Algorithms in Sequence Analysis

Ingo Ebersberger & Bardya Djahanschiri

Wo finde ich Informationen

Unterlagen (Folien, Infos, etc):

https://olat-ce.server.uni-frankfurt.de/olat/auth/RepositoryEntry/10780016646?0

http://applbio.biologie.uni-frankfurt.de/teaching/asa

user: TeachingAKE

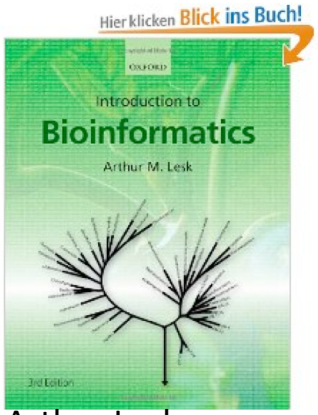
Passwort: TeachingAKE

slack

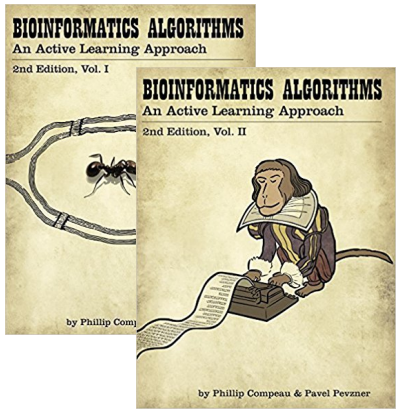
Grundlagen: (Folien, Übungsblätter, Infos, etc):

https://olat-ce.server.uni-frankfurt.de/olat/auth/RepositoryEntry/10770907140

Where can I read about the things I hear?

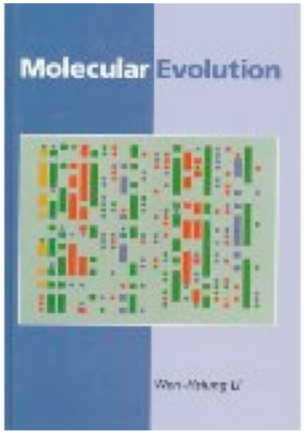


Arthur Lesk
Introduction to Bioinformatics (2008)
Oxford University Press

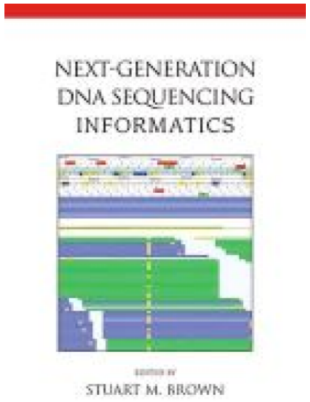


P. Compeau and P. Pevzner
Bioinformatics Algorithms (2016)
Active Learning Publishers, LLC

Dan Gusfield
Algorithms on Strings, Trees and Sequences (1997)
Cambridge University Press



W-H. Li
Molecular Evolution
Sinauer Associates



Stuart M. Brown
Next-Generation DNA sequencing Informatics
Cold Spring Harbor Laboratories

I will also identify the manuscripts describing the topics/algorithms addressed in this lecture

Where can I get additional information? (www.mygoblet.org)

Seminar

Modus 2021 – Vorbesprechung am 22. April (weitere Info folgt über slack)

Übung

- Im Anschluss an die Vorlesung
  - Pro Übung können 20 Punkte erreicht werden
  - Mit Hilfe der Übungspunkte können bis zu 10% der Klausurpunkte abgedeckt werden.
  - Copy&Paste Lösungen werden nicht anerkannt

Ablaufplan der Vorlesung

April	15.	Einführung in die Veranstaltung
	22.	Methoden der Hochdurchsatz-DNA Sequenzierung
	29.	Assemblierung von Genomen aus Whole Genome Shotgun Daten
Mai	06.	Die Verwendung von De Bruijn Graphen in Velvet
	13.	Feiertag (Christi Himmelfahrt)
	20.	Mapping & Referenz-basiertes Sequenz-Assembly
	27.	Algorithmen in der Analyse von RNA-Seq Daten
Juni	03.	Feiertag (Fronleichnam)
	10.	Sequenzalignments: schneller und/oder besser
	17.	Markov-Ketten und hidden Markov Modelle
	24.	Modelle und Algorithmen in der DNA Sequenzevolutions-Analyse
Juli	01.	Algorithmen in der Phylogenie-Rekonstruktion
	08.	Algorithmen in der Orthologensuche
	15.	Analyse von Clip- und Chip-Seq-Daten
August	05.	Klausur

Am Anfang steht die Sequenz

```
>hg19_dna
GAGGGTGGAGACGTCCTGGCCCCCGCCCCGCGTGCACCCCCAGGGGAGGC
CGAGCCCGCCGCCCGGCCCCGCGCAGGCCCCGCCCGGGACTCCCCTGCGG
TCCAGGCCGCGCCCCGGGCTCCGCGCCAGCCAATGAGCGCCGCCCGGCCG
GGCGTGCCCCCGCGCCCCAAGCATAAACCCTGGCGCGCTCGCGGCCCGGC
ACTCTTCTGGTCCCCACAGACTCAGAGAGAACCCACCATGGTGCTGTCTC
CTGCCGACAAGACCAACGTCAAGGCCGCCTGGGGTAAGGTCGGCGCGCAC
GCTGGCGAGTATGGTGCGGAGGCCCTGGAGAGGTGAGGCTCCCTCCCCTG
CTCCGACCCGGGCTCCTCGCCCGCCCGGACCCACAGGCCACCCTCAACCG
TCCTGGCCCCGGACCCAAACCCCACCCCTCACTCTGCTTCTCCCCGCAGG
ATGTTCCTGTCCTTCCCCACCACCAAGACCTACTTCCCGCACTTCGACCT
GAGCCACGGCTCTGCCCAGGTTAAGGGCCACGGCAAGAAGGTGGCCGACG
CGCTGACCAACGCCGTGGCGCACGTGGACGACATGCCCAACGCGCTGTCC
GCCCTGAGCGACCTGCACGCGCACAAGCTTCGGGTGGACCCGGTCAACTT
CAAGGTGAGCGGCGGGCCGGGAGCGATCTGGGTCGAGGGGCGAGATGGCG
CCTTCCTCGCAGGGCAGAGGATCACGCGGGTTGCGGGAGGTGTAGCGCAG
GCGGCGGCTGCGGGCCTGGGCCCTCGGCCCCACTGACCCTCTTCTCTGCA
CAGCTCCTAAGCCACTGCCTGCTGGTGACCCTGGCCGCCCACCTCCCCGC
CGAGTTCACCCCTGCGGTGCACGCCTCCCTGGACAAGTTCCTGGCTTCTG
TGAGCACCGTGCTGACCTCCAAATACCGTTAAGCTGGAGCCTCGGTGGCC
ATGCTTCTTGCCCCTTGGGCCTCCCCCCAGCCCCTCCTCCCCTTCCTGCA
CCCGTACCCCCGTGGTCTTTGAATAAAGTCTGAGTGGGCGGCAGCCTGTG
TGTGCCTGAGTTTTTTCCCTCAGCAAACGTGCCAGGCATGGGCGTGGACA
GCAGCTGGGACACACATGGCTAGAACCTCTCTGCAGCTGGATAGGGTAGG
AAAAGGCAGGGGCGGGAGGAGGGGATGGAGGAGGGAAAGTGGAGCCACCG
CGAAGTCCAGCTGGAAAAACGCTGGACCCTAGAGTGCTTTGA
```

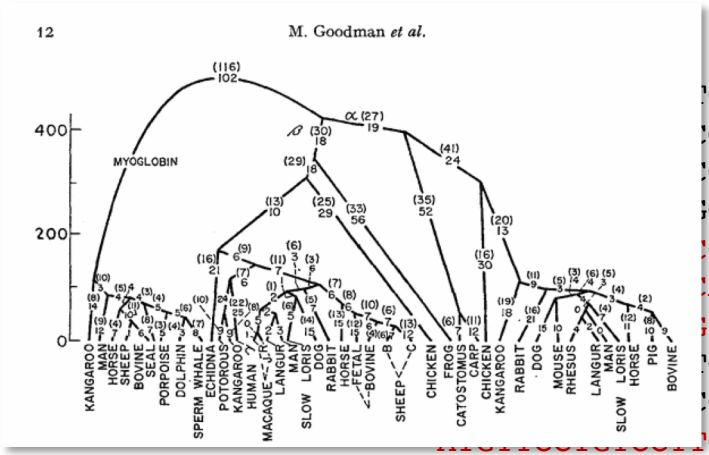
Am Anfang steht die Sequenz

```
>hg19_dna
GAGGGTGGAGACGTCCTGGCCCCCGCCCCGCGTGCACCCCCAGGGGAGGC
CGAGCCCGCCGCCCGGCCCCGCGCAGGCCCCGCCCGGGACTCCCCTGCGG
TCCAGGCCGCGCCCCGGGCTCCGCGCCAGCCAATGAGCGCCGCCCGGCCG
GGCGTGCCCCCGCGCCCCAAGCATAAACCCTGGCGCGCTCGCGGCCCGGC
ACTCTTCTGGTCCCCACAGACTCAGAGAGAACCCACCATGGTGCTGTCTC
CTGCCGACAAGACCAACGTCAAGGCCGCCTGGGGTAAGGTCGGCGCGCAC
GCTGGCGAGTATGGTGCGGAGGCCCTGGAGAGGTGAGGCTCCCTCCCCTG
CTCCGACCCGGGCTCCTCGCCCGCCCGGACCCACAGGCCACCCTCAACCG
TCCTGGCCCCGGACCCAAACCCCACCCCTCACTCTGCTTCTCCCCGCAGG
ATGTTCCTGTCCTTCCCCACCACCAAGACCTACTTCCCGCACTTCGACCT
GAGCCACGGCTCTGCCCAGGTTAAGGGCCACGGCAAGAAGGTGGCCGACG
CGCTGACCAACGCCGTGGCGCACGTGGACGACATGCCCAACGCGCTGTCC
GCCCTGAGCGACCTGCACGCGCACAAGCTTCGGGTGGACCCGGTCAACTT
CAAGGTGAGCGGCGGGCCGGGAGCGATCTGGGTCGAGGGGCGAGATGGCG
CCTTCCTCGCAGGGCAGAGGATCACGCGGGTTGCGGGAGGTGTAGCGCAG
GCGGCGGCTGCGGGCCTGGGCCCTCGGCCCCACTGACCCTCTTCTCTGCA
CAGCTCCTAAGCCACTGCCTGCTGGTGACCCTGGCCGCCCACCTCCCCGC
CGAGTTCACCCCTGCGGTGCACGCCTCCCTGGACAAGTTCCTGGCTTCTG
TGAGCACCGTGCTGACCTCCAAATACCGTTAAGCTGGAGCCTCGGTGGCC
ATGCTTCTTGCCCCTTGGGCCTCCCCCCAGCCCCTCCTCCCCTTCCTGCA
CCCGTACCCCCGTGGTCTTTGAATAAAGTCTGAGTGGGCGGCAGCCTGTG
TGTGCCTGAGTTTTTTCCCTCAGCAAACGTGCCAGGCATGGGCGTGGACA
GCAGCTGGGACACACATGGCTAGAACCTCTCTGCAGCTGGATAGGGTAGG
AAAAGGCAGGGGCGGGAGGAGGGGATGGAGGAGGGAAAGTGGAGCCACCG
CGAAGTCCAGCTGGAAAAACGCTGGACCCTAGAGTGCTTTGA
```

DNA/RNA/Protein?
DNA Replication?
Orientation of DNA?
Signal sequences in DNA?
Transcription and splicing?
Further mRNA processing?
Translation/Genetic Code?

```
>hg19_protein
MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLS
FPTTKTYFPHFDLSHGSAQVKGHGKKVADALTNAVAHV
DDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAA
HLPAEFTPAVHASLDKFLASVSTVLTSKYR
```

