

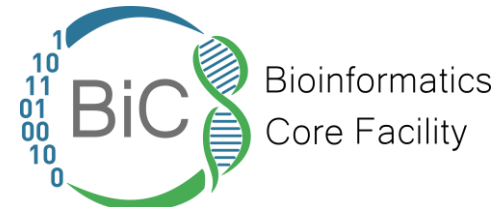
# Warum Menschen keine halben Bananen sind ...

... und sich Äpfel doch mit Birnen vergleichen lassen.

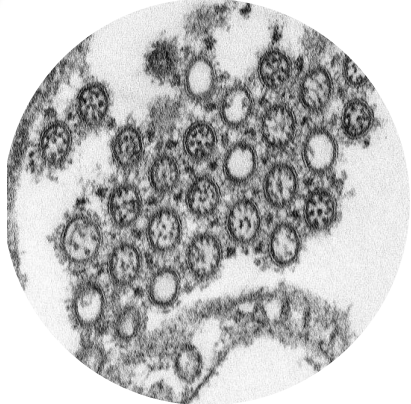
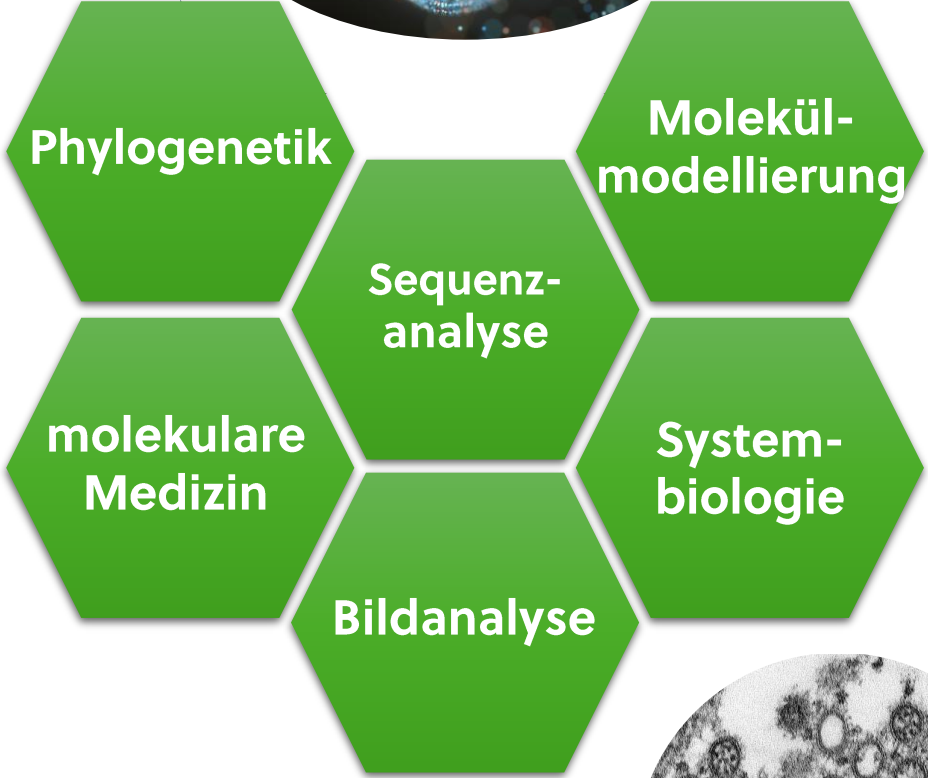
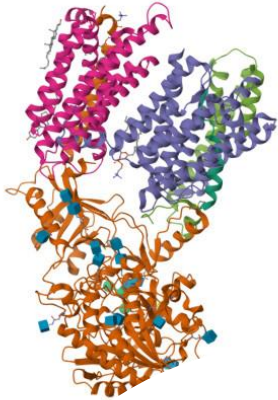
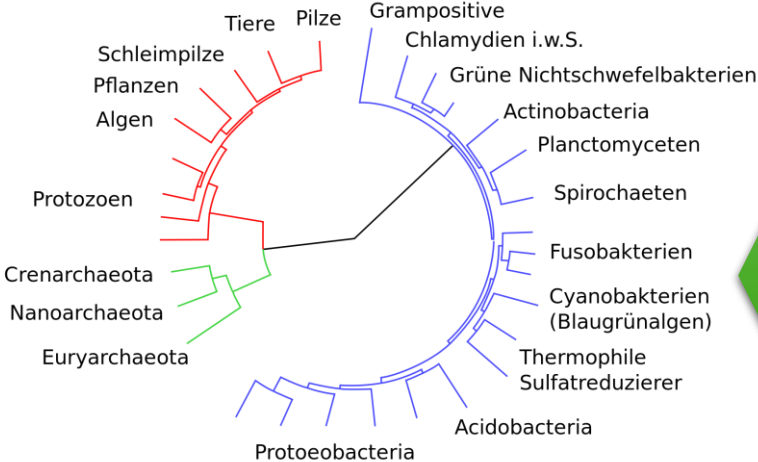
*Dr. rer. nat. Emanuel Barth*



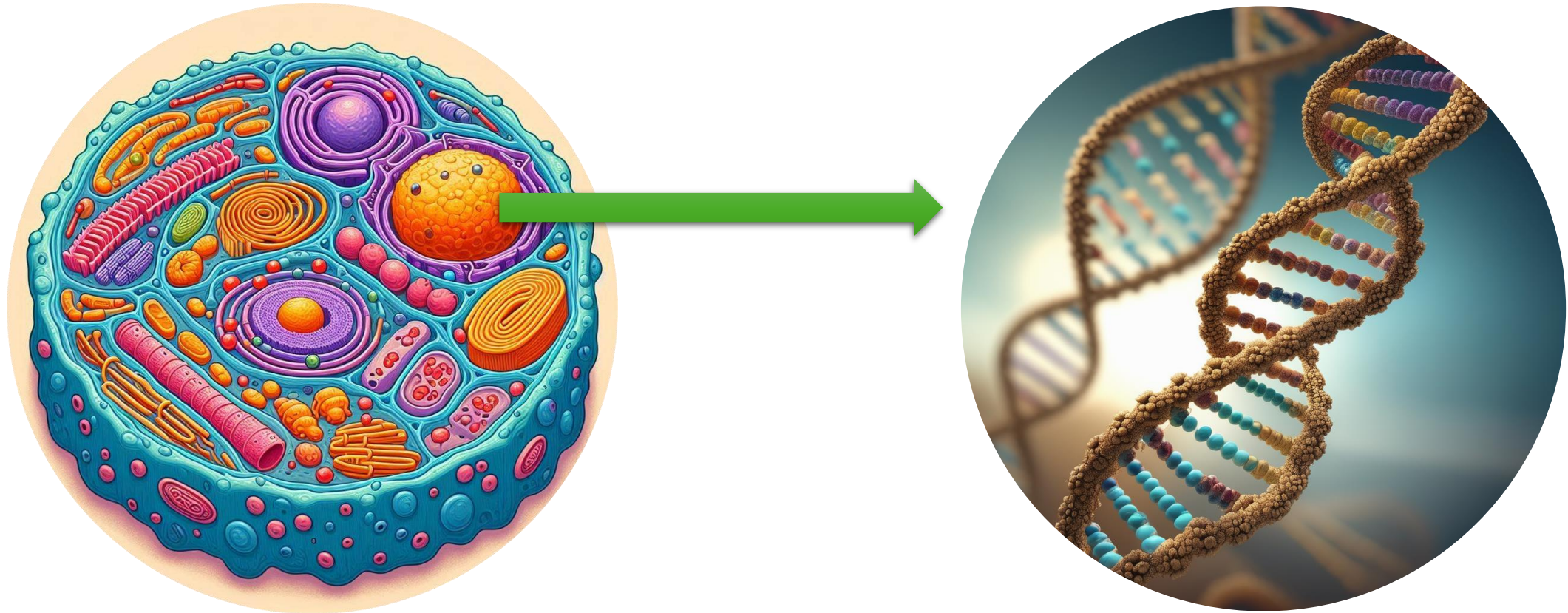
FRIEDRICH-SCHILLER-  
UNIVERSITÄT  
JENA



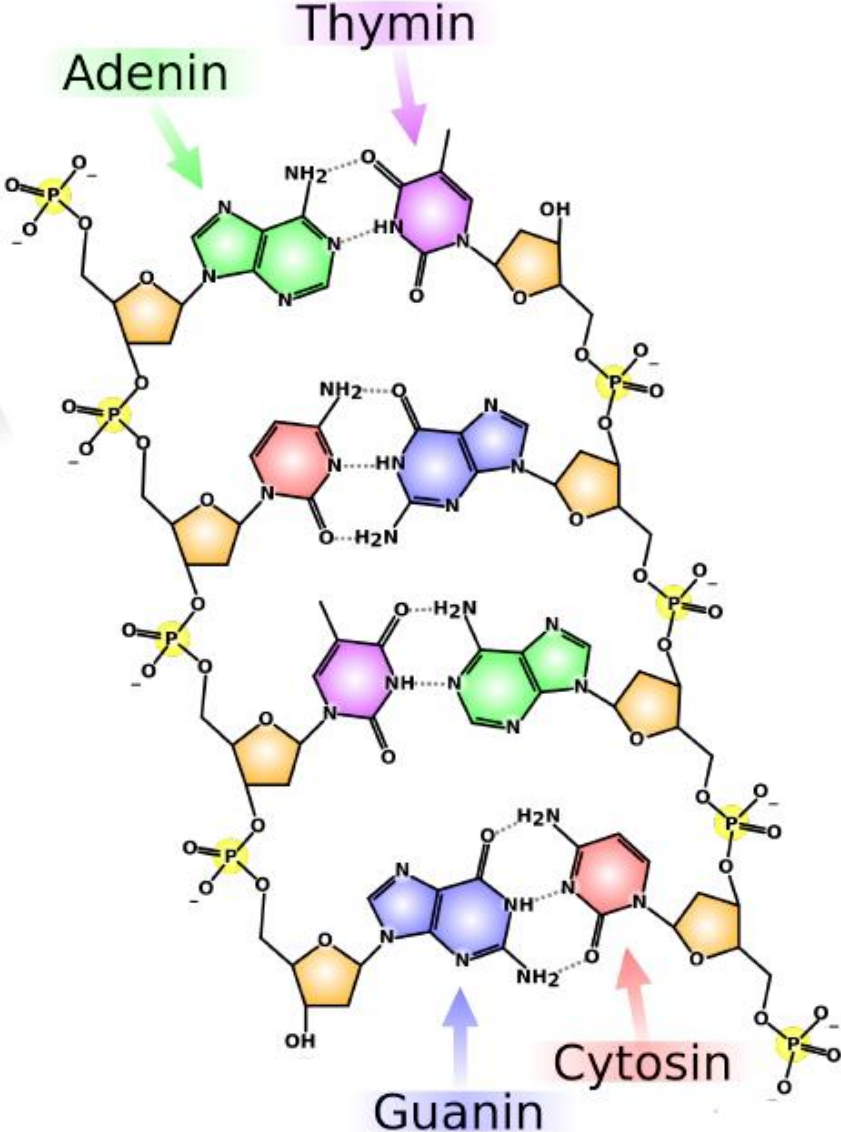
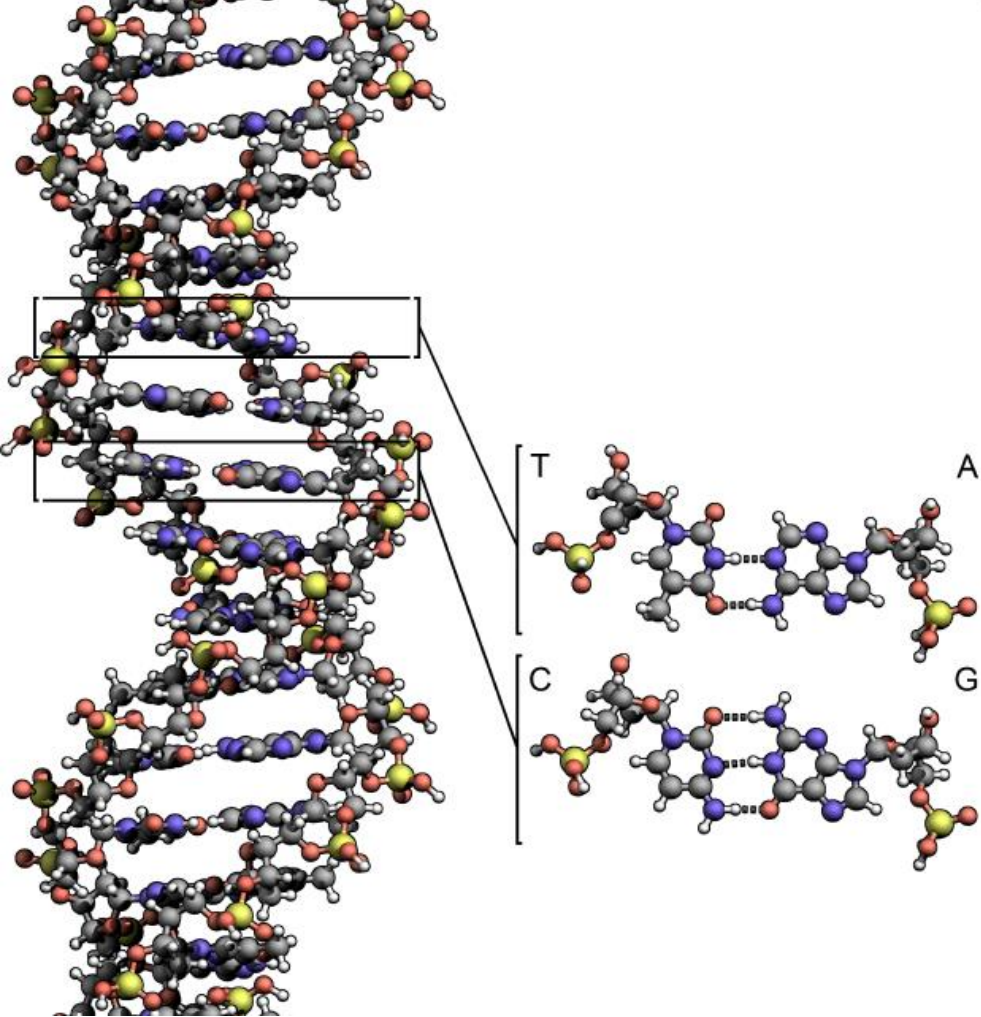
# Bioinformatik – eine Schnittstelle der Wissenschaften



