Computational approaches to study Genetics

遺伝学を研究するための計算アプローチ

2020/03/09

RIKEN
Dr. Jeffrey Fawcett

Arithmer Seminar

弊社社員によるスライドではなく、弊社にて平均週一回で開催されている、
「Arithmer Seminar」にて登壇頂いた外部の方々によるスライドです。

Computational approaches to study Genetics

Jeffrey Fawcett

(31 Oct 2019 @Arithmer)

RIKEN

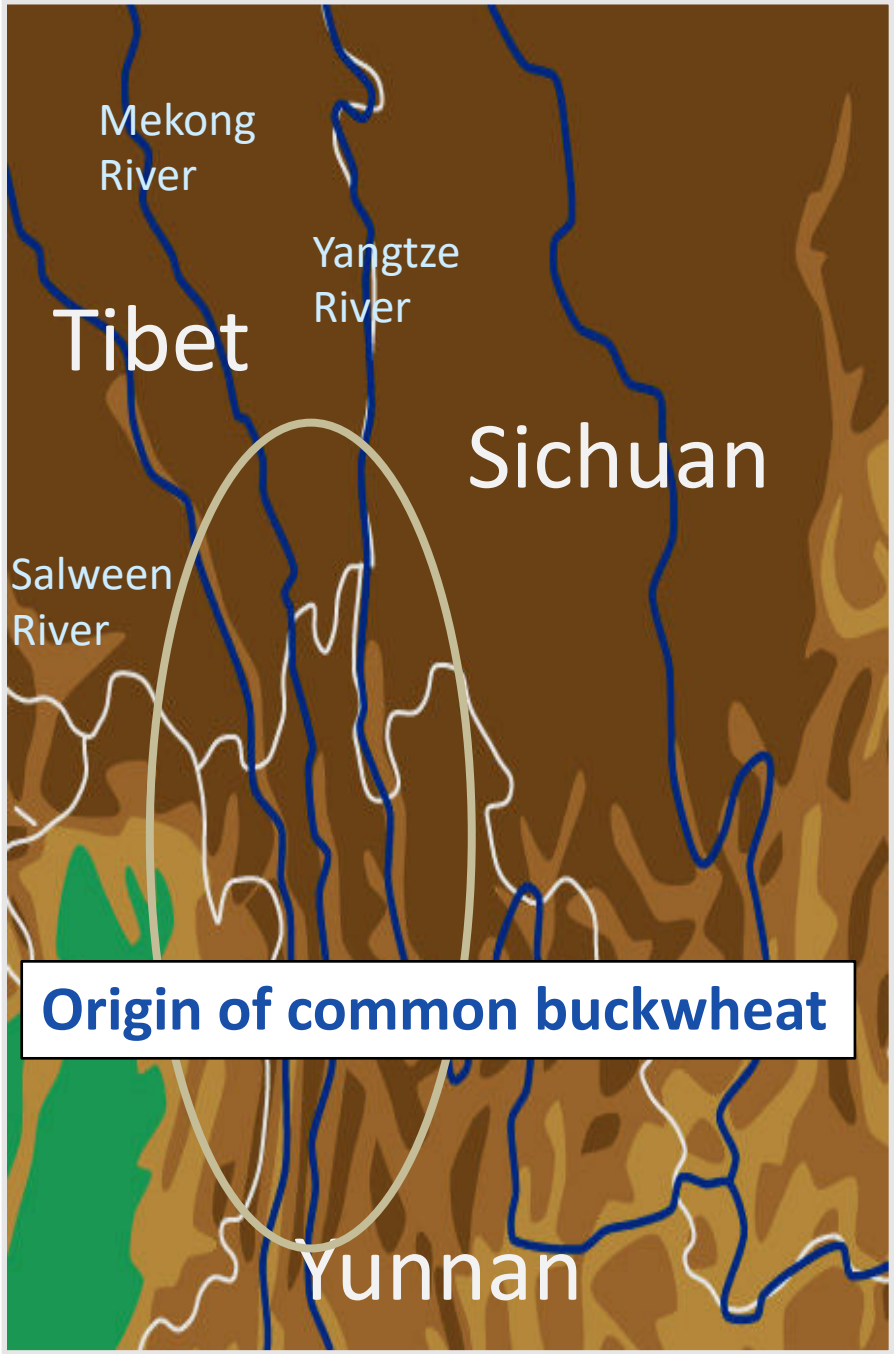
iTHEMS
RIKEN interdisciplinary Theoretical & Mathematical Sciences


梅里雪山（6,740m）
Mt. Meili Xueshan


CHINA
RUSSIA
0 500 km
0 300 miles
KAZAKHSTAN
MONGOLIA
Haerbin
Ürümqi
KYRGYZSTAN
Turpan
Shenyang
Kashgar
NORTH KOREA
BEIJING
Karakoram Highway
Yinchuan
Dalian
Under administration of China
Tianjin
SOUTH KOREA
Xining
Lanzhou
Zhengzhou
Qufu (1545m)
Tibet
Xi'an
Nanjing
Chengdu
Wuhan
Shanghai
NEPAL
Lhasa
Hangzhou
Mt Everest (8848m)
BHUTAN
INDIA
Fuzhou
Longgong Caves
BANGLADESH
INDIA
Kunming
Guangzhou
TAIWAN
Nanning
MYANMAR
Hong Kong
VIETNAM
Macau
PACIFIC OCEAN
BAY OF BENGAL
THAILAND
LAOS
Haikou





Fagopyrum esculentum ssp. *ancestrale*

Domestication

Fagopyrum esculentum ssp. *esculentum*

Yasuo Yasui
(Kyoto Uni)

Chengyun Li
(Yunnan Agricultural Uni)

Jeffrey Fawcett
(RIKEN iTHEMS)

Takanori Ohsako
(Kyoto Pref Uni)

Diane Lister
(Cambridge Uni)

Aim of Buckwheat Project

- ❑ Collect & maintain genetic resources of buckwheat-related species

  (many wild species are facing extinction due to development)

- ❑ Understand domestication process of buckwheat

- ❑ Identify genes important for breeding of buckwheat

"Extensive characterization of domestication-related genes in buckwheat by utilizing the genetic resource of Yunnan province, China"

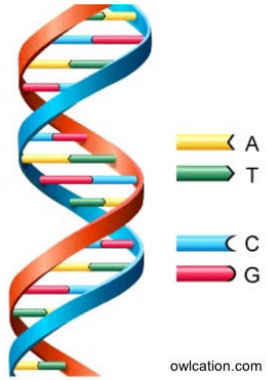
KAKENHI Fostering Joint International Research B, PI: Yasuo Yasui (Kyoto U)

中国雲南省の野生ソバ遺伝資源を活用した栽培化関連遺伝子の網羅的同定

科研費・ 国際共同研究強化B　研究代表者：安井康夫（京都大）

Why am I involved?

lots of DNA data to be generated



ATGGGCCAAGTTTTTGAAGTCTTAA
ATTTAAAAATCATATACACGTTGTA
AAAATGCGTAGGTTCATGAATGAAA
TCAGAATTTACCACACTTACTGAAA
ATGACTTGTGAGTTGTGATGGATAA
GTTTGATTAAATAAGAAAGGTAAAT
GCATATGGCTACAAAATGAAAGATT

- Huge advance in technology to generate large-scale DNA data
- Need for people with computational skills and knowledge of genetics (very few such people that work on buckwheat, horses, etc)

Main Research Interest

- Process of genetics/evolution responsible for creating the diversity of life
- Apply existing knowledge to (e.g. agronomically important) species

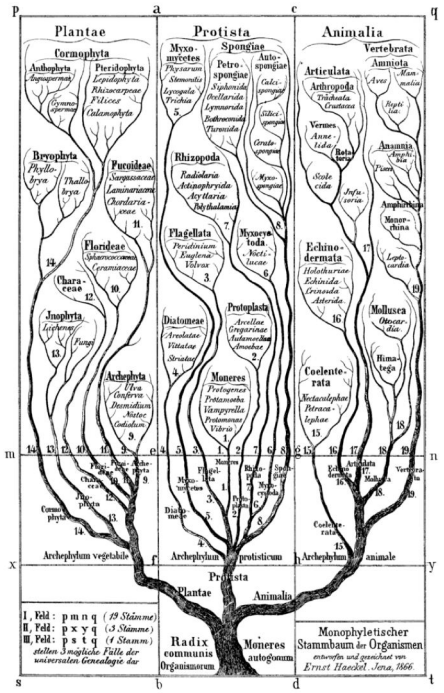
Outline

- ❑ Basic concepts of genetics and evolution
- ❑ What is written in the DNA?
- ❑ How can we know what's written in the DNA?
- ❑ How can we associate "genotype" with "phenotype"?
  - Research on Thoroughbred horses

History of Abstraction in Biology

Biology is… complex, diverse, changes over time

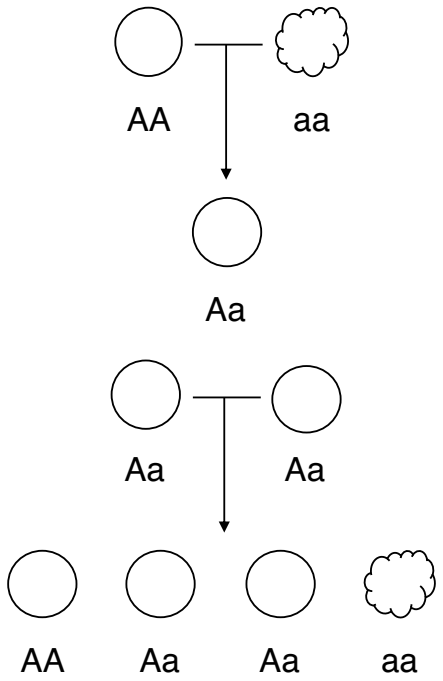
Evolution: “Tree of Life”



Branching process

“Difference” increases with time

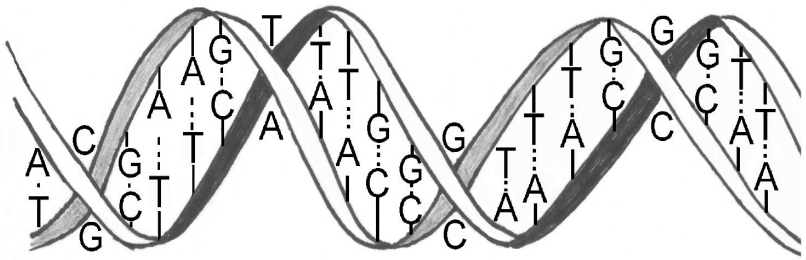
Mendelian Genetics



Traits as symbols (AA, Aa, aa)

