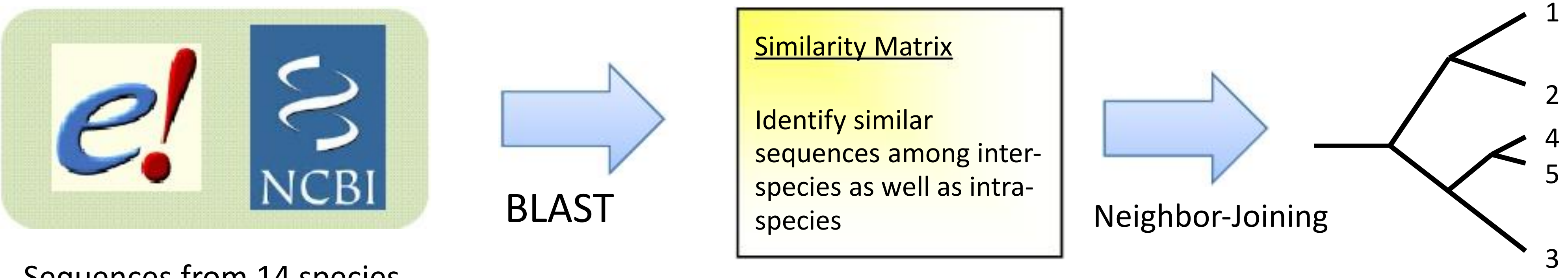
Question: Which Scoring Matrix would you like to use here?

A) PAM1
B) BLOSUM80
C) BLOSUM62
D) PAM2

C is the right answer, but B is also okay.

Either the answer is right or wrong, just go on.

Computationally screening for function divergent genes involved in early development regulation



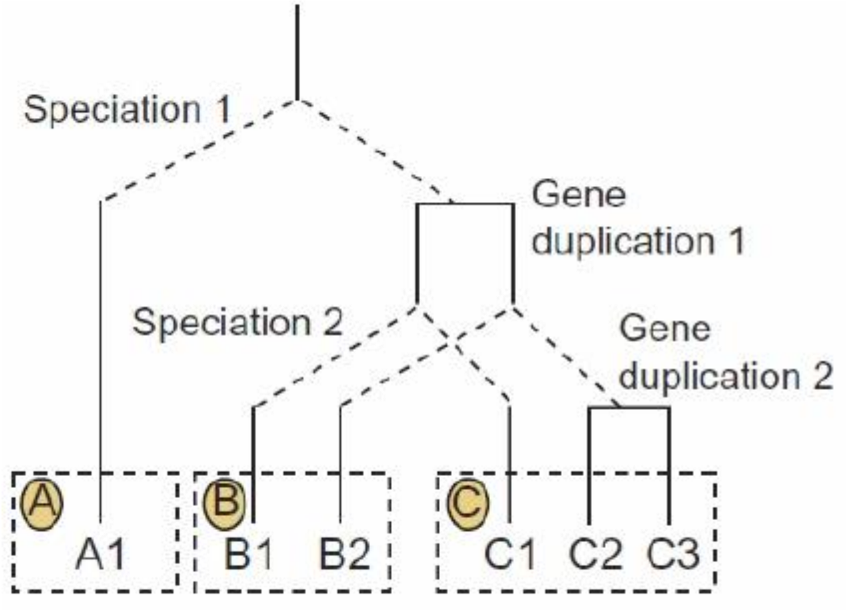
Sequences from 14 species

BLAST

Similarity Matrix

Identify similar sequences among inter-species as well as intra-species

Neighbor-Joining



(*Genome Biol.* 2:interactions1002)

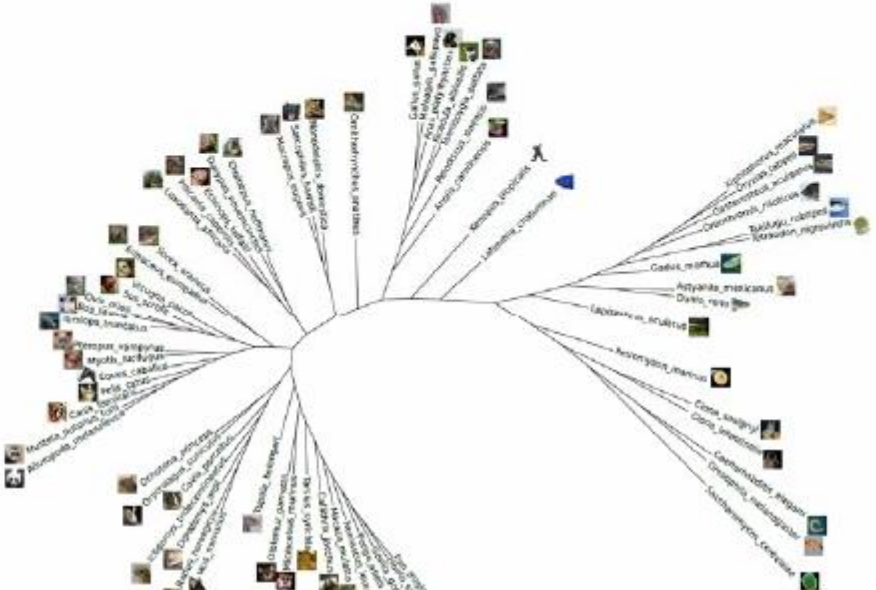
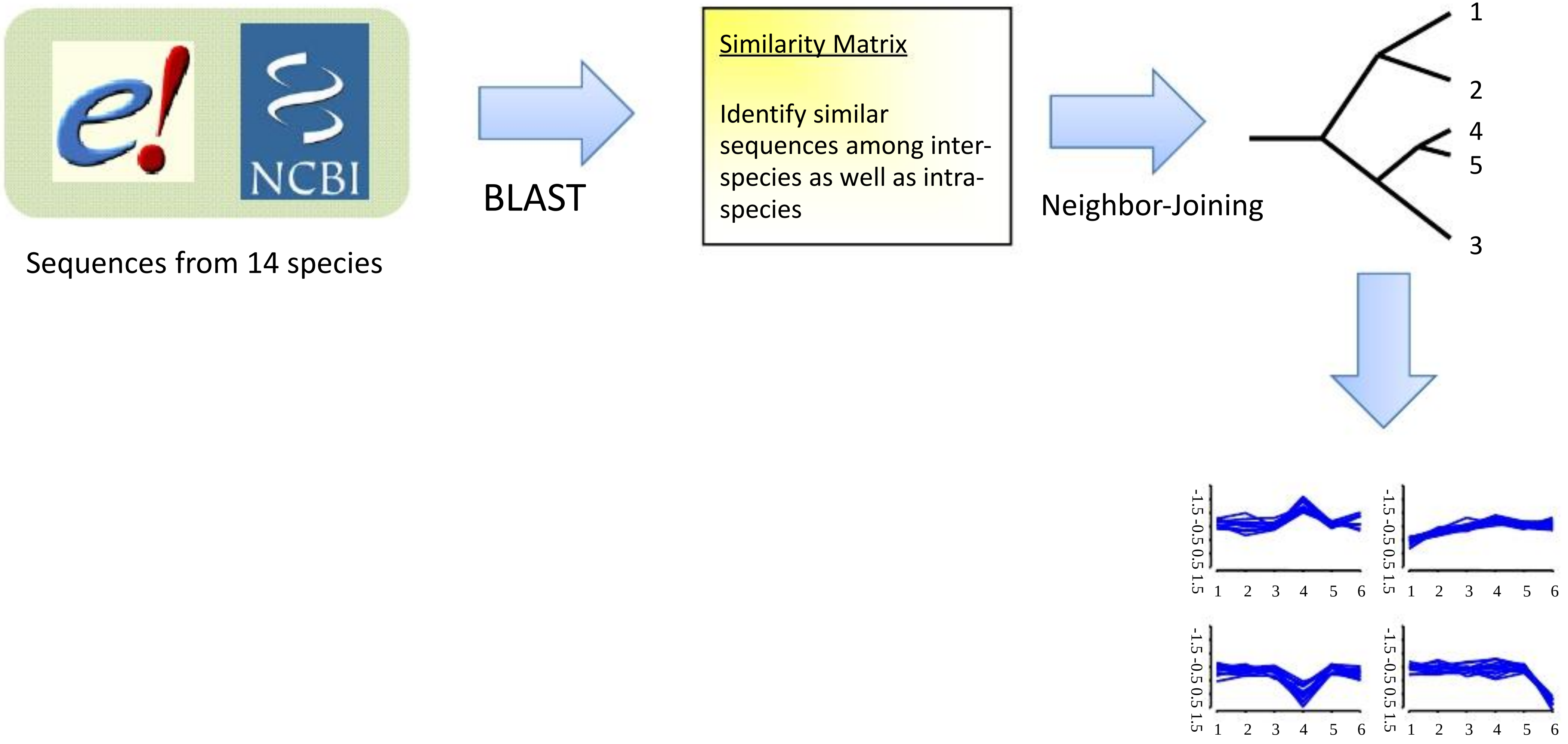


Image obtained using Dendroscope
Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks
D.H. Huson and C Scornavacca
Systematic Biology, 2012

Computationally screening for function divergent genes involved in early development regulation



Sequences from 14 species

BLAST

Similarity Matrix

Identify similar sequences among inter-species as well as intra-species

Neighbor-Joining

Expression Profiles

Insert an in-video survey here.

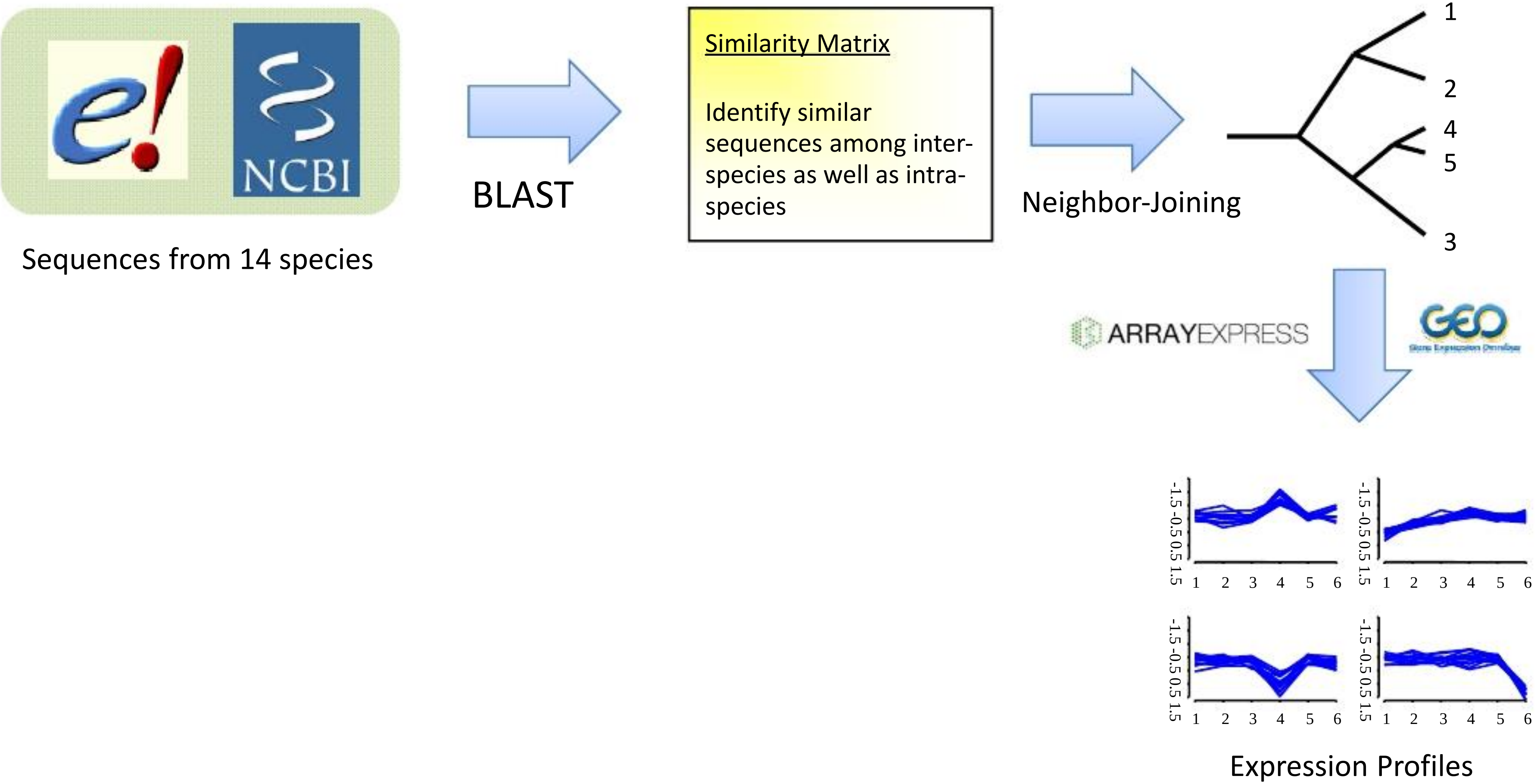
Question: Which database(s) you would NOT use here?

