

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

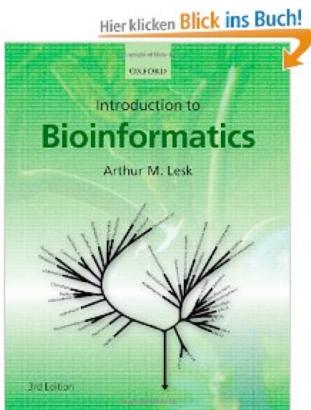
Passwort: TeachingAKE



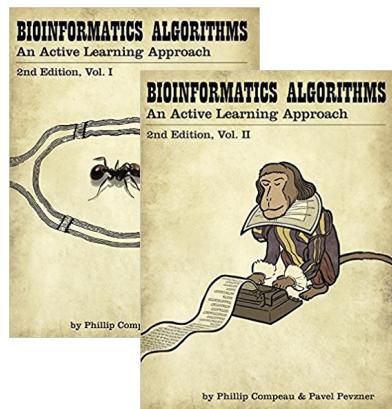
Grundlagen: (Folien, Übungsblätter, Infos, etc):

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

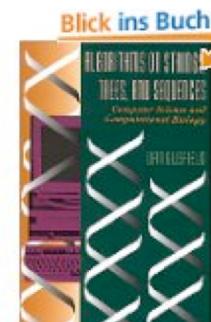
Where can I read about the things I hear?



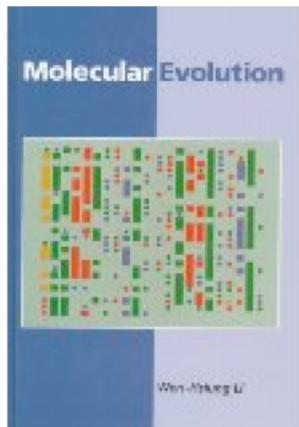
Arthur Lesk
Introduction to
Bioinformatics (2008)
Oxford University Press



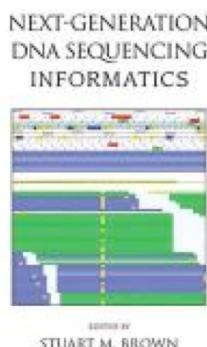
P. Compeau and P. Pevzner
Bioinformatics Algorithms
(2016)
Active Learning Publishers, LLC



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)

The screenshot shows the homepage of the GOBLET website (www.mygoblet.org). The header features the GOBLET logo (a stylized goblet icon) and the text "GOBLET Global Organisation for Bioinformatics Learning, Education & Training". Navigation links include "Training portal", "About us", and "Join us!". Below the header is a search bar with "Search..." and a "Search" button. A large black and white photograph of a group of people, identified as the "GOBLET Meeting 2015" held in Cape Town on November 20, 2015, is prominently displayed. Below the photo is a horizontal navigation bar with seven dots, likely for a slide show. The main content area includes sections for "Announcements" (listing events like SolBio International Conference 2016, Training needs for biocuration, Key Aspects of a Successful NGS Course, and An ELIXIR/GOBLET workshop), "Tweets" (a Twitter feed from @mygobletorg), and other informational pages.

Meistbesucht AK Ebersberger IngosTWiki Language Ganglia::Cluster R... AK-WIKI AK EbersbergerHe... Ingo Ebersberger |...

GOBLET
Global Organisation for Bioinformatics
Learning, Education & Training

Training portal About us Join us!

Become a contributor | Login Search... Search

GOBLET Meeting 2015
Cape Town, 20 November 2015

Announcements

- SolBio International Conference 2016
Fri, Apr 22 2016
- Training needs for biocuration
Wed, Apr 13 2016
- Key Aspects of a Successful NGS Course - a GOBLET initiative
Wed, Nov 4 2015
- An ELIXIR/GOBLET workshop: defining an e-learning lingua franca
Fri, Aug 21 2015

Tweets by @mygobletorg

GOBLET Retweeted
Sarah Morgan @SImorg Joint @ELIXIREurope @mygobletorg metagenomics material curation hackathon in full swing! Currently busy describing workflows @emblebi