Change of frequency as
a probabilistic process

DNA

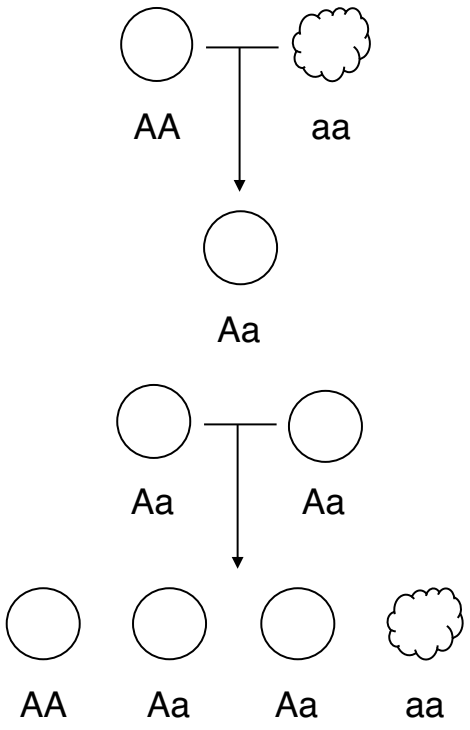


Forms of life as digital information,
“sequence” of letters

Basic concepts of genetics

- Parent and offspring are similar – genetic information is passed on

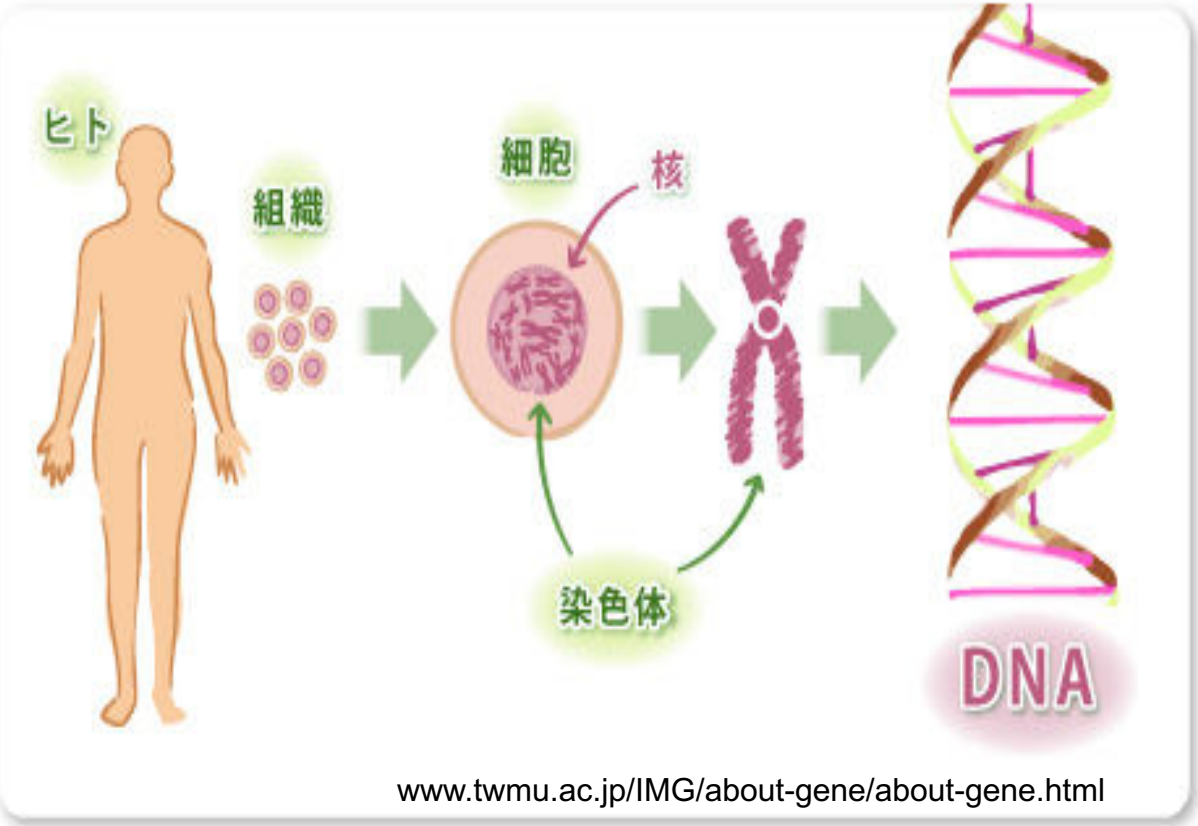
  What is passed on and how?



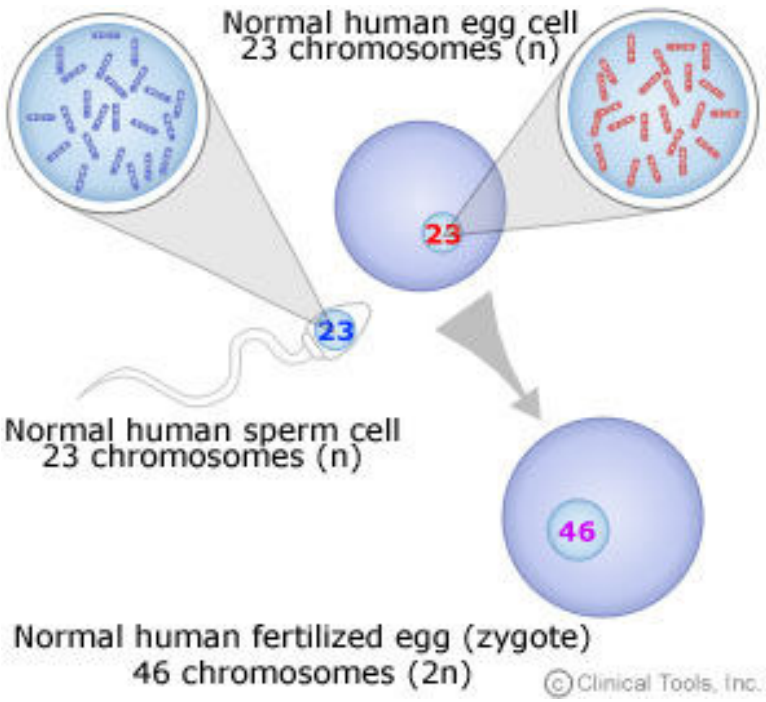
- Information of *A* and *a* is transmitted across generations without changing

- The "phenotype" (appearance) of *AA* and *Aa* are the same

Mechanism of Heredity



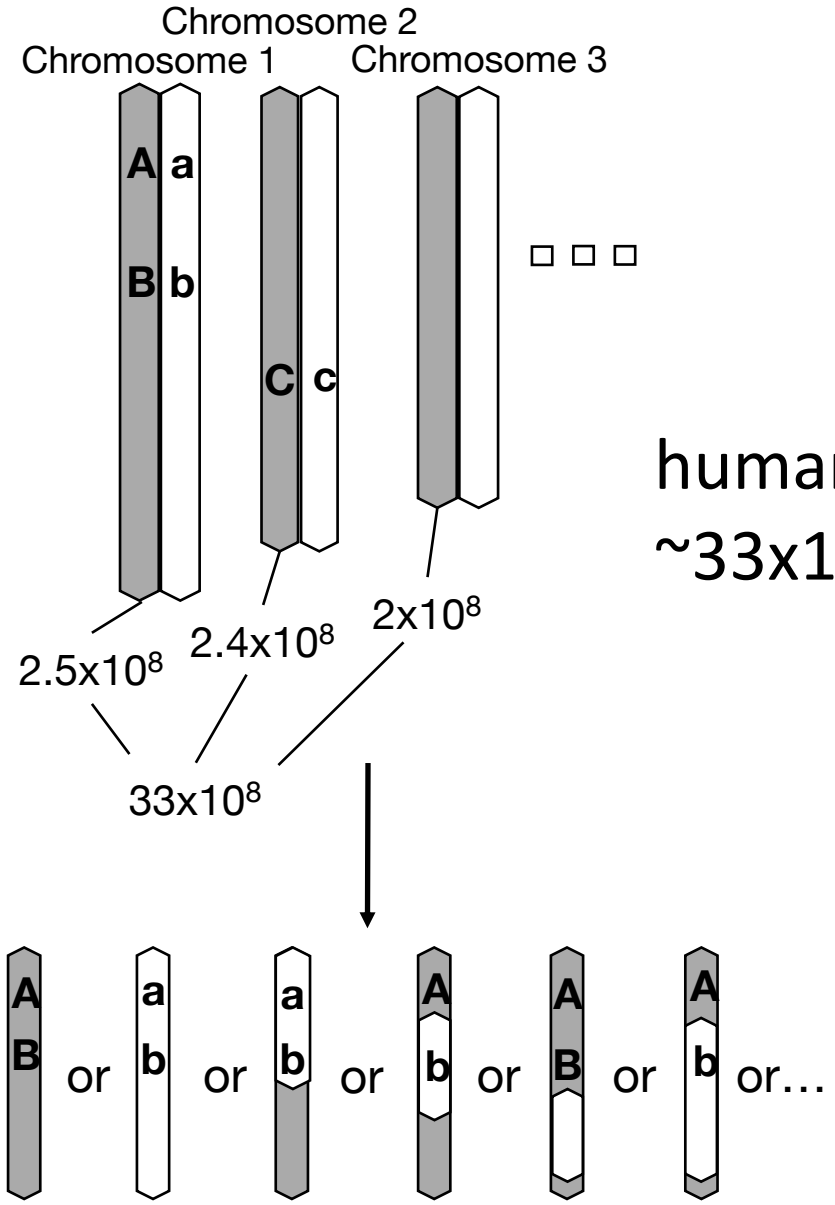
www.twmu.ac.jp/IMG/about-gene/about-gene.html



Genome: entire set of genetic information

Every human has 2 genomes (1 from each parent)