A) PDB
B) NCBI GEO
C) NCBI SRA
D) EBI ArrayExpress

A is the right answer.

Either the answer is right or wrong, just go on.

Computationally screening for function divergent genes involved in early development regulation



Sequences from 14 species

BLAST

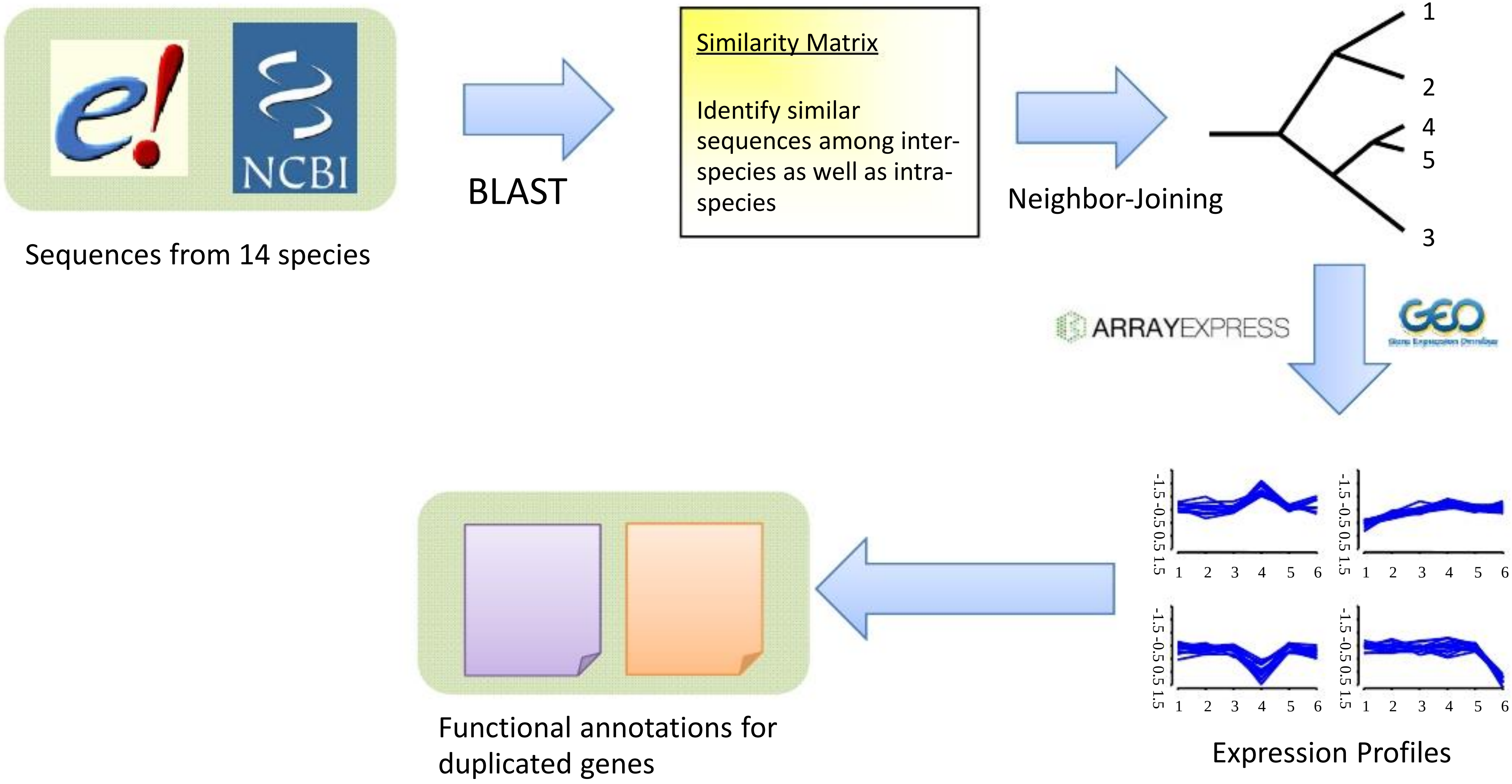
Similarity Matrix

Identify similar sequences among inter-species as well as intra-species

Neighbor-Joining

Expression Profiles

Computationally screening for function divergent genes involved in early development regulation



Sequences from 14 species

BLAST

Similarity Matrix

Identify similar sequences among inter-species as well as intra-species

Neighbor-Joining

Expression Profiles

Functional annotations for duplicated genes

Insert an in-video survey here.

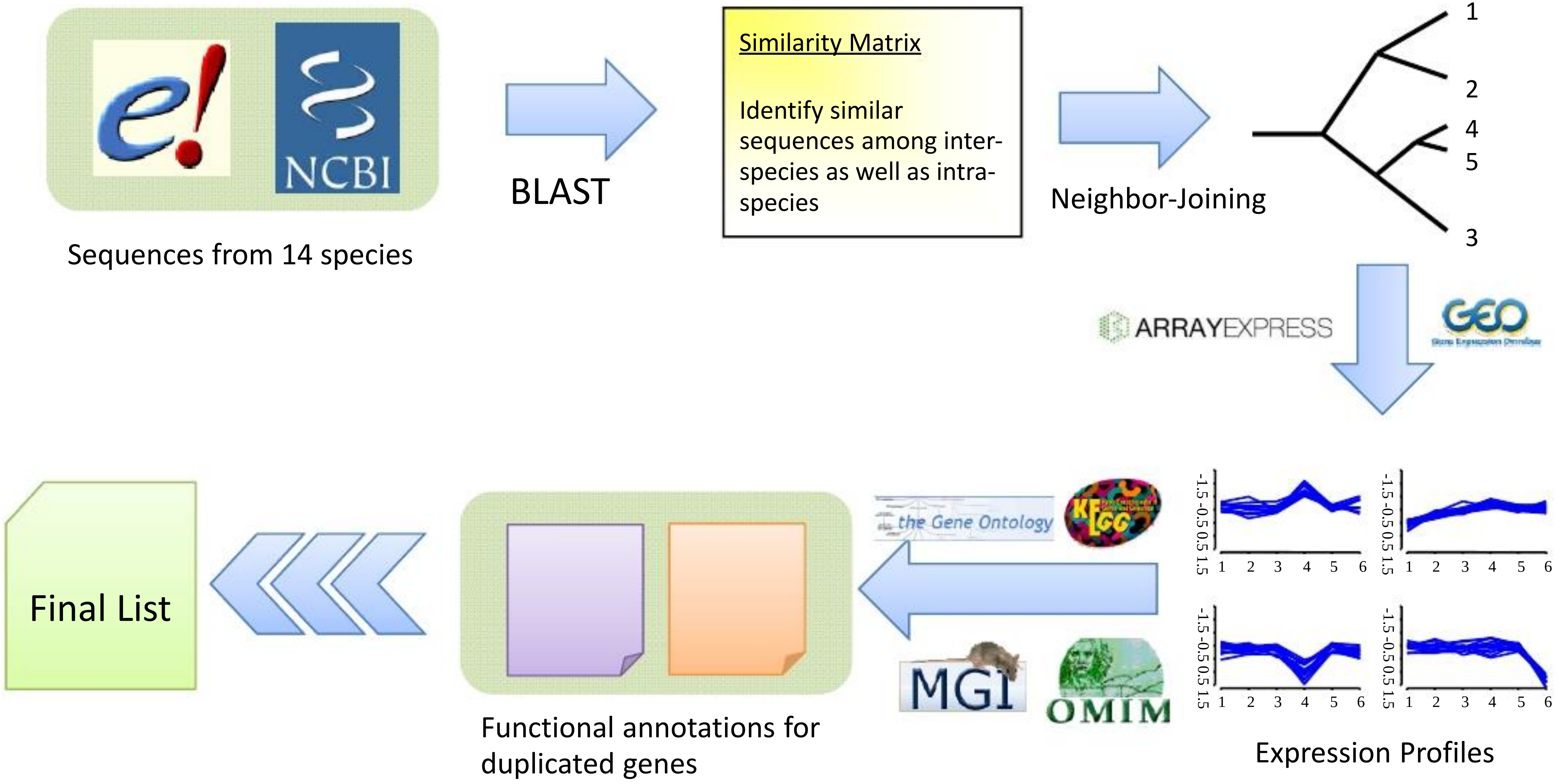
Question: Which database(s) would you like use here?

A) KEGG
B) Gene Ontology Annotation

Both A and B are correct

Either the answer is right or wrong, just go on.

Computationally screening for function divergent genes involved in early development regulation



Sequences from 14 species

BLAST

Similarity Matrix

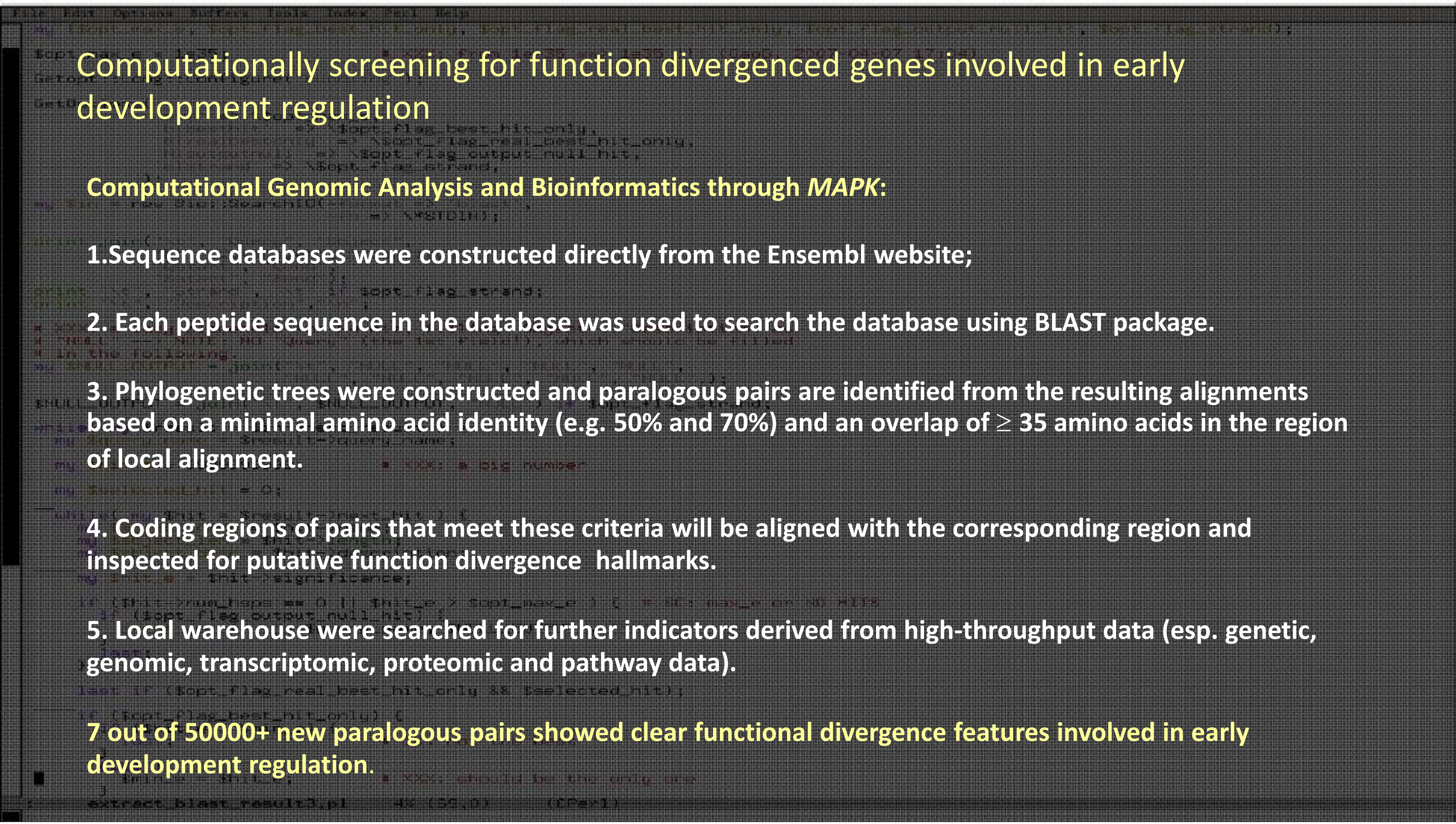
Identify similar sequences among inter-species as well as intra-species