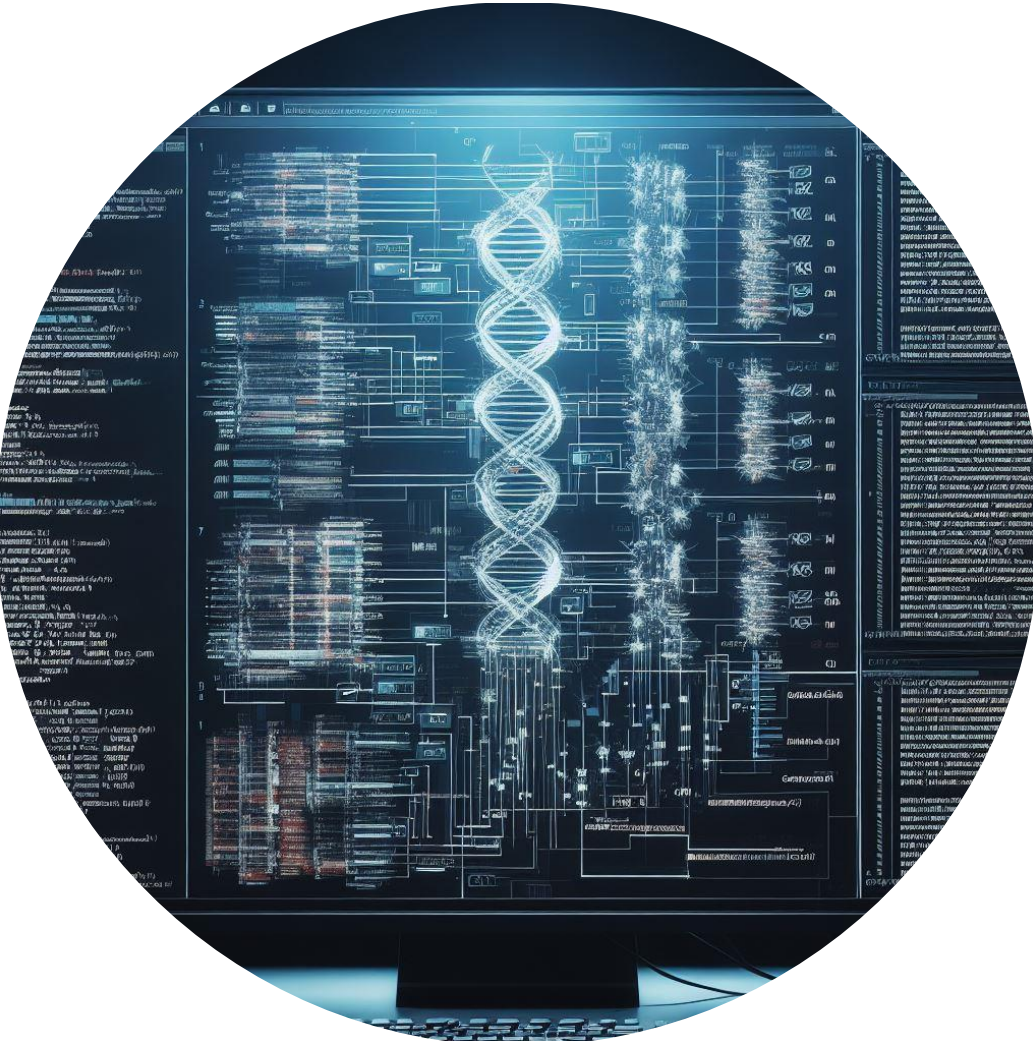
# DNA – Träger der Erbinformation



# DNA – Träger der Erbinformation



# DNA-Sequenzen aus Sicht der Bioinformatik



```

      AACGGATCC
    , ATGGCAATTGCGCTATA
  AATGGCTCCTAGCGCTACATGC
    TGAGTCGCAATGGTCCGAACCGTTA
  GCGGTAAACGGATCCGGGATATAAATGC
    CGTTTATGGCAATTTGAGCAAATGGCT
  GCGCTACATGCTGCATGCACCTGAGTCGCA
    GTCCGAACCGTTACGGCGATGCGGTAACGC
    CGGGATATAAATGCGCTACCGTTTATGG
  GCGCTATAGTACCGTGAGCAAATGGCT
    GCTACATGCTGCATGCACCTGAGTC
  GAACCGTTACGGCGATGTCCCG
    ATGGCAATTGCGCTATA
  TGGCCATT

```

## Schätzfrage:

Aus wie vielen "Buchstaben" besteht das menschliche Genom?

Online**TED**<sup>®</sup> LIVE

Ihre Teilnahmeoptionen (nur einmal erforderlich):

Webseite:  
**edu.onlinede.de**

Freischaltcode:  
**6048**

Teilnehmer: **0** / **100**

... oder scannen Sie den QR Code:

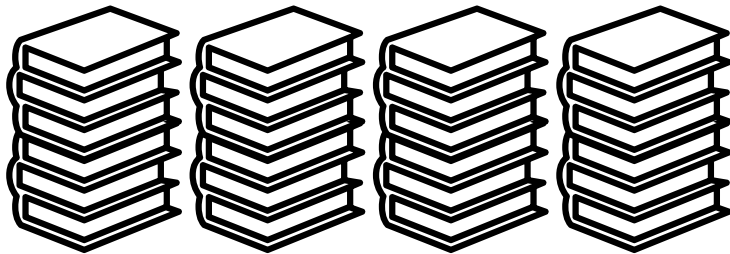
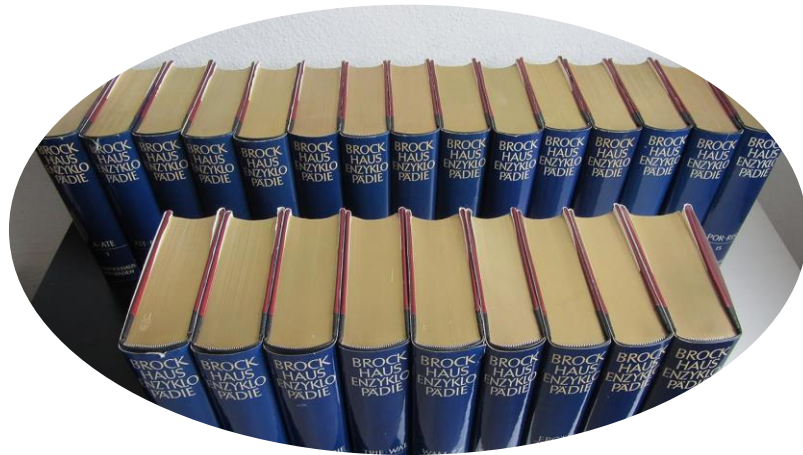


# Die Größe unseres Genoms

Brockhaus Lexikon (21. Auflage, 24 Bände)

~ 175 Millionen Buchstaben ( $175 * 10^6$ )

*167 Megabyte*



menschliches Genom (in einer Körperzelle)

~ 3 Milliarden Buchstaben ( $3 * 10^9$ )

*2.861 Megabyte = 2,79 Gigabyte*

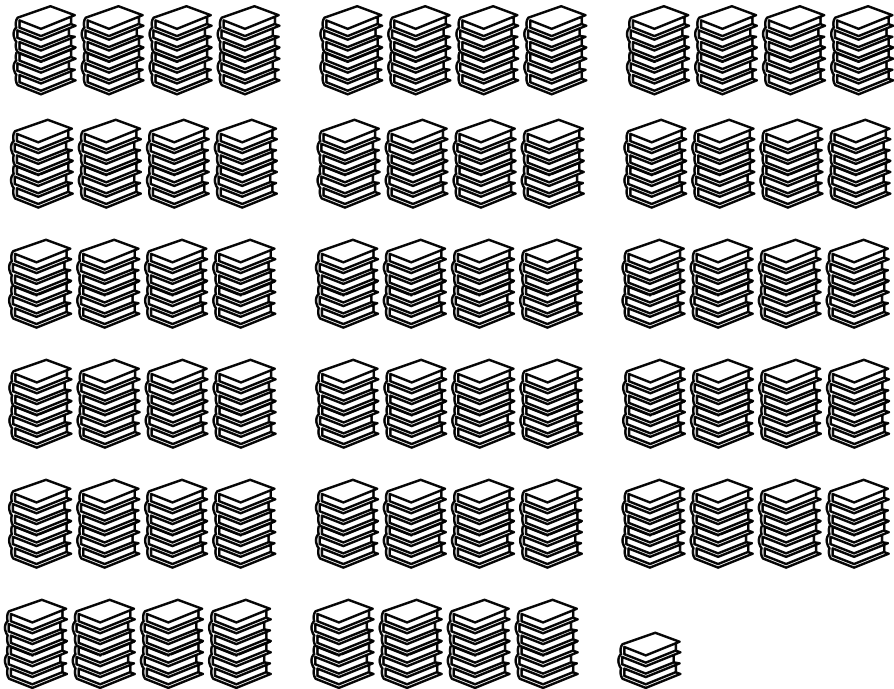


# Die Größe unseres Genoms

Brockhaus Lexikon (21. Auflage, 24 Bände)

~ 175 Millionen Buchstaben ( $175 * 10^6$ )

*167 Megabyte*



(413 Bücher)

menschliches Genom (in einer Körperzelle)

~ 3 Milliarden Buchstaben ( $3 * 10^9$ )

*2.861 Megabyte = 2,79 Gigabyte*



17,2 x



# Die Größe unseres Genoms

menschliches Genom (in einer Körperzelle)

~ 3 Milliarden Buchstaben ( $3 * 10^9$ )

*2.861 Megabyte = 2,79 Gigabyte*



Wikipedia (alle Artikel aller 316 Sprachen)

~ 250 Milliarden Buchstaben ( $250 * 10^{12}$ )

*232,83 Gigabyte*



# Die Größe unseres Genoms

menschliches Genom (in einer Körperzelle)  
~ 3 Milliarden Buchstaben ( $3 * 10^9$ )  
*2.861 Megabyte = 2,79 Gigabyte*



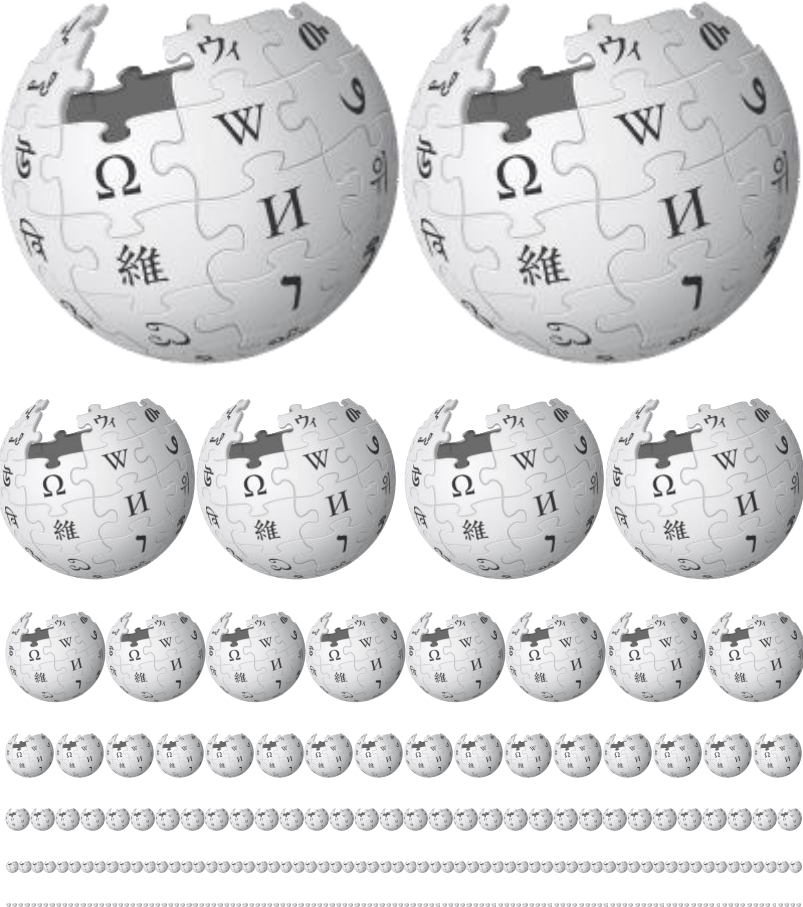
**83,3 x**

Wikipedia (alle Artikel aller 316 Sprachen)  
~ 250 Milliarden Buchstaben ( $250 * 10^{12}$ )  
*232,83 Gigabyte*



# Die Größe unseres Genoms

Wikipedia (alle Artikel aller 316 Sprachen)  
~ 250 Milliarden Buchstaben ( $250 * 10^{12}$ )  
*232,83 Gigabyte*



432.000.000.  
000.000 X

Ein Mensch: ~36 Billionen Zellen ( $36 * 10^{15}$ )  
→ Gesamtheit aller DNA eines Menschen:  
108 Quadrillion Buchstaben ( $108 * 10^{24}$ )  
*89,37 Yottabyte*



( $7 * 10^{22}$  Sterne im beobachtbaren Universum)

# Die Größe unseres Genoms

Alle im Internet verfügbaren Daten  
(Texte, Bilder, Videos, ...)  
*~200 Zettabyte*



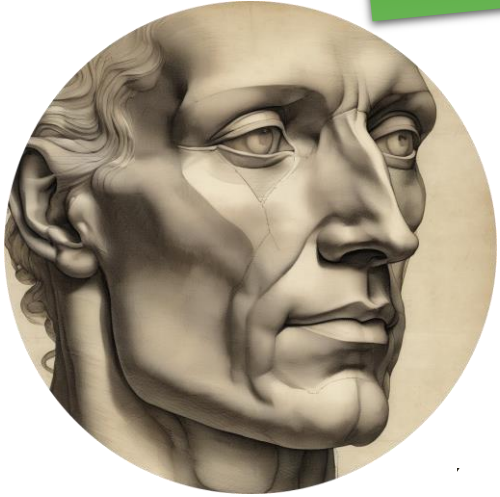
Ein Mensch: ~36 Billionen Zellen ( $36 * 10^{15}$ )  
→ Gesamtheit aller DNA eines Menschen:  
*89,37 Yottabyte*