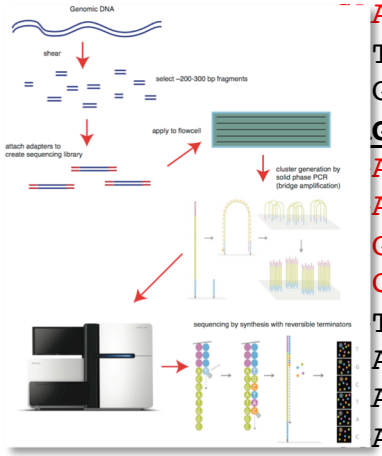
Am Anfang steht die Sequenz



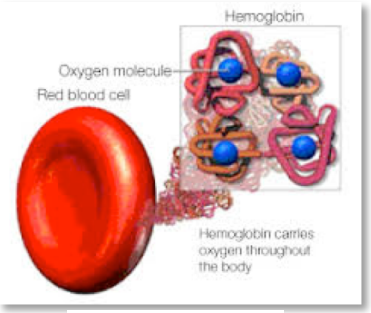
Evolutionäre Geschichte

Struktur

CCTGGCCCCCGCCCCGCGTGCACCCCCAG
GGCCCCGCGCAGGCCCCGCCCGGGACTCC
CGGGCTCCGCGCCAGCCAATGAGCGCCGC
CCCCAAGCATAAACCCTGGCGCGCTCGCG
CACAGACTCAGAGAGAACCCACC**ATG**GTG
AACGTCAAGGCCGCCTGGGGTAAGGTCGG
TGCGGAGGCCCTGGAGAG**GT**GAGGCTCCC
CCTCGCCCGCCCGGACCCACAGGCCACCC
CCAAACCCCACCCCTCACTCTGCTTCTCC
CCCCACCACCAAGACCTACTTCCCGCACT
TCTGCCCAGGTTAAGGGCCACGGCAAGAAGGTGGCCGA
CGCCGTGGCGCACGTGGACGACATGCCCAACGCGCTGT
GCCCTGAGCGACCTGCACGCGCACAAGCTTCGGGTGGACCCGGTCAACTT
AG**GT**GAGCGGCGGGCCGGGAGCGATCTGGGTCGAGGGGCGAGATGGCG
TTCCTCGCAGGGCAGAGGATCACGCGGGTTGCGGGAGGTGTAGCGCAG
GGCGGCTGCGGGCCTGGGCCCTCGGCCCCACTGACCCTCTTCTCTGCA
GCTCCTAAGCCACTGCCTGCTGGTGACCCTGGCCGCCCACCTCCCCGC
AGTTCACCCCTGCGGTGCACGCCTCCCTGGACAAGTTCCTGGCTTCTG
AGCACCGTGCTGACCTCCAAATACCGT**TAA**GCTGGAGCCTCGGTGG
GCTTCTTGCCCCTTGGGCCTCCCCCCAGCCCCTCCTCCCCTTCCTG
CGTACCCCCGTGGTCTTTG**AATAAA**GTCTGAGTGGGCGGCAGCCTG
TGCCTGAGTTTTTTCCCTCAGCAAACGTGCCAGGCATGGGCGTGGA
AGCTGGGACACACATGGCTAGAACCTCTCTGCAGCTGGATAGGGTA
AAGGCAGGGGCGGGAGGAGGGGATGGAGGAGGGAAAGTGGAGCCAC
AAGTCCAGCTGGAAAAACGCTGGACCCTAGAGTGCTTTGA



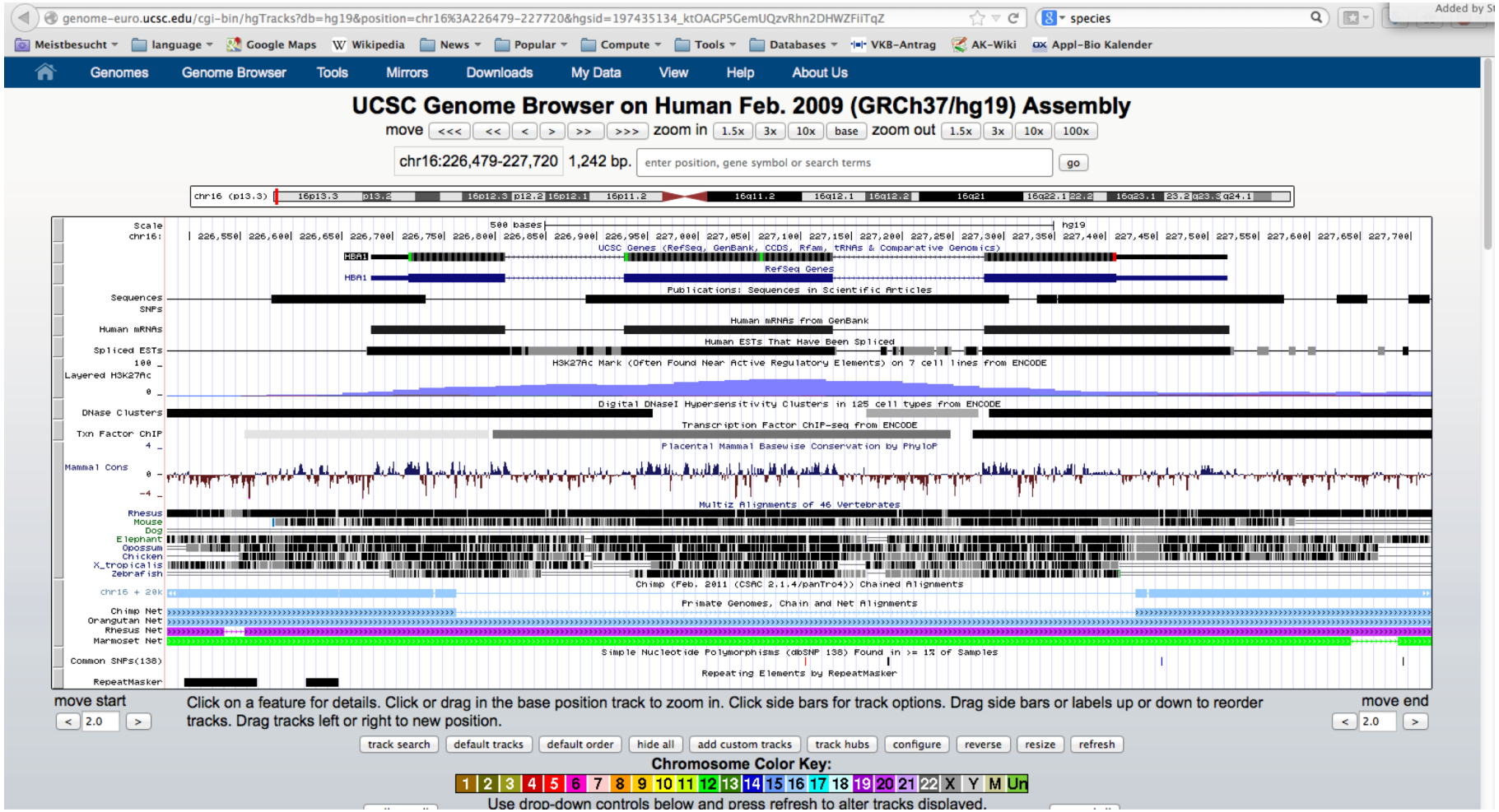
Daten-Historie



Funktion

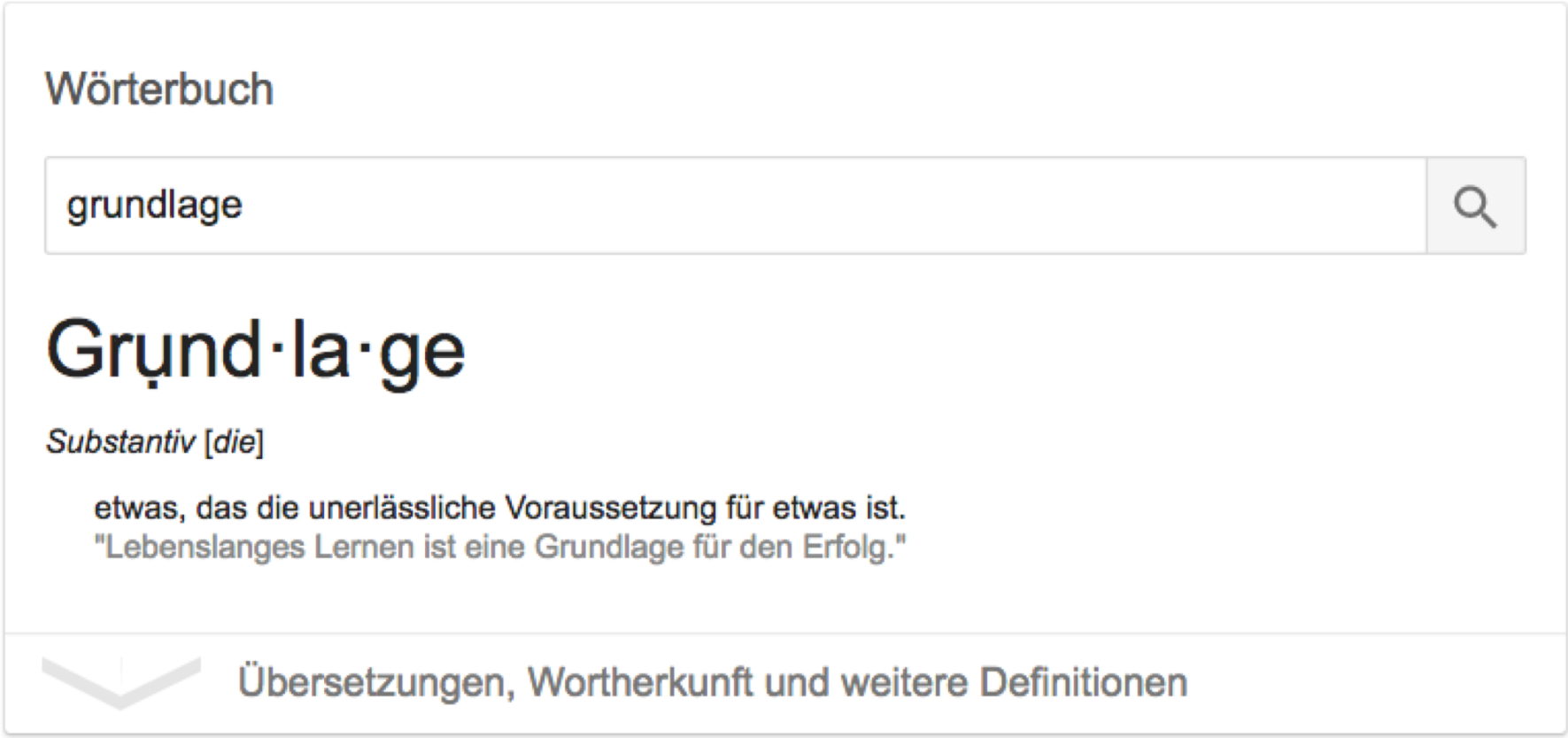
Tree from Goodman et al. J. Molec. Evol (1974) 3:1-48

Am Ende steht die Annotation



http://genome.ucsc.edu

Biologische Grundlagen



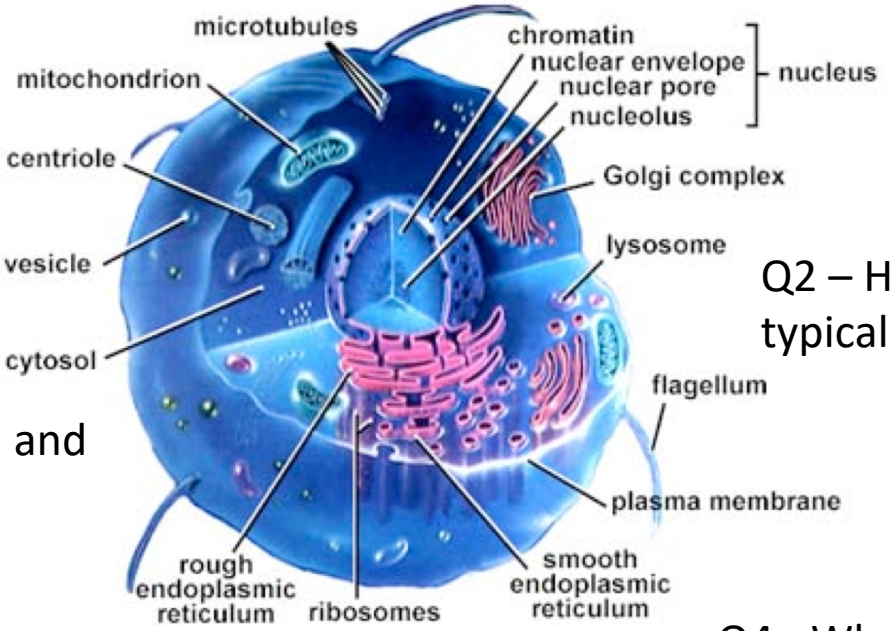
Grundlage 0: Characteristics of a eukaryotic cell

Q1 - How many different genomes do we have in a eukaryotic cell?