Neighbor-Joining

Expression Profiles

Functional annotations for duplicated genes

Final List



Computationally screening for function divergenced genes involved in early development regulation

Computational Genomic Analysis and Bioinformatics through *MAPK*:

1.Sequence databases were constructed directly from the Ensembl website;

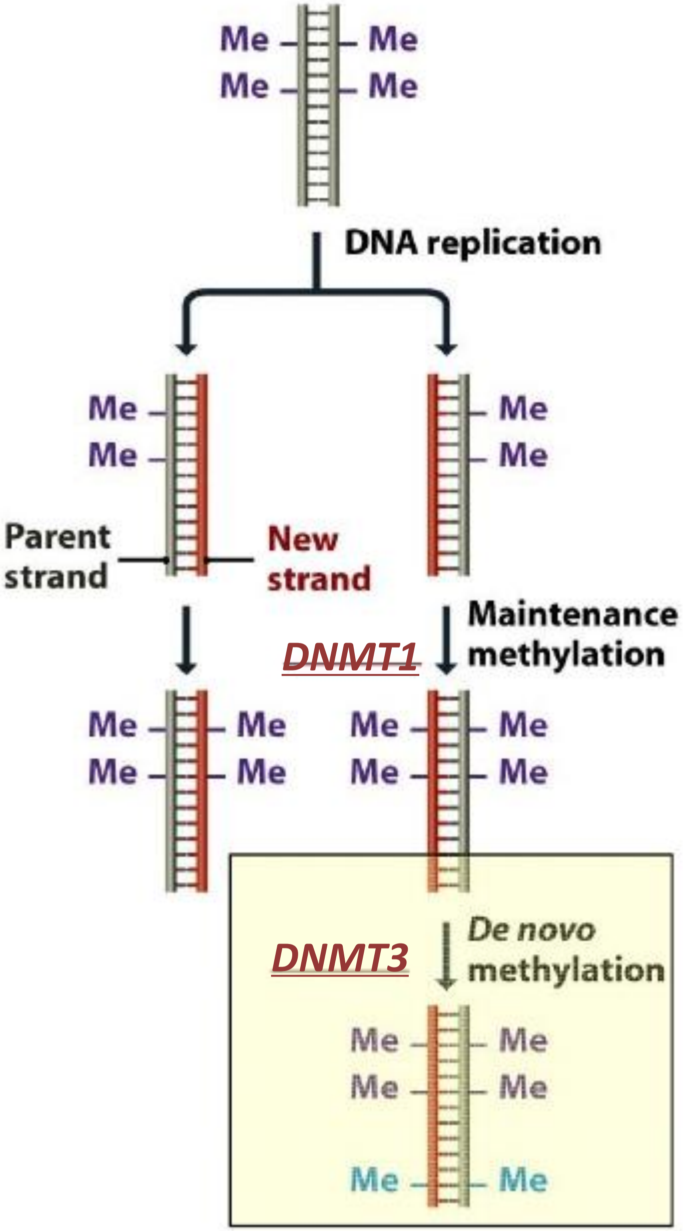
2. Each peptide sequence in the database was used to search the database using BLAST package.

3. Phylogenetic trees were constructed and paralogous pairs are identified from the resulting alignments based on a minimal amino acid identity (e.g. 50% and 70%) and an overlap of ≥ 35 amino acids in the region of local alignment.

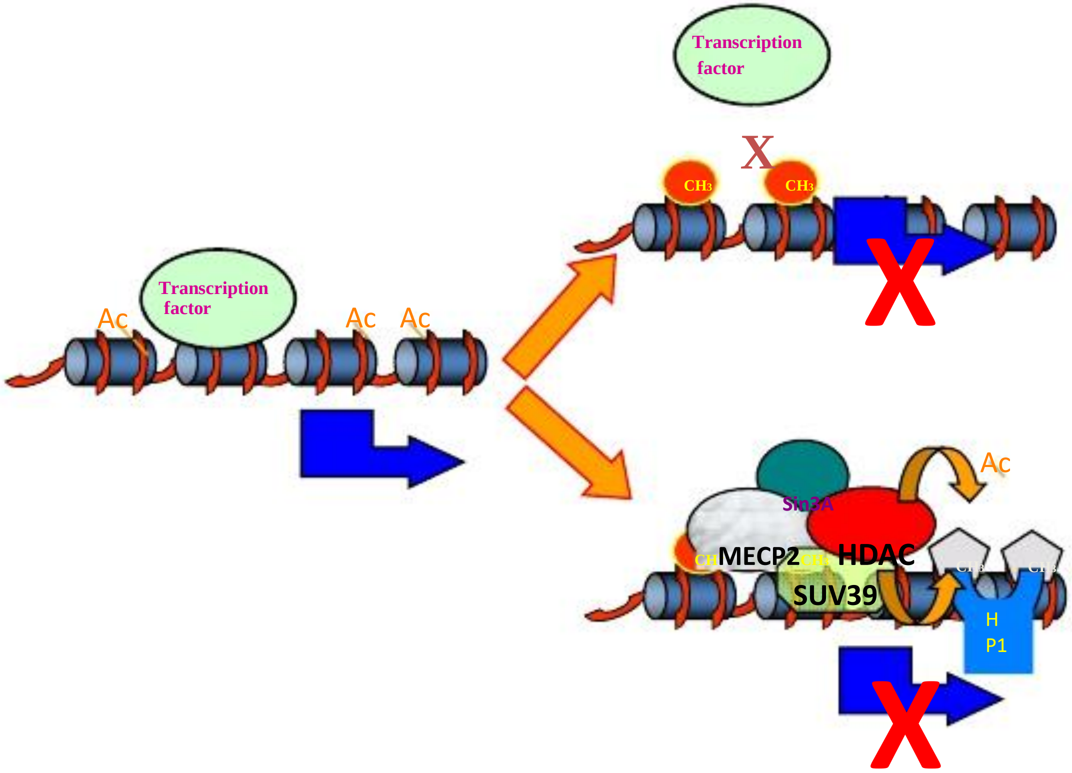
4. Coding regions of pairs that meet these criteria will be aligned with the corresponding region and inspected for putative function divergence hallmarks.

5. Local warehouse were searched for further indicators derived from high-throughput data (esp. genetic, genomic, transcriptomic, proteomic and pathway data).

7 out of 50000+ new paralogous pairs showed clear functional divergence features involved in early development regulation.

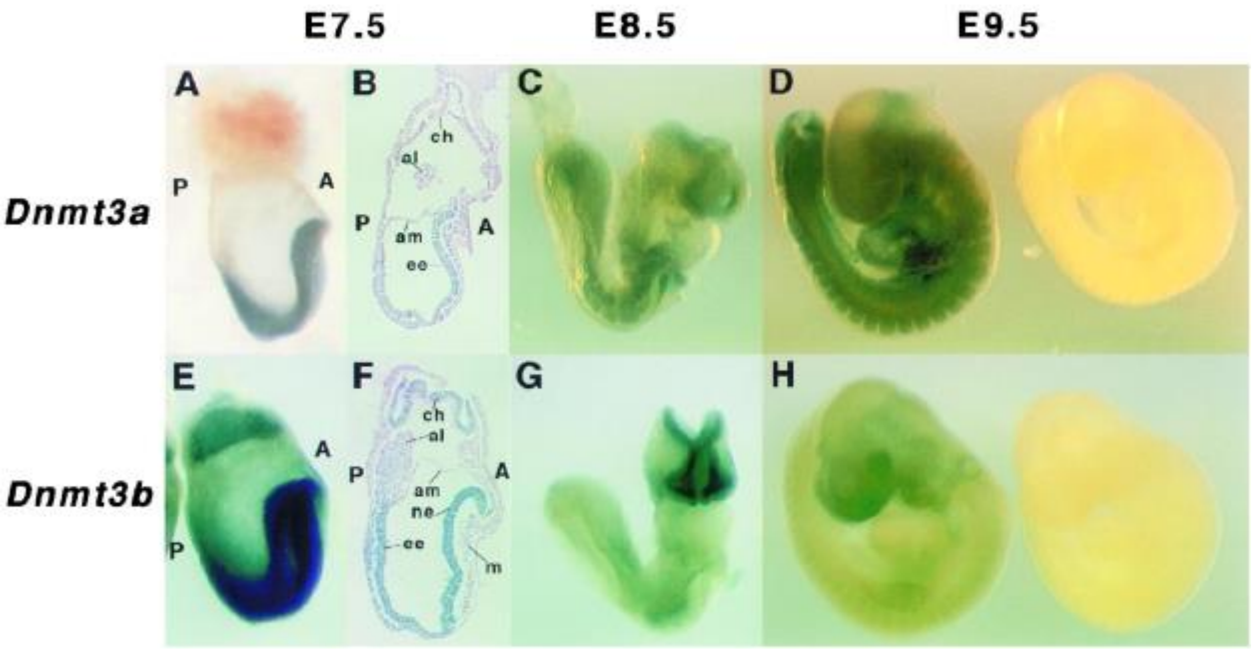


DNA methylation silences gene expression by two mechanisms



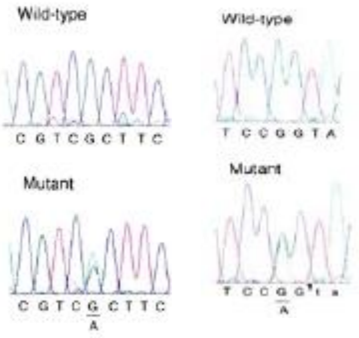
(Modified from Moshe Szyf'sslide)

DNMT3-induced methylation is critical for early mouse embryo development

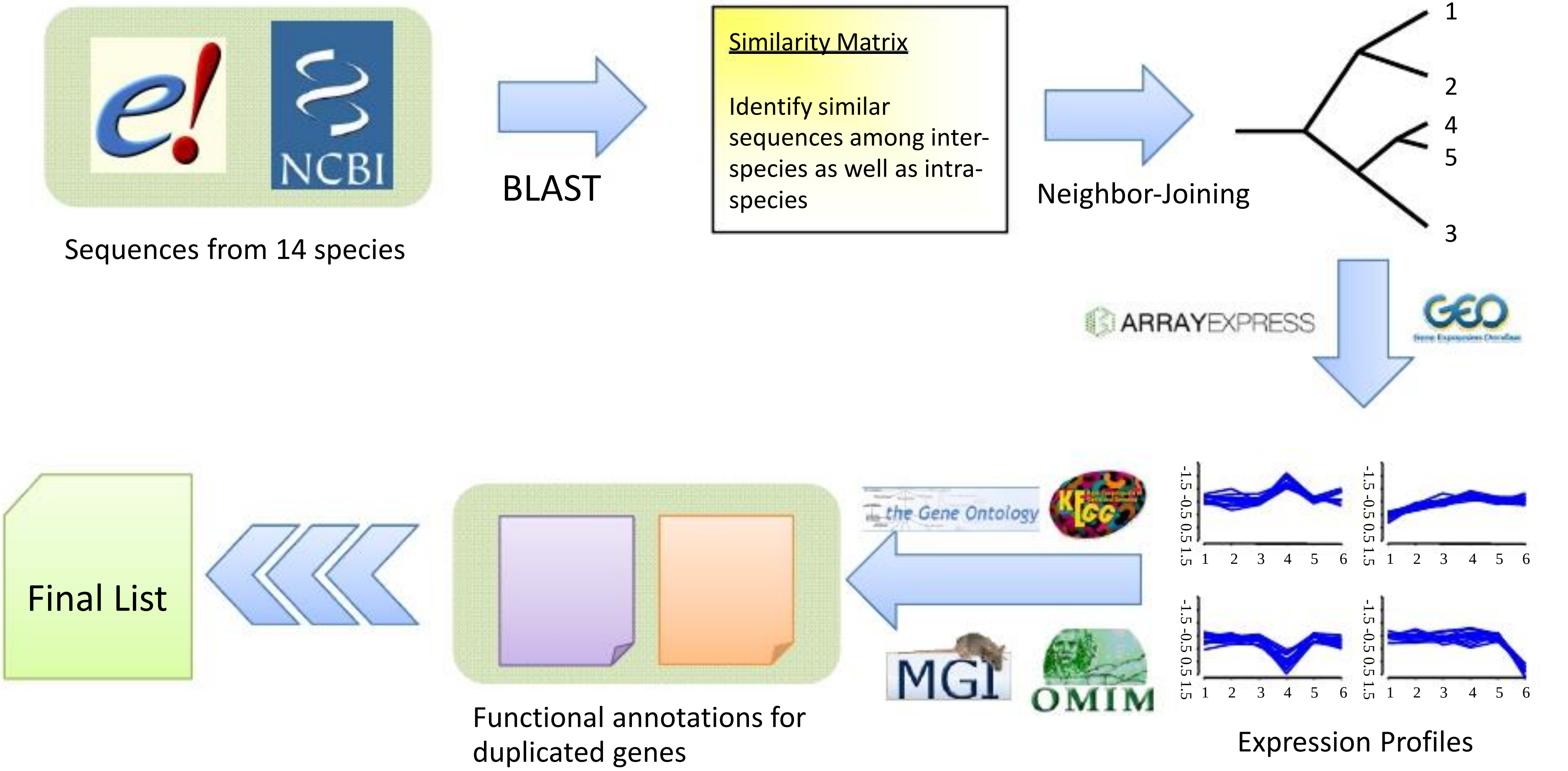


…and also for human ICF Syndrome

(http://www.newsoftheworld.co.uk/news/402211/Kiss-of-death-Grace-and-Luke-Hicklin-have-hereditary-ICF-syndrome.html)