4

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
GAGGGTGGAGACGTCTGGCCCCGCCGTGCACCCCCAGGGGAGGC
CGAGCCCCGCCGGCCCGCGCAGGGCCCGCCCAGGACTCCCTGC GG
TCCAGGCCGCCGGCTCCCGGCCAGCCAATGAGCGCCGCCGGCCG
GGCGTGCCTCGCCCAAGCATAAACCCCTGGCGCGCTCGCGGCCGGC
ACTCTTCTGGTCCCCACAGACTCAGAGAGAACCCACCATGGTGCTGTCTC
CTGCCGACAAGACCAACGTCAAGGCCCTGGGTAAGGTGGCGCAC
GCTGGCGAGTATGGTGGAGGCCCTGGAGAGGTGAGGCTCCCTCCCTG
CTCCGACCCGGCTCTCGCCGCCGACCCACAGGCCACCCCTAACCG
TCCTGGCCCCGGACCAAACCCACCCCTCACTCTGTTCTCCCCGAGG
ATGTTCTGTCCTCCCCACCAAGACCTACTTCCCGCACTCGACCT
GAGCCACGGCTCTGCCAGGTTAAGGCCACGGCAAGAAGGTGGCGACG
CGCTGACCAACGCCGTGGCGCACGTGGACGACATGCCAACCGCGCTGTCC
GCCCTGAGCGACCTGCACGCGACAAGCTTCGGGTGGACCCGGTCAACTT
CAAGGTGAGCGGCGGGCGGGAGCGATCTGGTCGAGGGCGAGATGGCG
CCTTCCTCGCAGGGCAGAGGATCACGCCGGTTGCCGGAGGTGTAGCGCAG
GCCGGCGCTGCCCTGGCCCTGCCCGCCACTGACCCCTTCTGCA
CAGCTCTAACGCCACTGCCGTGGTACCCCTGCCGCCACCTCCCCGC
CGAGTTCACCCCTGCCGTGCACGCCCTGGACAAGTCCTGGCTCTG
TGAGCACCGTGTGACCTCAAATACCGTTAACGCTGGAGCCTGGTGGCC
ATGCTTCTGCCCTGGCCTCCCCCAGCCCTCCTCCCTTGCA
CCCGTACCCCGTGGCTTGAAATAAGTCTGAGTGGCGGCAGCCTGTG
TGTGCCGTGAGTTTCCCTCAGCAAACGTGCCAGGCATGGCGTGGACA
GCAGCTGGACACACATGGCTAGAACCTCTGAGCTGGATAGGGTAGG
AAAAGGCAGGGCGGGAGGAGGGGATGGAGGGAGGGAAAGTGGAGCCACCG
CGAAGTCCAGCTGGAAAAACGCTGGACCC TAGAGTGCTTGAA
```

Am Anfang steht die Sequenz

```
>hg19_dna
GAGGGTGGAGACGTCTGGCCCCGCCGTGCACCCCCAGGGGAGGC
CGAGCCCGCCGCCGGCCCCGCAGGCCCCGCCGGACTCCCTGCGG
TCCAGGCCGCCGGCTCCGCCAGCCAATGAGGCCGCCGGCG
GGCGTCCCCCGGCCAAGCATAAACCCTGGCGCTCGCGGCCG
ACTCTTCTGGTCCCCACAGACTCAGAGAGAACCACCATGGTGCTGTCTC
CTGCCGACAAGACCAACGTCAAGGCCCTGGGTAAGGTCGGCGCAC
GCTGGCGAGTATGGTGCGGAGGCCCTGGAGAGGTGAGGCTCCCTCCCTG
CTCCGACCCGGCTCTGCCGCCGGACCCACAGGCCACCCCTAACCG
TCCTGGCCCCGGACCCAAACCCACCCCTCACTCTGCTTCTCCCGCAGG
ATGTTCTGTCTTCCCCACCAAGACCTACTTCCGCACTGACCT
GAGCCACGGCTGCCAGGTTAAGGCCACGGCAAGAAGGTGGCCAGC
CGCTGACCAACGCCGTGGCGCACGTGGACGACATGCCAACCGCCTGTCC
GCCCTGAGCGACCTGCACGCGACAAGCTCGGGTGACCCGGTCAACTT
CAAGGTGAGCGGGCCGGAGCGATCTGGTGAGGGGAGATGGCG
CCTTCCTCGCAGGGCAGAGGATCACGCGGGTTGCCGGAGGTGTAGCGCAG
GCCGGCGCTGCCGGCTGGGCCCTGCCGGCCACTGACCCCTCTCTGCA
CAGCTCCAAGCCACTGCCGTGGTGACCCTGCCGCCACCTCCCCGC
CGAGTTCACCCCTGCCGTGCACGCCCTGGACAAGTTCTGGCTCTG
TGAGCACCGTGTGACCTCAAATACCGTTAAGCTGGAGCCTGGTGGCC
ATGCTTCTTGCCTGGCCTCCCCCAGCCCTCCCTCCCTGCA
CCCGTACCCCGTGGCTTGAATAAAGTCTGAGTGGCGGCGAGCCTGTG
TGTGCCCTGAGTTTCCCTCAGCAAACGTGCCAGGCATGGCGTGGACA
GCAGCTGGGACACACATGGCTAGAACCTCTGCAGCTGGATAAGGTAGG
AAAAGGCAGGGCGGGAGGGAGGGATGGAGGAGGGAAAGTGGAGGCCACCG
CGAAGTCCAGCTGGAAAAACGCTGGACCCTAGAGTGCTTG
```

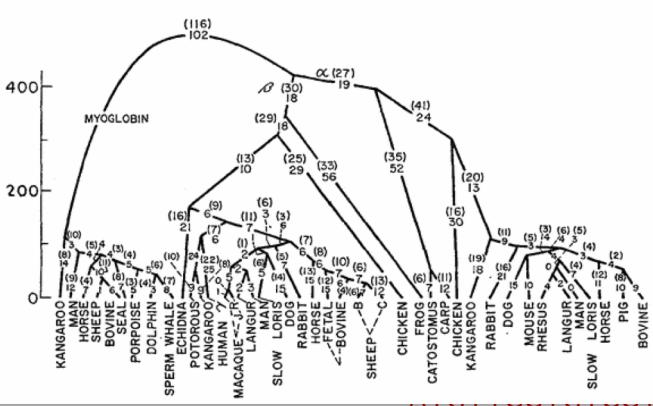
DNA/RNA/Protein?
DNA Replication?
Orientation of DNA?
Signal sequences in DNA?
Transcription and splicing?
Further mRNA processing?
Translation/Genetic Code?