Q3 - Why were the first analyses of DNA made with mtDNA?

Q9 - What is 'alternative splicing'?

Q6 - What does the endosymbiont hypothesis postulate?



Q2 – How many genes does a typical eukaryote have?

Q7 - Where are DNA, RNA, and proteins synthesized?

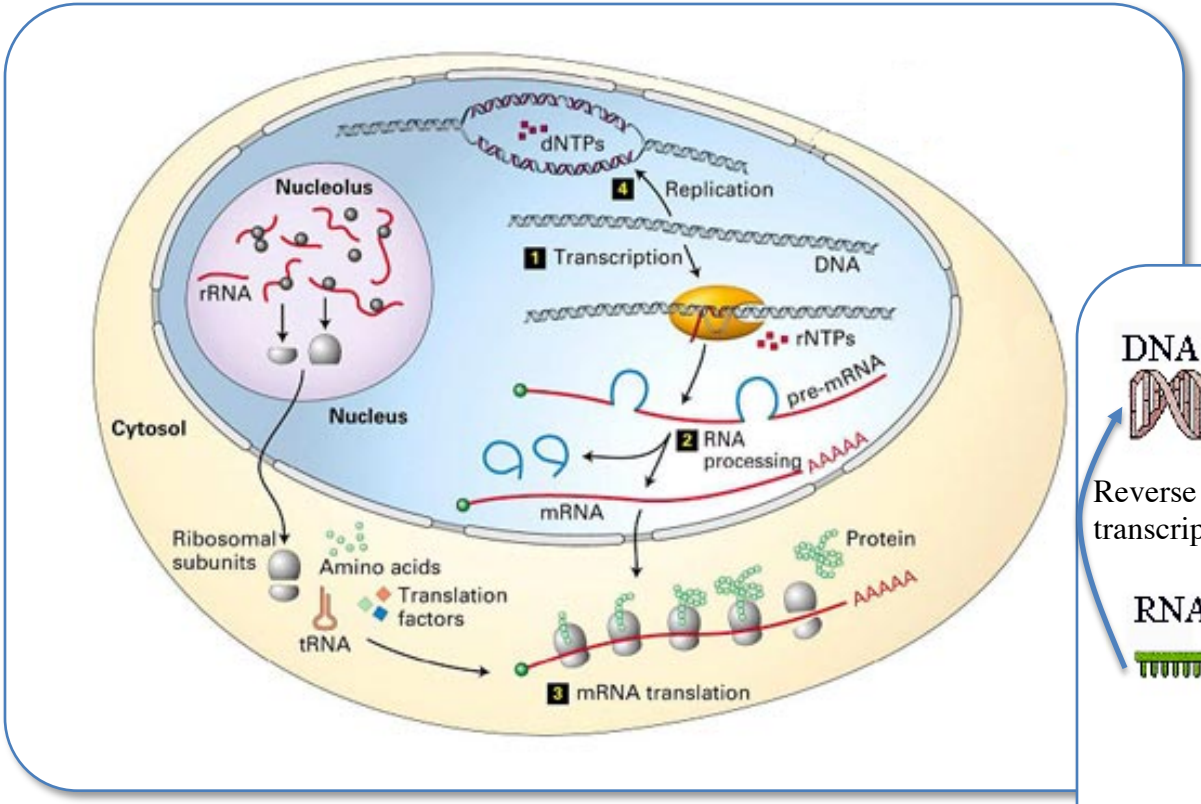
Q4 - What does a ribosome do?

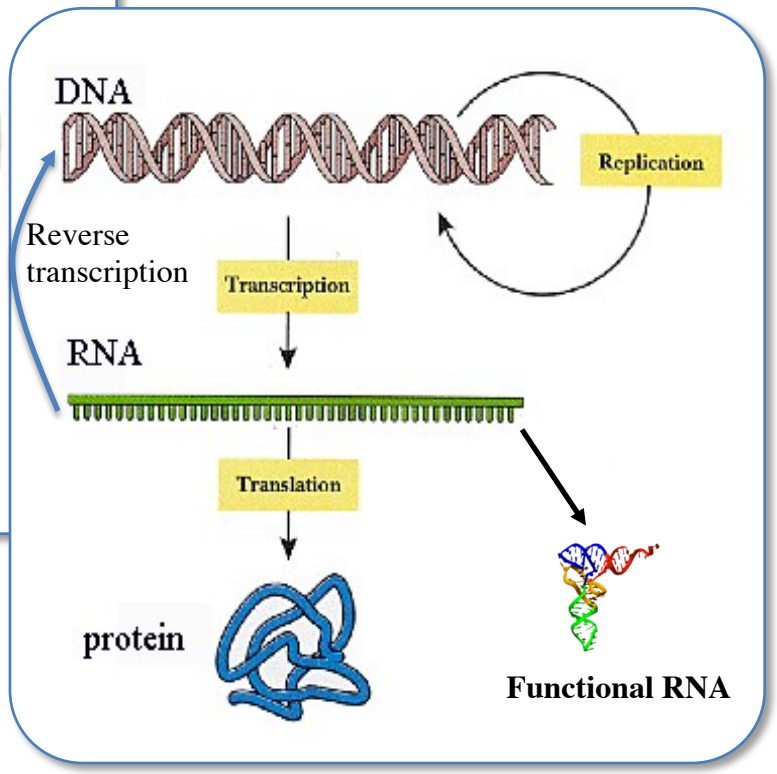
Q8 – What are exons, what are introns, and what characterizes the process 'splicing'?

Q5 - What does a mitochondrium do?

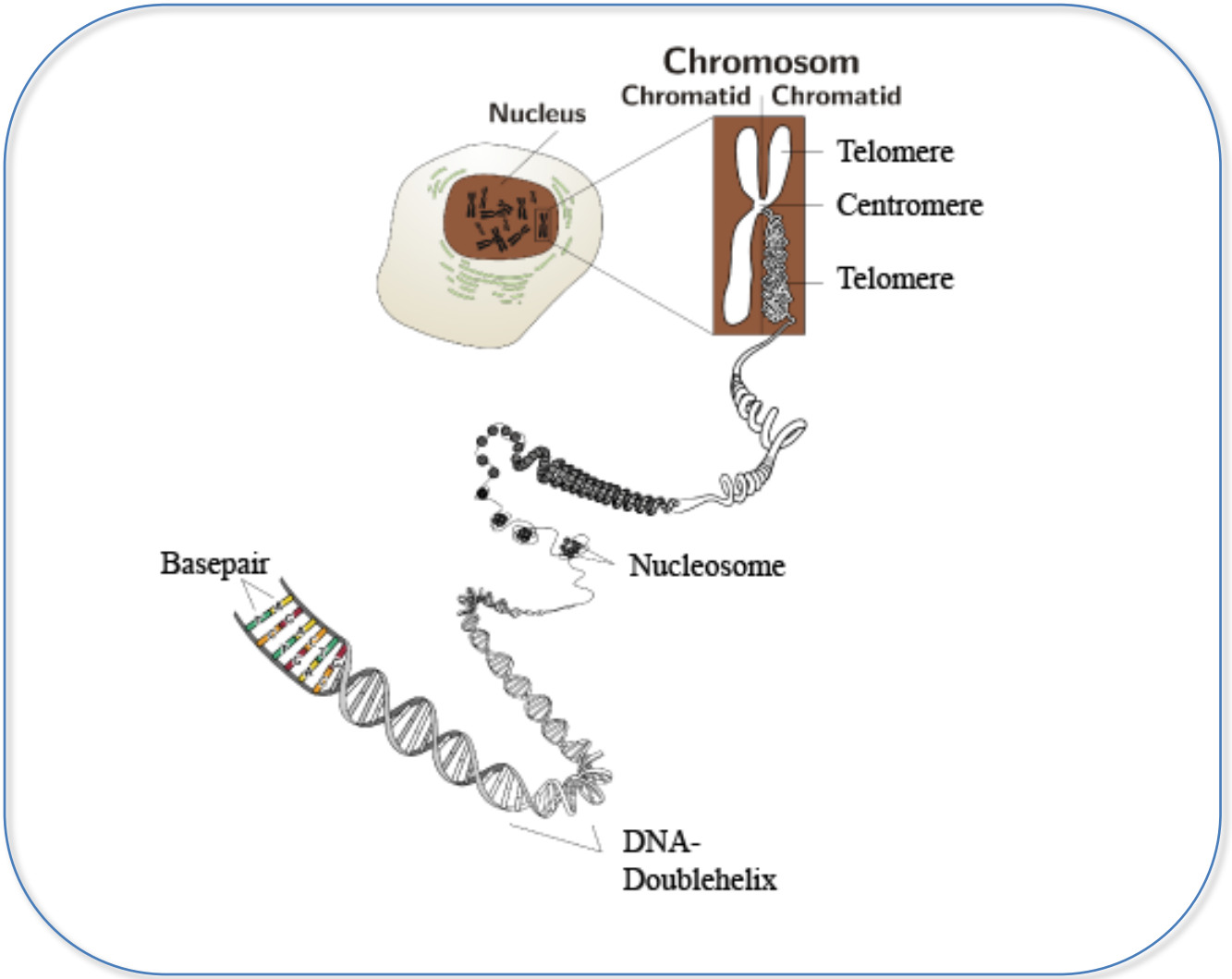
Q10 – What are the main differences to a bacterial (prokaryotic) cell?

Grundlage 1: Informationsfluss in einer Zelle (Das ‚zentrale Dogma' der Molekularbiologie)

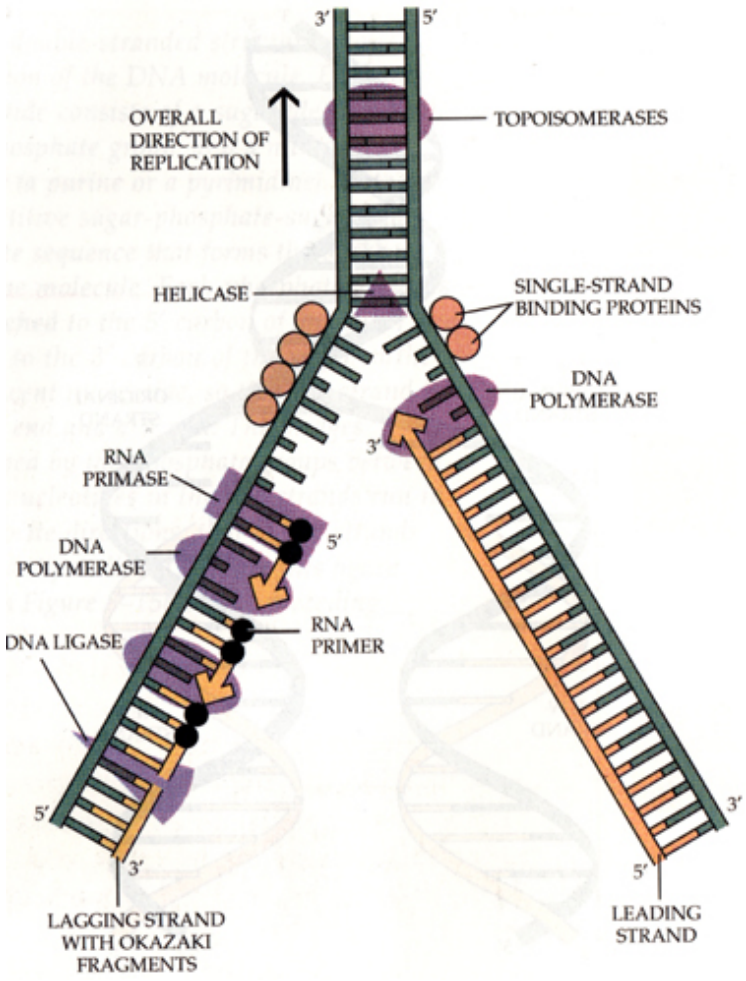




Grundlage 2: Organisation der genetischen Information in einer Zelle



Grundlage 3: DNA-Replikation ist semi-konservativ



Die 3 Grundprinzipien enzymatischer DNA Synthese:

- 5’➔3’
- template guided
- primed

Meselson, M. and Stahl, F.W. (1958) *The Replication of DNA in Escherichia Coli*. PNAS 44:671-682.