Computationally screening for function divergent genes involved in early development regulation



Sequences from 14 species

BLAST

Similarity Matrix

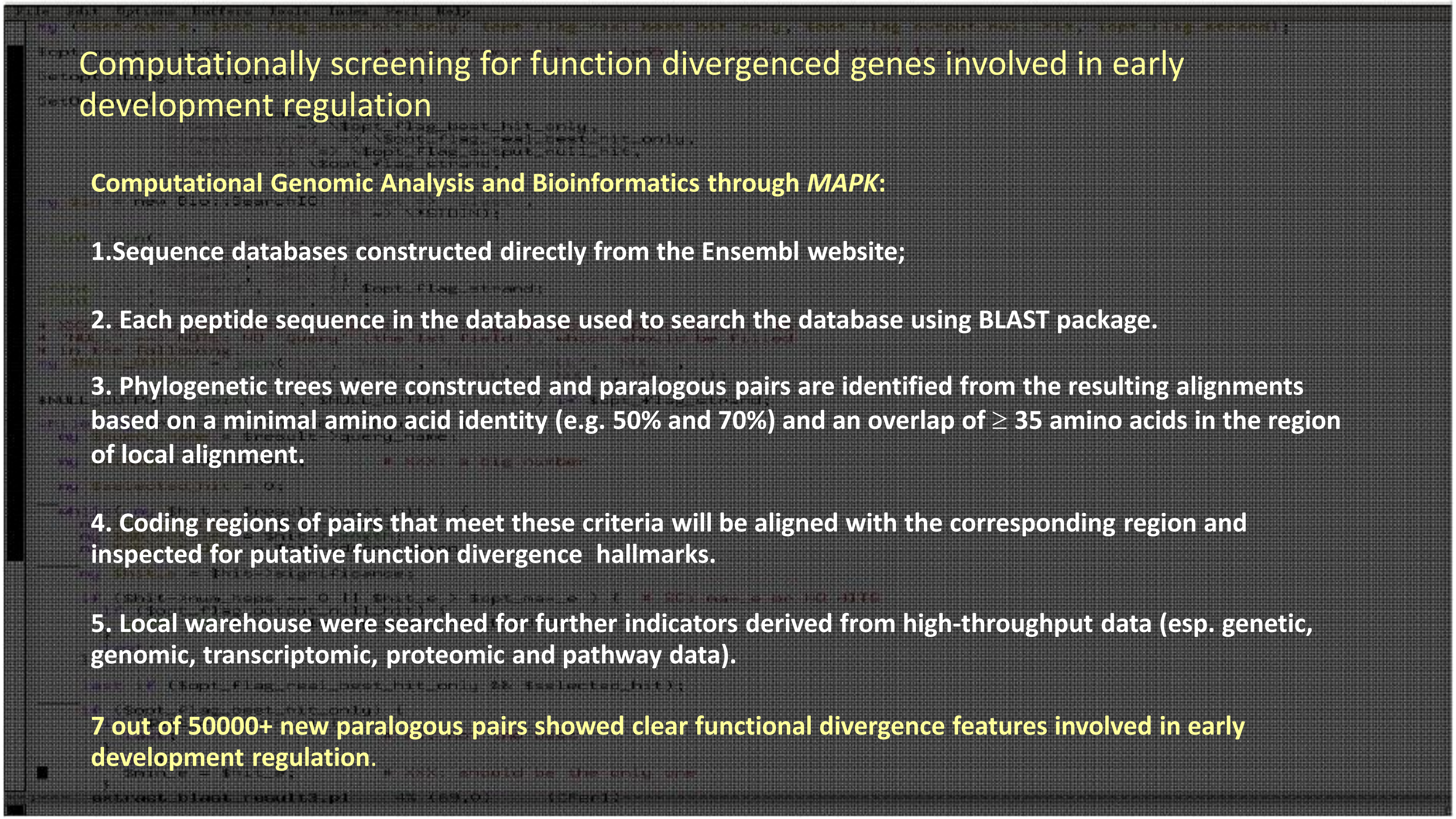
Identify similar sequences among inter-species as well as intra-species

Neighbor-Joining

Expression Profiles

Functional annotations for duplicated genes

Final List



Computationally screening for function divergenced genes involved in early development regulation

Computational Genomic Analysis and Bioinformatics through *MAPK*:

1.Sequence databases constructed directly from the Ensembl website;

2. Each peptide sequence in the database used to search the database using BLAST package.

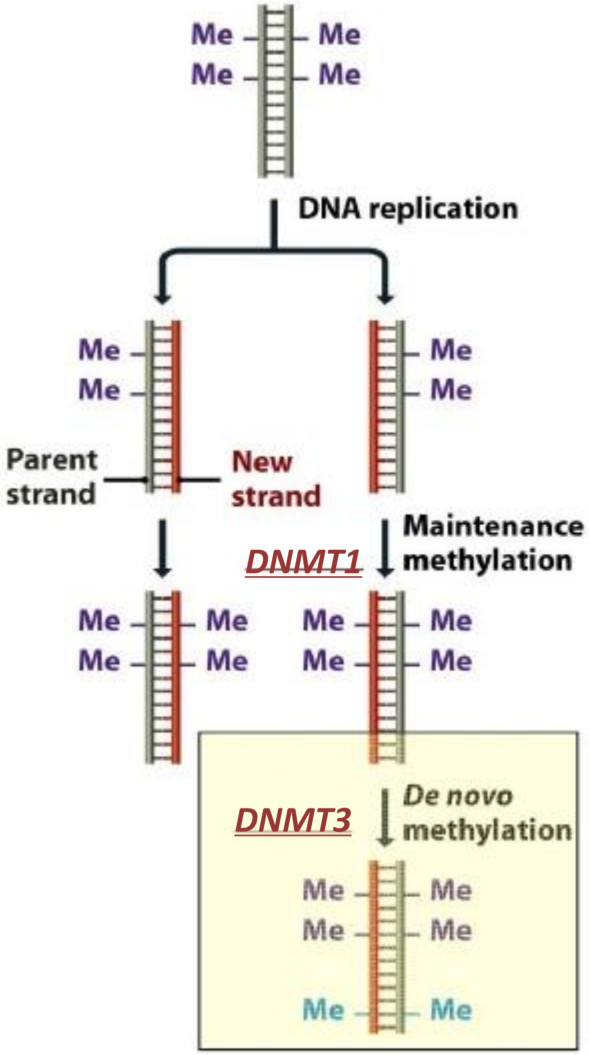
3. Phylogenetic trees were constructed and paralogous pairs are identified from the resulting alignments based on a minimal amino acid identity (e.g. 50% and 70%) and an overlap of ≥ 35 amino acids in the region of local alignment.

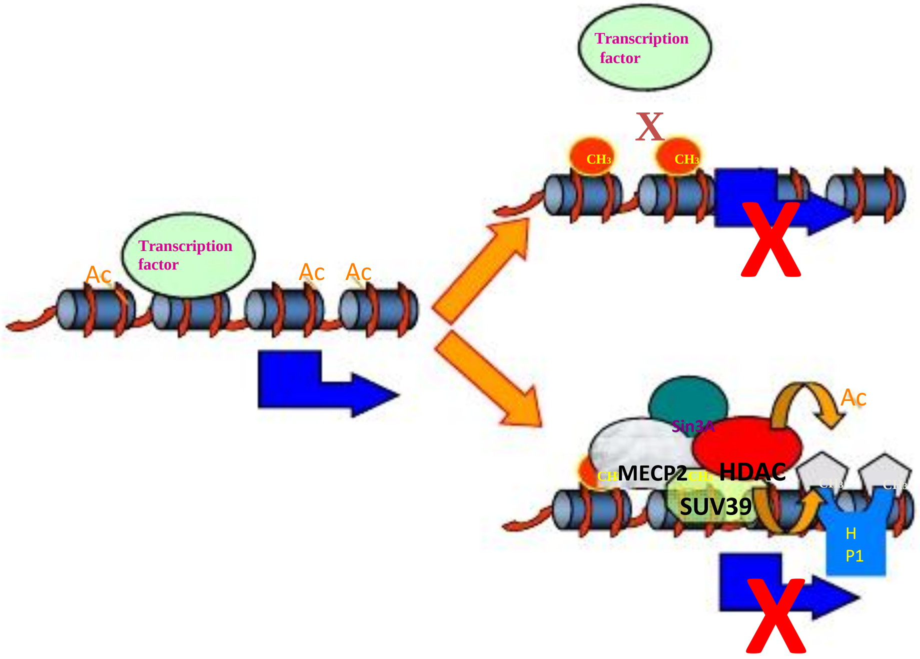
4. Coding regions of pairs that meet these criteria will be aligned with the corresponding region and inspected for putative function divergence hallmarks.

5. Local warehouse were searched for further indicators derived from high-throughput data (esp. genetic, genomic, transcriptomic, proteomic and pathway data).

7 out of 50000+ new paralogous pairs showed clear functional divergence features involved in early development regulation.

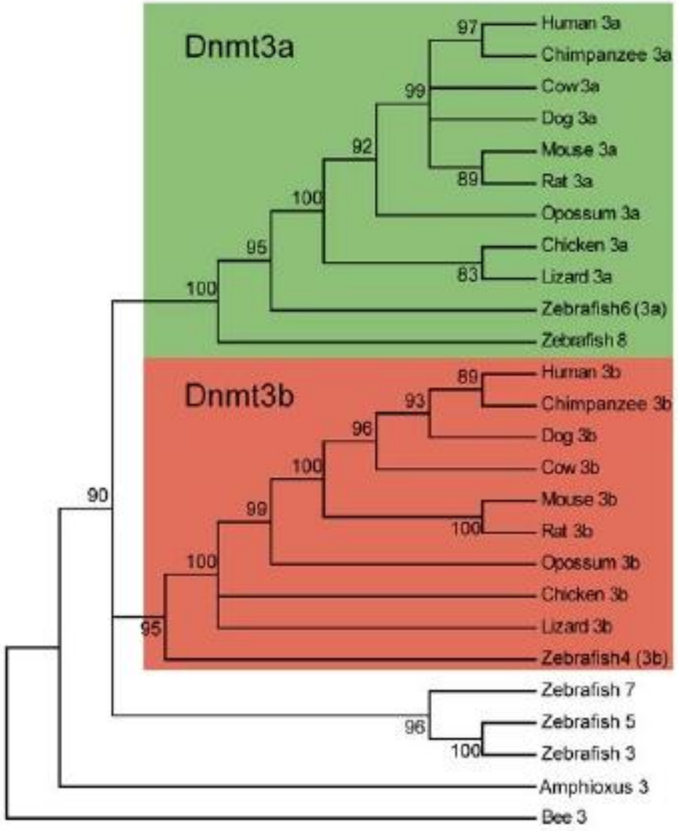
DNA methylation silences gene expression by two mechanisms





(Modified from Moshe Szyf's slide)

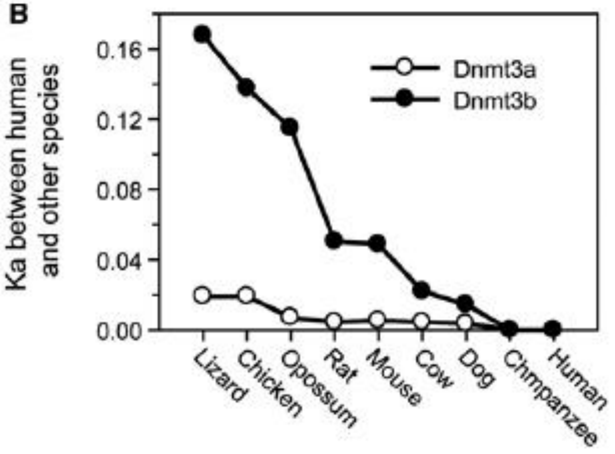
DNMT3 raised around the separation of vertebrates