**450 x**



# (angebliche) Ähnlichkeit zu anderen Lebewesen

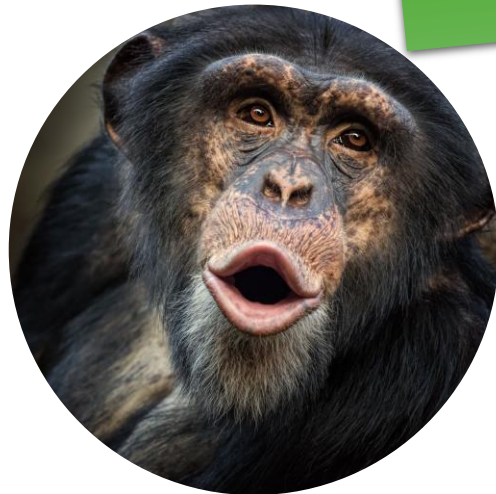
99,99%



85%



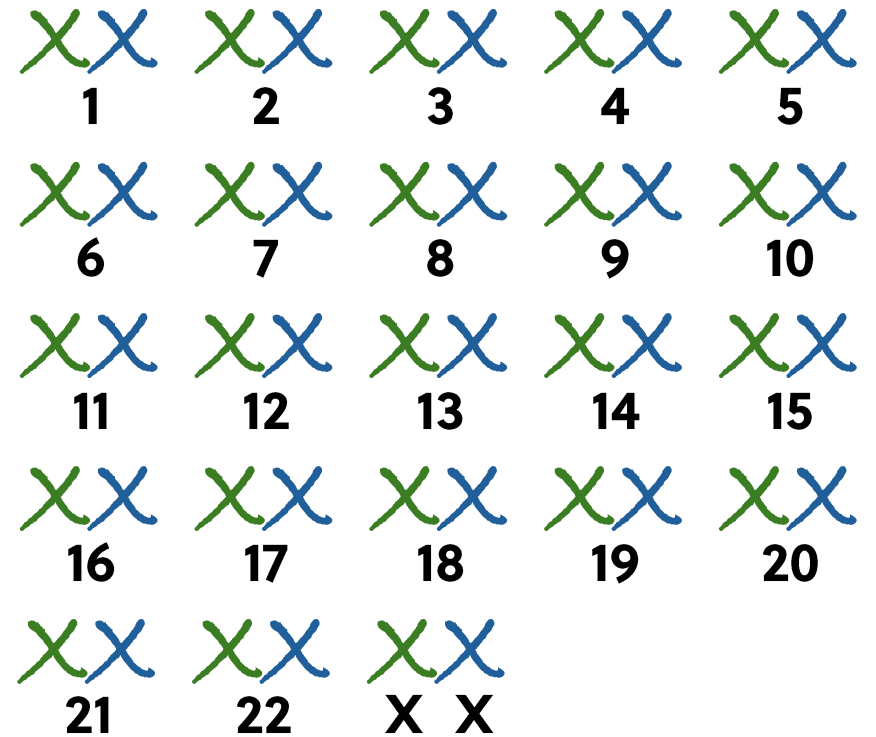
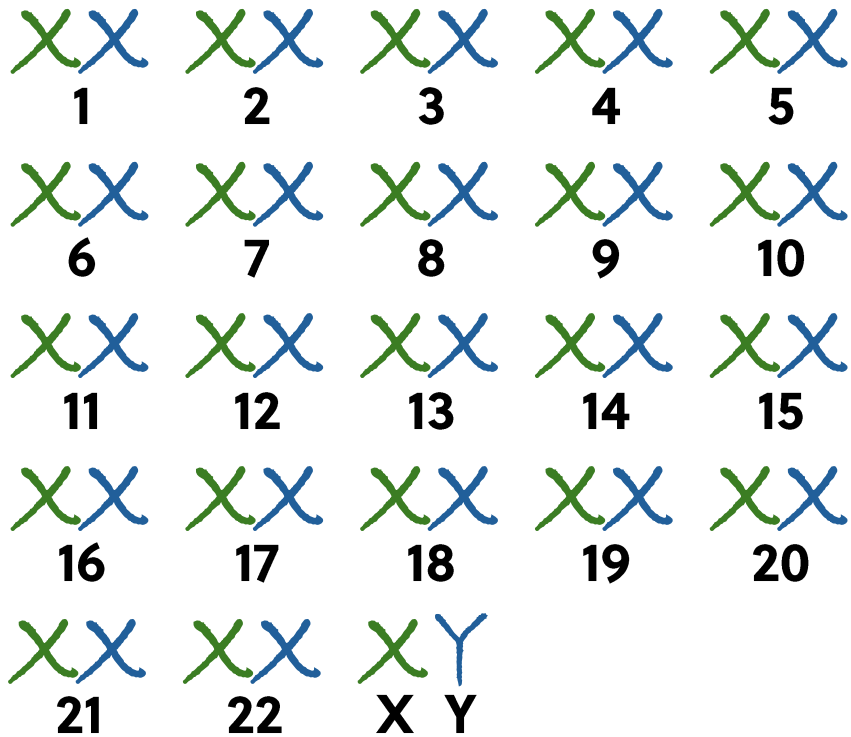
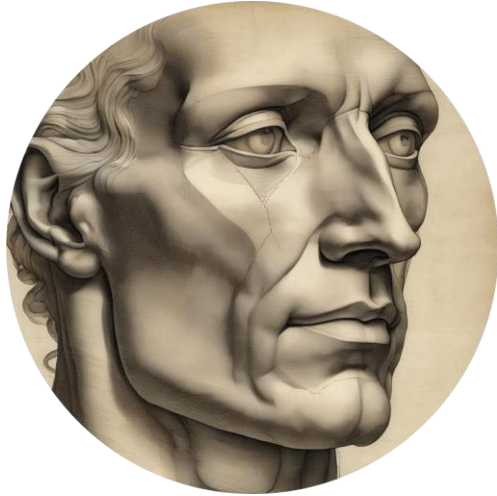
99%



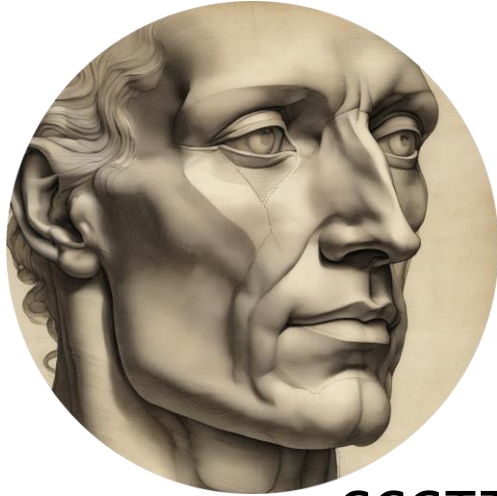
50%



# Wie bestimmt man die genomische Ähnlichkeit?



# Wie bestimmt man die genomische Ähnlichkeit?



DNA-Sequenz



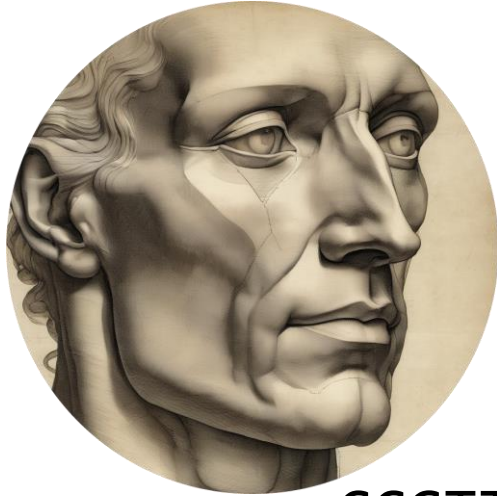
X	1	X	11	X	21
X	2	X	12	X	22
X	3	X	13	X	X
X	4	X	14		
X	5	X	15		
X	6	X	16		
X	7	X	17		
X	8	X	18		
X	9	X	19		
X	10	X	20		

CCGTTTGC GGTAACG  
GATCCCGGGATATAA  
TGCGTTTATGGCAAT  
TGCGCTATAGTACCG  
TGAGGAAAATGGCTC  
CTAGCGCTACATGCT  
GCATGCACCTGTGTC  
GCAATGGTCCGAACC  
GTTACGGCGATGCGG  
TAACGGATCCCGGGA  
...

CCGTTTGC GGTCACG  
GATCCCGGGATATAA  
TGCGTTTATGGCAAT  
TGCGCTATAGTACCG  
TGAGCAAAATGGCTC  
CTAGCGCTACATGCT  
GCATGCACCTGAGTC  
GCAATGGTCCGAACC  
GTTACGGGGATGCGG  
TAACGGATCCCGGGA  
...

X	1	X	11	X	21
X	2	X	12	X	22
X	3	X	13	X	X
X	4	X	14		
X	5	X	15		
X	6	X	16		
X	7	X	17		
X	8	X	18		
X	9	X	19		
X	10	X	20		

# Wie bestimmt man die genomische Ähnlichkeit?



DNA-Sequenz



X	1	X	11	X	21
X	2	X	12	X	22
X	3	X	13	X	X
X	4	X	14		
X	5	X	15		
X	6	X	16		
X	7	X	17		
X	8	X	18		
X	9	X	19		
X	10	X	20		

CCGTTTGC~~A~~GGTAACG  
GATCCCGGGATATAA  
TGCGTTTATGGCAAT  
TGCGCTATAGTACCG  
TGAGGAAAATGGCTC  
CTAGCGCTACATGCT  
GCATGCACCTGIGTC  
GCAATGGTCCGAACC  
GTTACGGCGATGCGG  
TAACGGATCCCGGGA  
...

CCGTTTGC~~C~~GGTCACG  
GATCCCGGGATATAA  
TGCGTTTATGGCAAT  
TGCGCTATAGTACCG  
TGAGCAAAATGGCTC  
CTAGCGCTACATGCT  
GCATGCACCTGAGTC  
GCAATGGTCCGAACC  
GTTACGGGGATGCGG  
TAACGGATCCCGGGA  
...

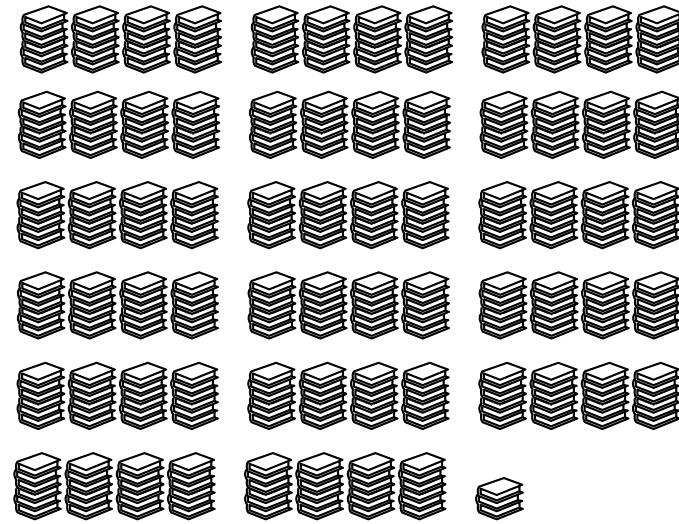
X	1	X	11	X	21
X	2	X	12	X	22
X	3	X	13	X	X
X	4	X	14		
X	5	X	15		
X	6	X	16		
X	7	X	17		
X	8	X	18		
X	9	X	19		
X	10	X	20		



# Wie bestimmt man die genomische Ähnlichkeit?



3 Milliarden Buchstaben



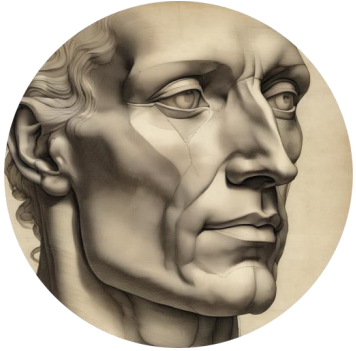
(413 Bücher)



300 Millisekunde pro Buchstabe  
→ 28,5 Jahre

0,001 Millisekunden pro Buchstabe  
→ 50 Minuten

# Wie entstehen Unterschiede im Genom?



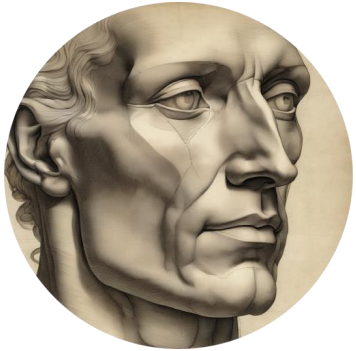
CCGTTTGCGGT AACGGATCCCGGGATATAATGCGTTTATGG

Punktmutation



CCGTTTGCGGT CACGGATCCCGGGTAATGCGTTTATGGCCA

# Wie entstehen Unterschiede im Genom?



CCGTTTGC GGT AACGGATCCCGGG ATAT AATGCGTT TATGG

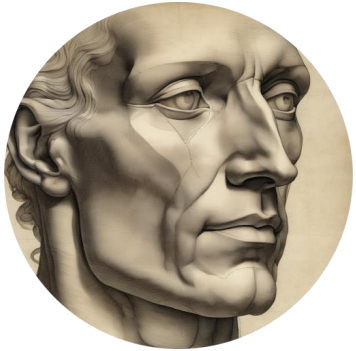
Punktmutation



CCGTTTGC GGT CACGGATCCCGGG TAAT GCGTTTAT GGCCA

15 Punktmutationen?

# Wie entstehen Unterschiede im Genom?



CCGTTTGC GGT AACGGATCCCGGGATA **TAATGCGTTTATGG**

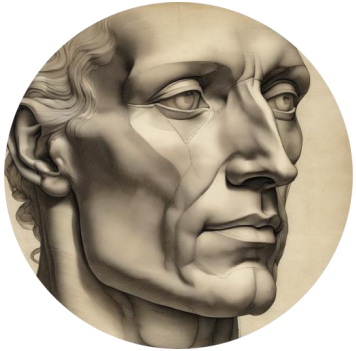


CCGTTTGC GGT CACGGATCCCGGG **TAATGCGTTTATGGCCA**

15 Punktmutationen?

→ Nein! Verschiebungen im Genom durch  
Verlust/Hinzufügen von DNA-Sequenzen

# Wie entstehen Unterschiede im Genom?



CCGTTTGC GGT AACGGATCCCGGG ATATAATGCGTTTATGG ---

"Lücke"



CCGTTTGC GGT CACGGATCCCGGG --- TAATGCGTTTATGG CCA



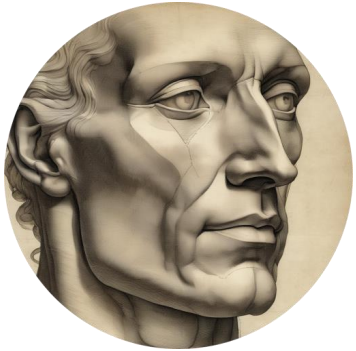
15 Punktmutationen?

→ Nein! Verschiebungen im Genom durch  
Verlust/Hinzufügen von DNA-Sequenzen

# Die korrekte Berechnung eines DNA-Sequenzvergleichs

## 1. Idee: naiver Ansatz

Probiere alle möglichen Verschiebungen bei zwei gegebenen DNA-Sequenzen aus!  
Finde die Verschiebung mit möglichst vielen Übereinstimmungen  
und möglichst wenigen Unterschieden!



C



C

3 Möglichkeiten

C  
C

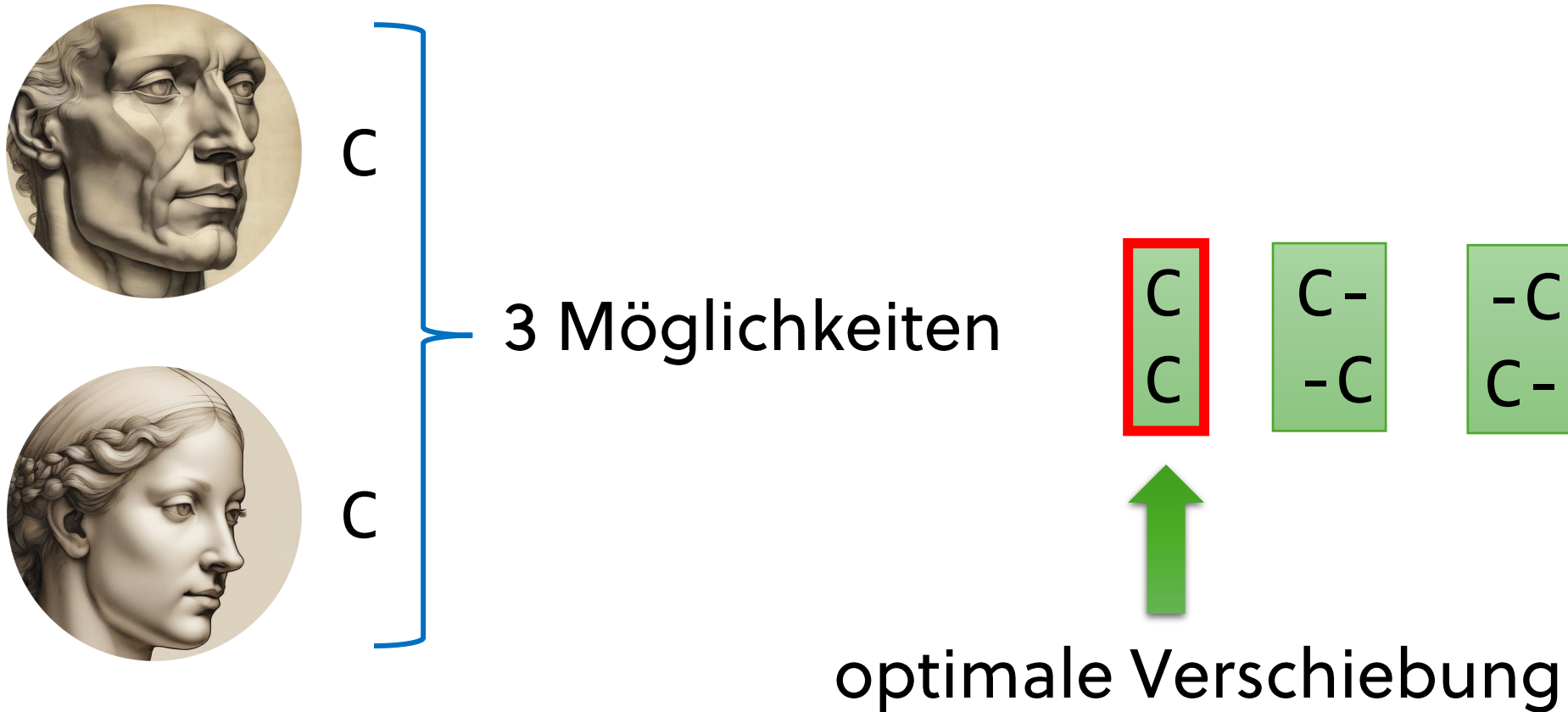
C -  
- C

- C  
C -

# Die korrekte Berechnung eines DNA-Sequenzvergleichs

## 1. Idee: naiver Ansatz

Probiere alle möglichen Verschiebungen bei zwei gegebenen DNA-Sequenzen aus!  
Finde die Verschiebung mit möglichst vielen Übereinstimmungen  
und möglichst wenigen Unterschieden!