Grundlage 4: Die molekularen Prinzipien der enzymatischen DNA Synthese

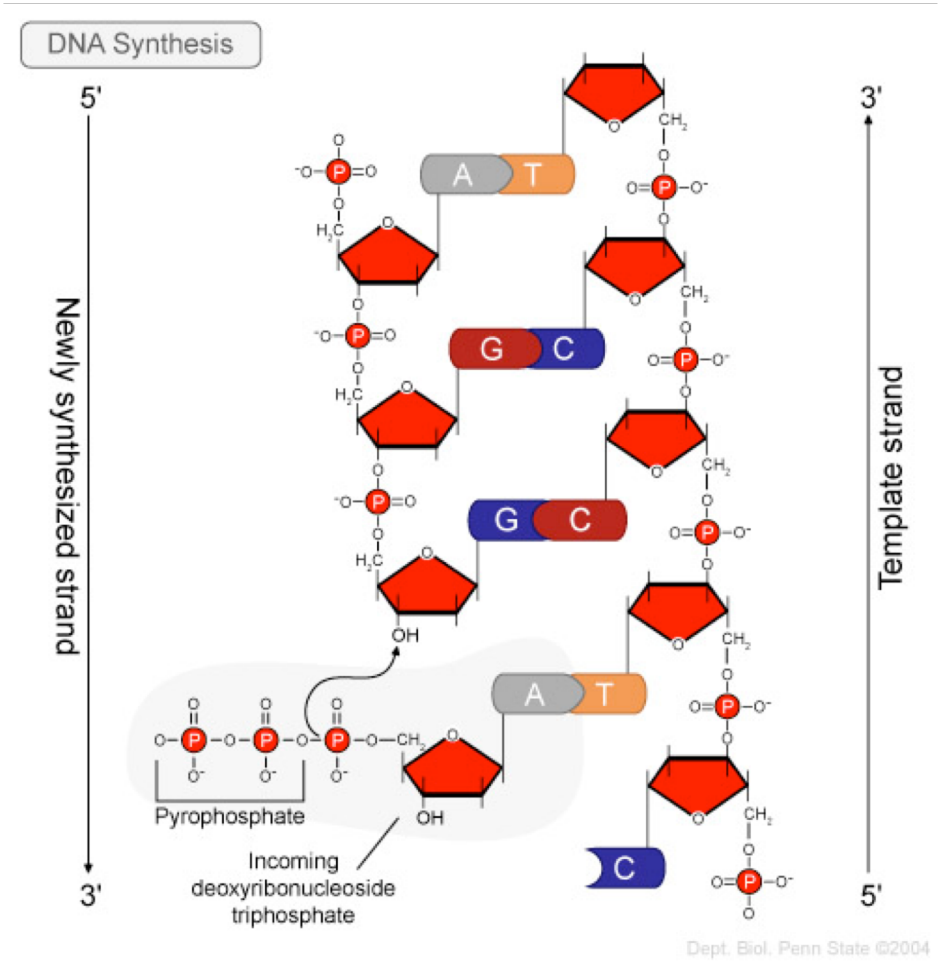
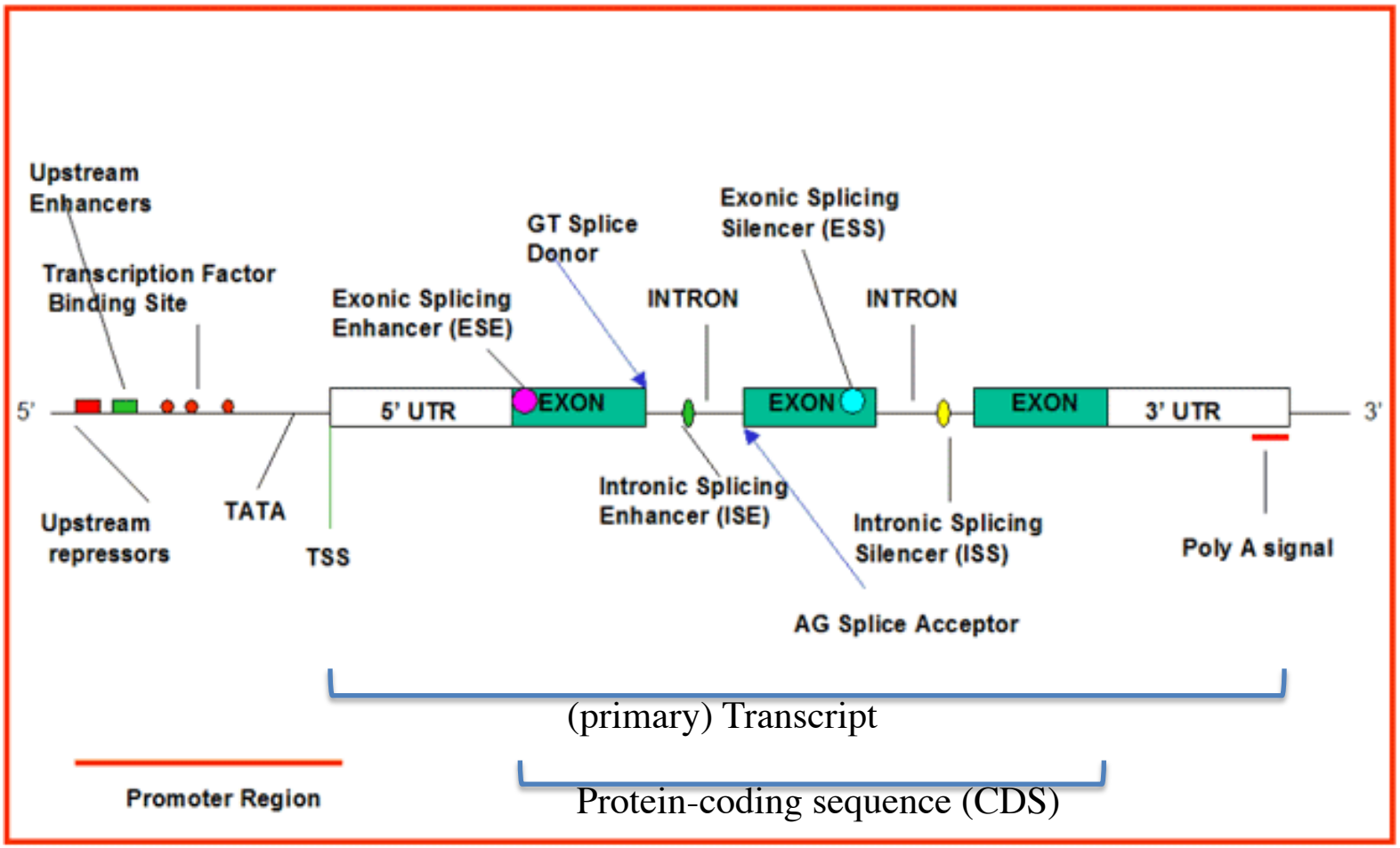


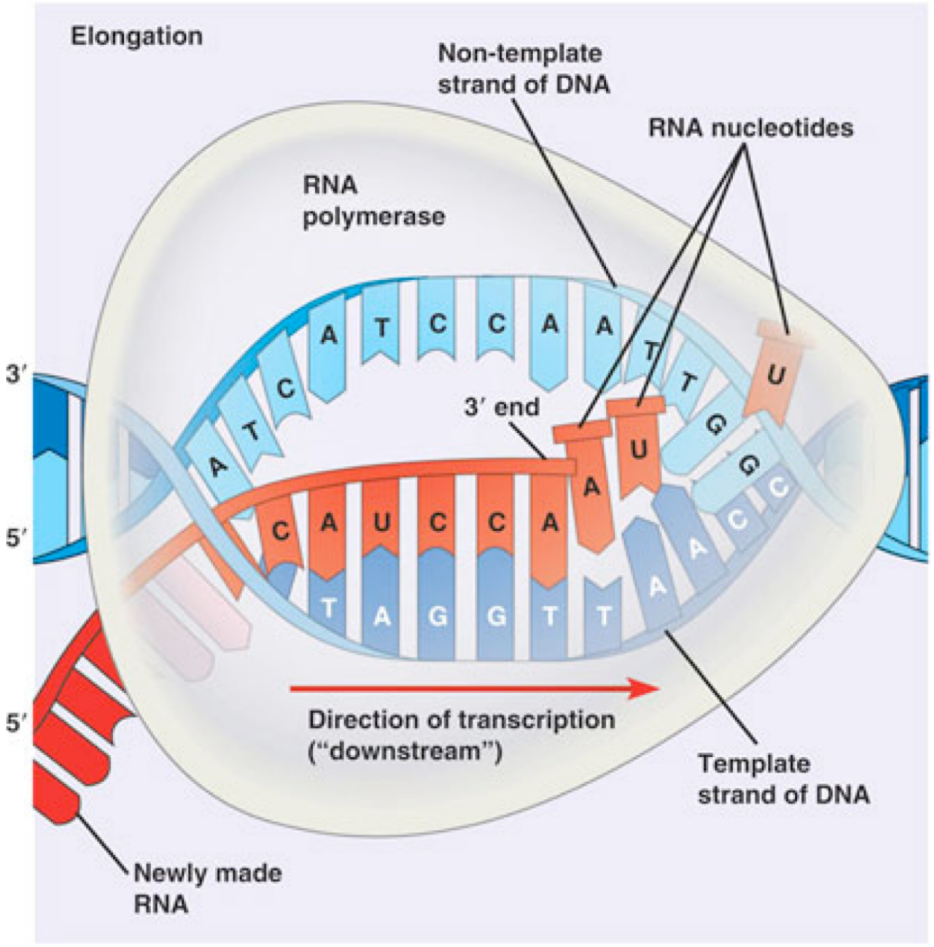
Image from: https://wikispaces.psu.edu/display/Biol230WFall09/DNA+Replication+and+Repair

Grundlage 5: Ein typisches eukaryotisches Protein-kodierendes Gen besteht aus unterschiedlichen funktionellen Bausteinen