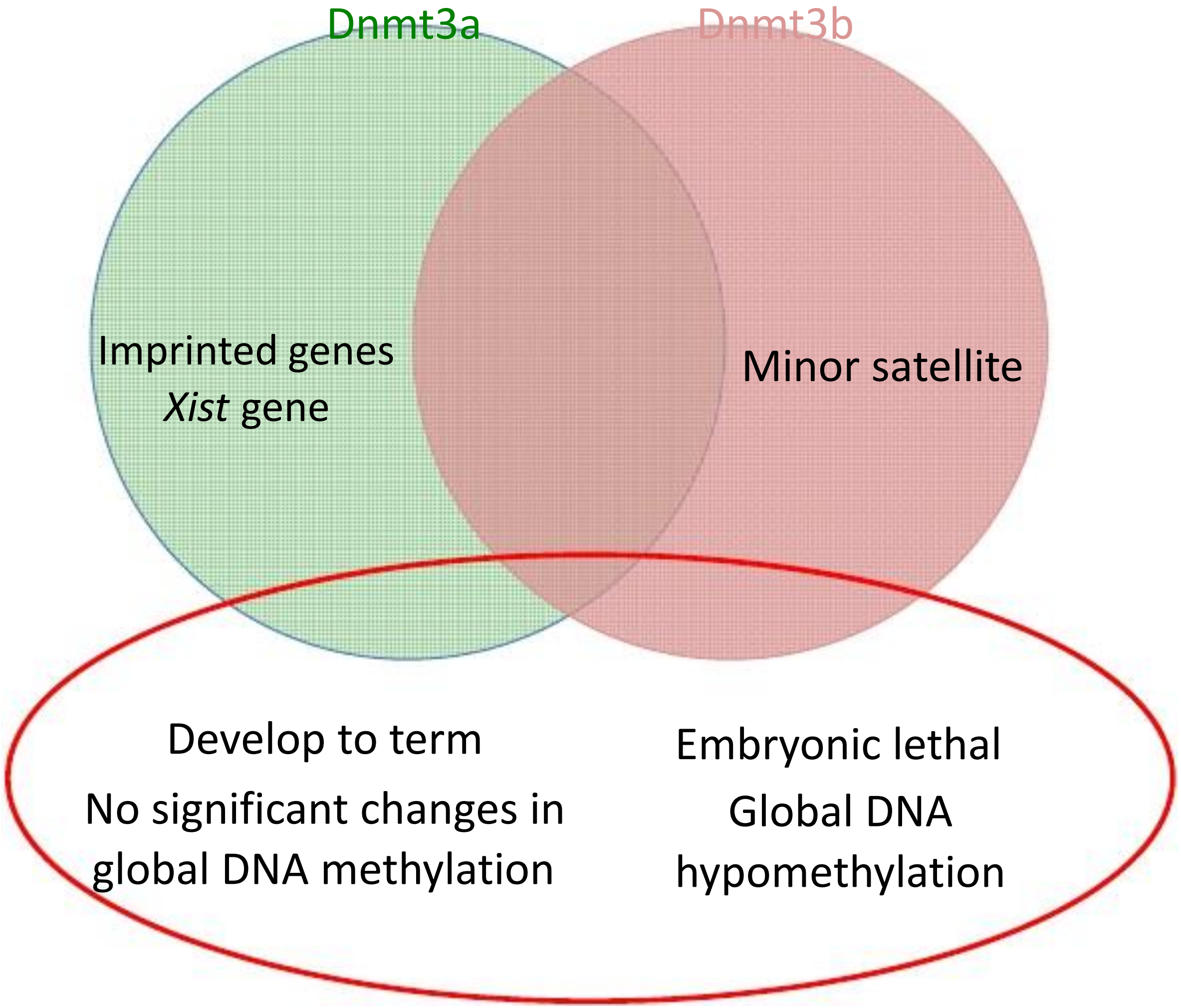


(Shen *et al*. *Nucle Acids Res.* 2010)

A

	Synonymous			Non-Synonymous		
	S_{3a}	S_{3b}	P-value	N_{3a}	N_{3b}	P-value
Human	221	228	0.377	294	343	0.025*
Chimpanzee	222	227	0.413	294	343	0.025*
Dog	227	236	0.340	298	342	0.041*
Cow	230	222	0.366	299	347	0.029*
Mouse	220	228	0.342	300	345	0.040*
Rat	227	229	0.460	298	340	0.049*
Opossum	247	257	0.319	298	357	0.011*
Chicken	226	208	0.196	307	350	0.047*
Lizard	231	229	0.463	305	353	0.031*





Dnmt3a

Dnmt3b

Specific targets

Imprinted genes
Xist gene

Minor satellite

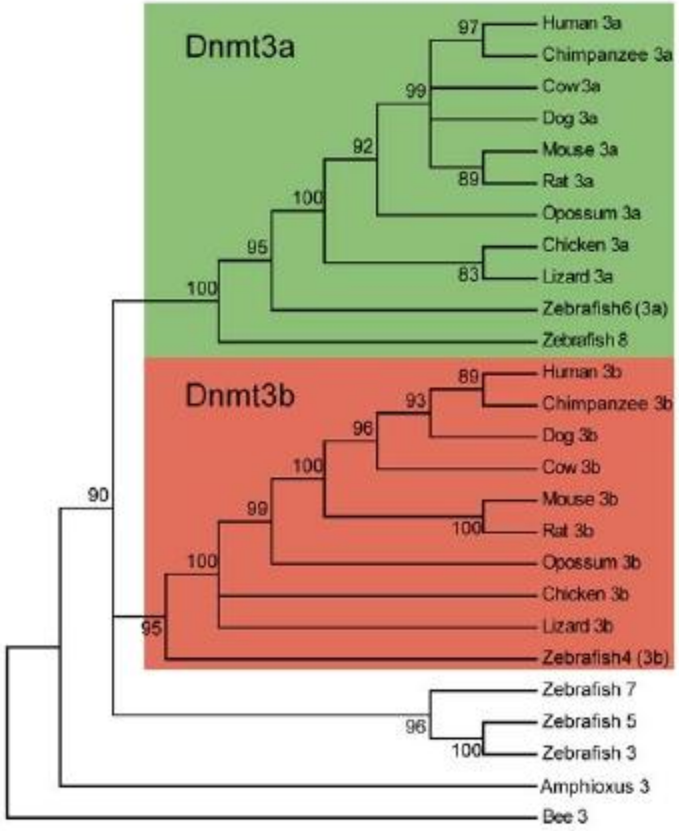
Phenotype of KO mice

Develop to term
No significant changes in global DNA methylation

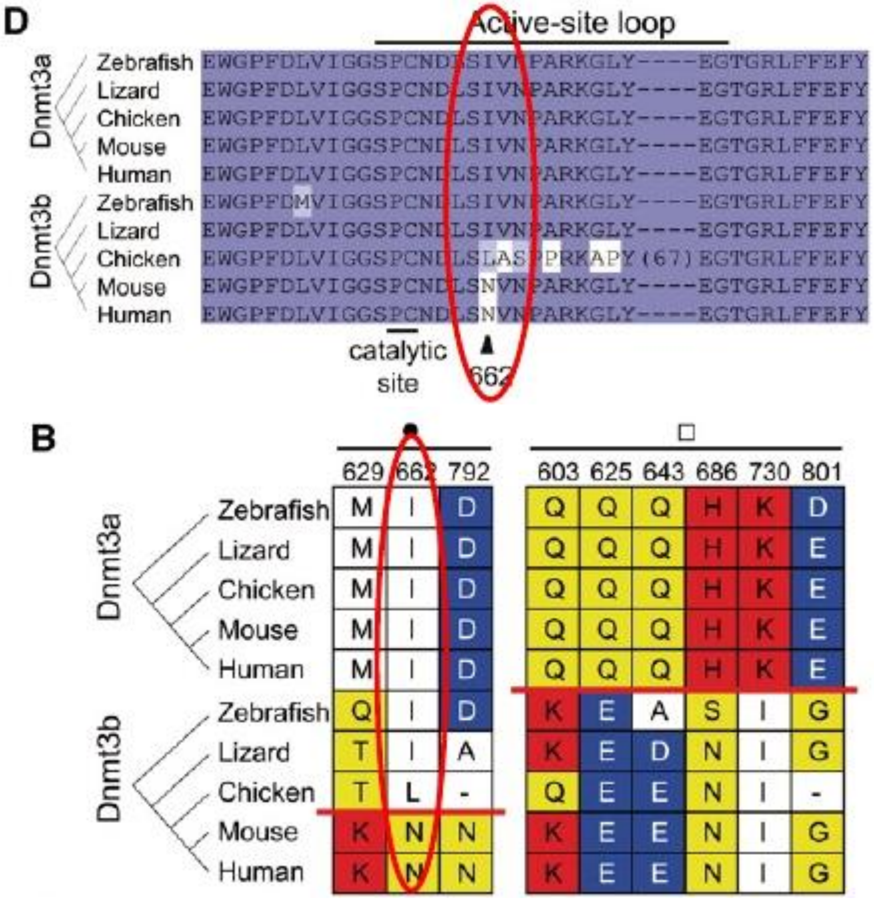
Embryonic lethal
Global DNA hypomethylation

(Source: Li Shen, 2010)

A mammalian DNMT3b-specific amino acid change appeared near catalytic site



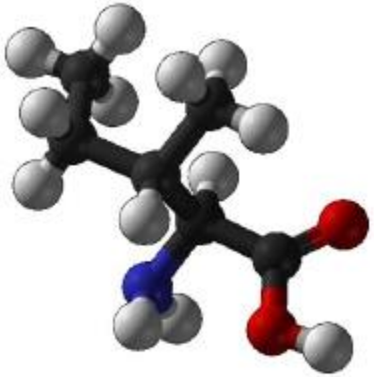
(Shen *et al*. *Nucle Acids Res.* 2010)



Dnmt3a	Zebrafish	EWGPFDLVIGGSPCNDLSIVNPARKGLY----EGTGRLFFEFY
	Lizard	EWGPFDLVIGGSPCNDLSIVNPARKGLY----EGTGRLFFEFY
	Chicken	EWGPFDLVIGGSPCNDLSIVNPARKGLY----EGTGRLFFEFY
	Mouse	EWGPFDLVIGGSPCNDLSIVNPARKGLY----EGTGRLFFEFY
	Human	EWGPFDLVIGGSPCNDLSIVNPARKGLY----EGTGRLFFEFY
Dnmt3b	Zebrafish	EWGPFDMVIGGSPCNDLSIVNPARKGLY----EGTGRLFFEFY
	Lizard	EWGPFDLVIGGSPCNDLSIVNPARKGLY----EGTGRLFFEFY
	Chicken	EWGPFDLVIGGSPCNDLSLASPPRKAPY(67)EGTGRLFFEFY
	Mouse	EWGPFDLVIGGSPCNDLSNVNPARKGLY----EGTGRLFFEFY
	Human	EWGPFDLVIGGSPCNDLSNVNPARKGLY----EGTGRLFFEFY

		629	662	792	603	625	643	686	730	801
Dnmt3a	Zebrafish	M	I	D	Q	Q	Q	H	K	D
	Lizard	M	I	D	Q	Q	Q	H	K	E
	Chicken	M	I	D	Q	Q	Q	H	K	E
	Mouse	M	I	D	Q	Q	Q	H	K	E
	Human	M	I	D	Q	Q	Q	H	K	E
Dnmt3b	Zebrafish	Q	I	D	K	E	A	S	I	G
	Lizard	T	I	A	K	E	D	N	I	G
	Chicken	T	L	-	Q	E	E	N	I	-
	Mouse	K	N	N	K	E	E	N	I	G
	Human	K	N	N	K	E	E	N	I	G