# Die korrekte Berechnung eines DNA-Sequenzvergleichs

## 1. Idee: naiver Ansatz

Probiere alle möglichen Verschiebungen bei zwei gegebenen DNA-Sequenzen aus!  
Finde die Verschiebung mit möglichst vielen Übereinstimmungen  
und möglichst wenigen Unterschieden!



CA



TC

13 Möglichkeiten

		C-A	C-A	CA--	--CA
		-TC	TC-	--TC	TC--
CA	CA	CA-	CA-	-CA-	C--A
TC	TC	-TC	T-C	T--C	-TC-
		-CA	-CA	C-A-	-C-A
		TC-	T-C	-T-C	T-C-



# Die korrekte Berechnung eines DNA-Sequenzvergleichs

## 1. Idee: naiver Ansatz

Probiere alle möglichen Verschiebungen bei zwei gegebenen DNA-Sequenzen aus!  
Finde die Verschiebung mit möglichst vielen Übereinstimmungen  
und möglichst wenigen Unterschieden!



CA



TC

13 Möglichkeiten

		C-A -TC	C-A TC-	CA-- --TC	--CA TC--
CA	TC	CA- -TC	CA- T-C	-CA- T--C	C--A -TC-
		-CA TC-	-CA T-C	C-A- -T-C	-C-A T-C-



optimale Verschiebung

# Die korrekte Berechnung eines DNA-Sequenzvergleichs

## 1. Idee: naiver Ansatz

Probiere alle möglichen Verschiebungen bei zwei gegebenen DNA-Sequenzen aus!  
Finde die Verschiebung mit möglichst vielen Übereinstimmungen  
und möglichst wenigen Unterschieden!



Sequenzlänge N



Sequenzlänge N

$$\sum_{k=0}^N \binom{N+k}{k} \binom{N}{k}$$

Anzahl möglicher  
Lücken in Sequenz 2

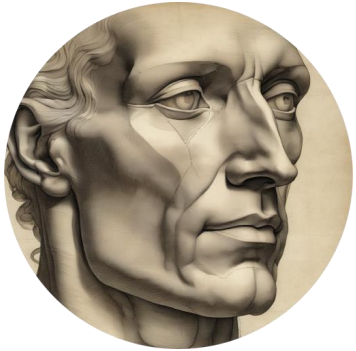
Möglichkeiten

Anzahl möglicher  
Lücken in Sequenz 1

# Die korrekte Berechnung eines DNA-Sequenzvergleichs

## 1. Idee: naiver Ansatz

Probiere alle möglichen Verschiebungen bei zwei gegebenen DNA-Sequenzen aus!  
Finde die Verschiebung mit möglichst vielen Übereinstimmungen  
und möglichst wenigen Unterschieden!



$$\sum_{k=0}^N \binom{N+k}{k} \binom{N}{k}$$

### Möglichkeiten:

$$N = 10 \rightarrow \text{ca. } 10.000$$

$$N = 15 \rightarrow \text{ca. } 100.000.000$$

$$N = 20 \rightarrow \text{ca. } 10^{11}$$

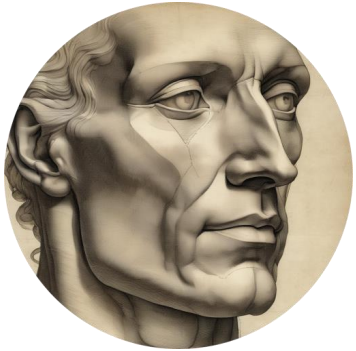
$$N = 500 \rightarrow \text{ca. } 10^{299}$$

$$N = 3.000.000.000 \rightarrow ???$$

# Die korrekte Berechnung eines DNA-Sequenzvergleichs

## 2. Idee: dynamische Programmierung

Verkleinere das Problem und nutze die Lösung des kleineren Problems um das größere Problem zu lösen.



GTGCACA



TCACTTA

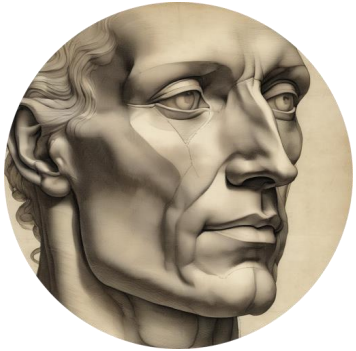
$$D(i, j) = \min \begin{cases} D(i-1, j) & + k(u_i, -) \\ D(i, j-1) & + k(-, v_j) \\ D(i-1, j-1) & + k(u_i, v_j) \end{cases} \quad k(u_i, v_j) = \begin{cases} 0, & \text{wenn } u_i = v_j \\ 1, & \text{sonst} \end{cases}$$

	G	T	G	C	A	C	A
T							
C							
A							
C							
T							
T							
A							

# Die korrekte Berechnung eines DNA-Sequenzvergleichs

## 2. Idee: dynamische Programmierung

Verkleinere das Problem und nutze die Lösung des kleineren Problems um das größere Problem zu lösen.



GTGCACA



TCACTTA

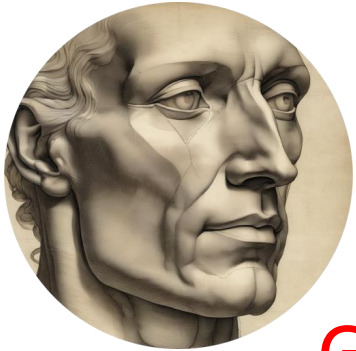
$$D(i, j) = \min \begin{cases} D(i-1, j) & + k(u_i, -) \\ D(i, j-1) & + k(-, v_j) \\ D(i-1, j-1) & + k(u_i, v_j) \end{cases} \quad k(u_i, v_j) = \begin{cases} 0, & \text{wenn } u_i = v_j \\ 1, & \text{sonst} \end{cases}$$

	G	T	G	C	A	C	A
T	1	1	2	3	4	5	6
C	2	2	2	2	3	4	5
A	3	3	3	3	2	3	4
C	4	4	4	3	3	2	3
T	5	4	5	4	4	3	3
T	6	5	5	5	5	4	4
A	7	6	6	6	5	5	<b>4</b>

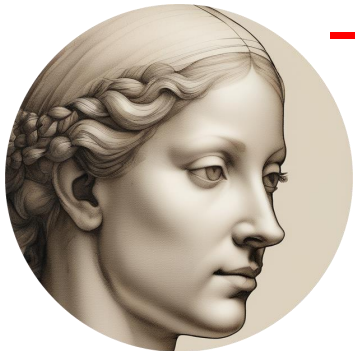
# Die korrekte Berechnung eines DNA-Sequenzvergleichs

## 2. Idee: dynamische Programmierung

Verkleinere das Problem und nutze die Lösung des kleineren Problems um das größere Problem zu lösen.



GTGCAC--A  
-T-CACTTA



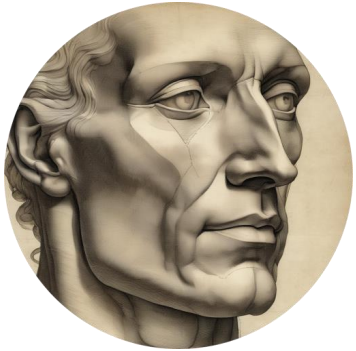
$$D(i, j) = \min \begin{cases} D(i-1, j) & + k(u_i, -) \\ D(i, j-1) & + k(-, v_j) \\ D(i-1, j-1) & + k(u_i, v_j) \end{cases} \quad k(u_i, v_j) = \begin{cases} 0, & \text{wenn } u_i = v_j \\ 1, & \text{sonst} \end{cases}$$

	G	T	G	C	A	C	A
T	1	1	2	3	4	5	6
C	2	2	2	2	3	4	5
A	3	3	3	3	2	3	4
C	4	4	4	3	3	2	3
T	5	4	5	4	4	3	3
T	6	5	5	5	5	4	4
A	7	6	6	6	5	5	4

# Die korrekte Berechnung eines DNA-Sequenzvergleichs

## 2. Idee: dynamische Programmierung

Verkleinere das Problem und nutze die Lösung des kleineren Problems um das größere Problem zu lösen.



Sequenzlänge N



Sequenzlänge N

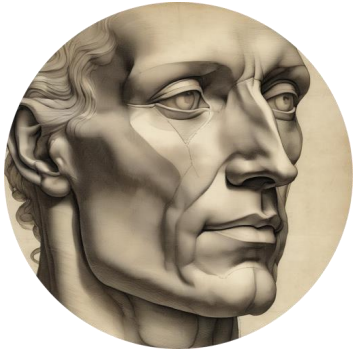
$N * N = N^2$  Rechenschritte

	G	T	G	C	A	C	A
T	1	1	2	3	4	5	6
C	2	2	2	2	3	4	5
A	3	3	3	3	2	3	4
C	4	4	4	3	3	2	3
T	5	4	5	4	4	3	3
T	6	5	5	5	5	4	4
A	7	6	6	6	5	5	4

# Die korrekte Berechnung eines DNA-Sequenzvergleichs

## 2. Idee: dynamische Programmierung

Verkleinere das Problem und nutze die Lösung des kleineren Problems um das größere Problem zu lösen.



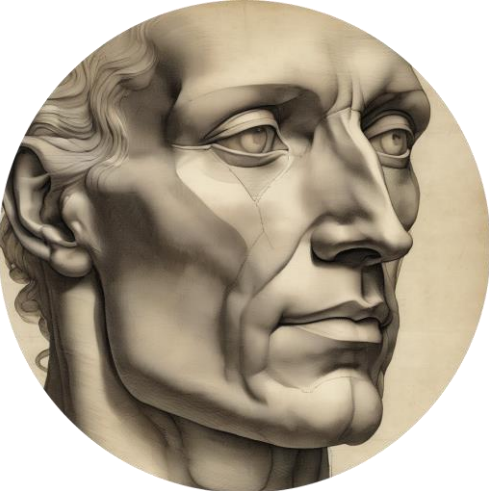
$3.000.000.000^2$  Rechenschritte  $\approx$  285.388 Jahre



0,001 Millisekunden  
pro Rechenschritt



# Bringen wir die Biologie zurück in das Problem

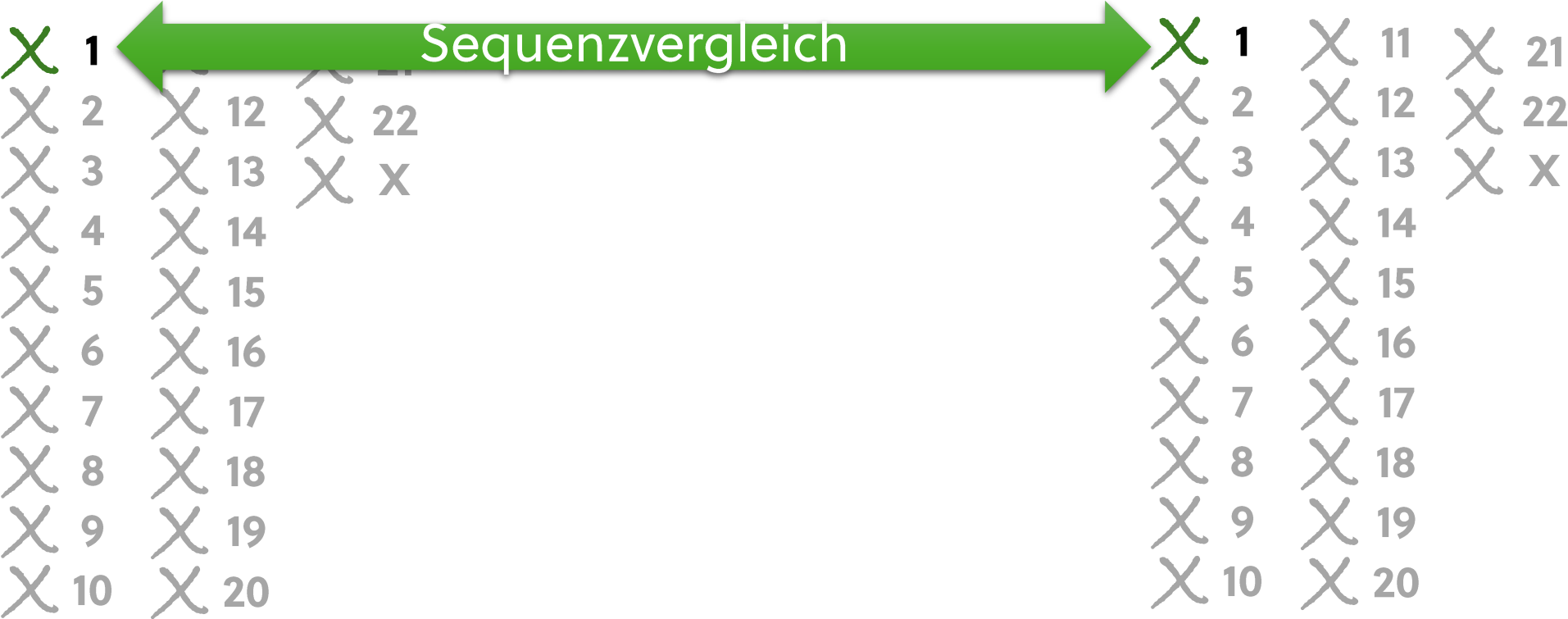
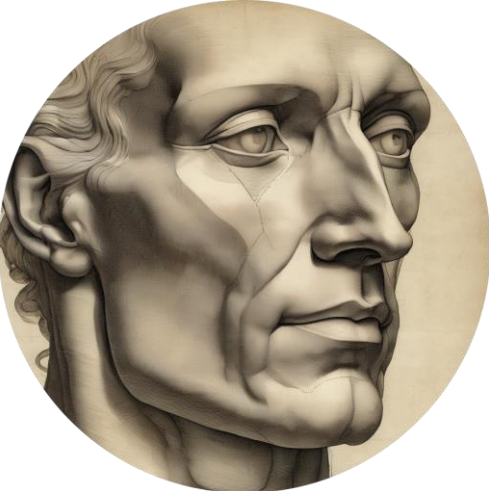