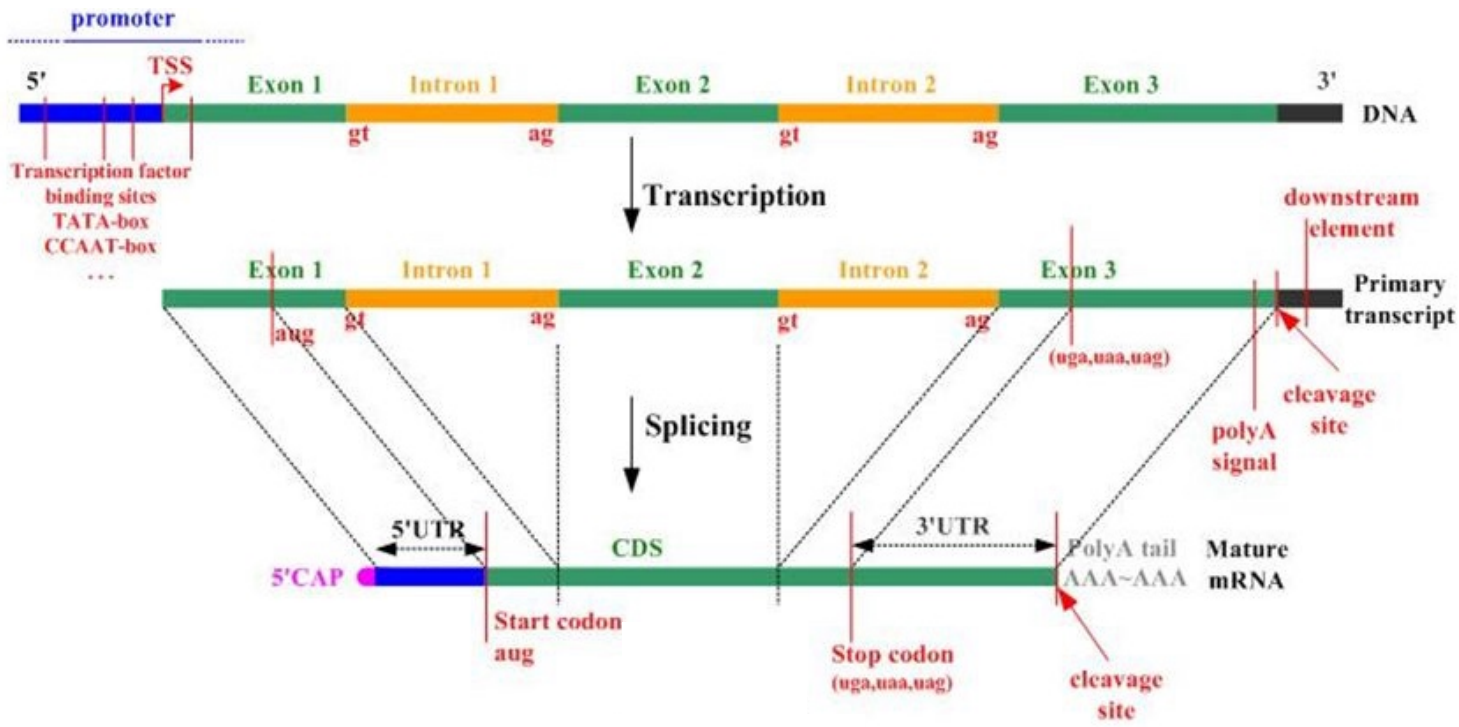
TSS: Transcription start site

Grundlage 6: Im Rahmen der **Transkription** wird die in der DNA kodierte Information auf ein mobiles Molekül kopiert, die **messenger RNA (mRNA)**



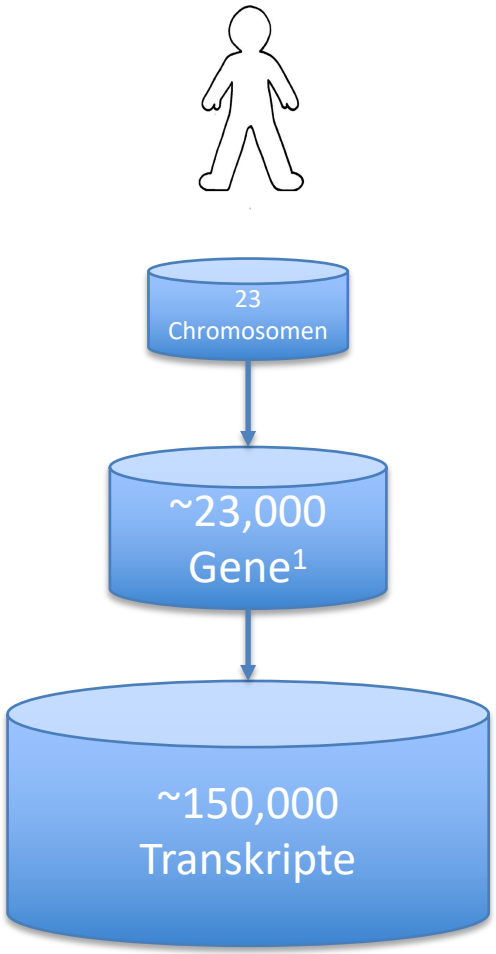
*make sure to get the orientation and direction of synthesis right

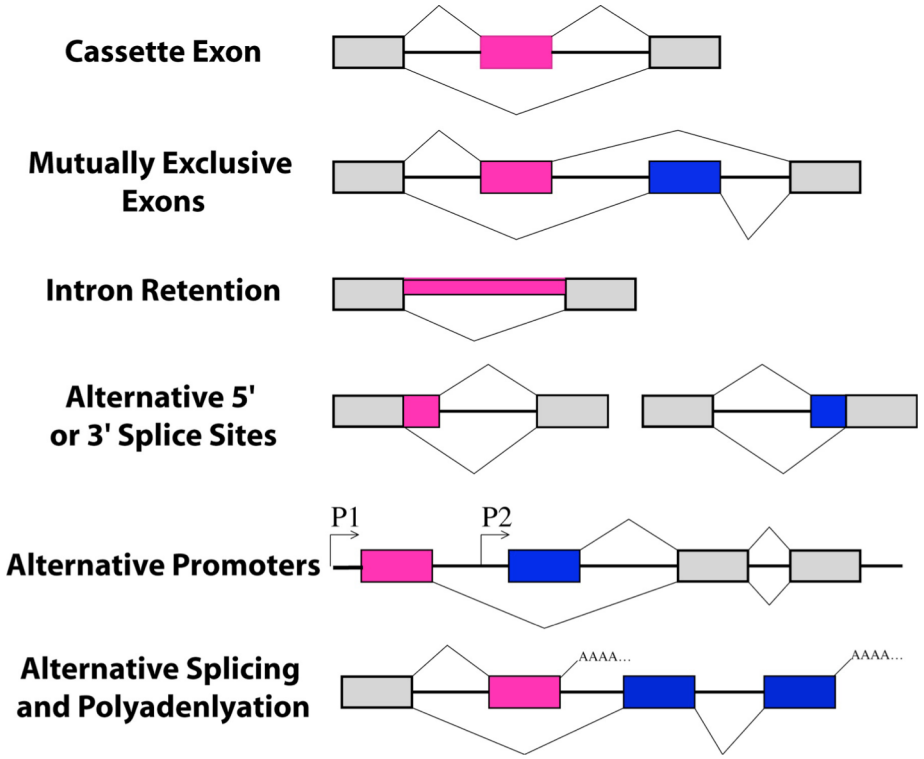
Aus dem Gen auf Ebene der genomischen DNA wird über ein primäres Transkript die reife mRNA gebildet



Adapted in part from
http://online.itp.ucsb.edu/online/infobio01/burge/

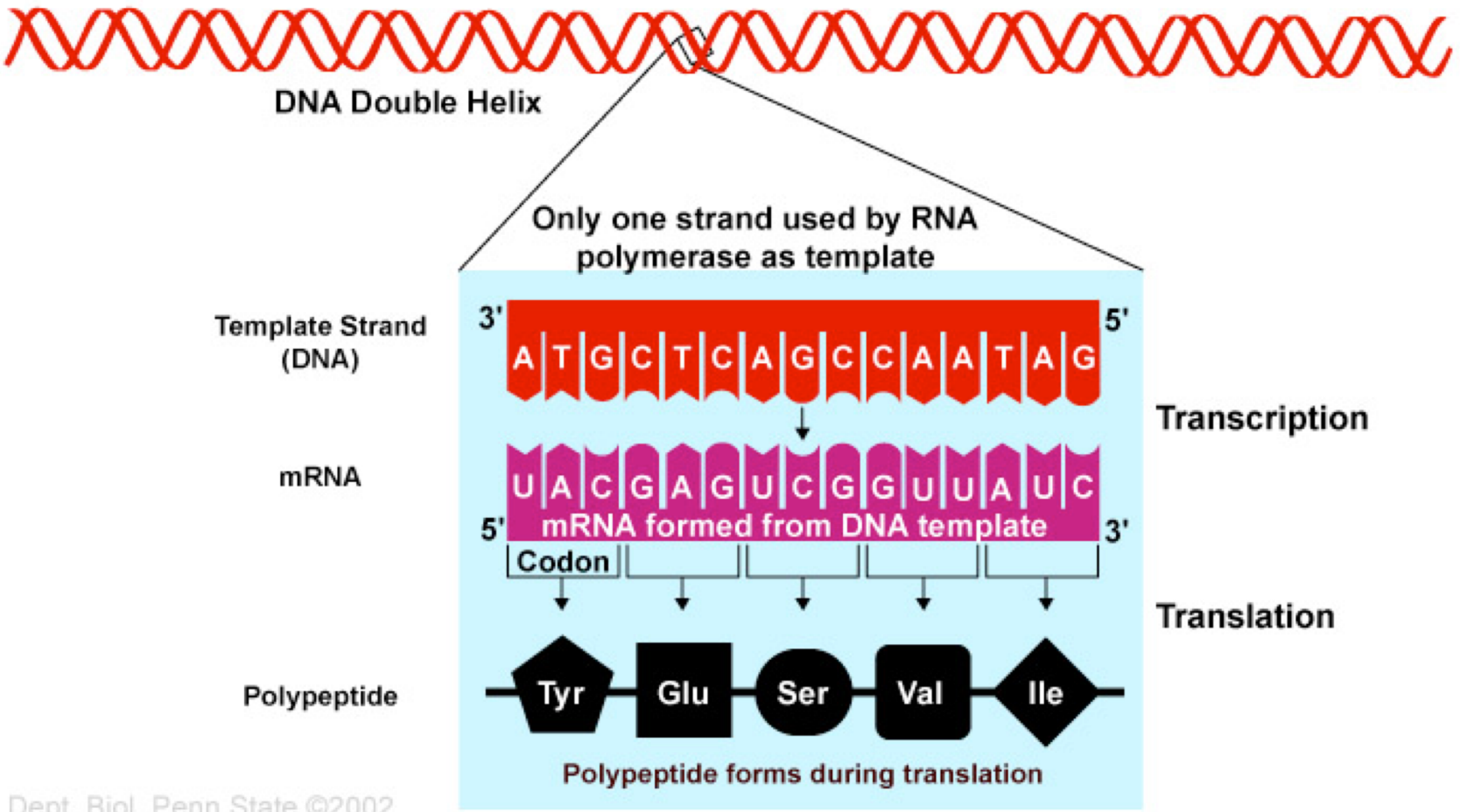
Unterschiedliche Exon-Usage kann aus einem Gen viele Genprodukte generieren