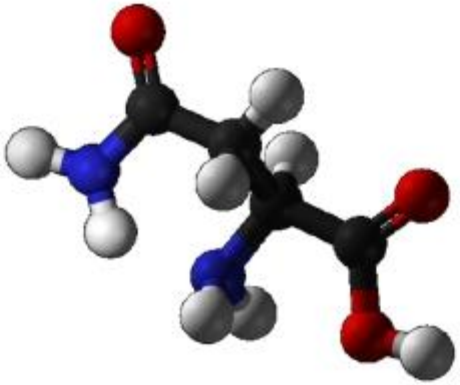
Isoleucine (I)
异亮氨酸



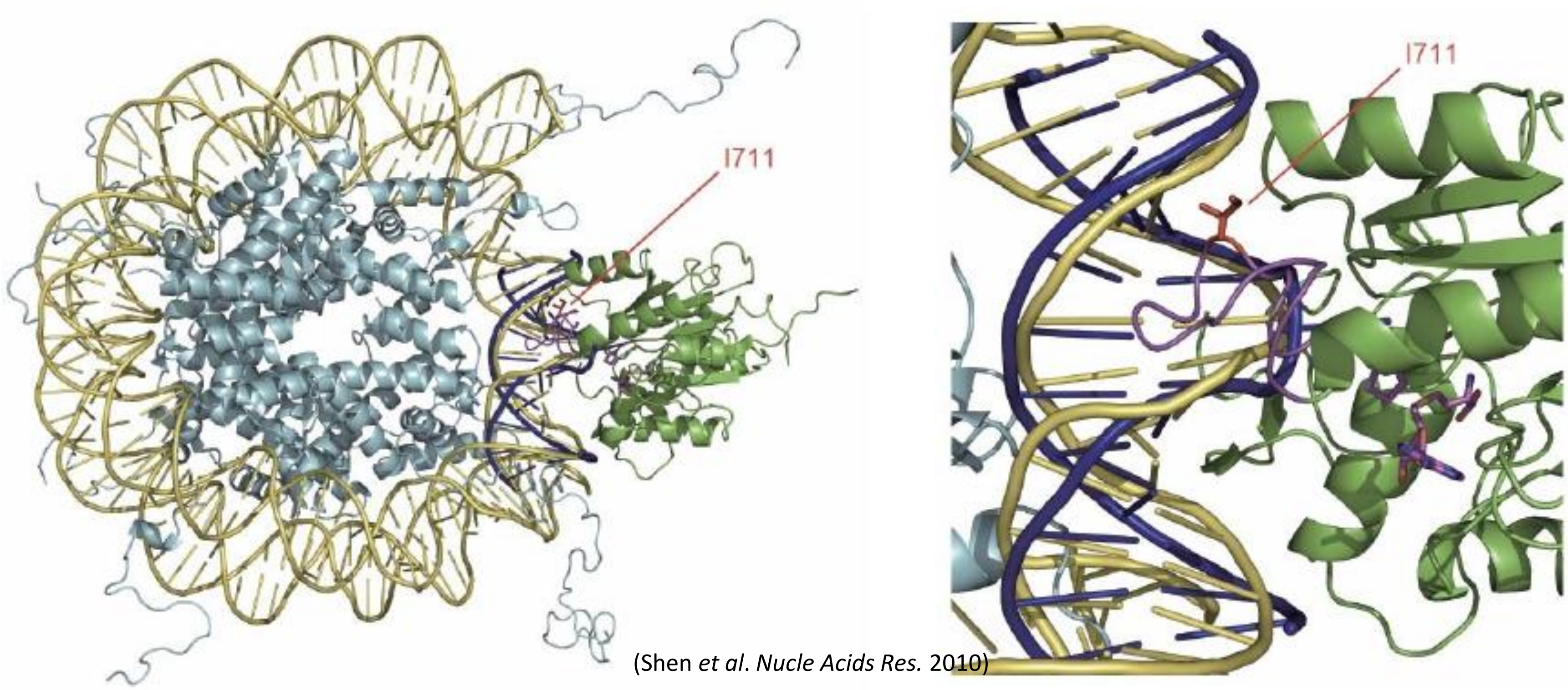
$C_4H_8N_2O_3$

- non-Polar,
- Hydropathy index = 4.5
- pI at 25° C = 6.04

Asparagine (N)
天冬酰胺

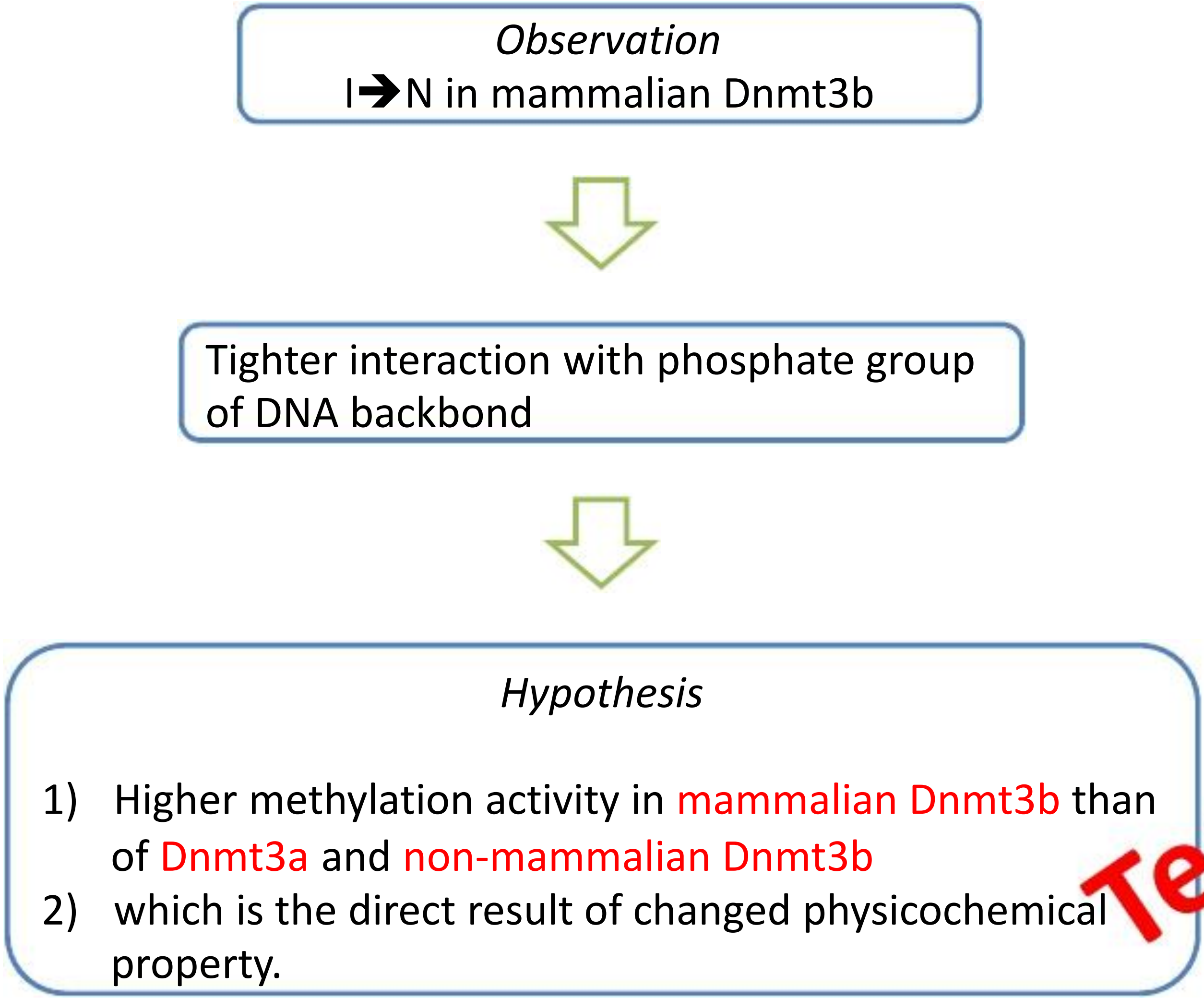


- Polar,
- Hydropathy index = −3.5
- pI at 25° C = 10.76



(Shen *et al*. *Nucle Acids Res.* 2010)

Structural analysis suggested that the I→N results in a tighter enzyme-DNA interaction



Observation

I➔N in mammalian Dnmt3b

Tighter interaction with phosphate group of DNA backbond

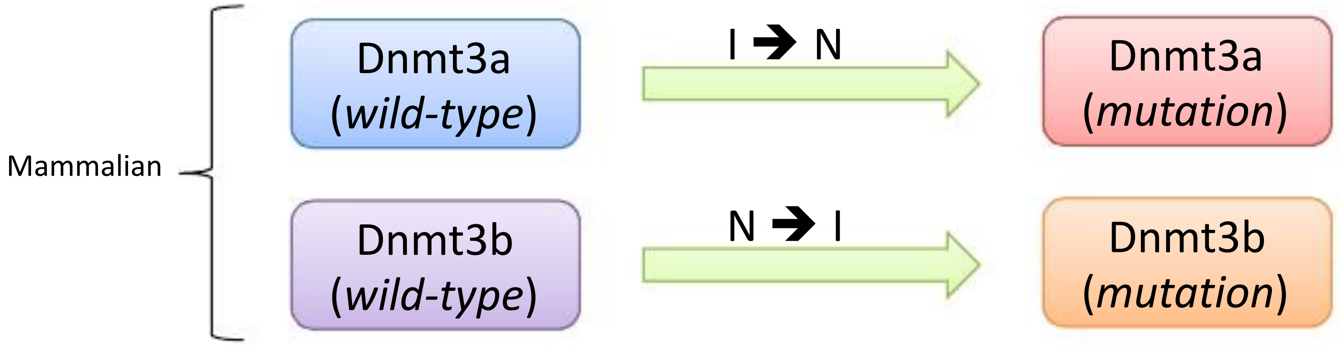
Hypothesis

1) Higher methylation activity in mammalian Dnmt3b than of Dnmt3a and non-mammalian Dnmt3b
2) which is the direct result of changed physicochemical property.

Test!

Dr. Li Shen
Research Fellow, Harvard

Test the hypothesis with wet experiences



Mammalian

Dnmt3a (*wild-type*)

I ➔ N

Dnmt3a (*mutation*)

Dnmt3b (*wild-type*)

N ➔ I

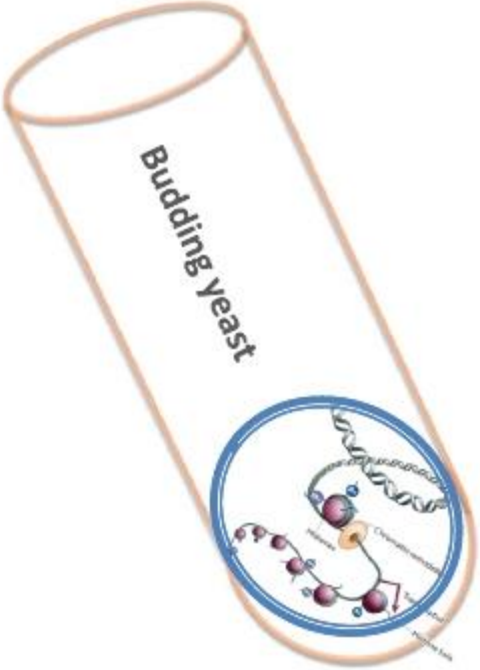
Dnmt3b (*mutation*)

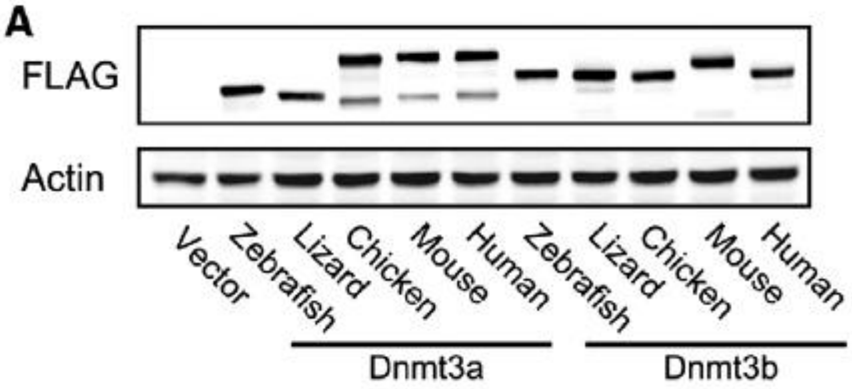
“Site-directed mutagenesis is a molecular biology method that is used to make specific and intentional changes to the DNA sequence of a gene and any gene products. Also called site-specific mutagenesis or oligonucleotide-directed mutagenesis, it is used for investigating the structure and biological activity of DNA, RNA, and protein molecules, and for protein engineering. With decreasing costs of oligonucleotide synthesis, artificial gene synthesis is now occasionally used as an alternative to site-directed mutagenesis.” (Source: wikipedia.org)

endogenous DNMTs ??