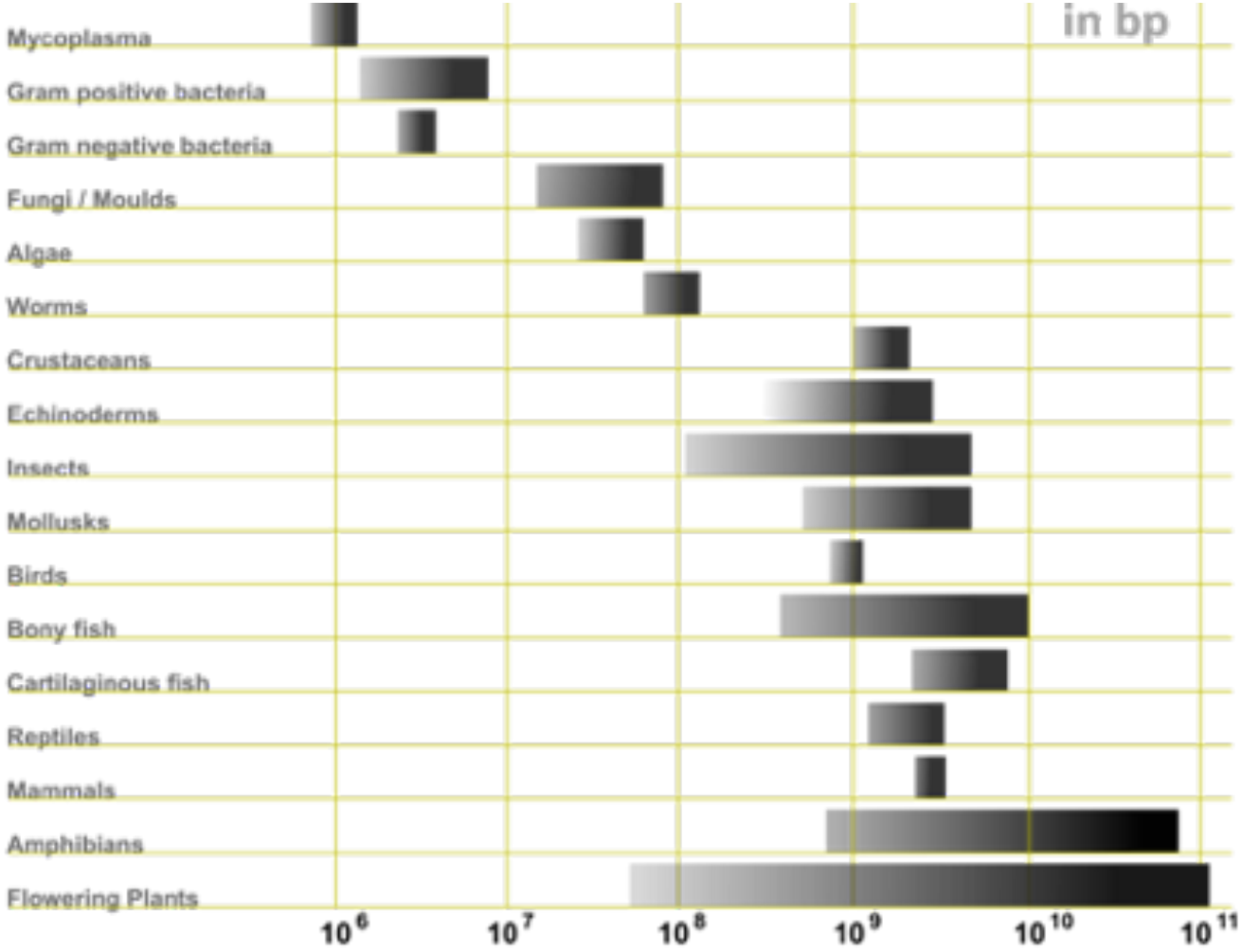
Mechanism of Heredity



human: 23x2 chromosomes (22x2 + XX or XY),
~33×10^8 (3 billion) (x2) nucleotides

Large variation in genome size

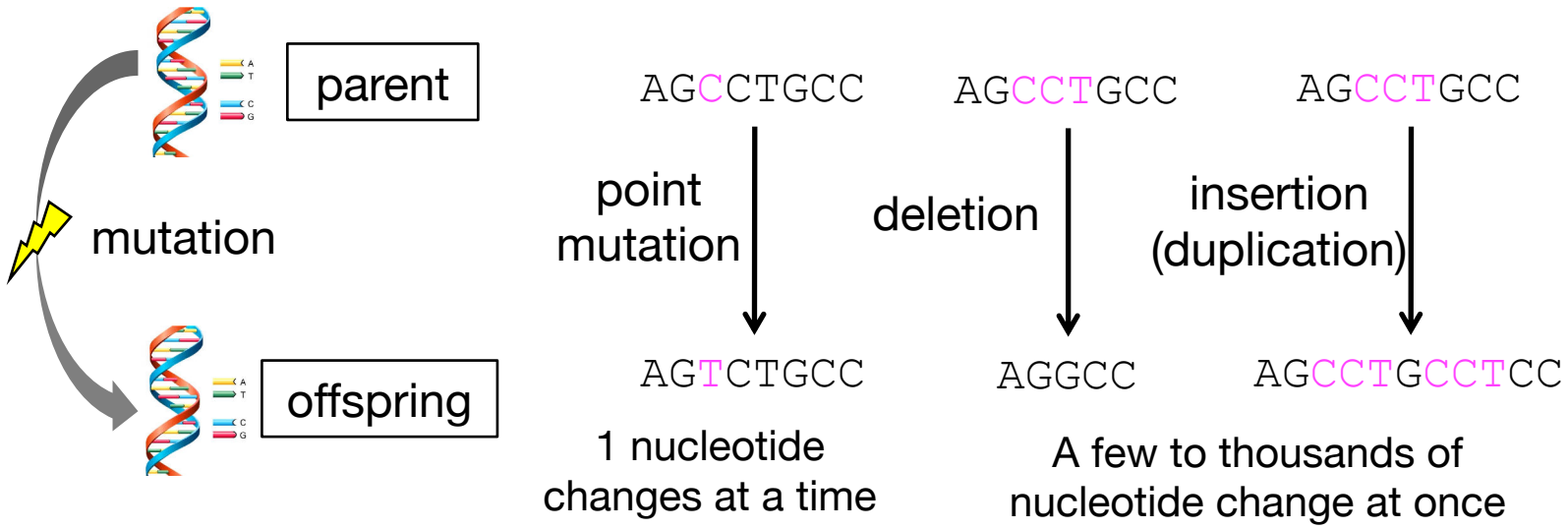
Species	#Nucleotides ($x10^8$)
Budding yeast	0.121
Fruit fly	1.75
Fugu	3.9
Rice	4.5
Maize	25
Mice	27
Human	33
Onion	150
Grasshopper	650
Lungfish	1300
Canopy plant (キヌガサソウ)	1490



en.wikipedia.org

Mutation

"Replication" cannot generate diversity, genetic information changes

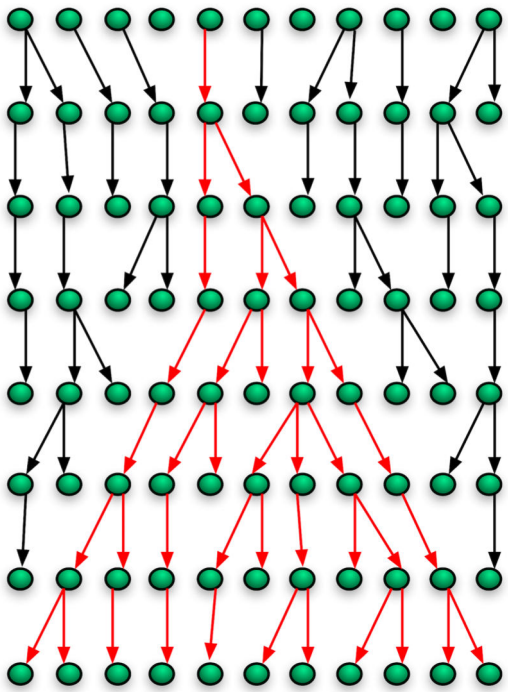


~100 mutations per generation in humans

~$1/10^8$ point mutations per nucleotide per generation

Evolution

- Heredity
- Mutation => some of them change the “phenotype”
- “Population process” (competition, “struggle for survival”)



evolutionnews.org

Many individuals do not produce any offspring

We can’t observe mutations in those individuals

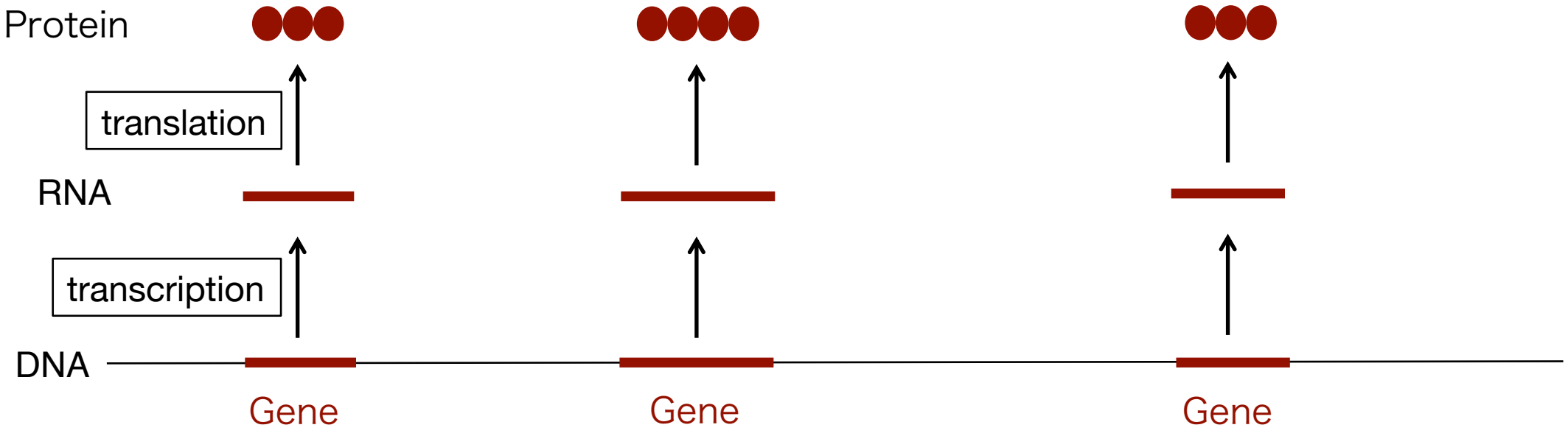
Do all individuals have equal chance to produce offspring?

Did the mutation change the chance to produce offspring?

Outline

- ❑ Basic concepts of genetics and evolution
- ❑ What is written in the DNA?
- ❑ How can we know what's written in the DNA?
- ❑ How can we associate "genotype" with "phenotype"?
  - Research on Thoroughbred horses

What is written in the genome?



All cells contain the same set of genes

The amount and timing of RNA/proteins produced differ

(very complicated process)

DNA -> RNA (transcription)

A (Adenine) -> A (Adenine)

G (Guanine) -> G (Guanine)

C (Cytosine) -> C (Cytosine)

T (Thymine) -> U (Uracil)

RNA -> Amino Acid (translation)

- 20 kinds of amino acids

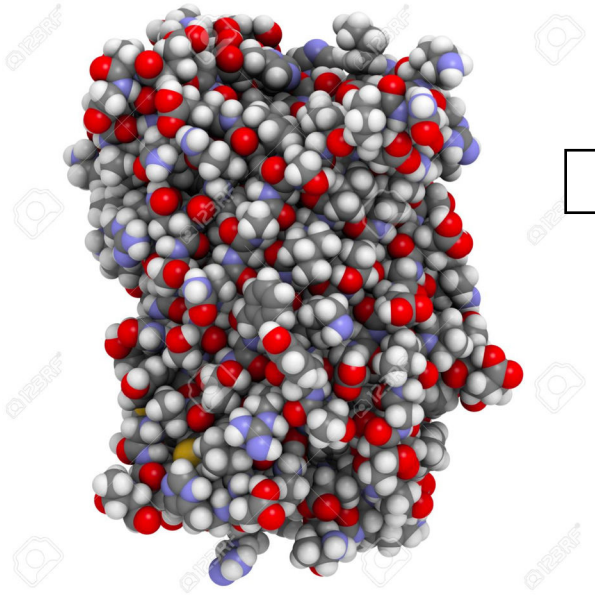
Second Letter

1st letter	U		C		A		G		3rd letter
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG	Stop	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lys	AGA	Arg	A
	AUG	Met	ACG		AAG		AGG		G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu	GGA		A
	GUG		GCG		GAG		GGG		G