X	1	X	11	X	21
X	2	X	12	X	22
X	3	X	13	X	X
X	4	X	14		
X	5	X	15		
X	6	X	16		
X	7	X	17		
X	8	X	18		
X	9	X	19		
X	10	X	20		

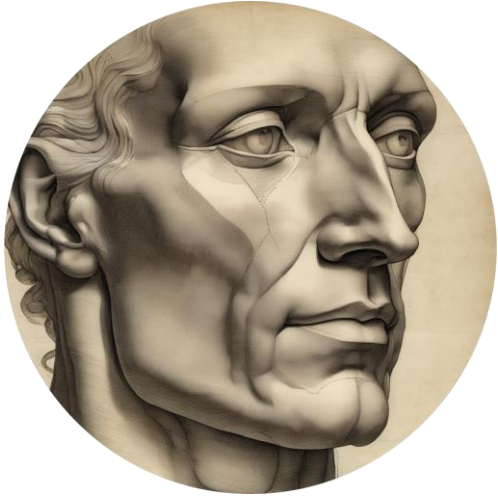


X	1	X	11	X	21
X	2	X	12	X	22
X	3	X	13	X	X
X	4	X	14		
X	5	X	15		
X	6	X	16		
X	7	X	17		
X	8	X	18		
X	9	X	19		
X	10	X	20		

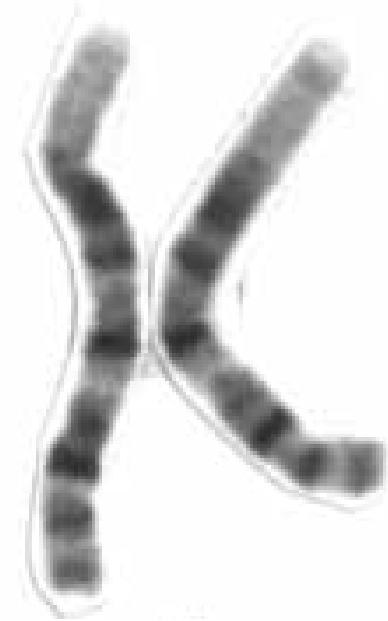
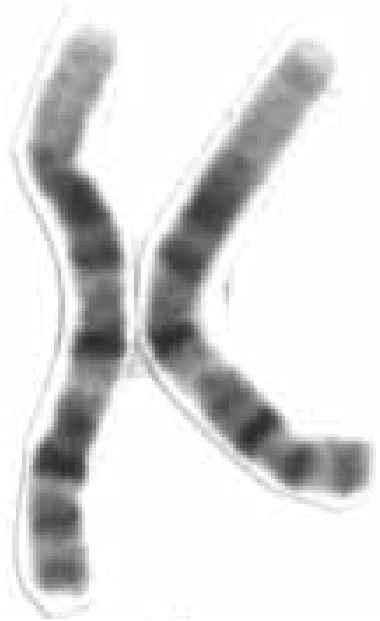
# Bringen wir die Biologie zurück in das Problem



# Bringen wir die Biologie zurück in das Problem

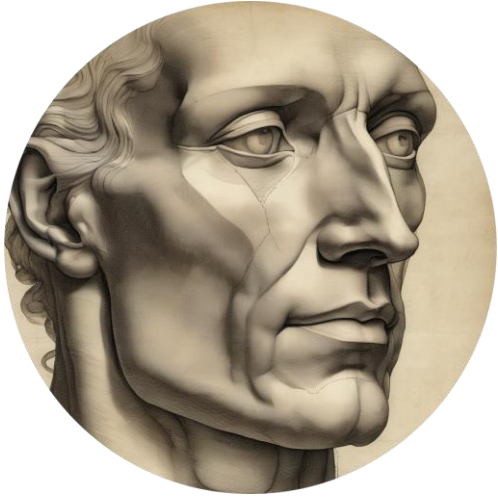


Chromosom 1  
249 Millionen Buchstaben

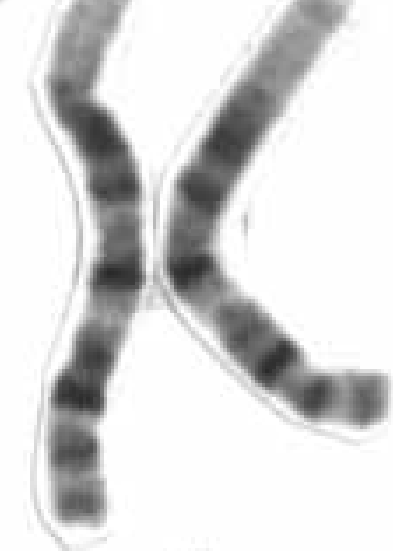
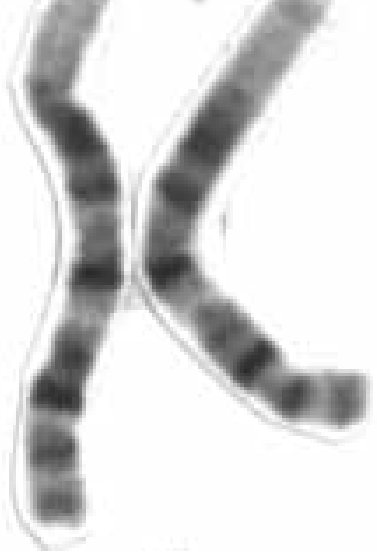


$249.000.000^2$  Rechenschritte  
 $\approx 1.966$  Jahre

# Bringen wir die Biologie zurück in das Problem

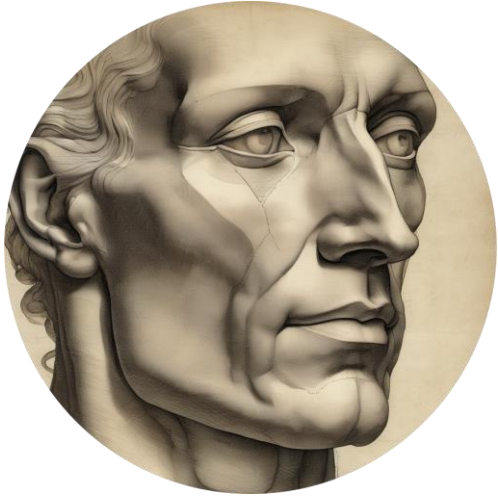


Chromosom 1  
249 Millionen Buchstaben

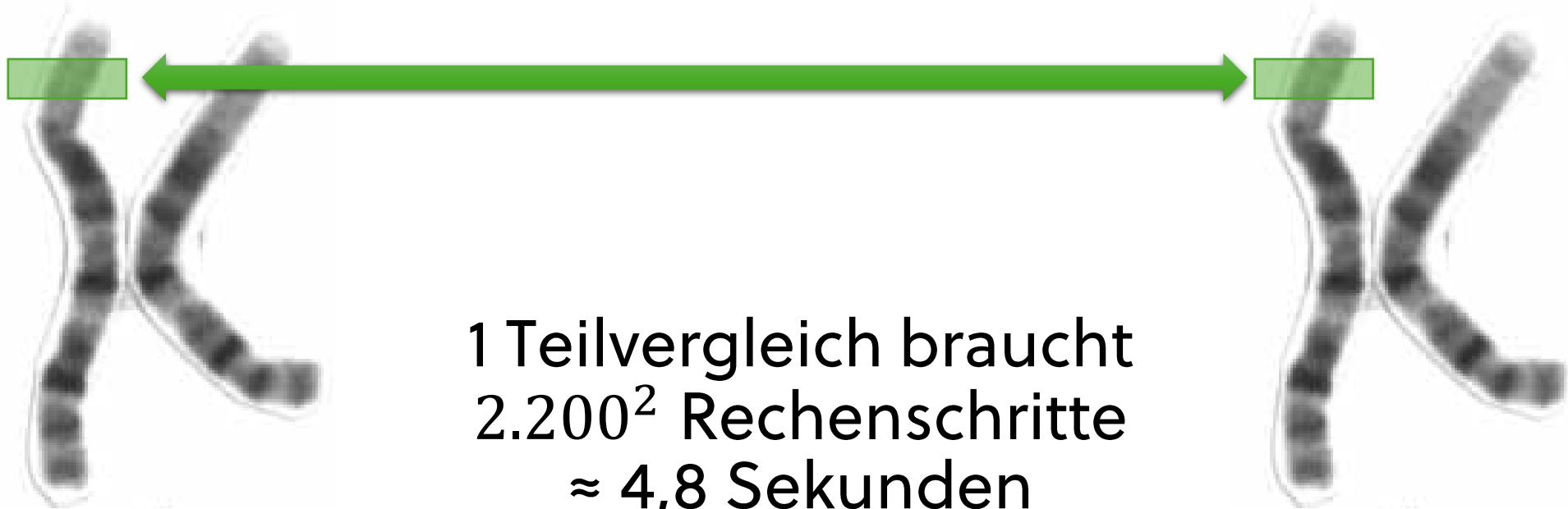


1 Teilvergleich braucht  
 $2.200^2$  Rechenschritte  
 $\approx 4,8$  Sekunden

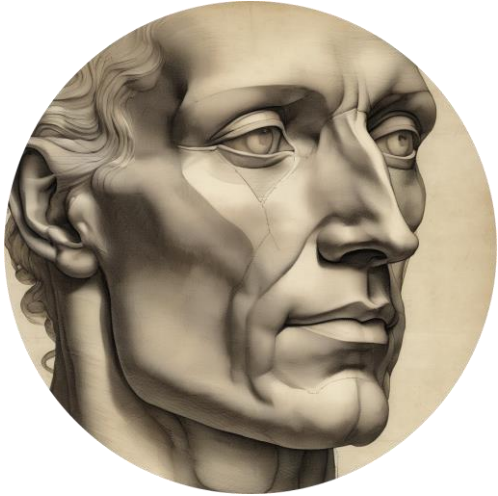
# Bringen wir die Biologie zurück in das Problem



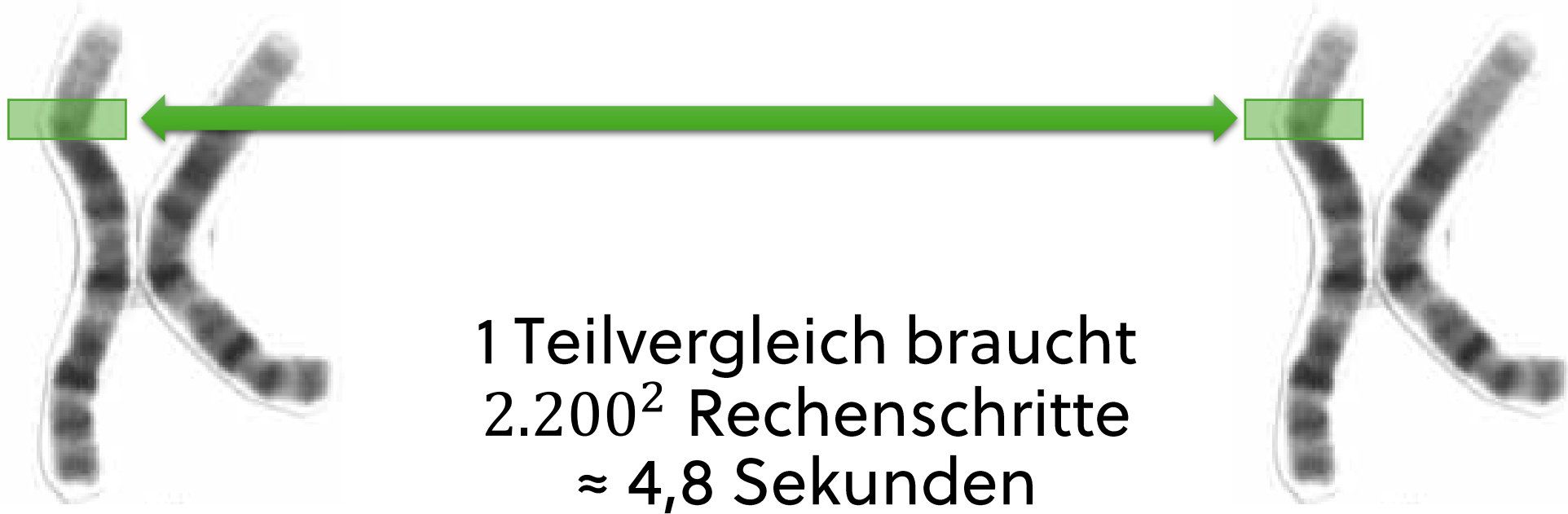
Chromosom 1  
249 Millionen Buchstaben



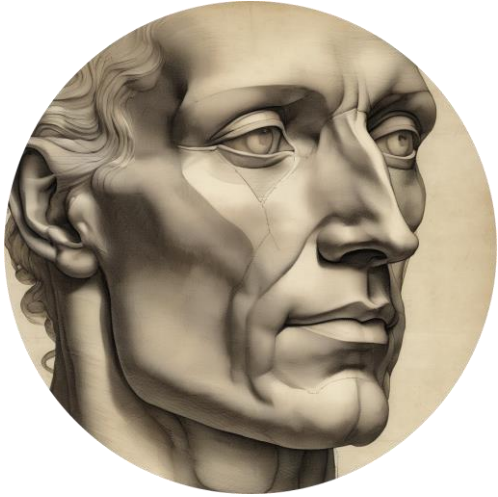
# Bringen wir die Biologie zurück in das Problem



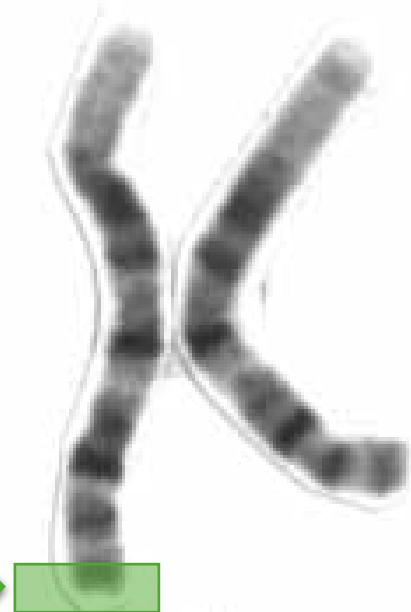
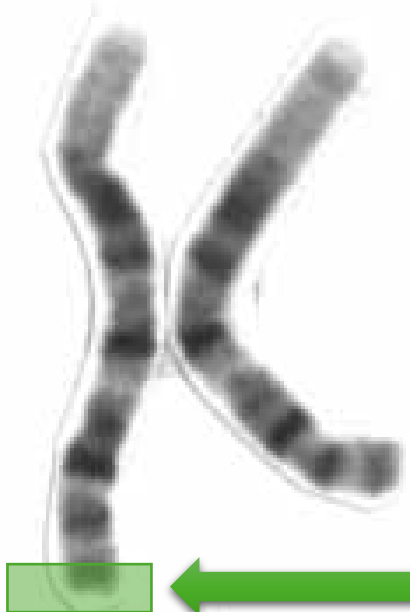
Chromosom 1  
249 Millionen Buchstaben



# Bringen wir die Biologie zurück in das Problem



Chromosom 1  
249 Millionen Buchstaben



Insgesamt:  
131.182 Teilvergleiche  
≈ 6,3 Tage

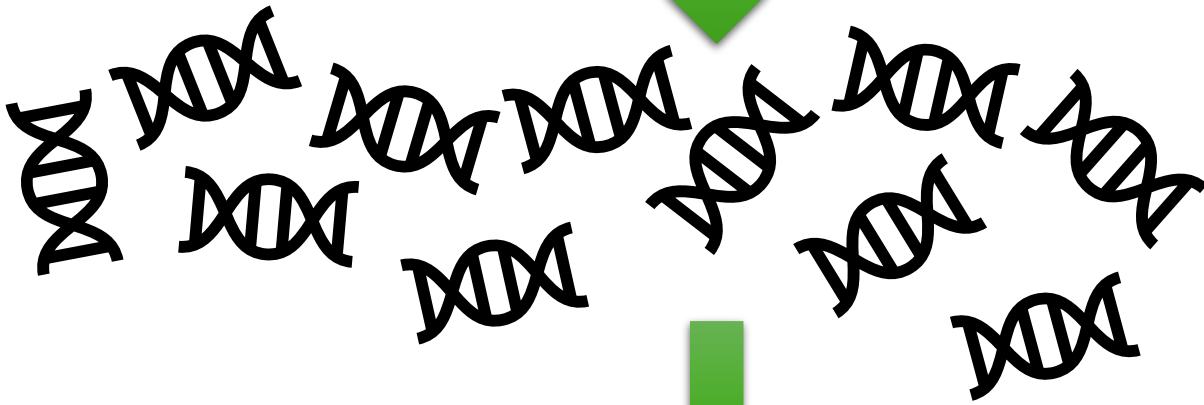


Woher kommen eigentlich die Genom-Sequenzen?