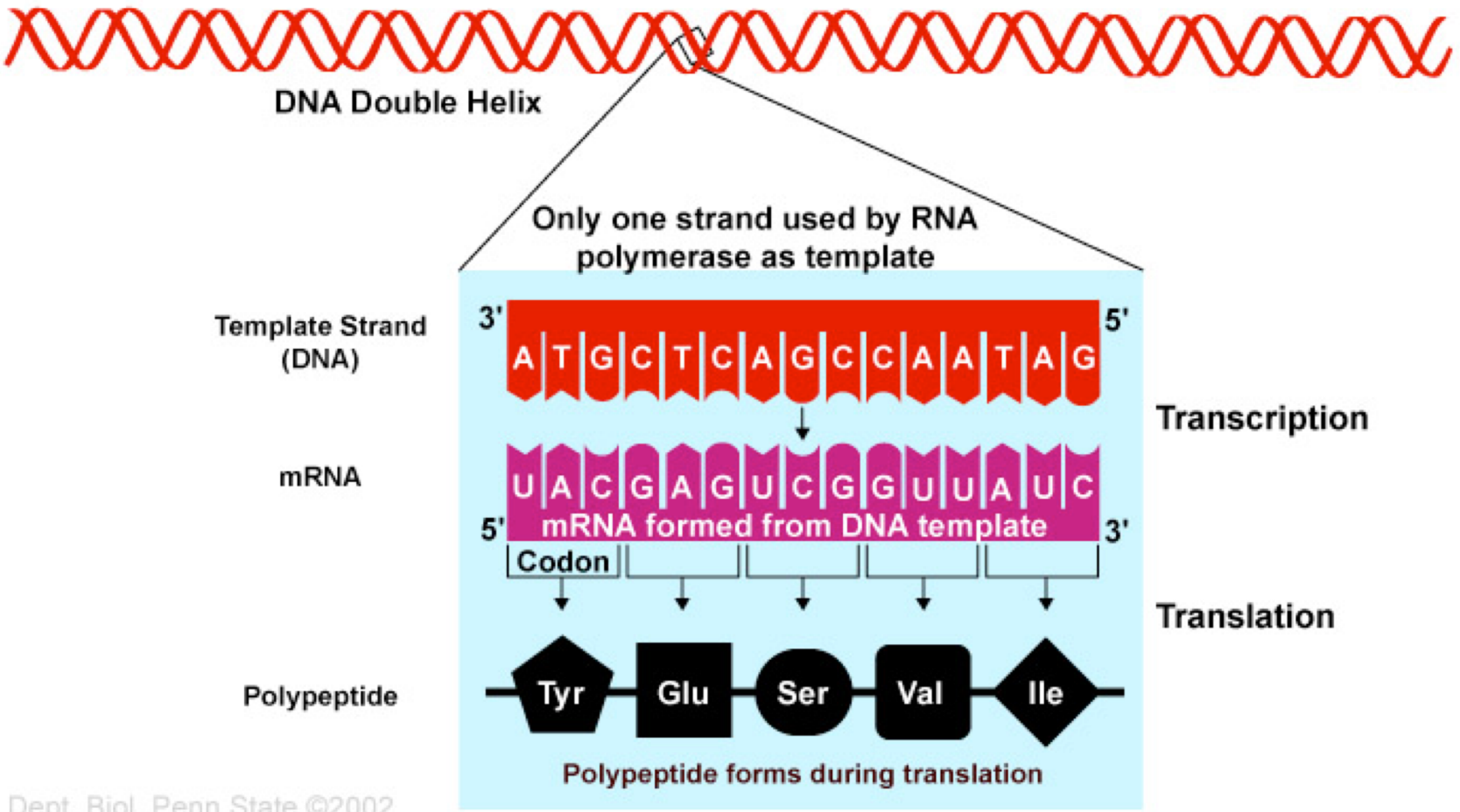


1. Protein-kodierende „known genes“

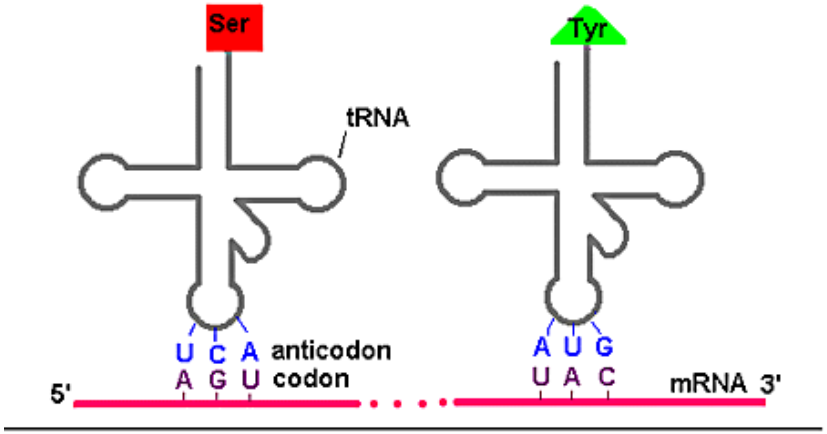
Grundlage 7: Die Information ist in Form von Basentriplets (Codons) abgelegt



Die Information ist in Form von Basentriplets (Codons) abgelegt



Der genetische Code…
ist degeneriert* (und es gibt mehr als nur den Standard-Code!!)



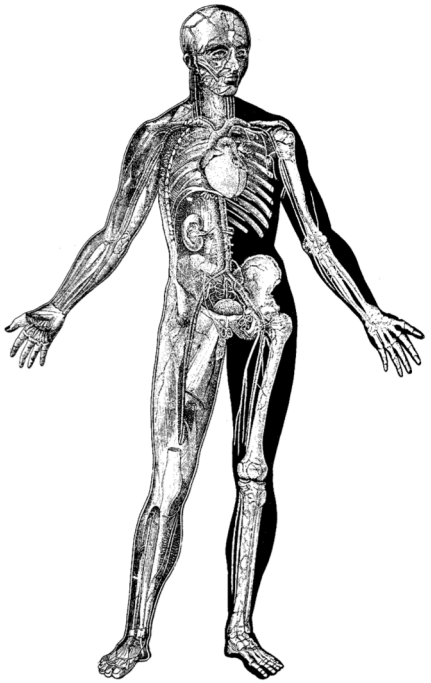
2nd base in codon

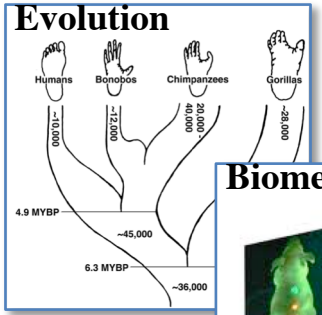
1st base in codon	U	C	A	G	3rd base in codon
U	Phe Phe Leu Leu	Ser Ser Ser Ser	Tyr Tyr STOP STOP	Cys Cys STOP Trp	U C A G
C	Leu Leu Leu Leu	Pro Pro Pro Pro	His His Gln Gln	Arg Arg Arg Arg	U C A G
A	Ile Ile Ile Met	Thr Thr Thr Thr	Asn Asn Lys Lys	Ser Ser Arg Arg	U C A G
G	Val Val Val Val	Ala Ala Ala Ala	Asp Asp Glu Glu	Gly Gly Gly Gly	U C A G

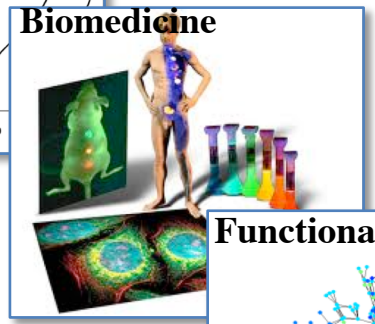
Merke:
Die Existenz unterschiedlicher genetischer Codes wird bei der in-silico Translation häufig übersehen. Daraus resultieren falsch vorhergesagte Proteinsequenzen!

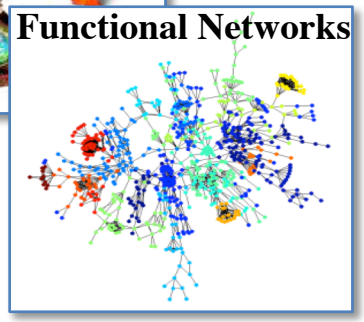
*manche Aminosäuren werden durch mehr als ein Triplet kodiert

Ziel bioinformatischer Sequenzanalyse









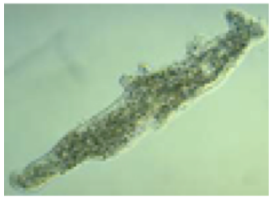
etc…

Wir möchten die funktionellen, regulatorischen und evolutionären Netzwerke in einem Organismus entschlüsseln und verstehen

In den überwiegenden Fällen arbeiten wir mit DNA Sequenzen oder Information, die wir von DNA Sequenzen abgeleitet haben.

Die Genomgrößen einer Reihe von Arten

Homo sapiens
3,200,000,000



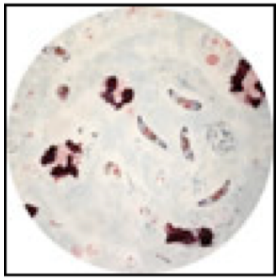
Amoeba proteus
290,000,000,000

Amoeba dubia
670,000,000,000

Bufo bufo
6,900,000,000

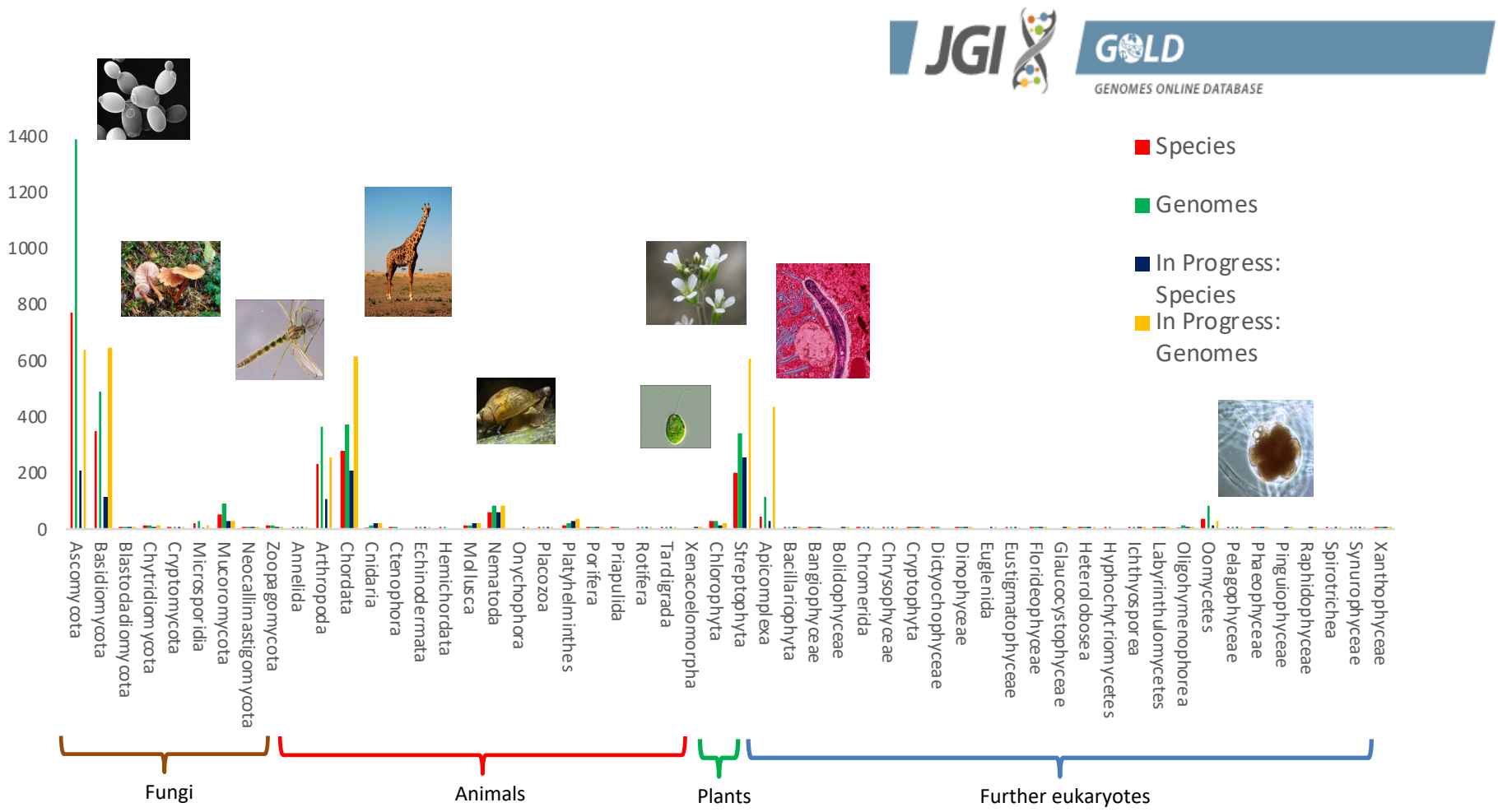
Rhinolophus ferrumequinum
1,929,400,000

Human immunodeficiency virus type 1
19,750



Plasmodium falciparum
25,000,000

Was wurde bis 2018 sequenziert?