redundancy in genetic code

Proteins as amino acid sequences

Protein



www.123rf.com

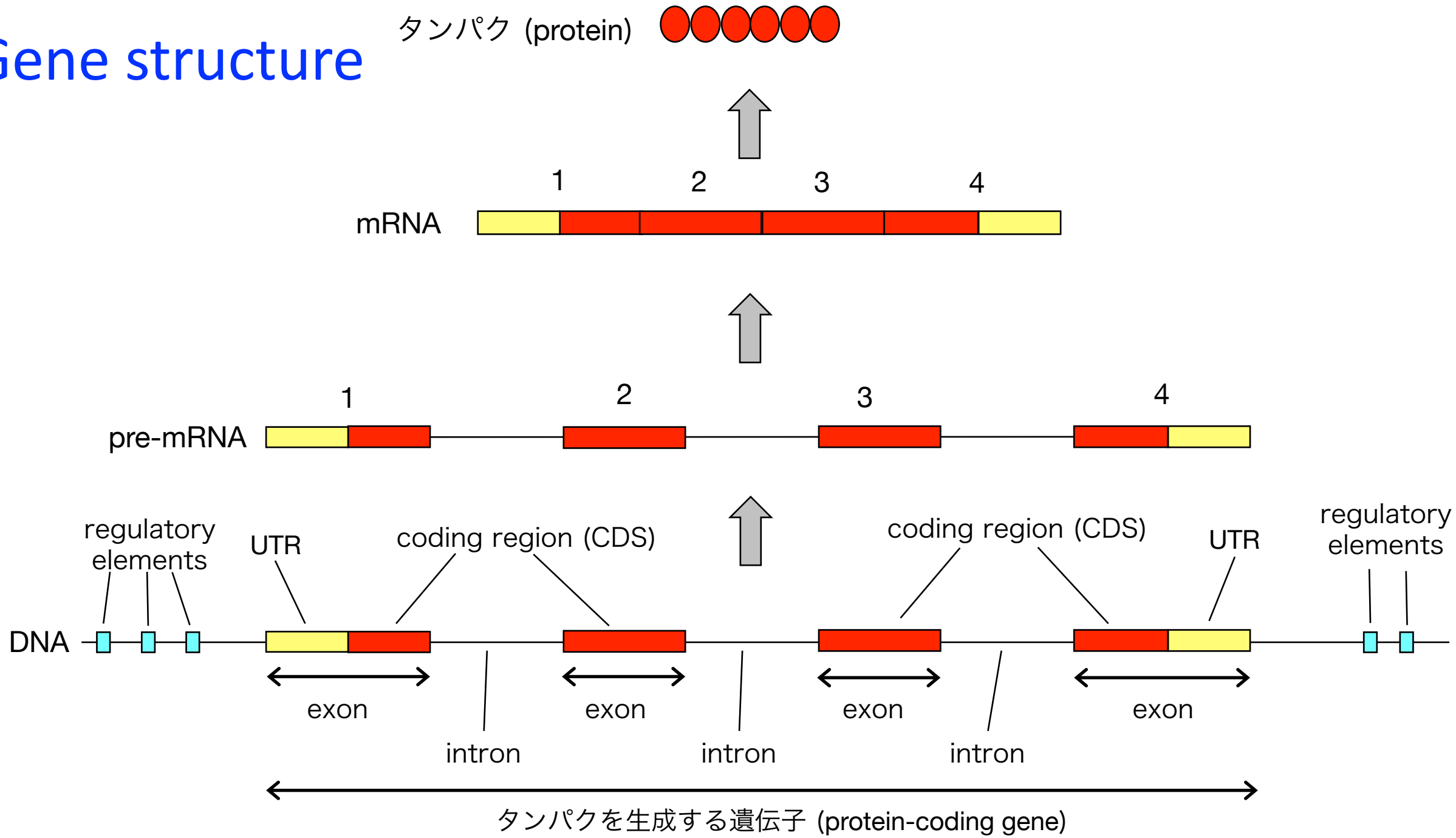
Amino Acid sequence of Histone H4 gene

```
Human          YEETRGVLKVFLENVIRDAVTYTEHAKRKTVTAMDVVYAL
Fruitfly       YEETRGVLKVFLENVIRDAVTYTEHAKRKTVTALDVVYAL
Tomato         YEETRGVLKIFLENVIRDSVTYTEHARRKTVTAMDVVYAL
Volvox         YEETRTVLKNFLENVIRDSVTYTEHARRKTVTAMDVVYAL
Budding Yeast  YEEVRAVLKSFLESVIRDSVTYTEHAKRKTVTSLDVVYAL
```

- The “same” genes are similar across species
- Similar sequence -> similar function

➢ Many genes share high similarity across distantly related species

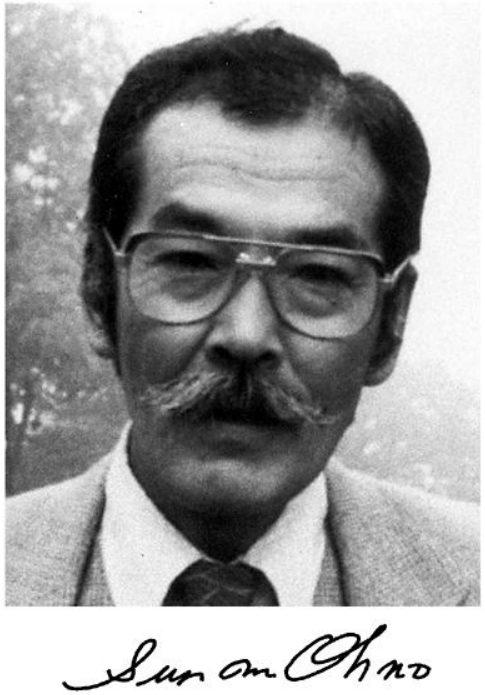
Gene structure

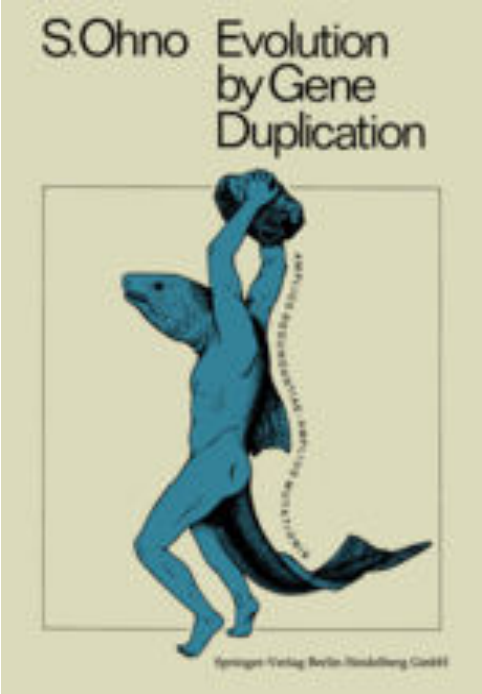


Number of genes are not so different across species

Species	#Nucleotides ($x10^8$)	#Genes
Budding yeast	0.121	6,000
Fruit fly	1.75	17,000
Fugu	3.9	28,000
Rice	4.5	40,000
Maize	25	32,000
Mice	27	23,000
Human	33	21,000
Onion	150	?
Grasshopper	650	?
Lungfish	1300	?
Canopy plant （キヌガサソウ）	1490	?

Evolution by Gene Duplication





- most genes were created by duplication of another existing gene
- mutation can create a new function while keeping the original function

What is written in the genome?

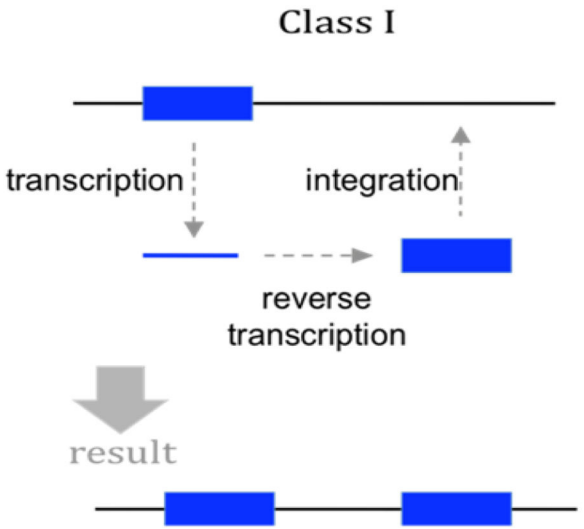
- Human genome

  protein-coding sequences: 1-2%

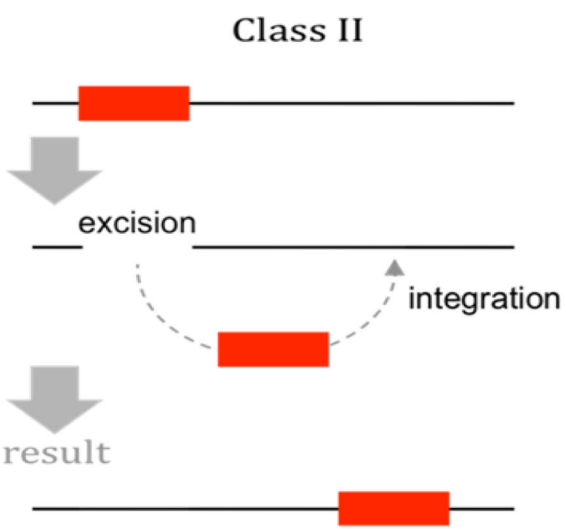
  regulatory sequences: ~5-10%

  Transposable Elements: >60%

Retrotransposon



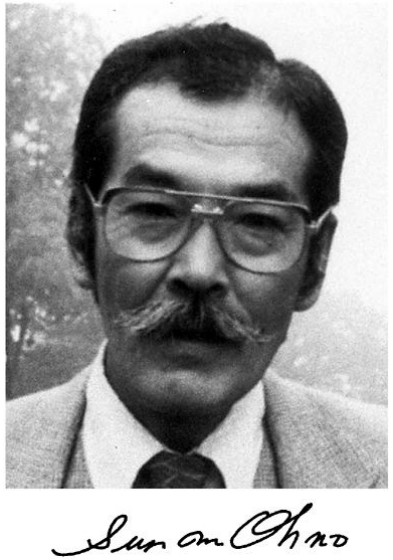
DNA transposon



http://cubocube.com

- major reason for genome size variation
- virus-like parasites that amplify
- have their own “genes”
- often harmful, sometimes beneficial

A lot of unnecessary information in the genome!?



"So much *Junk DNA* in our Genome"
(Susumu Ohno, 1972)

We all acquire ~100 new mutations but most of us are fine (i.e. most mutations are harmless)

"Junk DNA" should be removed during evolution!?!?

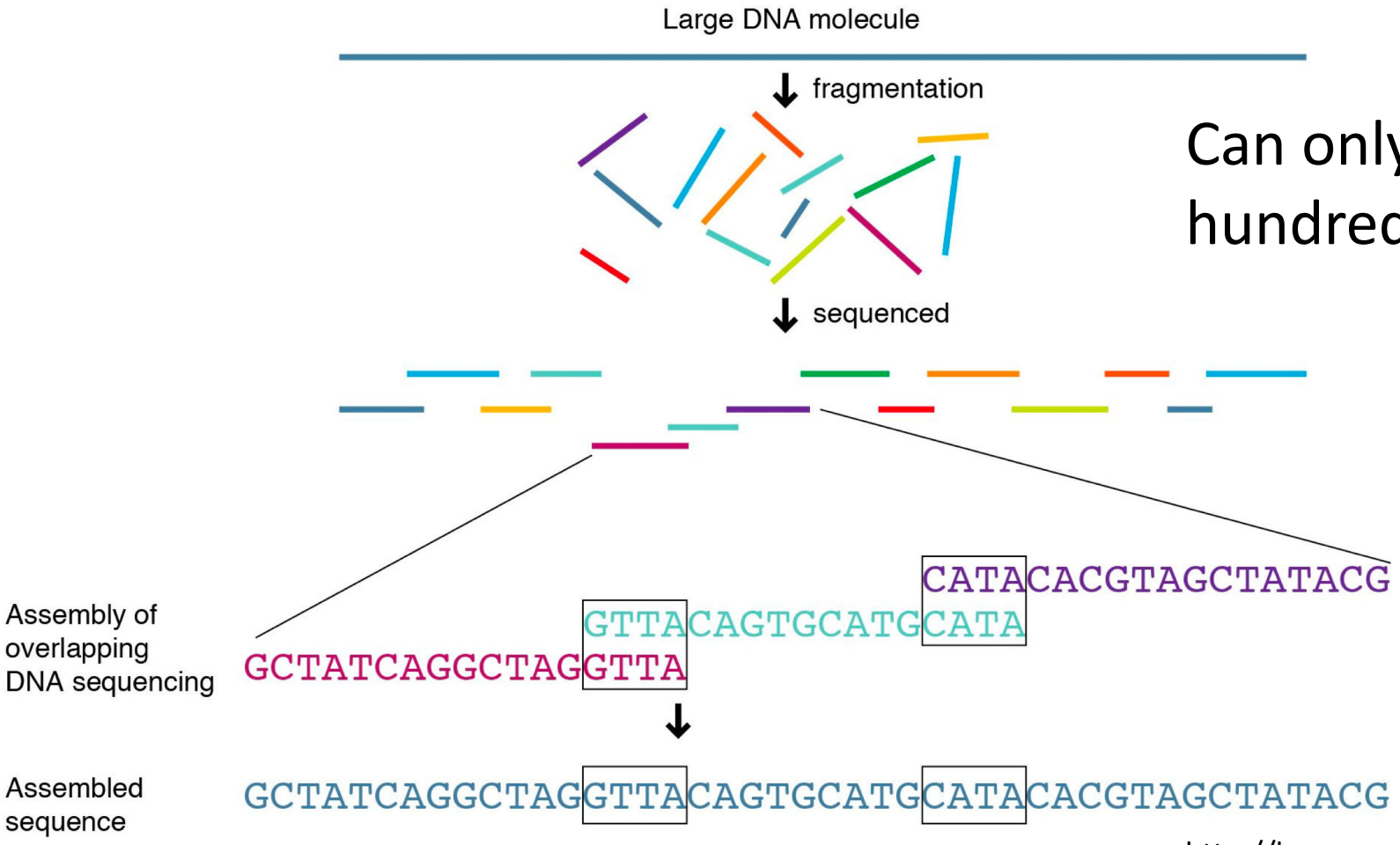
- each "junk" is removed but lots of "junk" are being generated
- they are parasites that try to survive themselves

Outline

- ❑ Basic concepts of genetics and evolution
- ❑ What is written in the DNA?
- ❑ How can we know what's written in the DNA?
- ❑ How can we associate "genotype" with "phenotype"?
  - Research on Thoroughbred horses

How do we know the sequence of the genome?