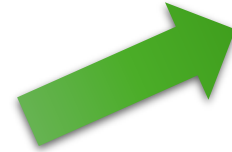




# Woher kommen eigentlich die Genom-Sequenzen?



CCGTA GTCAG GCACT ACTGC CCTGA  
ATGCA TTTAT CAGAA GTAGT TTATG



CCTGA TTTAT CCGTA GCACT ATGCA CAGAA GTCAG TTATG GTAGT ACTGC

# Woher kommen eigentlich die Genom-Sequenzen?



Human Genome  
Project

1990 - 2001



Kosten: ~ 3 Milliarden Euro



## 1. Generation

- 500-800 lange DNA-Stücke
- bis zu 20.000 Basen/Tag

# Woher kommen eigentlich die Genom-Sequenzen?

## 2. Generation

- 50-350 lange DNA-Stücke
- bis zu 3 Billionen Basen/Tag
- Kosten: ~ 19.000 Euro



## 1. Generation

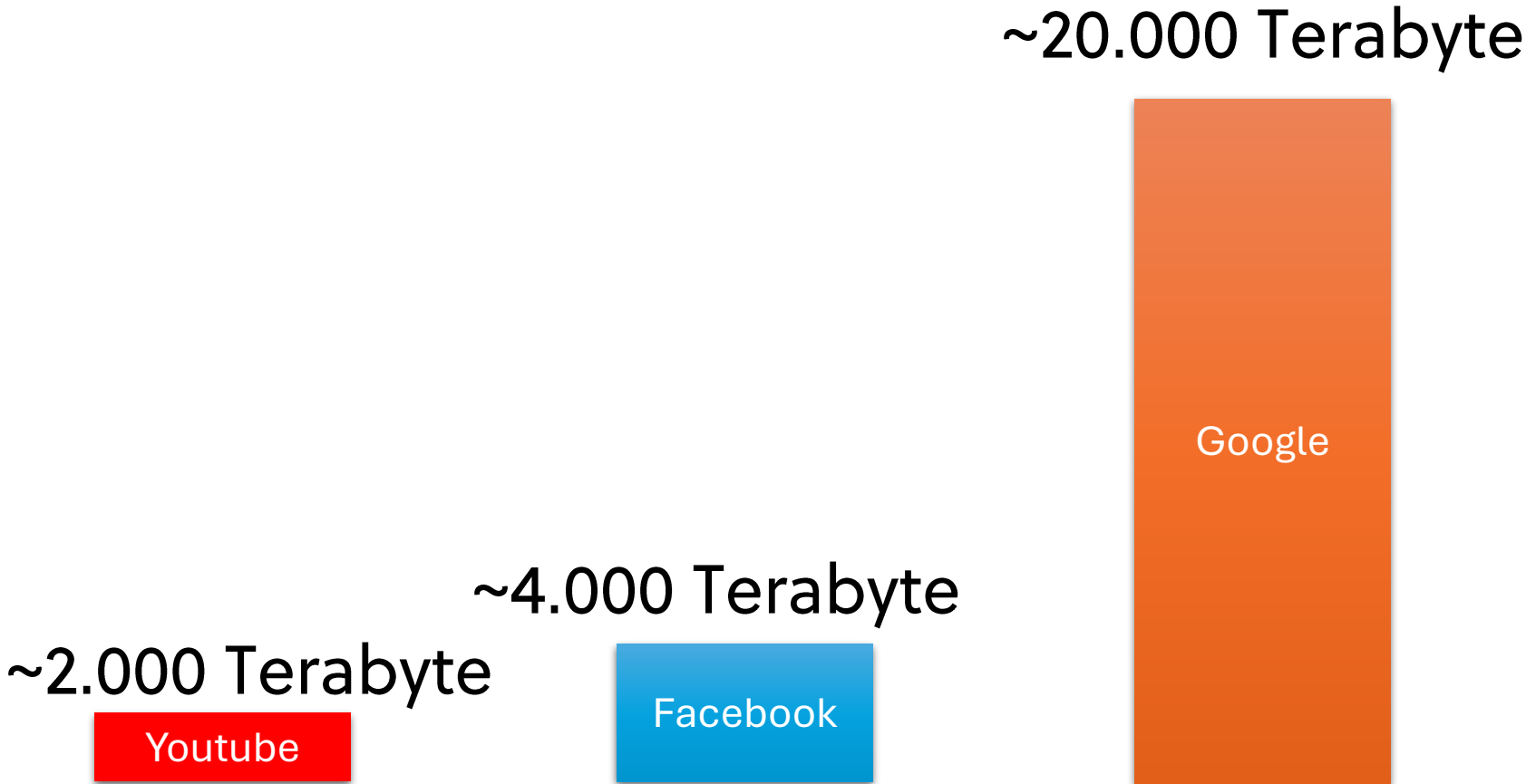
- 500-800 lange DNA-Stücke
- bis zu 20.000 Basen/Tag
- Kosten: ~ 3 Milliarden Euro



## 3. Generation

- >5.000 lange DNA-Stücke
- bis zu 20 Milliarden Basen/Tag
- Kosten: ~ 800 Euro

# Produzierte Datenmenge



# Produzierte Datenmenge pro Tag

~200.000 Terabyte



~20.000 Terabyte



~2.000 Terabyte



~4.000 Terabyte



# Was ist nun eigentlich mit der Banane?

*Musa acuminata*



Von Ichwarsnur - Eigenes Werk, CC BY-SA 4.0  
<https://commons.wikimedia.org/w/index.php?curid=53329538>

*Musa balbisiana*



Von Ruestz - Eigenes Werk, CC BY-SA 3.0  
<https://commons.wikimedia.org/w/index.php?curid=824363>



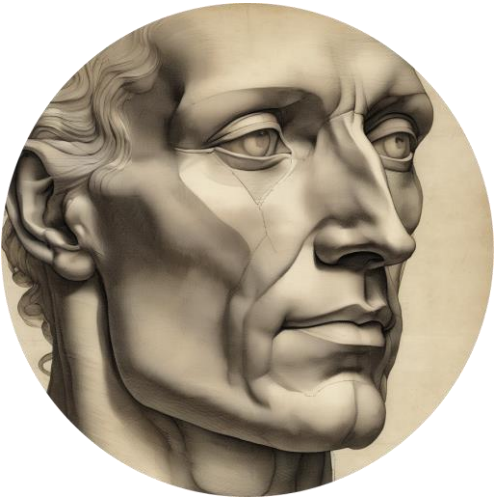
*Musa x paradisiaca*

# Was ist nun eigentlich mit der Banane?



50%

Sequenzvergleich



11 Chromosomen



0,5 Milliarden  
Buchstaben

23 Chromosomen



3,0 Milliarden  
Buchstaben

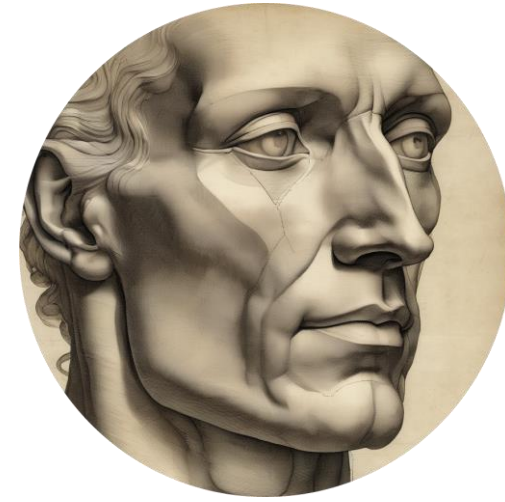


# Was ist nun eigentlich mit der Banane?



??%

Sequenzvergleich



11 Chromosomen



0,5 Milliarden  
Buchstaben

23 Chromosomen



3,0 Milliarden  
Buchstaben





# Was steht eigentlich im Genom?

Gen

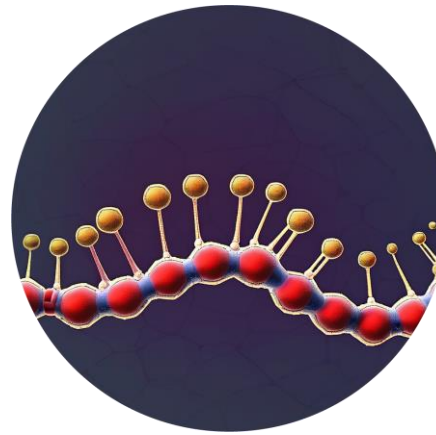
CCGTATGCGGTACCGATGACGTCA



DNA

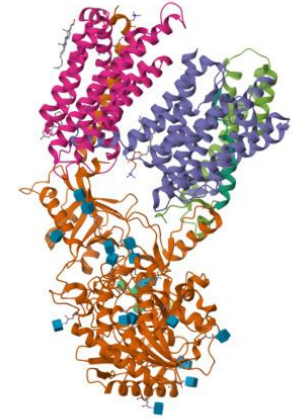
Gentranskript

AUGCGGUACCGAUGA



RNA

Genprodukt



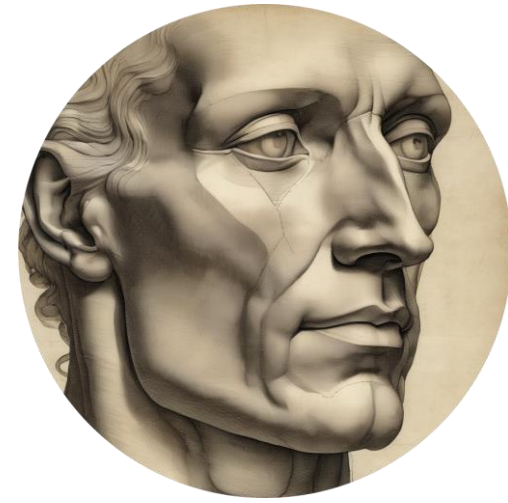
Protein



# Vergleich der proteinkodierenden Gene



~30.700 Gene



~20.100 Gene

Ø Länge:  
2.200 Buchstaben

**Gen B1**

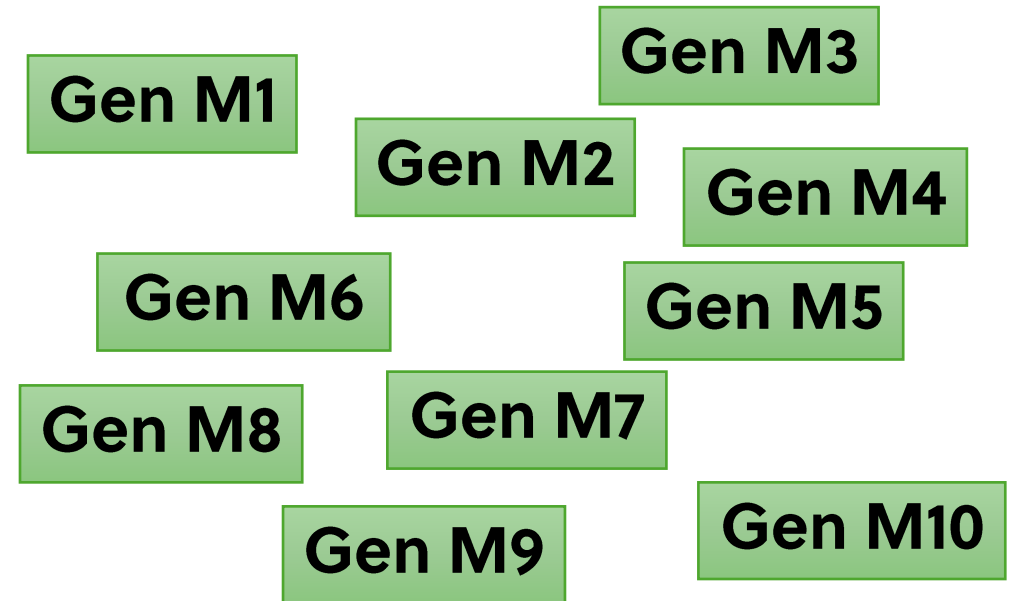
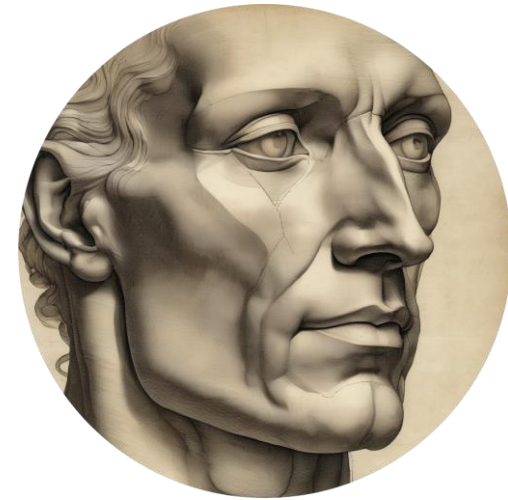
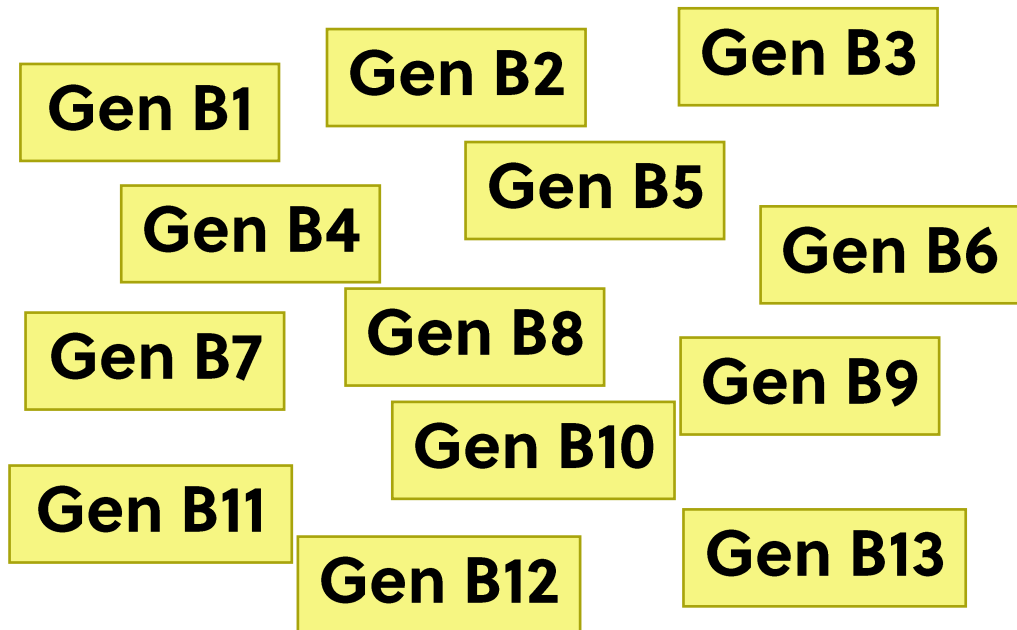
ATGCGGTACCGATGA