Womit wir bald rechnen können

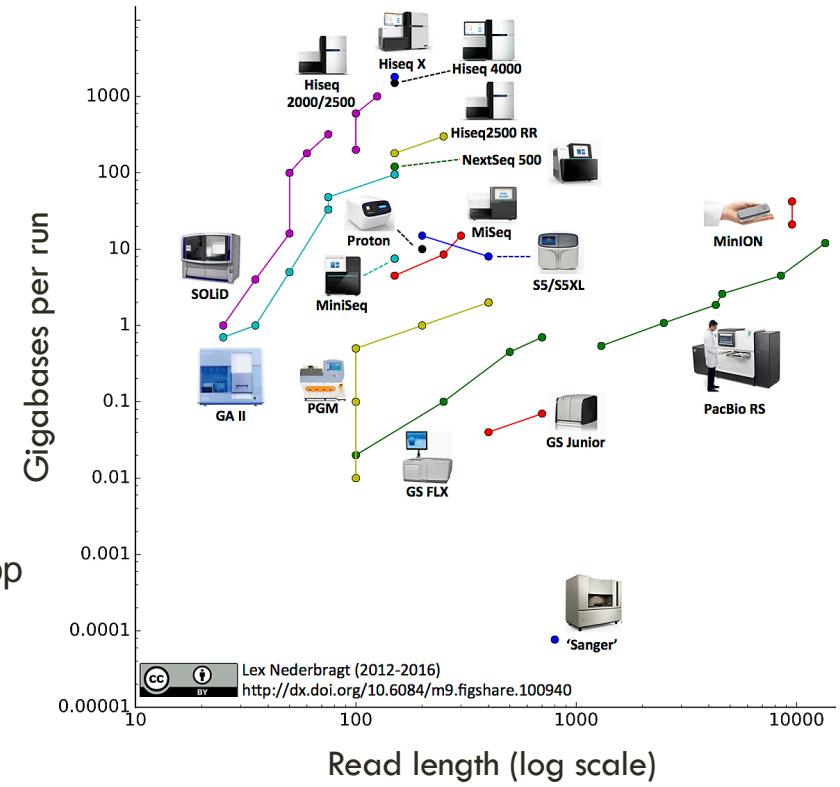
(…)For the first time in history, it is possible to efficiently ***sequence the genomes of all known species****, and to use genomics to help discover the remaining 80 to 90 percent of species that are currently hidden from science.*

HOW DO WE SEQUENCE DNA?

1st generation (1977)
- **Sanger** method: Sequencing by synthesis
- **Maxam-Gilbert** method: chemical sequencing

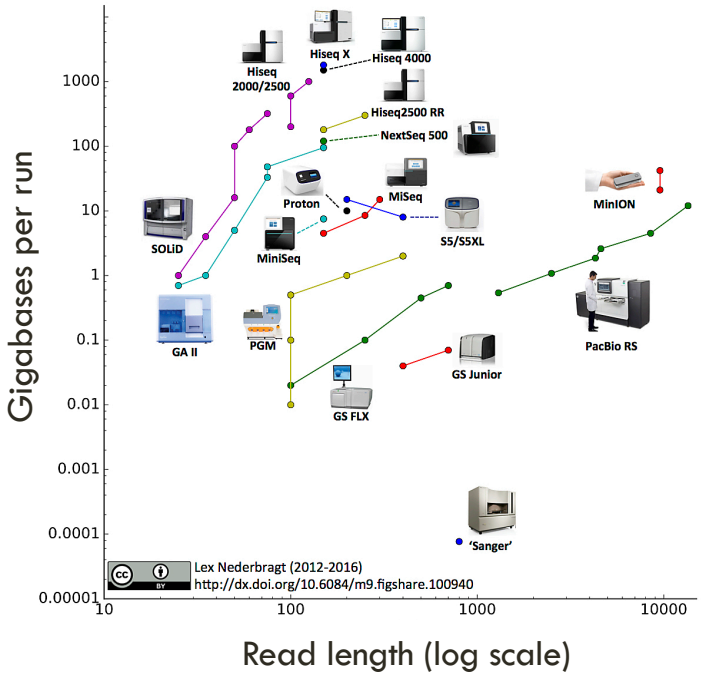
2nd generation ("next generation"; 2005)
- **454** - pyrosequencing
- **SOLiD** – sequencing by ligation
- **Illumina** – sequencing by synthesis
- **Ion Torrent** – ion semiconductor
- **Pac Bio** – Single Molecule Real-Time sequencing, 1000 bp

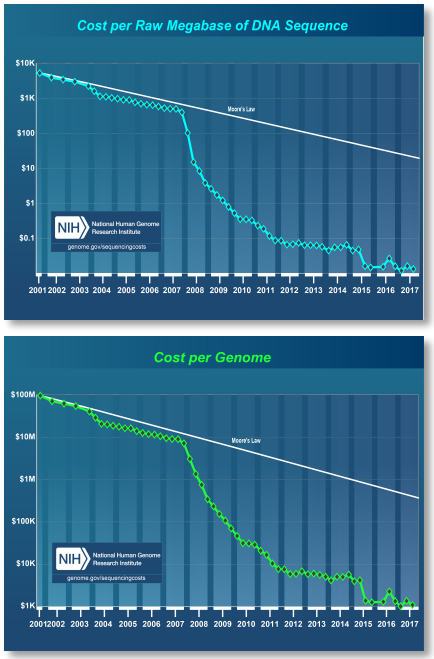
3rd generation (2015)
- **Pac Bio** – SMRT, Sequel system, 20,000 bp
- **Nanopore** – ion current detection
- **10X Genomics** – novel library prep for Illumina



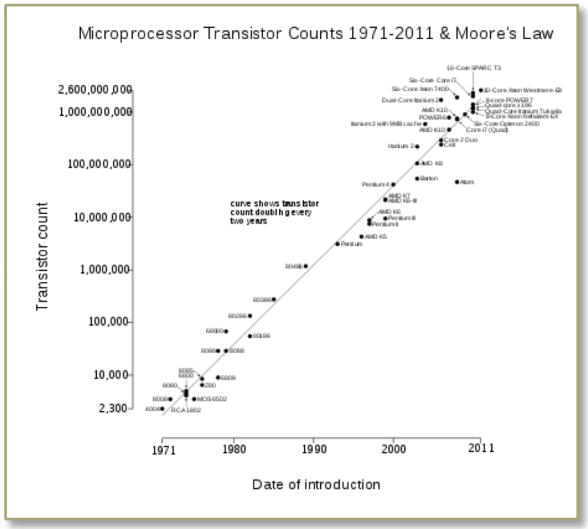
Del Angel et al. (2018) F1000 Research 7(ELIXIR):148

SEQUENCE DATA GROWS FASTER THAN COMPUTER POWER





Source: https://www.genome.gov/sequencingcostsdata/



Source: wikipedia

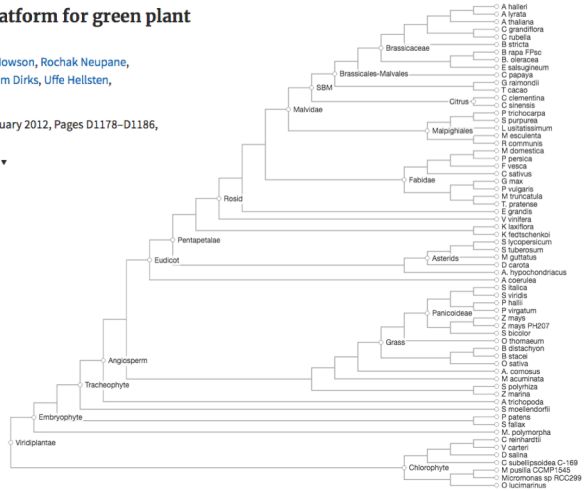
Die Sequenzierung gesamter Genome – Wo findet man die Daten?

Phytozome: a comparative platform for green plant genomics

David M. Goodstein, Shengqiang Shu, Russell Howson, Rochak Neupane, Richard D. Hayes, Joni Fazo, Therese Mitros, William Dirks, Uffe Hellsten, Nicholas Putnam ... Show more

Nucleic Acids Research, Volume 40, Issue D1, 1 January 2012, Pages D1178–D1186,
https://doi.org/10.1093/nar/gkr944
Published: 22 November 2011 **Article history**

[Phytozome]



[Flybase]

[Wormbase]

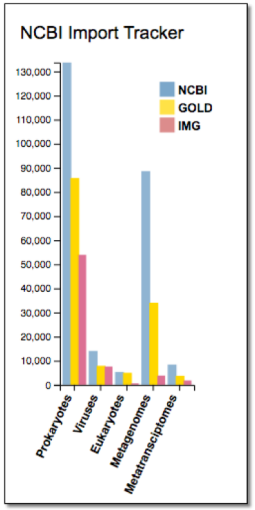
WormBase Version: WS263

About Directory Tools Downloads

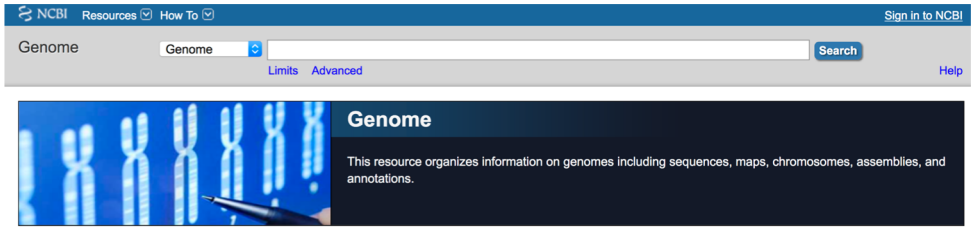
[JGI GOLD]

JGI GOLD
GENOMES ONLINE DATABASE

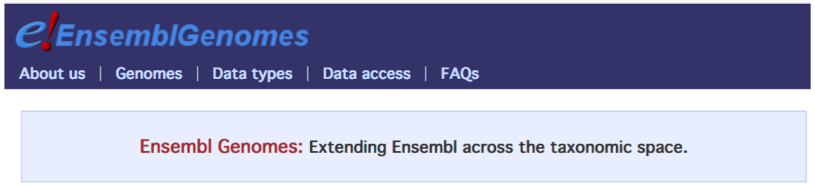
Studies	32,279
Biosamples	45,958
Sequencing Projects	195,918
Analysis Projects	154,307
Organisms	297,344



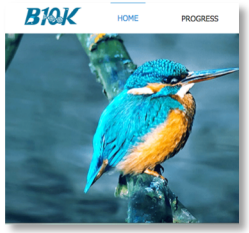
[NCBI Genome]



[EnsemblGenomes]

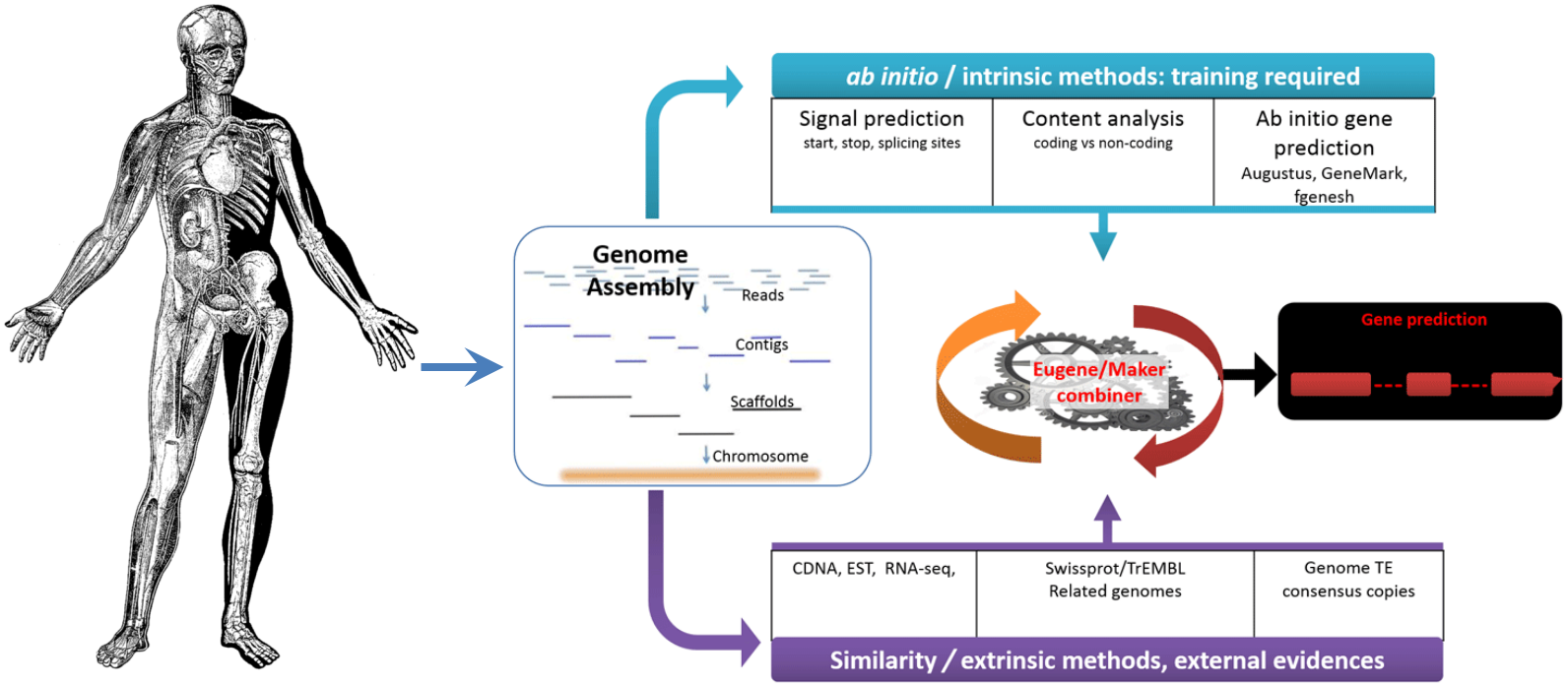


Large scale genome sequencing projects – Why?