- Genome sequencing and assembly

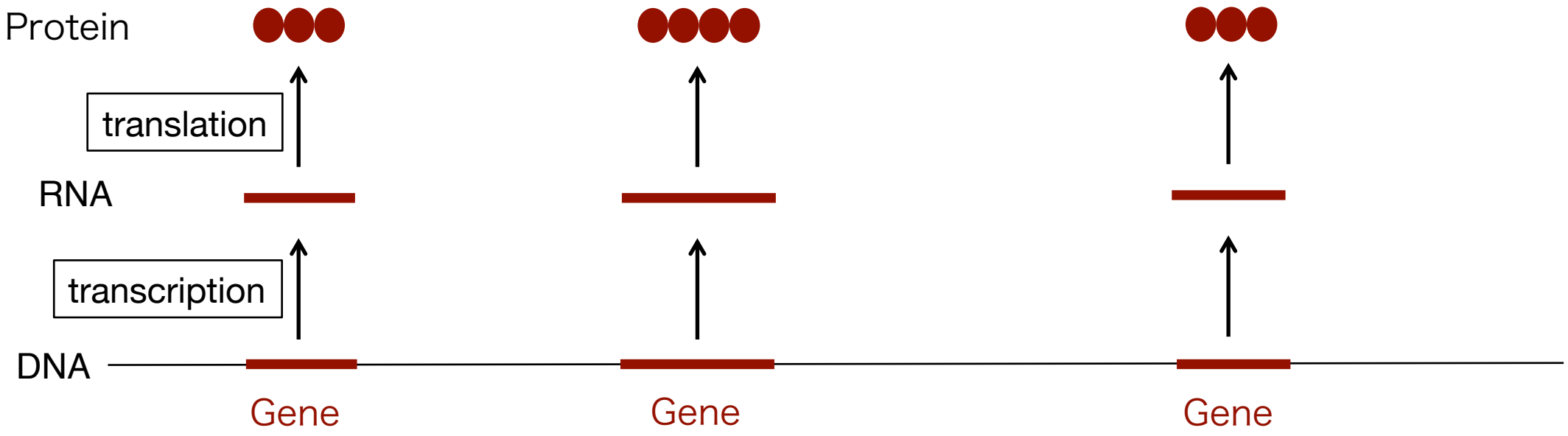


Can only sequence a few hundred nucleotides at once

Difficult because of many similar sequences

http://knowgenetics.org

How can we “find” the genes?



Extract RNA, sequence, and “map” them to the genome

Not so simple because of many similar genes and the presence of introns

How can we “find” the genes?

- Histone H4 gene

```
Human         MSGRGKGGKGLGKGGAKRHRKVLRDNIQGITKPAIRRLARRGGVKRISGLIYEETRGVLKVFLENVIRDAVTYTEHAKRKTVTAMDVVYALKRQGRTLYGFGG
Fruitfly      MTGRGKGGKGLGKGGAKRHRKVLRDNIQGITKPAIRRLARRGGVKRISGLIYEETRGVLKVFLENVIRDAVTYTEHAKRKTVTALDVVYALKRQGRTLYGFGG
Tomato        MSGRGKGGKGLGKGGAKRHRKVLRDNIQGITKPAIRRLARRGGVKRISGLIYEETRGVLKIFLENVIRDSVTYTEHARRKTVTAMDVVYALKRQGRTLYGFGG
Volvox        MSGRGKGGKGLGKGGAKRHRKVLRDNIQGITKPAIRRLARRGGVKRISGLIYEETRTVLKNFLENVIRDSVTYTEHARRKTVTAMDVVYALKRQGRTLYGFGG
Budding Yeast MSGRGKGGKGLGKGGAKRHRKILRDNIQGITKPAIRRLARRGGVKRISGLIYEEVRAVLKSFLESVIRDSVTYTEHAKRKTVTSLDVVYALKRQGRTLYGFGG
```

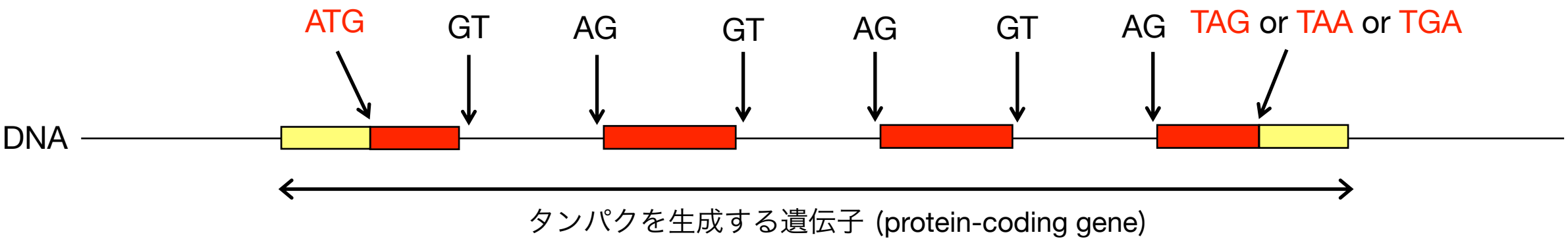
Search for similar sequences in the genome

Most genes in 1 species have similar genes in other species

Similar genes are likely to have similar functions

How can we “find” the genes?

Predict genes based on their features



Nucleotide composition is statistically different

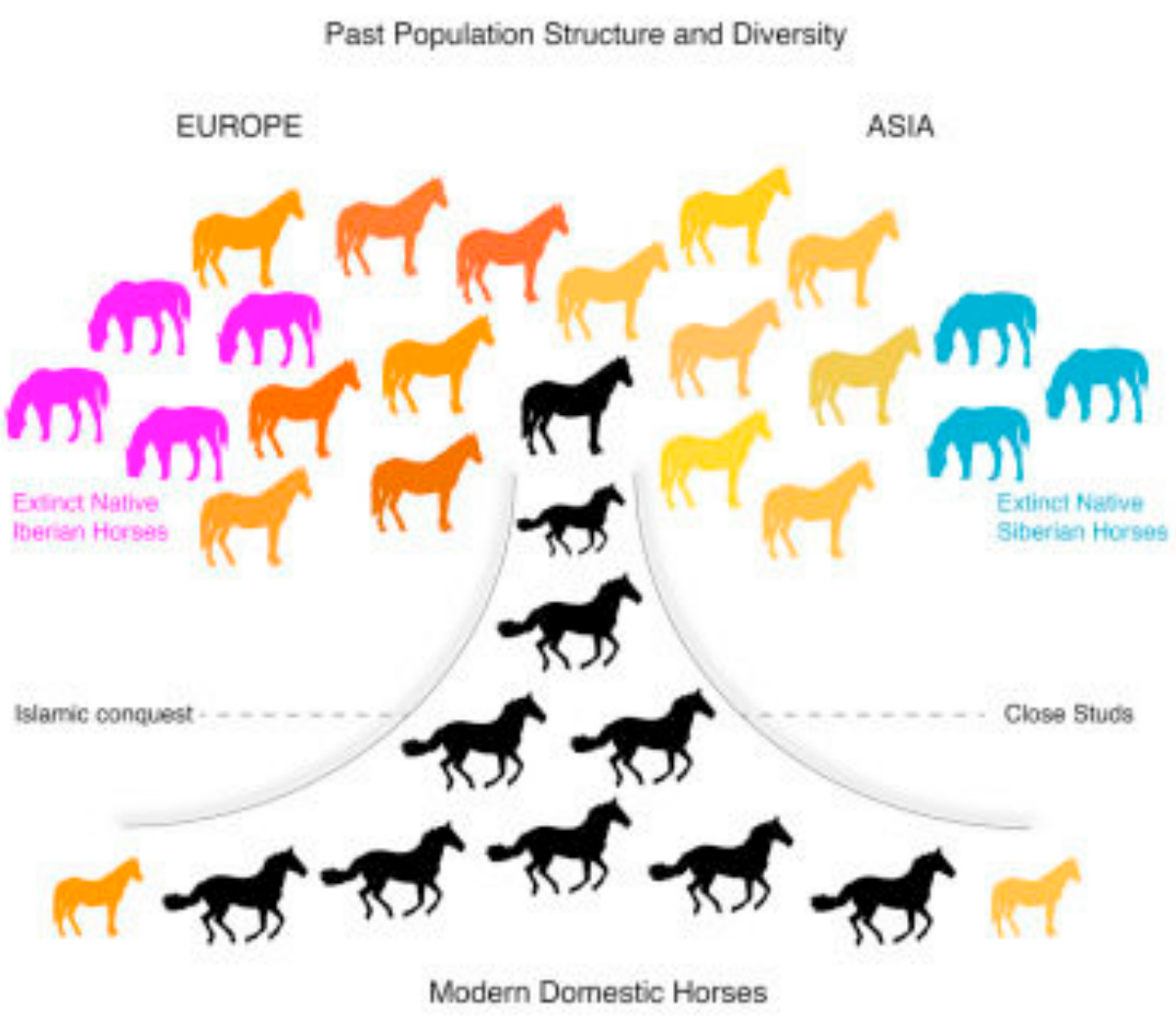
protein-coding sequence vs non-coding sequences, introns

GT-AG of intron vs GT-AG not associated with introns

Outline

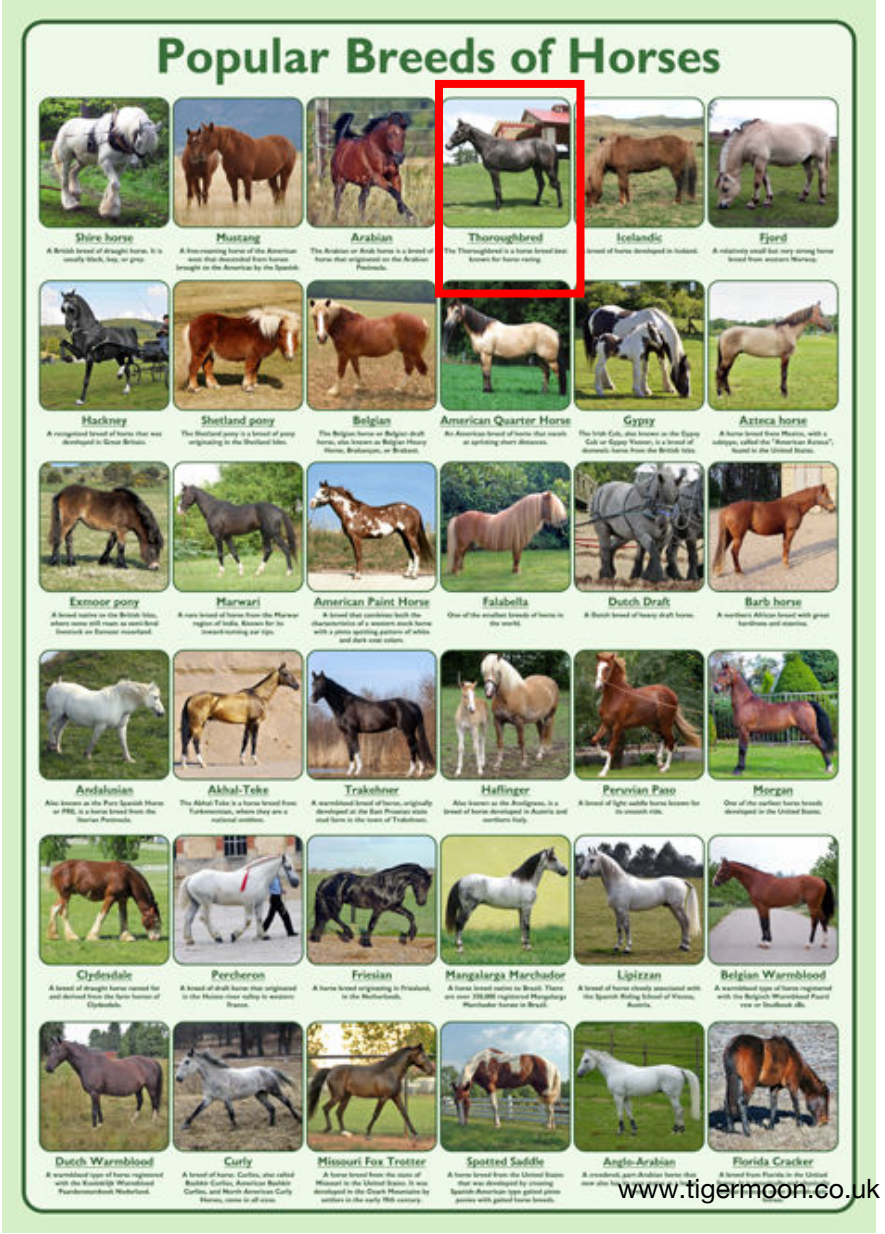
- ❑ Basic concepts of genetics and evolution
- ❑ What is written in the DNA?
- ❑ How can we know what's written in the DNA?
- ❑ How can we associate "genotype" with "phenotype"?
  - Research on Thoroughbred horses