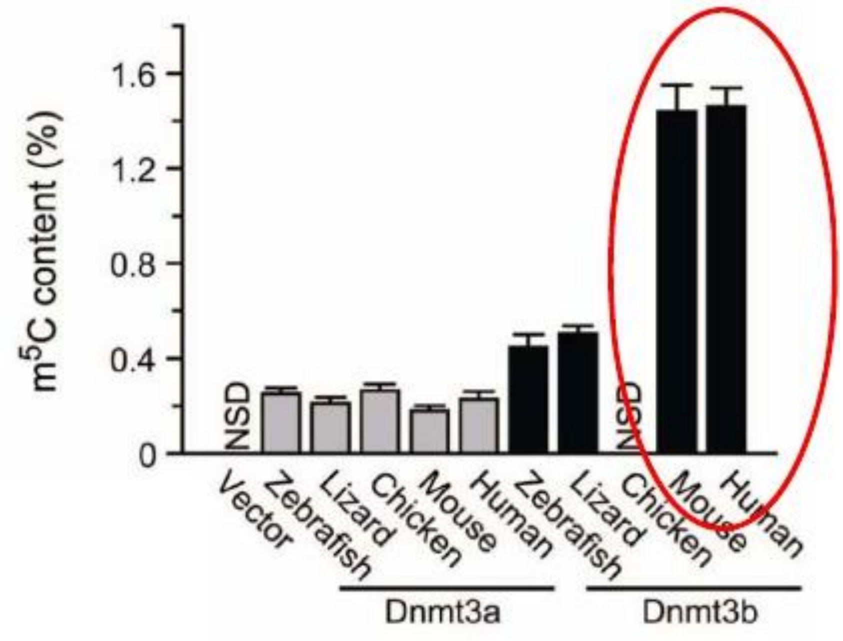
Using budding yeast as the "*in vivo* test tube"

Epigenetic features	mammals	budding yeast
Chromatin	Yes	Yes§
Histone acetylation	Yes	Yes§
H3K4	Yes	Yes§
H3K36	Yes	Yes§
H3K79	Yes	Yes§
SWI/SNF complexes	Yes	Yes§
CHD1 ATPase	Yes	Yes§
SWR1 ATPase	Yes	Yes§
ISWI ATPase	Yes	Yes§
Endogenous methylation	Yes	No



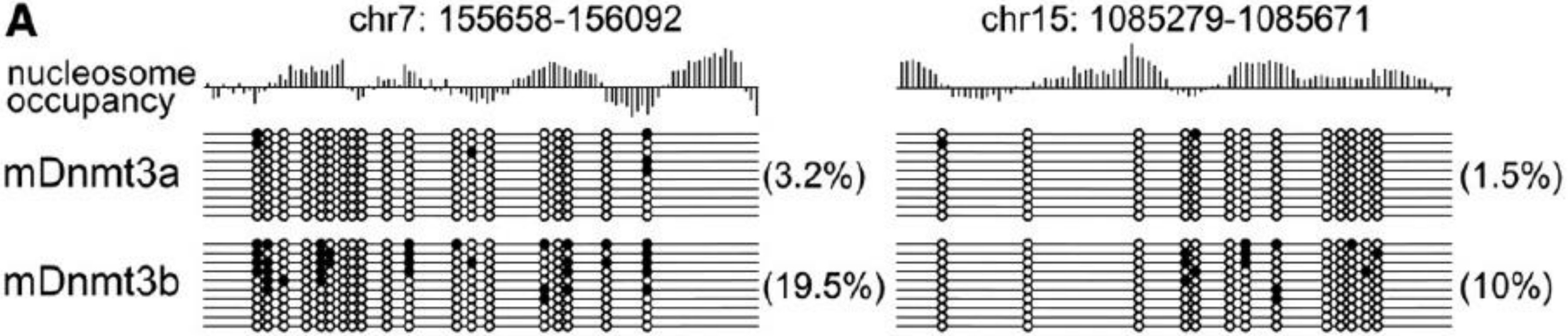


(Shen *et al*. *Nucle Acids Res.* 2010)

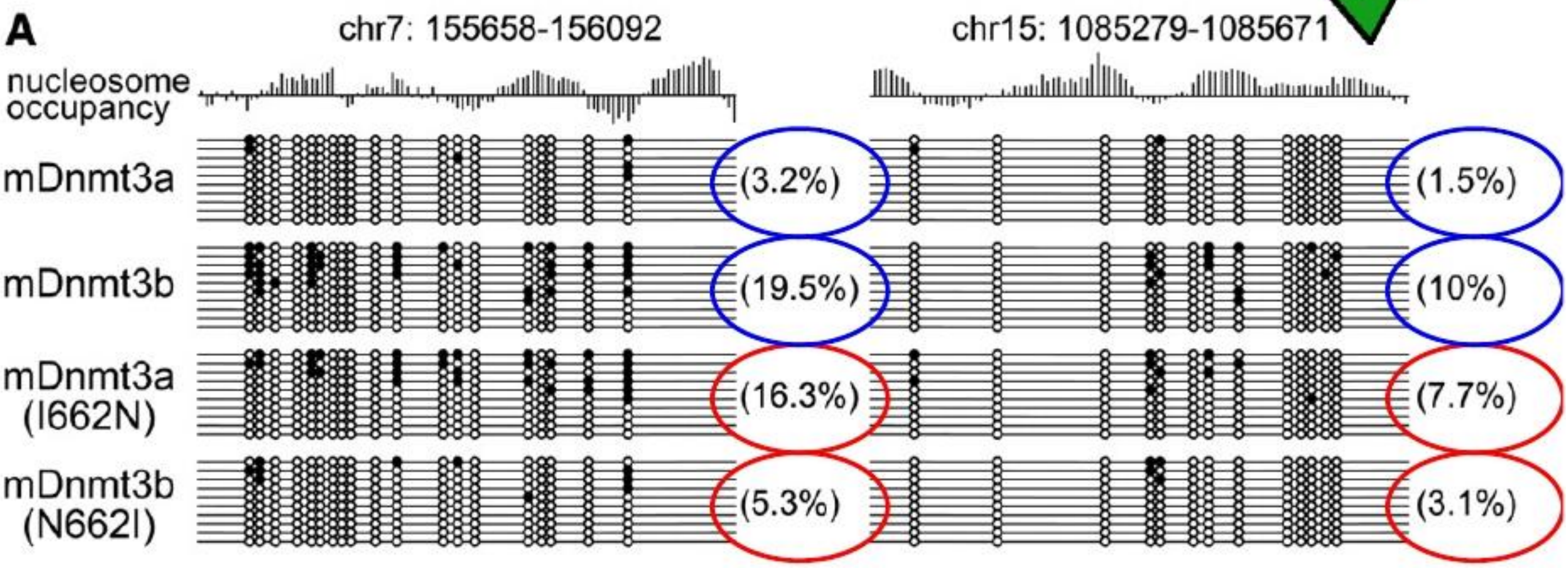


Mammalian Dnmt3b posses higher chromatin DNA methylation activity than Dnmt3a and non-mammalian Dnmt3b

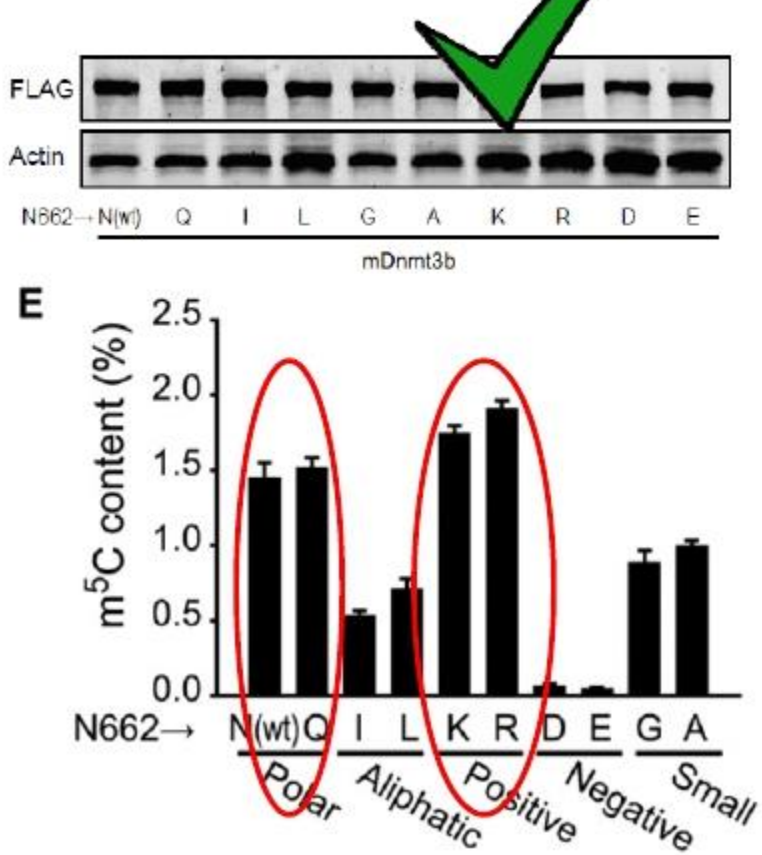
Would I662N substitution accounts for the increased nucleosome DNA methylation activity in mammalian DNMT3b?

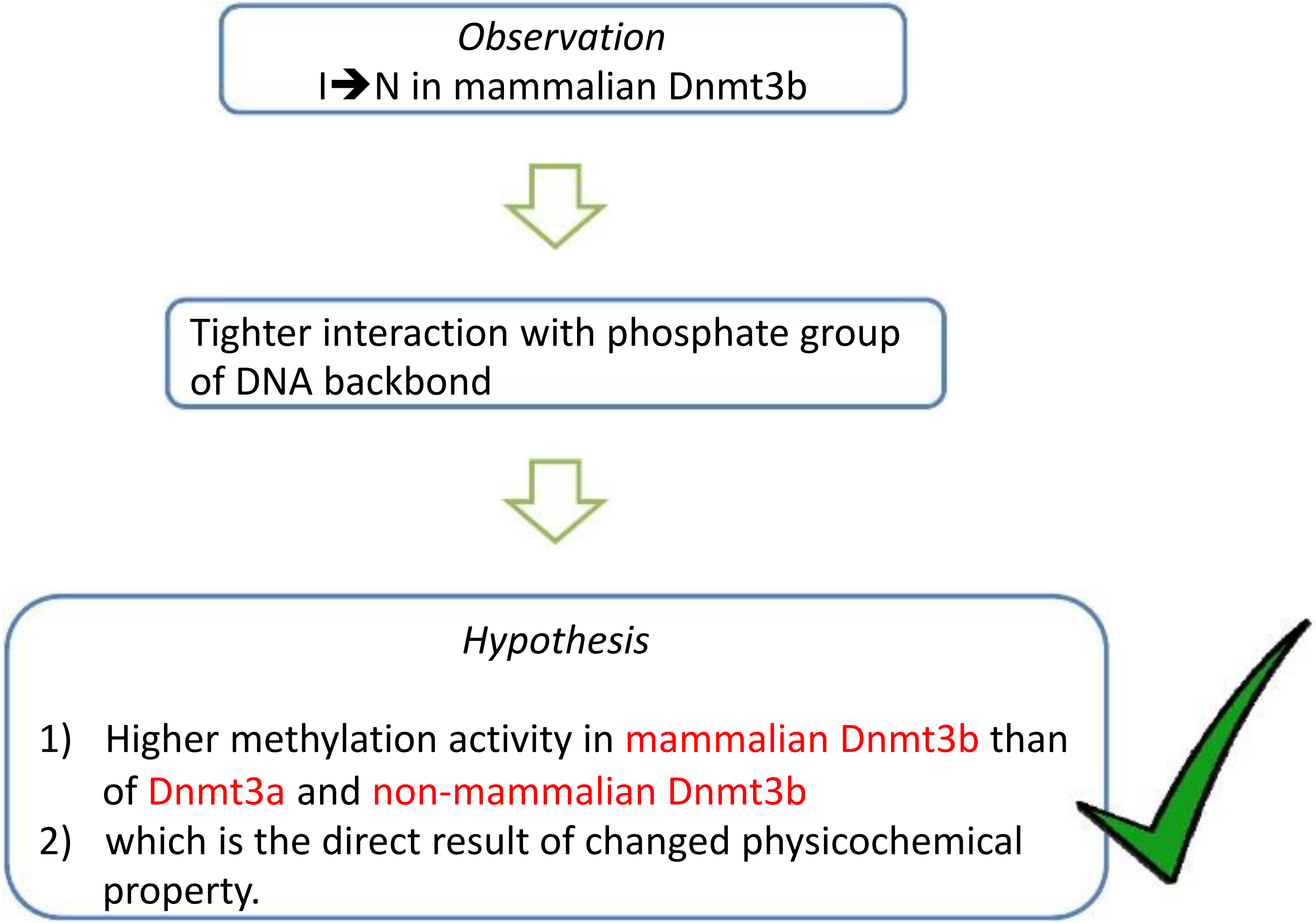


Would I662N substitution accounts for the increased nucleosome DNA methylation activity in mammalian DNMT3b?



Does the changed physicochemical property matter?