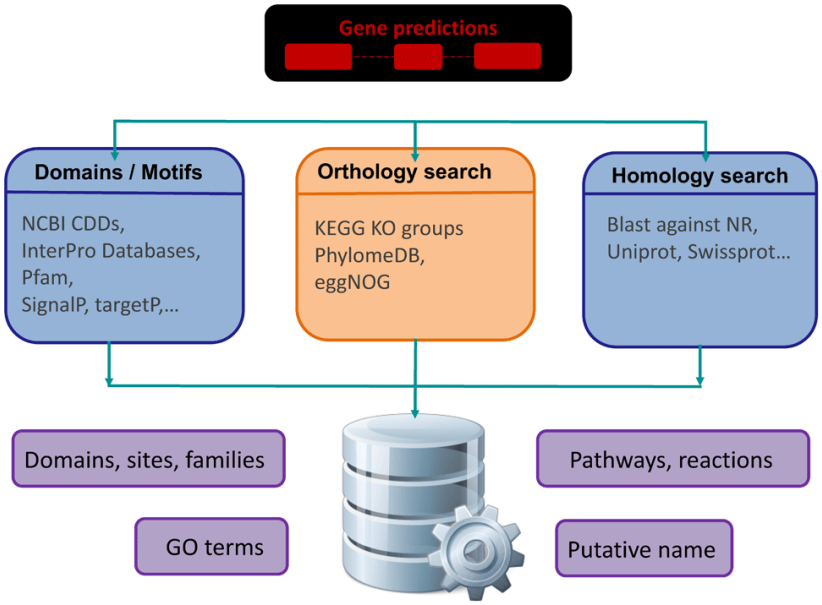
- Access to the genome sequence of related organisms / species
- Generate catalogues of
  - the encoded protein-coding and RNA genes
  - transposable elements
  - Structural variation
  - Genetic diversity within populations
  - Metabolic pathways -> natural compounds
- Assessing the metabolic capacities of species
- Understanding the link between genotype and phenoype
- Reconstruct evolutionary events both on organismic and molecular level
- ...

The prediction of genes from genome assemblies



Del Angel et al. (2018) F1000 Research 7(ELIXIR):148

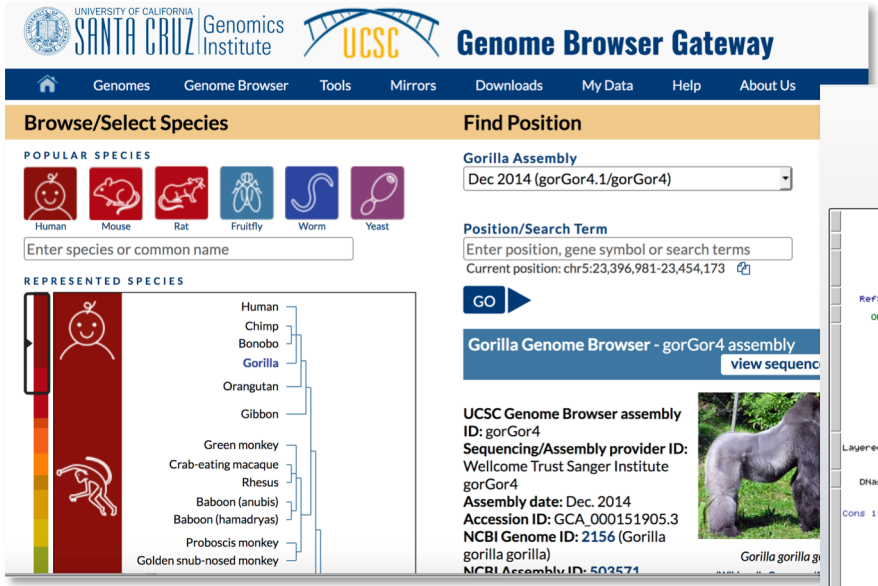
Functional annotation of Genes *in silico*

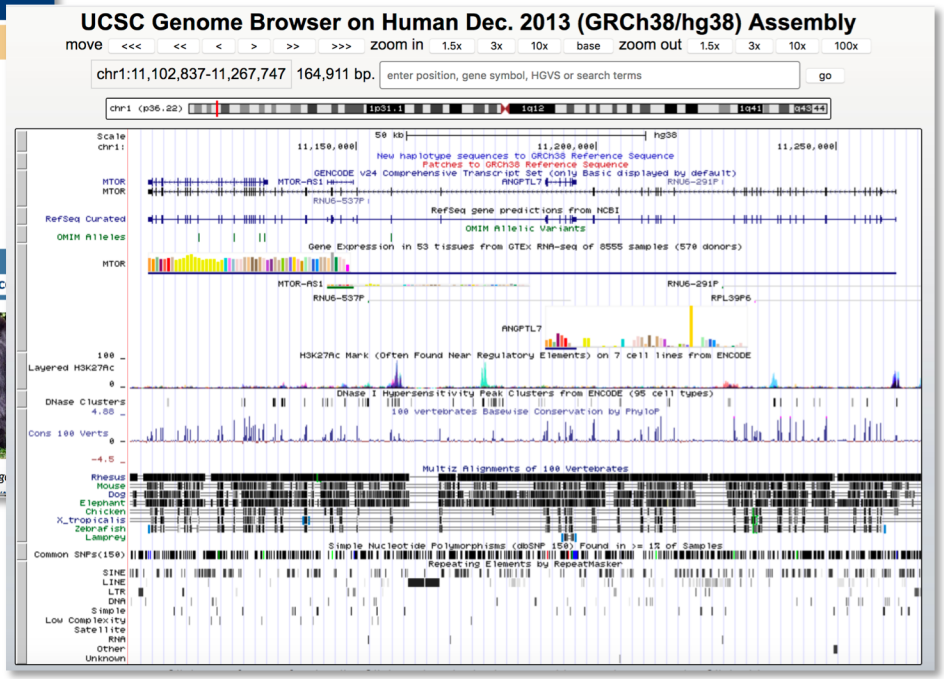


Del Angel et al. (2018) F1000 Research 7(ELIXIR):148

1. **Significant** sequence similarity (sequence conservation)
2. Orthology relationships – descendants of the same gene in the last common ancestral species
3. Conserved genomic position (positional homologs/orthologs)
4. Similar/identical domain architectures
5. Similar/identical 3D structures
6. Agreeing (conserved) expression patterns
7. Interaction partners present / conservation of interaction networks

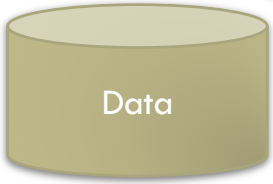
GENOMES AND THEIR ANNOTATION – THE UCSC GENOME BROWSER (HTTPS://GENOME.UCSC.EDU)





WHAT TO CONSIDER IN A TYPICAL COMPARATIVE GENOMICS ANALYSIS[1]?

Complete
Representative
Error?
- Noise?
- Systematic?



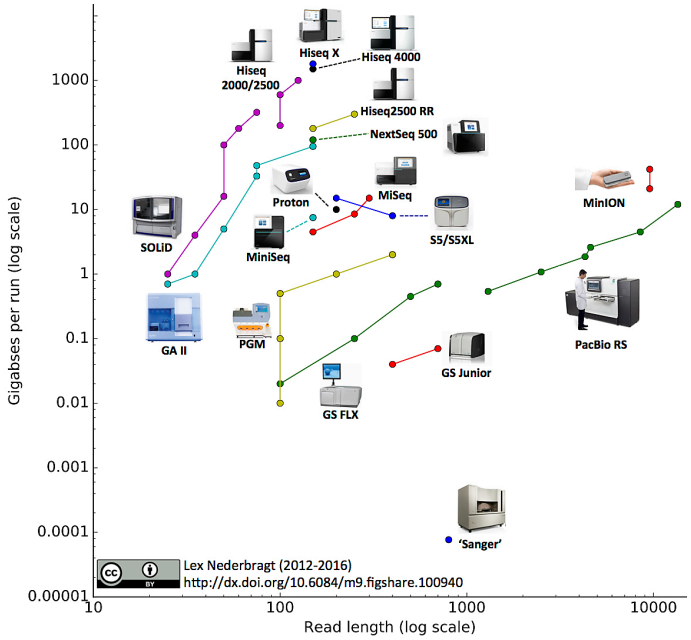
Model 1
Assumption 1
Algorithm 1

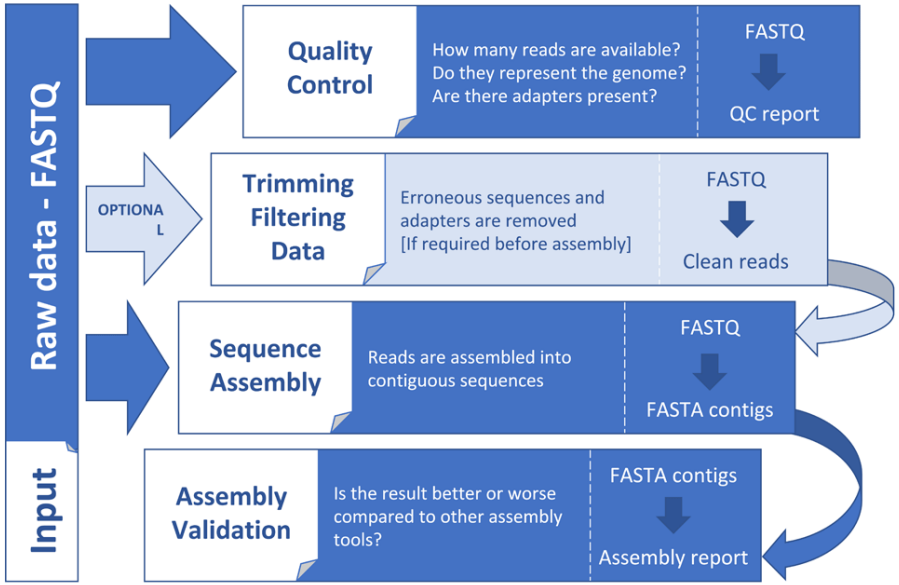
Which assumption?
What is the fit of the model?
Optimality of solution
- Locally optimal?
- Globally optimal?
- And what does 'optimal' mean?

Model 2
Assumption 2
Algorithm 2

[1] And actually in every bioinformatics analysis

DNA SEQUENCING TECHNOLOGIES





Del Angel et al. (2018) F1000 Research 7(ELIXIR):148