Domestication of Horses (~5500 yrs ago)



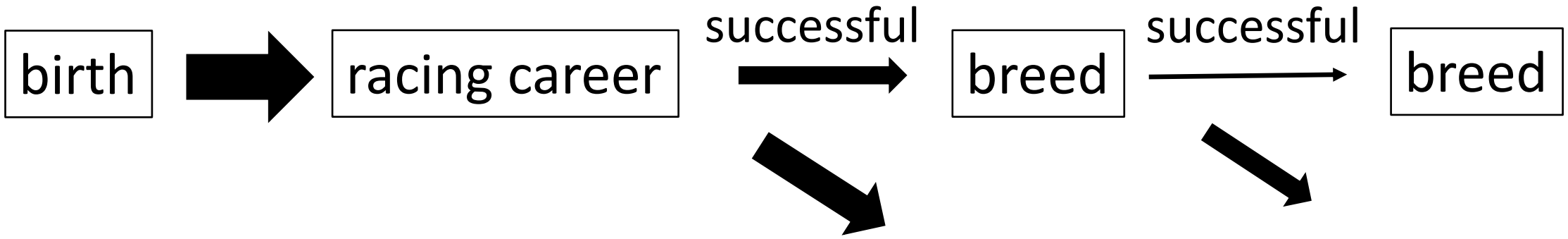
Fages et al. Cell 2019

Many diverse breeds established after domestication



- Thoroughbreds
  - Originated in 18th Century
    (3 “founder” stallions)
  - Horses that win races are selected to breed in many different countries

Selective breeding of Thoroughbreds in Japan



- ~7000 horses are born and registered at JRA (Japan Racing Association)
- Only 10-20 males per generation (>50% of females) are selected for breeding

➢ Thoroughbreds are much faster than other horse breeds due to selective breeding for 20-30 generations

=> but, genetic information has not been utilized yet

Many differences between wild and cultivated Buckwheat

Fagopyrum esculentum ssp. *ancestrale*

Domestication

Loss of seed shattering

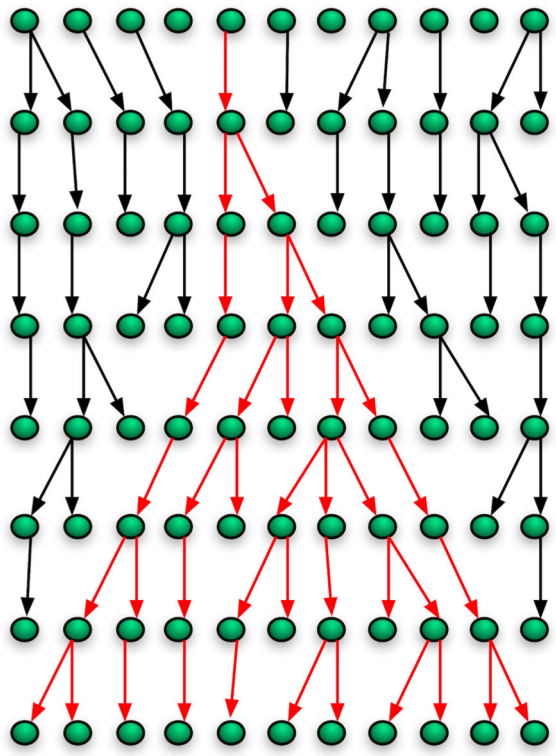
Loss of seed dormancy

Erect growth

Larger seeds

Fagopyrum esculentum ssp. *esculentum*

Process of domestication

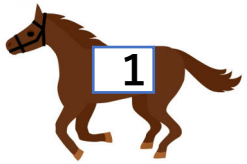


evolutionnews.org

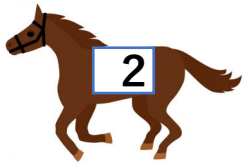
- Mutations occur that result in a desirable trait
  (might already exist in wild, or occur afterwards)
- Humans preferentially select and breed those individuals
- Can we identify the mutations selected by humans?
  (should speed up the selective breeding)
- Can we identify other useful mutations/genes that could be useful for breeding?

Variation in the DNA of each individual

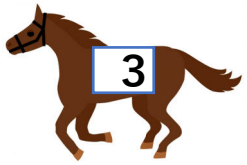
Single Nucleotide Polymorphism (SNP)



ATGTACGTGAGTCAGTACGATGCAGATCTATGAAGACATCCGA...
ATGTACGTGAGTCAGTACGATGCAGATGTATGAAGAAATCCGA...



ATGTACGTGAGTCAGGACGATGCAGATGTATGAAGACATCCGA...
ATGTACGCGAGTCAGGACGATGCAGATGTATGAAGACATCCGA...

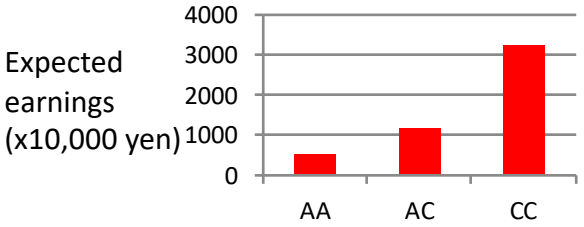


ATGTACGCGAGTCAGTACGATGCAGATGTATGAAGAAATCCGA...
ATGTACGCGAGTCAGGACGATGCAGATCTATGAAGAAATCCGA...

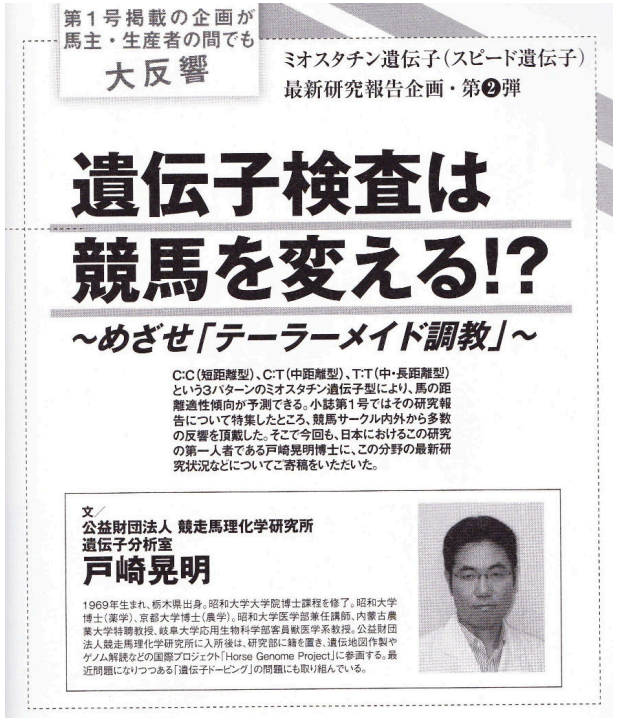
SNP of Myostatin gene

TT: Long distance
CT: Middle distance
CC: Short distance

SNP of Gene X related to racing ability



Example of variation (SNP) associated with racing ability



第1号掲載の企画が馬主・生産者の間でも
大反響

ミオスタチン遺伝子（スピード遺伝子）
最新研究報告企画・第❷弾

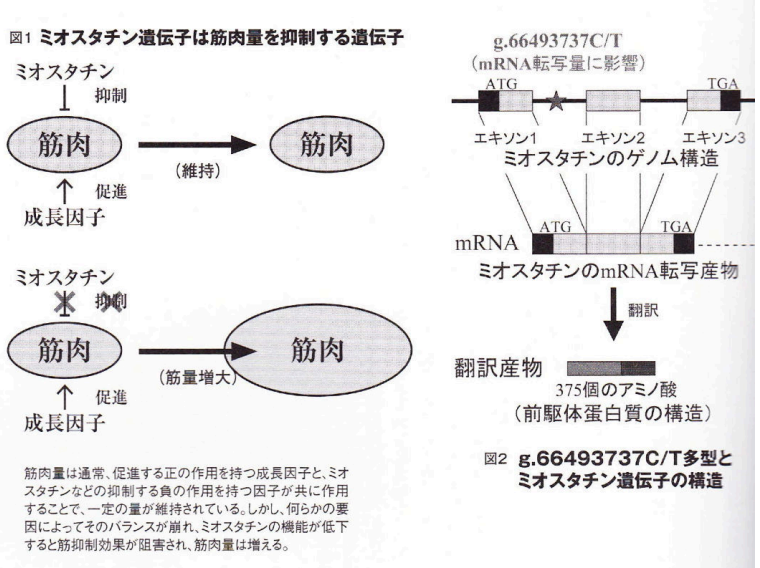
遺伝子検査は競馬を変える!?