Sequenzvergleich

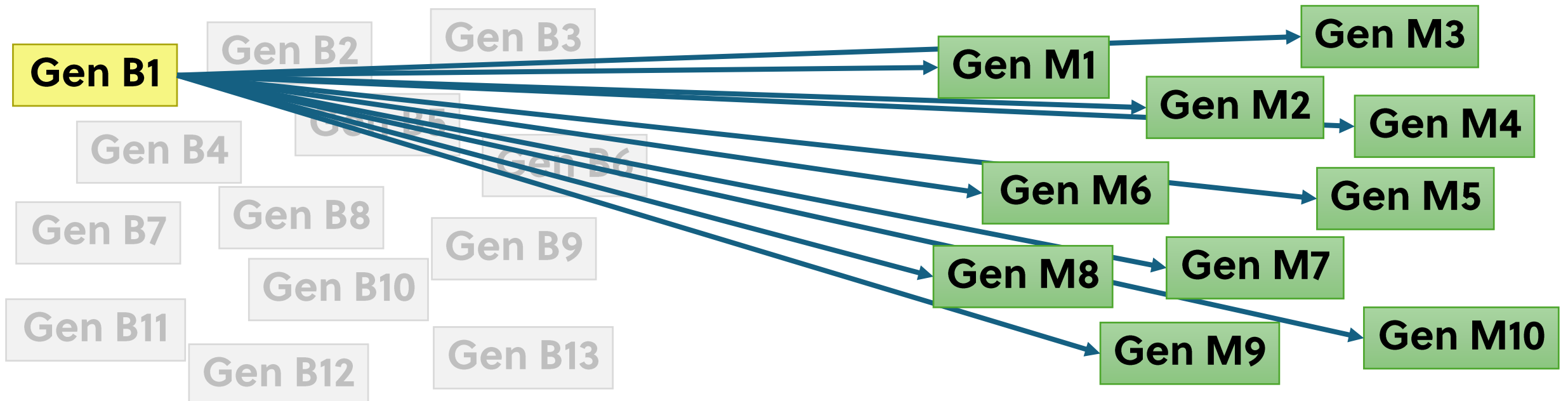
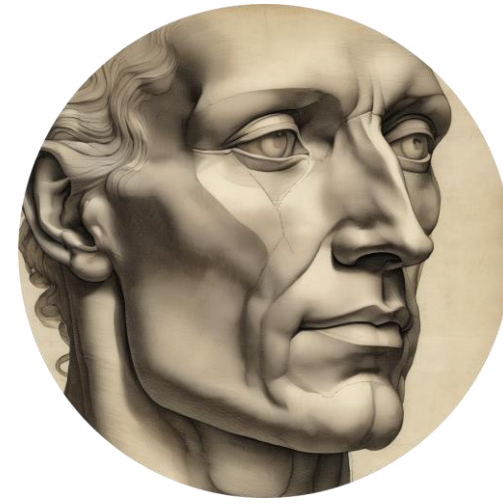
**Gen M1**

ATGTGCAGATGCTAA

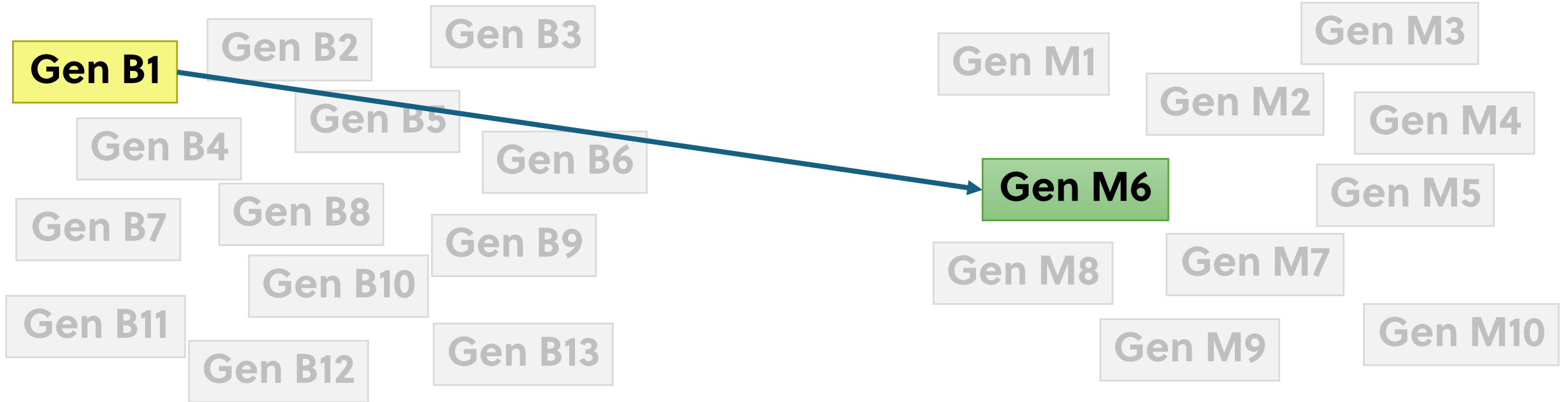
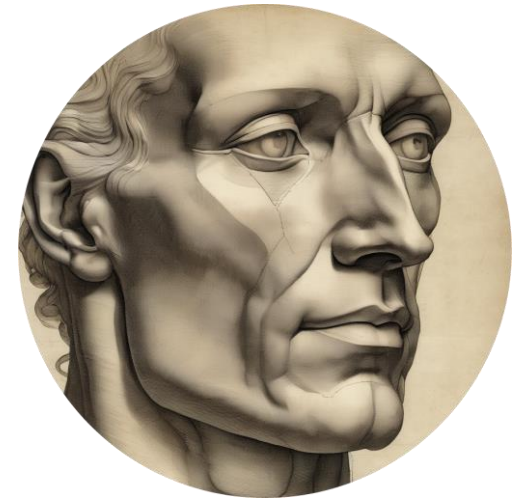
# Vergleich der proteinkodierenden Gene



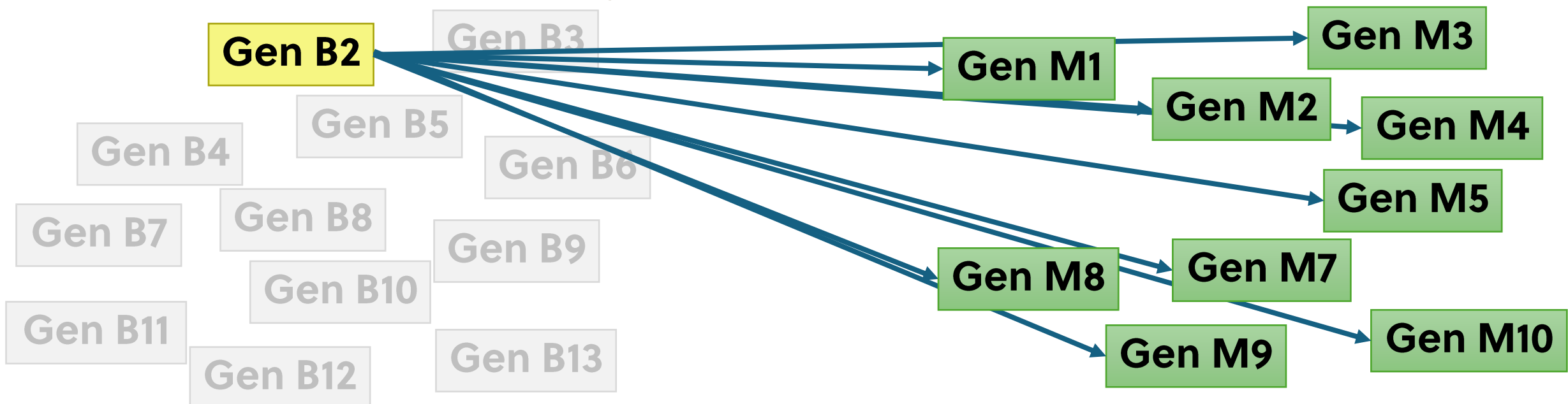
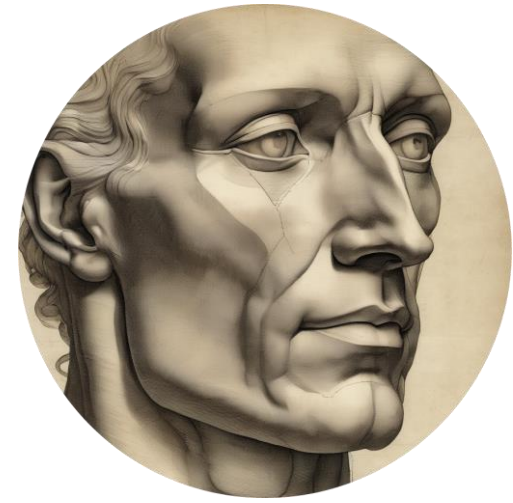
# Vergleich der proteinkodierenden Gene



# Vergleich der proteinkodierenden Gene



# Vergleich der proteinkodierenden Gene



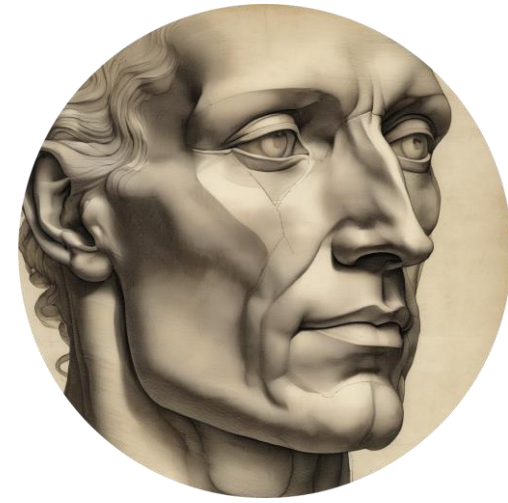
# Vergleich der proteinkodierenden Gene



**Gen B1**

78%

**Gen M6**



**Gen B2**

Gen B3

Gen B4

Gen B5

Gen B6

Gen B7

Gen B8

Gen B9

Gen B10

Gen B11

Gen B12

Gen B13

Gen M1

Gen M3

Gen M2

Gen M4

Gen M5

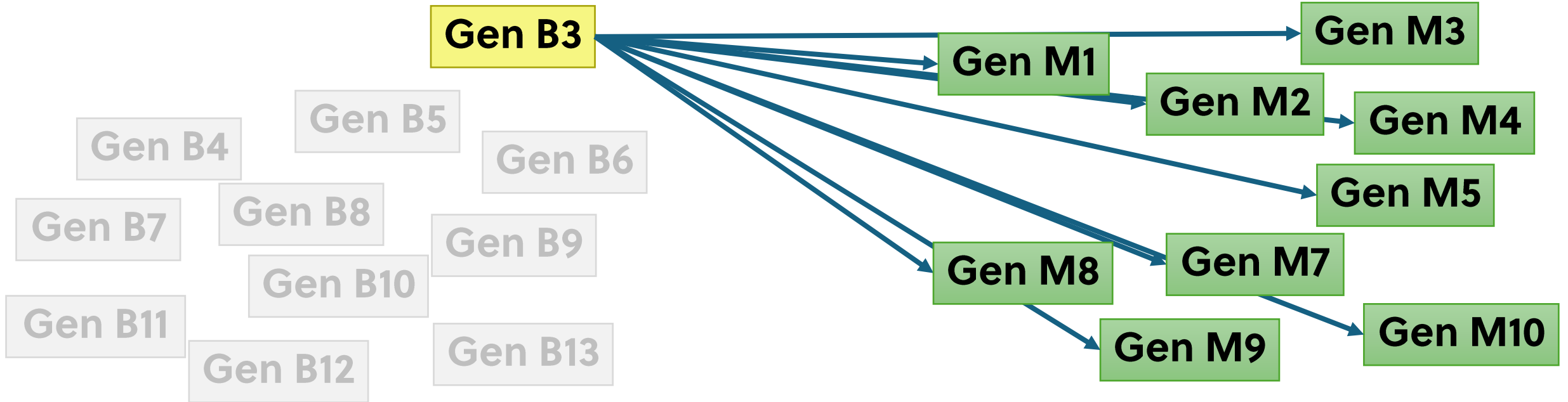
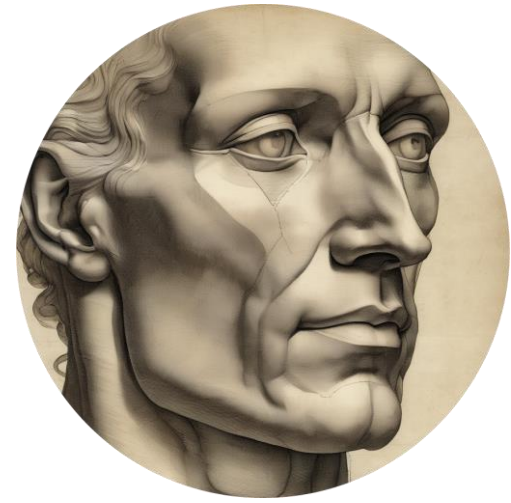
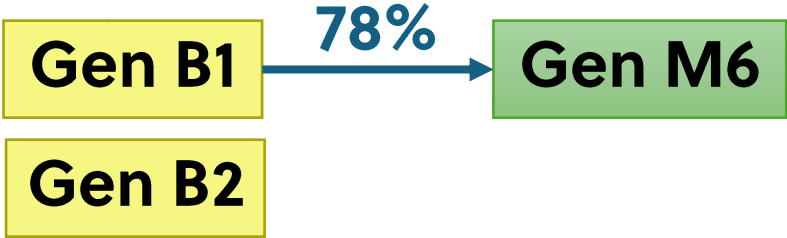
Gen M8

Gen M7

Gen M9

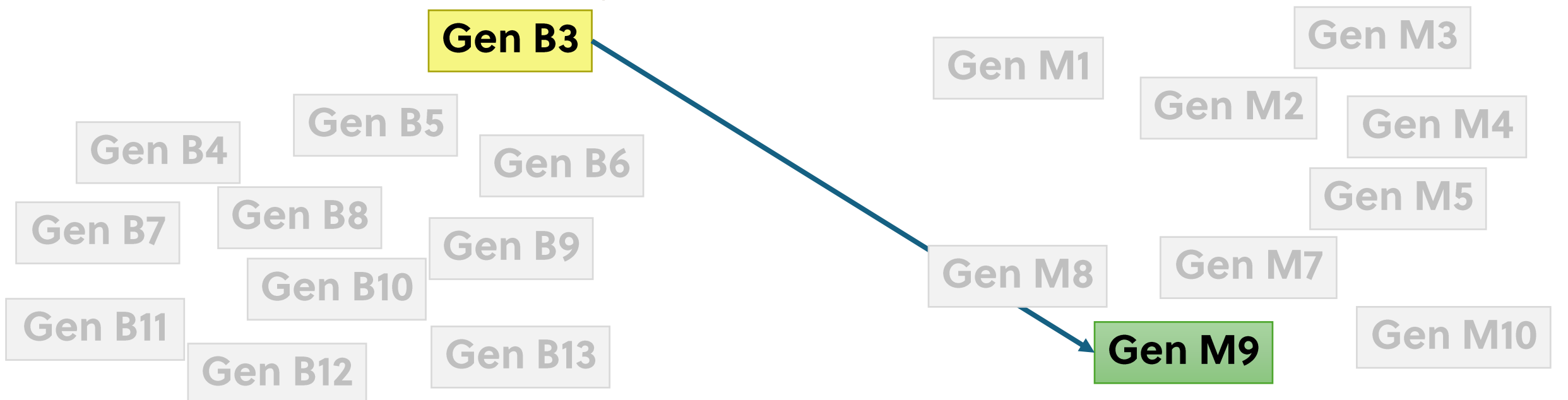
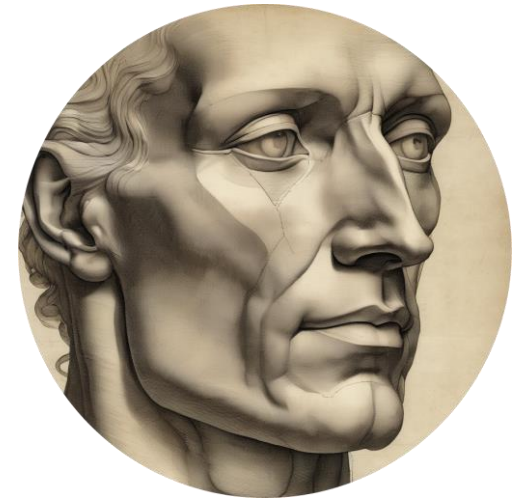
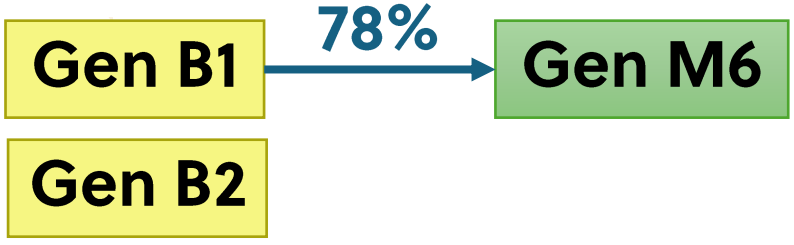
Gen M10