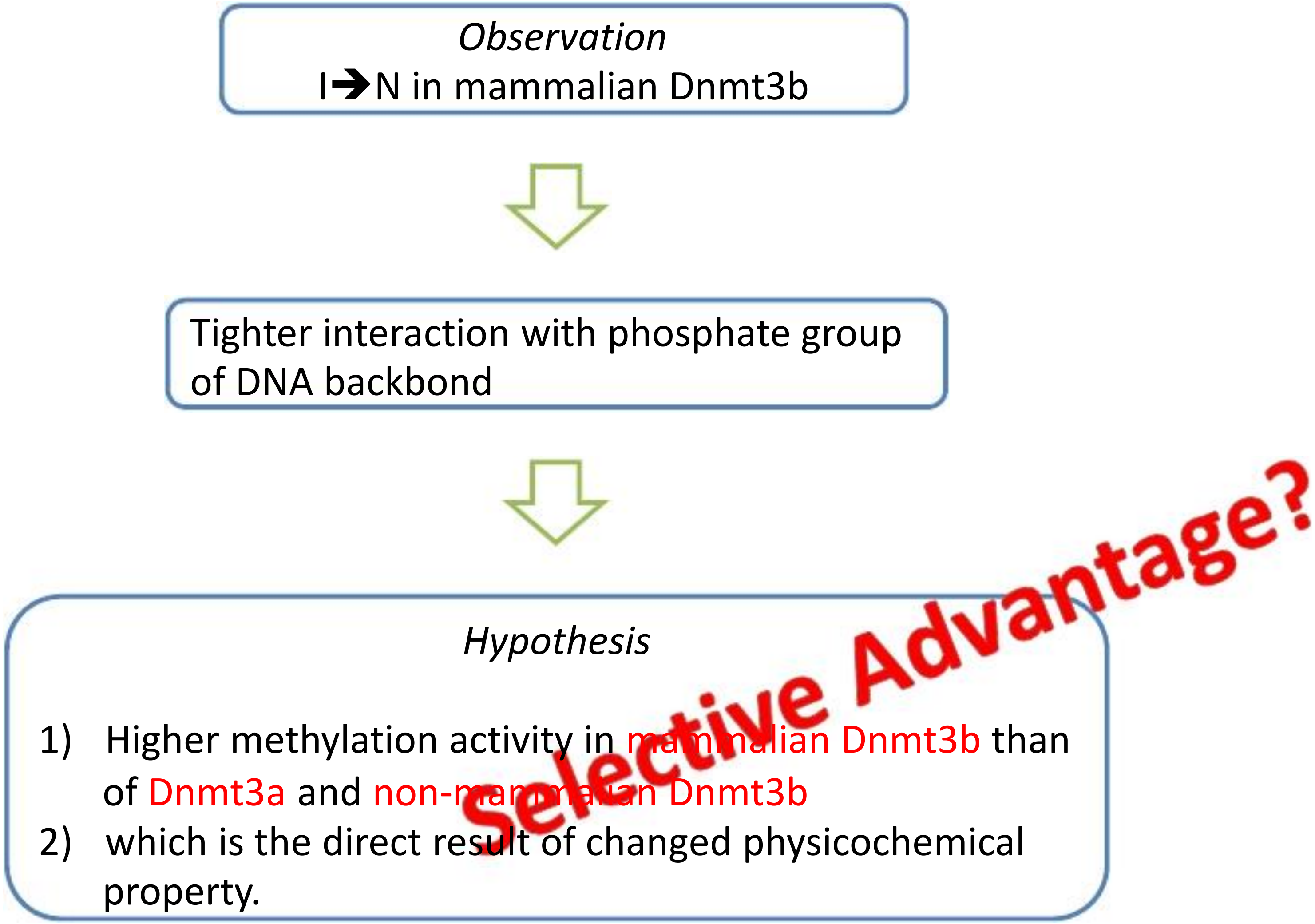
Observation

I➔N in mammalian Dnmt3b

Tighter interaction with phosphate group of DNA backbond

Hypothesis

1) Higher methylation activity in mammalian Dnmt3b than of Dnmt3a and non-mammalian Dnmt3b
2) which is the direct result of changed physicochemical property.



Observation

I→N in mammalian Dnmt3b

Tighter interaction with phosphate group of DNA backbond

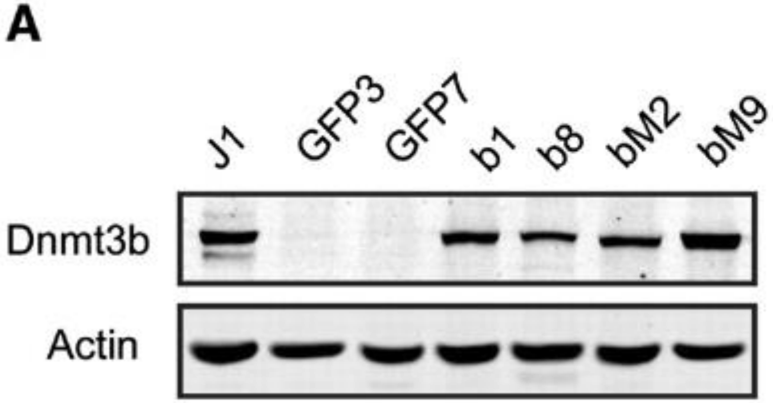
Hypothesis

1) Higher methylation activity in mammalian Dnmt3b than of Dnmt3a and non-mammalian Dnmt3b
2) which is the direct result of changed physicochemical property.

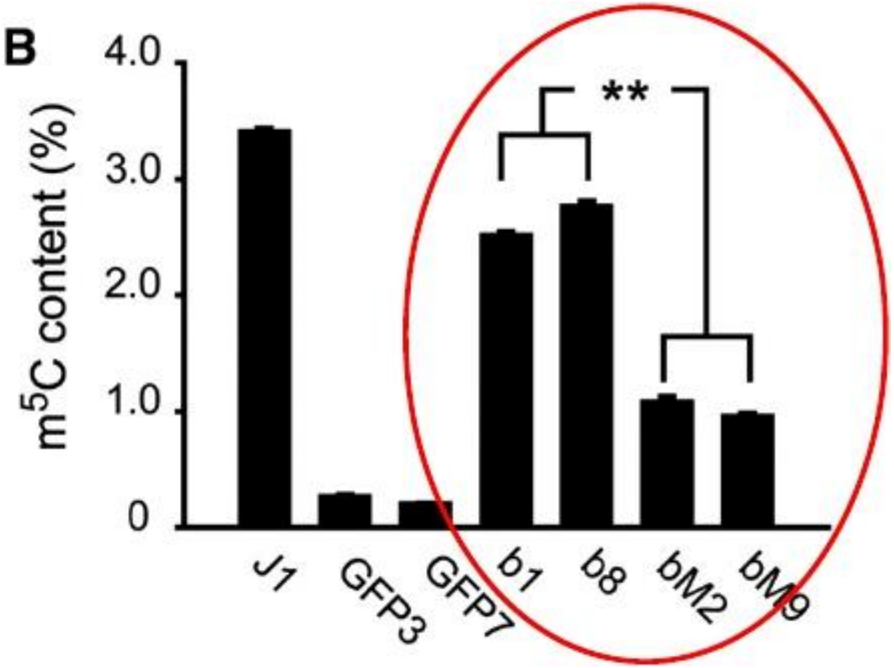
Selective Advantage?

A (hypothesis) connection of the chromatin DNA methylation activity of Dnmt3b with the density of repetitive sequences in the genome?

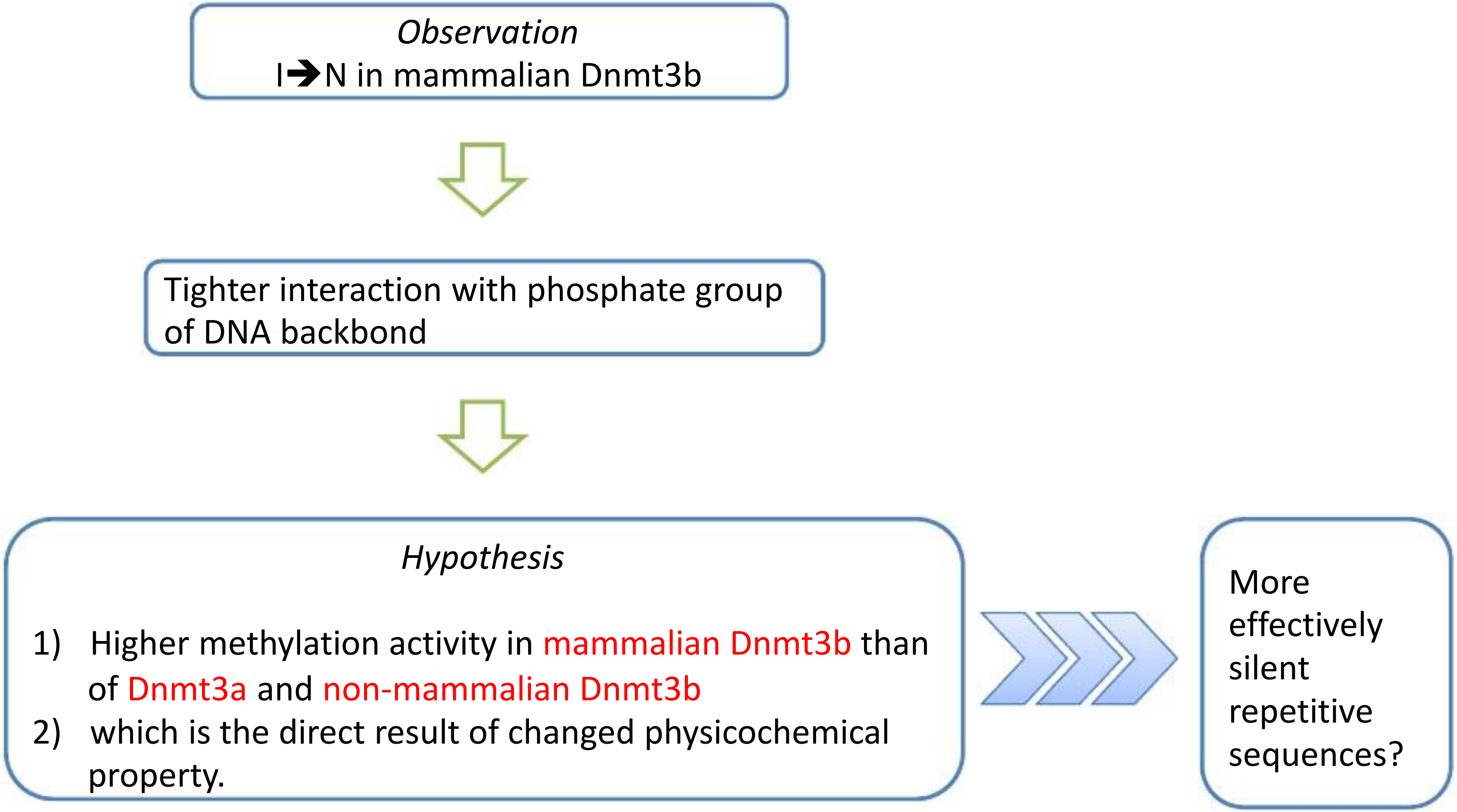
	Percentage of repeats in the genome	Chromatin DNA methylation activity of Dnmt3b
	~ 40-50%	highest
~	~10%	inactive



GFP: GFP3, GFP7
mDnmt3b: b1, b8
mDnmt3b(N662I): bM2, bM9



The substitution (I662N) is crucial for mammalian Dnmt3b to efficiently methylate repetitive sequences in mammalian cells



Observation

I➔N in mammalian Dnmt3b

Tighter interaction with phosphate group of DNA backbond

Hypothesis

1) Higher methylation activity in mammalian Dnmt3b than of Dnmt3a and non-mammalian Dnmt3b
2) which is the direct result of changed physicochemical property.

More effectively silent repetitive sequences?

Summary

- Evolution-guided bioinformatics analysis successfully identified interesting genes involved in early development regulation showed clear functional novelty during evolution, and also provided strong hints for the key substitution, its biochemical effort, and the eventually functional significance.

- Key single substitution could result in significant functional novelty and help novel gene (re-)wired itself into existing circuits.

- An integrated, genome-scale bioinformatic analysis combined with targeted experimental assay is effective in studying complex biological system.

Bioinformatics: an interdisciplinary field that develop and apply computer and computational technologies to study biomedical questions

- As a technology, bioinformatics is a powerful technology to manage, search, and analyze big data in life sciences.
- As a methodology, bioinformatics is a top-down, holistic, data-driven, genome-wide, and systems approach that generates new hypotheses, find new patterns, and discover new functional elements.

Bioinformatics: Introduction and Methods

Computer Science Department, Southwest University

Thank you