～めざせ「テーラーメイド調教」～

C:C（短距離型）、C:T（中距離型）、T:T（中・長距離型）という3パターンのミオスタチン遺伝子型により、馬の距離適性傾向が予測できる。小誌第1号ではその研究報告について特集したところ、競馬サークル内外から多数の反響を頂戴した。そこで今回も、日本におけるこの研究の第一人者である戸崎晃明博士に、この分野の最新研究状況などについてご寄稿をいただいた。

文／
公益財団法人 競走馬理化学研究所
遺伝子分析室
戸崎晃明

1969年生まれ、栃木県出身。昭和大学大学院博士課程を修了。昭和大学博士（薬学）、京都大学博士（農学）。昭和大学医学部兼任講師、内蒙古農業大学特聘教授、岐阜大学応用生物科学部客員獣医学系教授。公益財団法人競走馬理化学研究所に入所後は、研究部に籍を置き、遺伝地図作製やゲノム解読などの国際プロジェクト「Horse Genome Project」に参画する。最近問題になりつつある「遺伝子ドーピング」の問題にも取り組んでいる。



資料B ミオスタチン遺伝子型別の距離適性傾向

ミオスタチン	筋量傾向	戸崎博士統計	UCD統計
C/C型	やや多い	1000～1800m	1000～1600m
C/T型	普通	1200～2000m	1400～2400m
T/T型	やや少ない	1800m以上	2000m以上

SNP in myostatin gene affects optimum racing distance

出典：サラBLOOD Vol. 1&2

Example of variation (SNP) associated with racing ability

【天皇賞・秋】春連覇のフェノーメノ、遺伝子検査では中距離型だった！

2014年10月28日6時0分　スポーツ報知

注目されるＧ１で、牡馬、牝馬の強豪古馬が

３２００メートルの天皇賞・春を２連覇し
ルのＧ１でもタイトルをつかむことができる
などから、推測する材料を与えてくれるが、
スピード遺伝子検査が、大きなヒントを与え

春秋合わせて区切りの１５０回を迎える伝統の一戦に、“春の盾”を連覇したフェノーメノが、テイエムオペラオー以来史上２頭目の天皇賞３勝目を目指しエントリーした。実績から長距離向きの印象が強いが、ＤＮＡレベルから距離適性を推測する「エクイノム・スピード遺伝子検査」で、「中距離型」との判

【共同通信杯】イスラボニータ完勝“実に強い”ダービー見えた！

イスラボニータ（右）が共同通信杯を制して３連勝で皐月賞に向かう

クラシックの栄冠が見えた。降雪で代替開催（東京６日目）となった「第４８回共同通信杯」が２４日、東京競馬場で行われ、１番人気、蛯名正義（４４）騎乗のイスラボニータが直線で抜け出して完勝。昨年のいちょうＳ、東京スポーツ杯２歳Ｓに続く３連勝で東のクラシック最有力候補に浮上した。今後は皐月賞（４月２０日、中山）へ直行する。【レース結果】

放牧明けで臨んだ３歳始動戦。同師は「ダービーがピークになるように調整しているが、２歳時に比べて首さしがグッとたくましくなった」と成長ぶりに目を細めた。昨年暮れには放牧先の社台ファーム（北海道千歳）から朗報が届けられた。競走馬理化学研究所が実施している遺伝子のＤＮＡ距離鑑定を受けたところ、ダービーの２４００メートルまで守備範囲との評価。吉田代表は「２４００メートルはやってみないと分からないが、少なくとも２０００メートルまでは問題ない。シンから丈夫でキリッとした馬。これは走るぞ」。次走・皐月賞を経てダービー（６月１日、東京）へ。「さあ、これからが本当の勝負だ」。栗田博師は念願のクラシック獲りへ再び顔を紅潮させた。

1200m好走馬の次走が2200m、一見無茶な選択の裏に遺伝子検査/吉田竜作マル秘週報

2014年01月29日(水)18時00分　　注目数：21人

◆科学はサラブレッドの世界をどう変えていくのか

12日の京都芝外2200メートルを舞台にした3歳未勝利戦でのこと。結果は1番人気ディルガ→2番人気アグリッパーバイオというごく無難な決着に終わった一方で、ある画期的な試みが実践に移されていた。

エイシンソルティー(牝)は西園キュウ舎の当世代のトップバッターを務めた馬。デビューは6月の阪神(ダ1200メートル)で0秒2差3着。2戦目(中京芝1200メートル)にはクビ差の2着まで詰め寄り、初勝利は目前と思われたが…。その後も詰めの甘さはなかなか解消せず、前走前の時点での成績は[0-1-3-1]。とはいえ、あと一れまでと同じ短距離戦を選択するところだろう。が、西園調教頭の舞台を選択。実に前走から1000メートルもの距離延長と

刃でも」ってノリでこの手の極端な条件変更をするケースもあらかに違うケース。実は“科学的な根拠”がこの1000メートル延

れているが、これが形となって表れたのが昨年の英国クラシッルジャー調教師が擁するドーンアプローチだ。圧倒的なスピー入ってきたのは英ダービー。しかし研究者でもあるボルジャスピード遺伝子を強く受け継いでおり、英ダービーは適距離と究成果通りなら、愛馬は距離の壁に泣く、何ともシニカルな状ま激しく折り合いを欠き、大惨敗を喫した。

エイシンソルティーも栗東トレセンの診療所で「スピード遺伝が向く」という判定が出た。これを受けたオーナーの指示もあけだ。

Using genetic information for Thoroughbred breeding

- Can evaluate the genetic potential of each horse

  => more informed choice of which horse to keep and which to discard

- Can identify the best breeding (male x female) combination

  => more informed choice of which male and female to mate

Genome-wide SNP analysis of Japanese Thoroughbred racehorses

Jeffrey A. Fawcett[1,2]*, Fumio Sato[3], Takahiro Sakamoto[1], Watal M. Iwasaki[1], Teruaki Tozaki[4], Hideki Innan[1]*

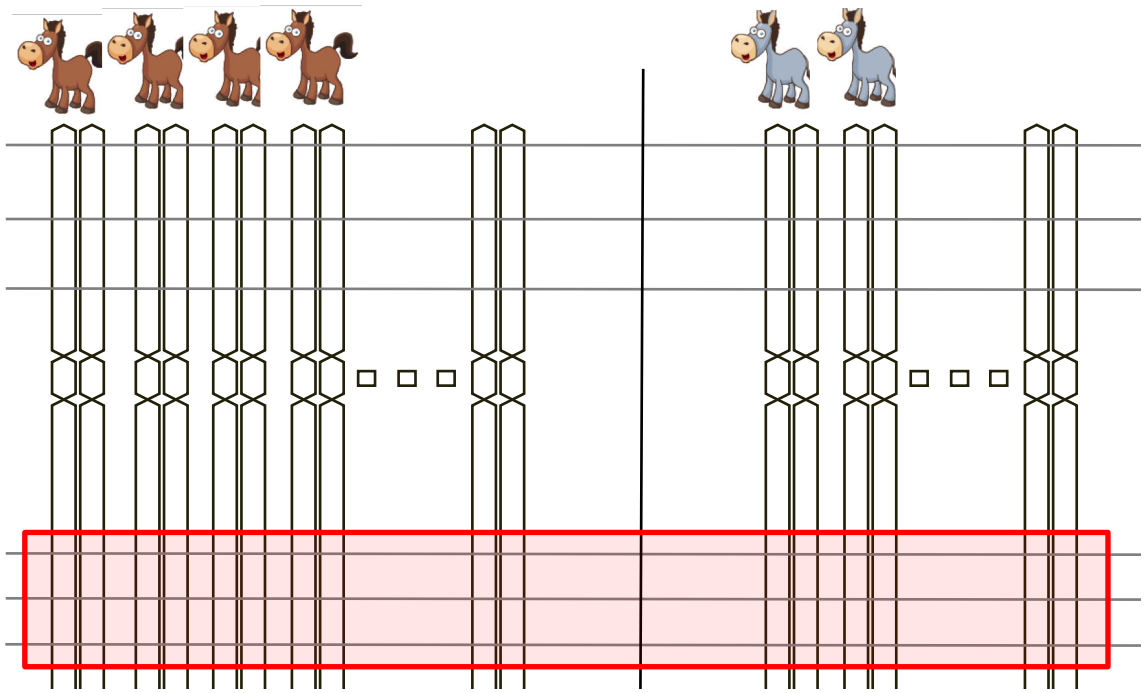
Kit to survey 600,000 SNPs available for horses

- ~400 Thoroughbred horses from JRA (Currently extended to ~1000 samples)
- Collect blood samples, extract DNA, identify SNPs
- Can we identify genetic variation associated with variation in traits?
- Can we identify genetic variation associated with variation in racing ability?

Genome-wide Association Study (GWAS)

Horses with trait A (30)

Horses without trait A (15)



A:40 T:20 A:20 T:10

G:31 C:29 G:14 C:16

T:5 G:55 T:4 G:26

G:30 C:30 G:2 C:28

T:29 A:31 T:1 A:29

A:32 T:28 A:3 T:27

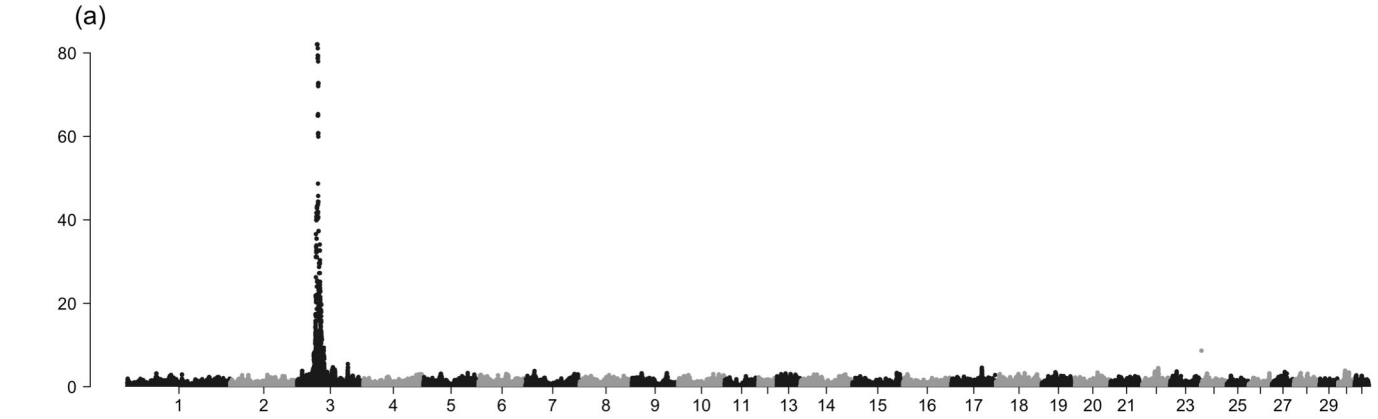
Genetic factor responsible for difference in trait A should be present

Strong association between nucleotide frequency and difference in trait

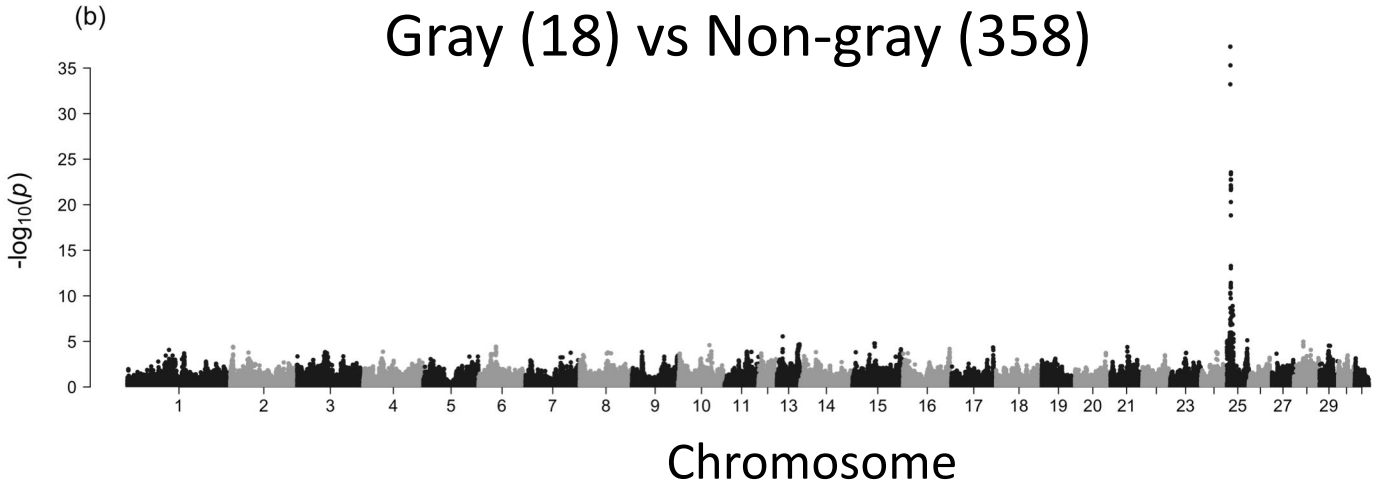
SNPs in neighboring region are “linked”

GWAS works with coat color genes (proof of concept)