# Vergleich der proteinkodierenden Gene

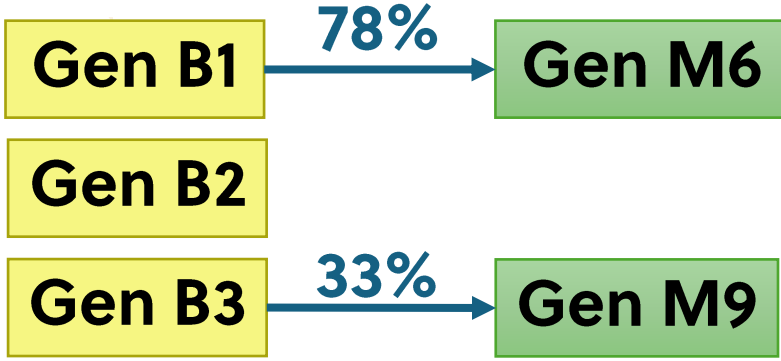




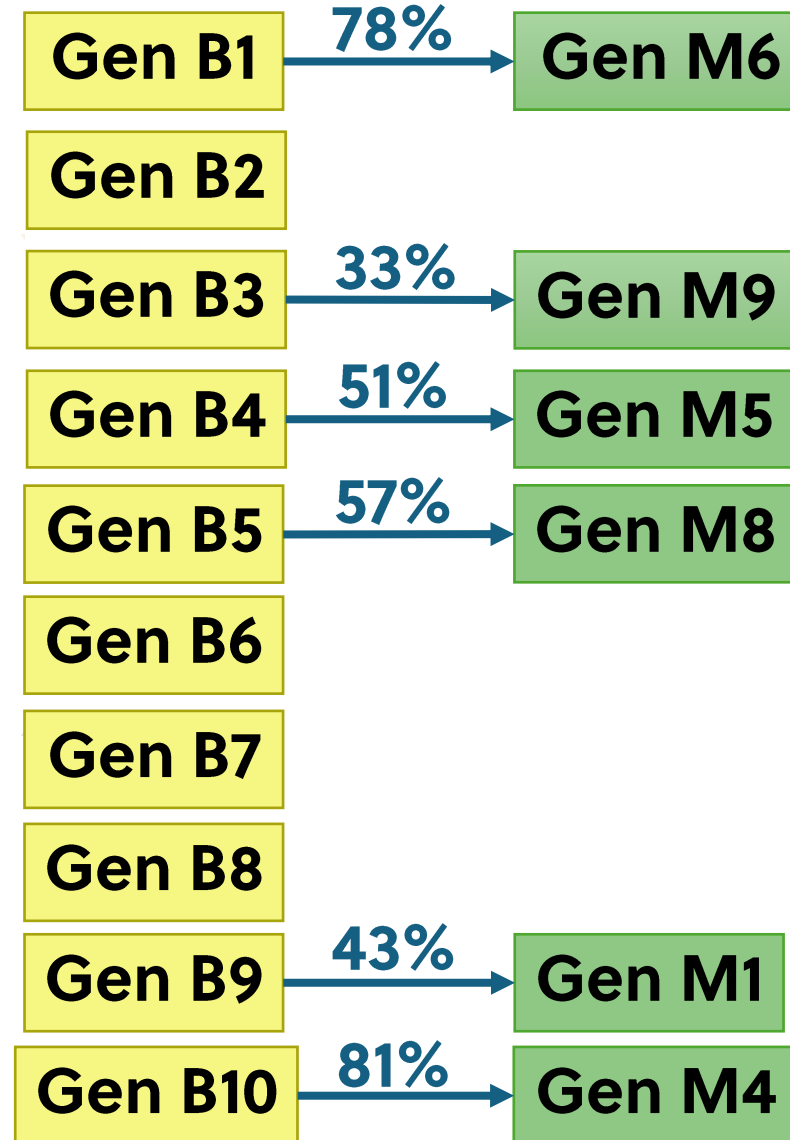
# Vergleich der proteinkodierenden Gene



# Vergleich der proteinkodierenden Gene



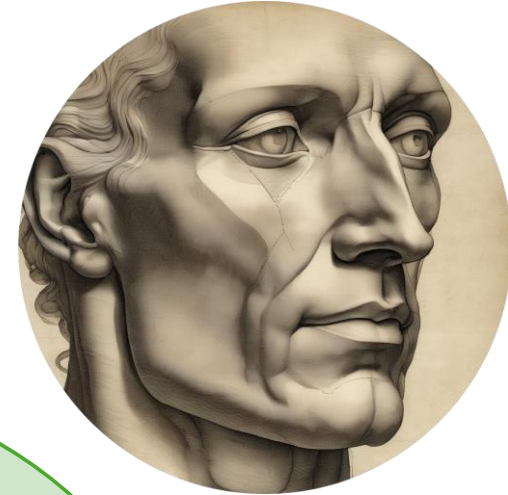
# Vergleich der proteinkodierenden Gene



# Vergleich der proteinkodierenden Gene

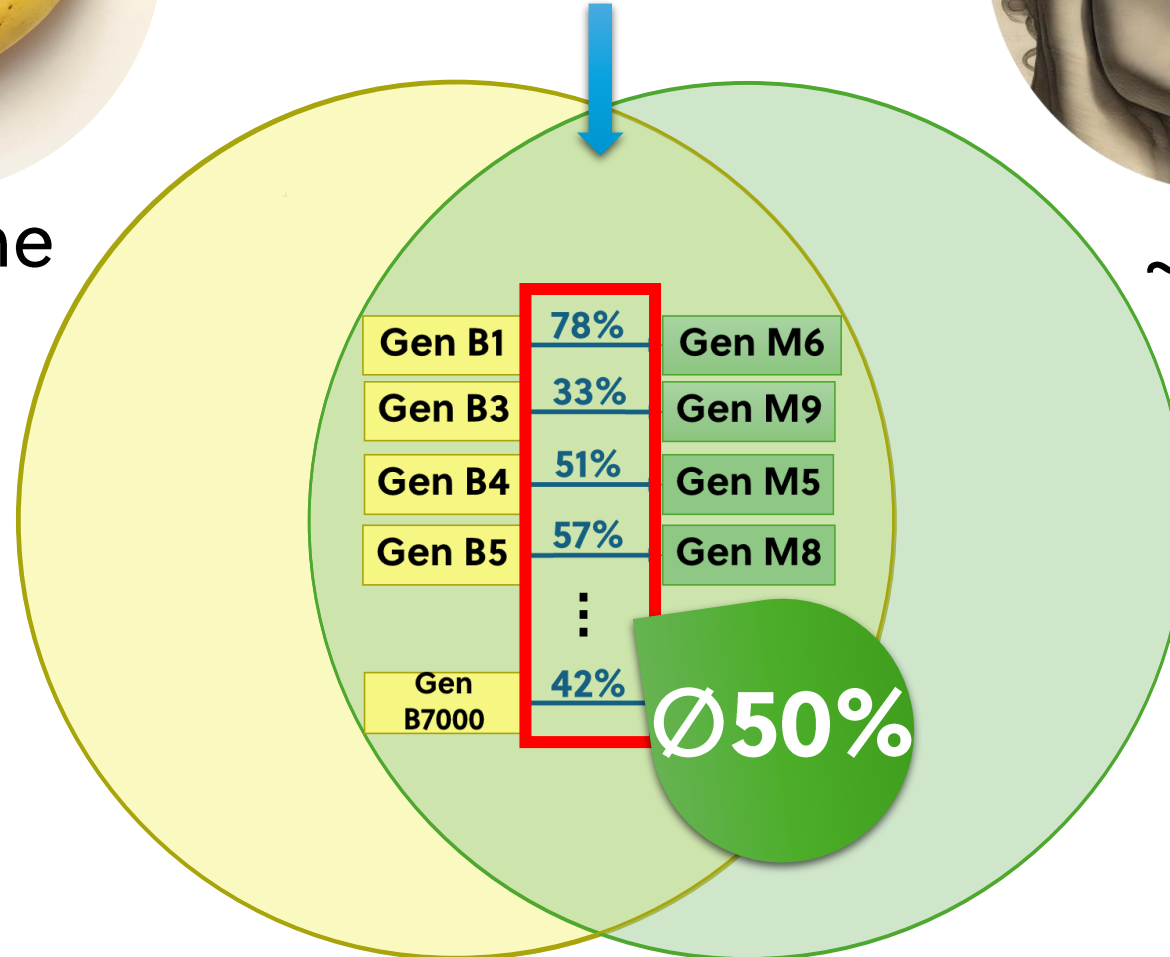


~30.700 Gene



~7.000 Gene

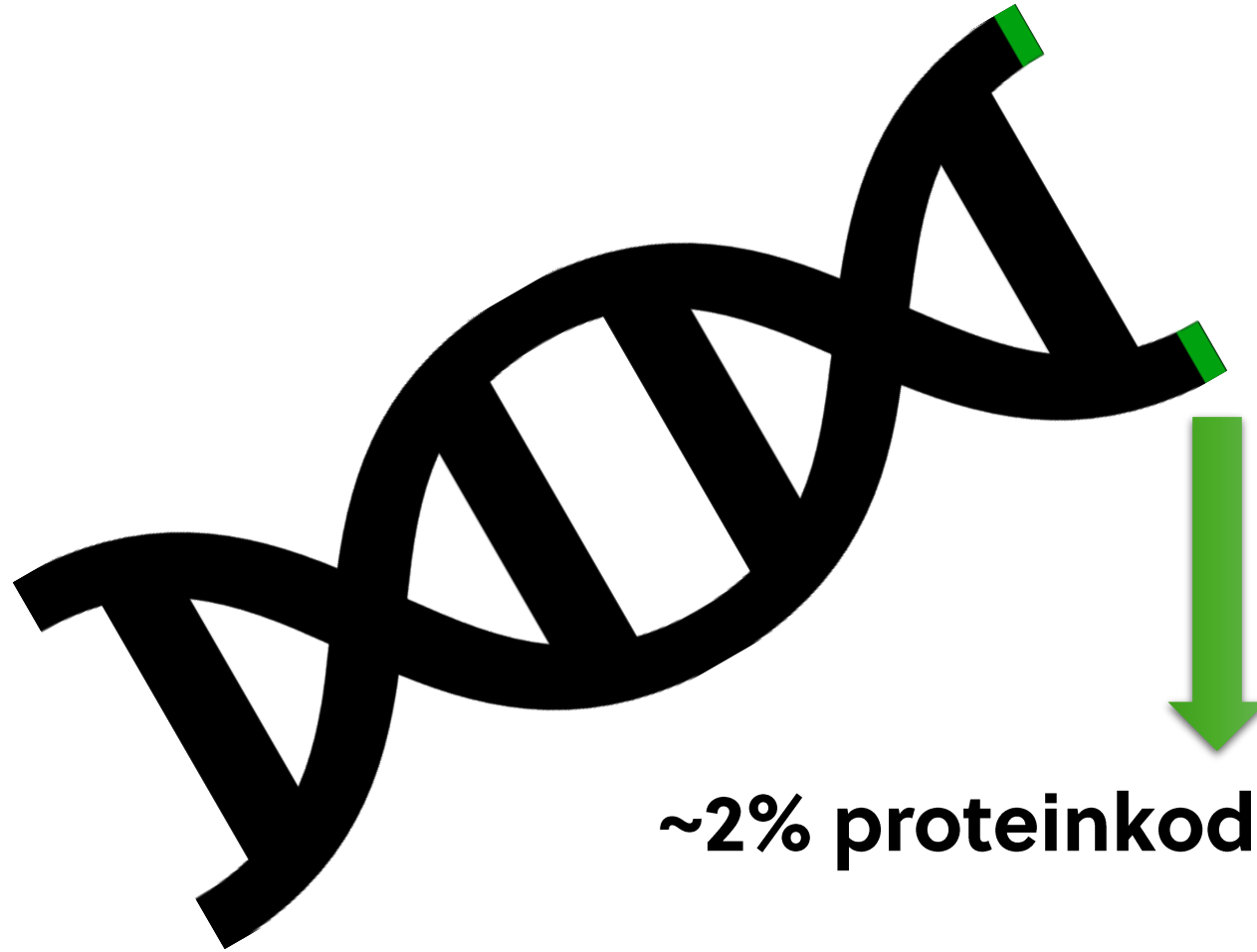
~20.100 Gene



# Anteil proteinkodierender Gene im Genom



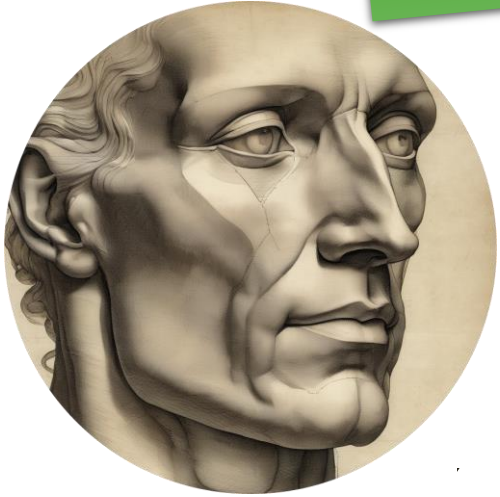
# Anteil proteinkodierender Gene im Genom



**~2% proteinkodierende Gene**

# Ähnlichkeit basierend auf vergleichbaren Gensequenzen

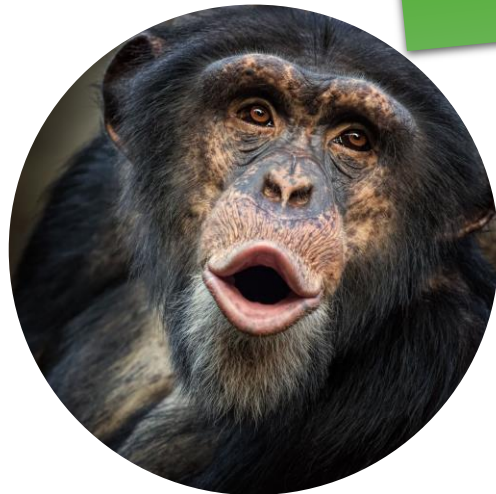
99,99%



85%



99%



50%



# Ähnlichkeit basierend auf den vergleichbaren Teilen des **Genom**





# Äpfel mit Birnen vergleichen



*Malus domestica*



17 Chromosomen



0,7 Milliarden  
Buchstaben

~36.000 Gene

??%



*Pyrus communis*



17 Chromosomen



0,6 Milliarden  
Buchstaben

~36.000 Gene

# Äpfel mit Birnen vergleichen



*Malus domestica*



17 Chromosomen



0,7 Milliarden  
Buchstaben

~36.000 Gene



*Pyrus communis*



17 Chromosomen

0,6 Milliarden  
Buchstaben

~36.000 Gene



~97%



# Äpfel mit Birnen vergleichen



*Malus domestica*



17 Chromosomen



0,7 Milliarden  
Buchstaben

~36.000 Gene



*Pyrus communis*



17 Chromosomen



0,6 Milliarden  
Buchstaben

~36.000 Gene

~85%