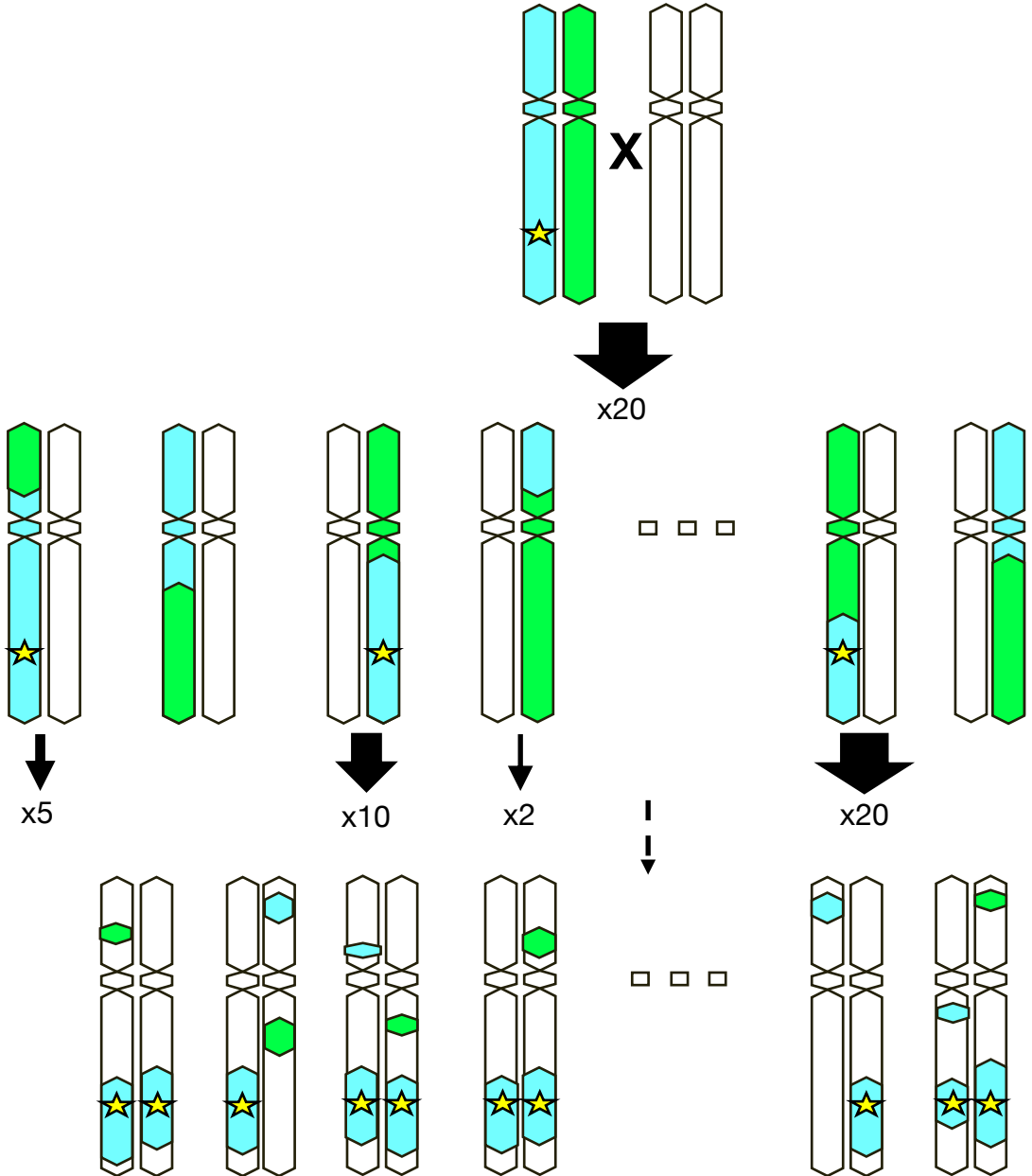
Chestnut (94) vs Non-chestnut (258)



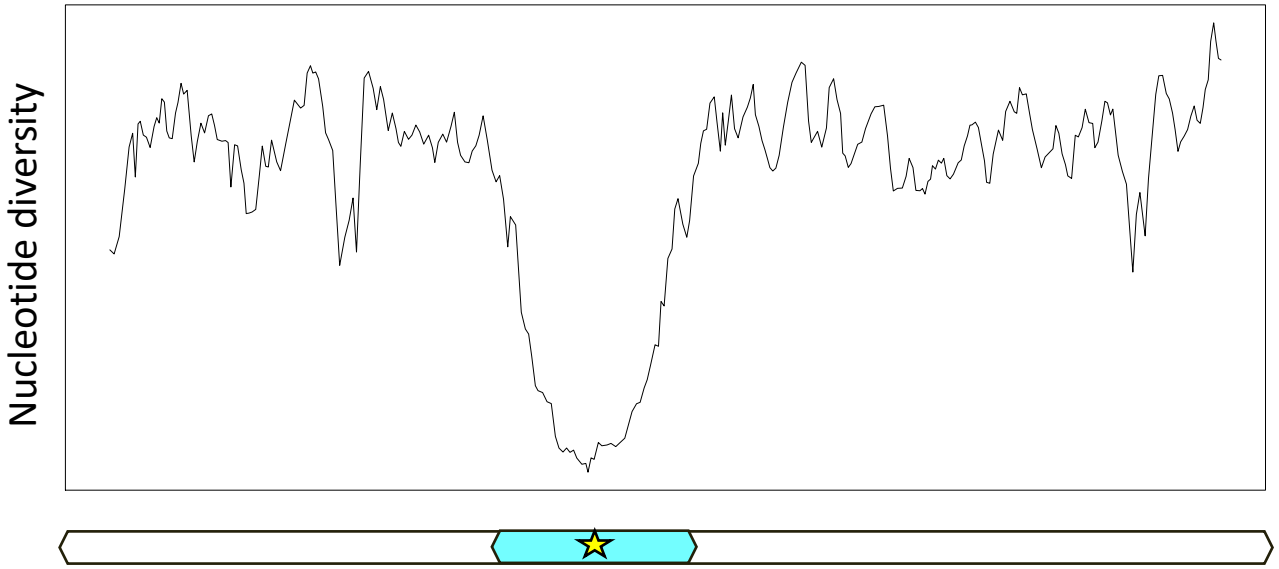
Gray (18) vs Non-gray (358)



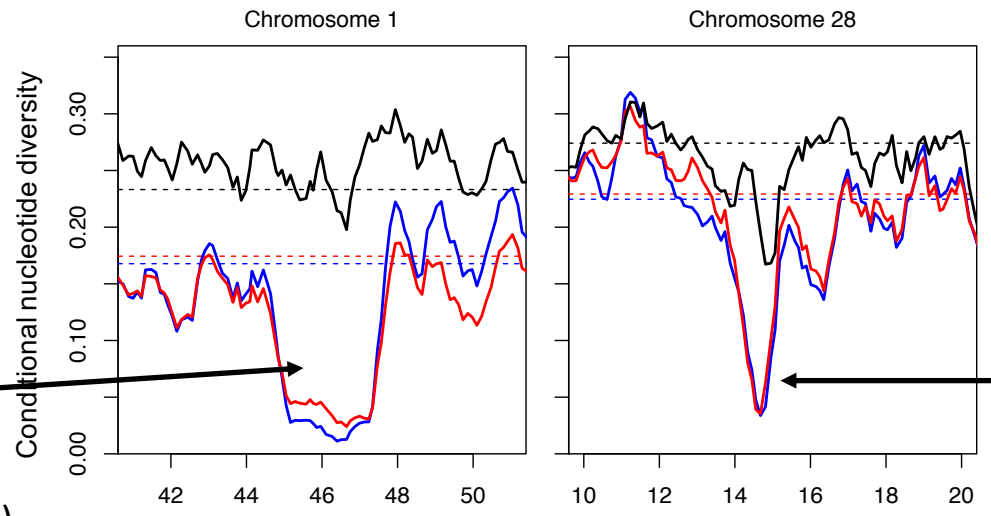
Genomic signatures of selection



- Region containing the mutation has reduced diversity



Regions of reduced diversity regions in Thoroughbreds



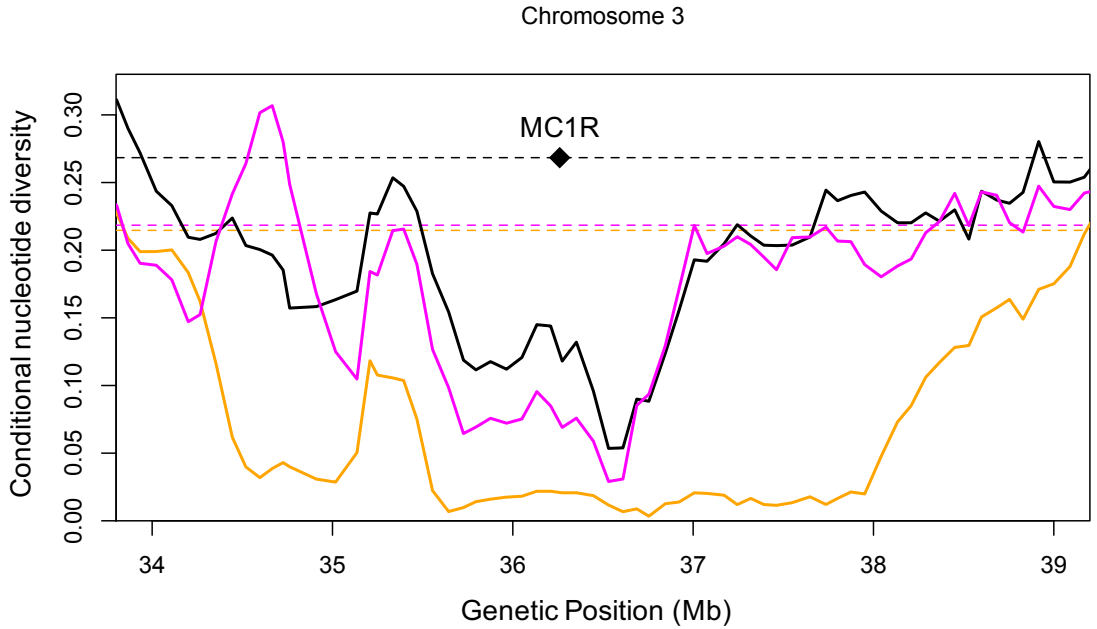
Thoroughbred (This study)
Thoroughbred (Schaefer et al. 2017)
Non-thoroughbred (Schaefer et al. 2017)

PCDH15

(hearing and balance functions)

KITLG

(fertility, neural cell development)



Chestnut Thoroughbred (This study)
Non-chestnut Thoroughbred (This study)
Non-thoroughbred (Schaefer et al. 2017)

Computational approaches to identify important regions

- ❑ Genome-wide association studies

  very powerful when there is a targeted, simple trait

- ❑ Genomic signature of artificial/natural selection

  can use also for complex traits and where there is no specific trait

- Difficult to pinpoint actual causative mutation/gene
- Can identify candidate regions/genes that can be tested experimentally

Summary

- We can explain the process creating the diversity of life in a common mathematical/computational framework
- We can apply the same approaches to study many different species (e.g. horses, buckwheat)
- Computational/mathematical approaches are becoming more important with the increase of data
- More advanced mathematics are needed to go from “sequences” to complex networks, shapes, 3D structures etc?

Acknowledgements

- Thoroughbred project

  Hideki Innan, Takahiro Sakamoto, Watal M Iwasaki (SOKENDAI), Fumio Sato (JRA Hidaka), Teruaki Tozaki (Lab of Racing Chemistry), Takao Suda (Journalist), etc

- Buckwheat project

  Yasuo Yasui (Kyoto U), Chengyun Li (Yunnan Agri U), Takanori Ohsako (Kyoto Pref U), Hideki Hirakawa (Kazusa DNA Res), etc