```
>hg19_protein
MVLSPADKTNVKAAGWKVGAHAGEYGAEALERMFSL
FPTTKTYFPHFDLSHGSAQVKGHGKKVADALTNAVAHV
DDMPNALSASLDLHAHKLRVDPVNFKLLSHCLLVTLAA
HLPAEFTPASLDKFLASVSTVLTSKYR
```

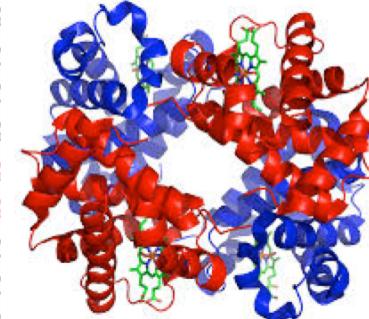
Am Anfang steht die Sequenz

12

M. Goodman et al.

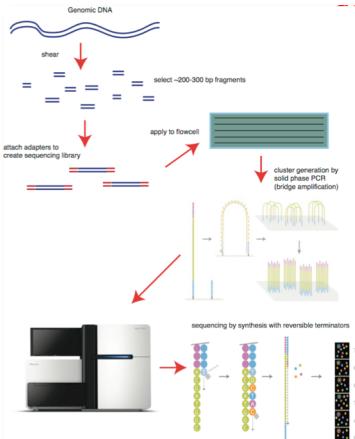


CCTGGCCCCCGCCCCGCGTGCACCCCAAG
GGCCCCCGCGCAGGCCCGCCCGGGACTCC
CGGGCTCCCGGCCAGCCAATGAGCGCCGC
CCCCAAGCATAAACCTGGCGCGCTCGCC
CACAGACTCAGAGAGAACCCACCATGGTG
AACGTCAAGGCCGCTGGGTAAGGTCGG
TGCGGAGGCCCTGGAGAGGTGAGGCTCCC
CCTCGCCCGCCCGACCCACAGGCCACCC
CCAAACCCACCCCTCACTCTGCTTCTCC
CCCCACCAAGACCTACTTCCGCACITCGA
CTGCCAGGTTAAGGCCACGGCAAGAAGGTGGCGA
GCCGTGGCGCACGTGGACGACATGCCAACCGCGCTGT
GCCCTGAGCGACCTGCACGCGACAAGCTTCGGGTGGACCCGGTCAACTT
AGGTGAGCGGCGGGCGGGAGCGATCTGGTCGAGGGCGAGATGGCG
TTCCTCGCAGGGCAGAGGATCACGCGGGTTGCGGGAGGTGTAGCGCAG
GGCGGCTGCGGGCTGGCCCTGGCCCCACTGACCCCTTCTCTGCA
GTCCTAAGCCACTGCCTGCTGGTACCCCTGCCGCCACCTCCCCGC
AGTTCACCCCTGCGGTGCACGCCCTGGACAAGTCCTGGCTTCTG
AGCACCGTGTGACCTCAAATACCGTTAAGCTGGAGCCTCGGTGG
GCTTCTTGCCTTGGCCTCCCCCAGCCCCCTCCTCCCCTCCTG
CGTACCCCGTGGCTTTGAATAAAGTCTGAGTGGCGGCAGCCTG
TGCCTGAGTTTCTCAGCAAACGTGCCAGGCATGGCGTGG
AGCTGGACACACATGGCTAGAACCTCTCTGCAGCTGGATAGGGTA
AAGGCAGGGCGGGAGGAGGGGATGGAGGAGGGAAAGTGGAGGCCAC
AAGTCCAGCTGGAAAAACGCTGGACCCCTAGAGTGCTTGA

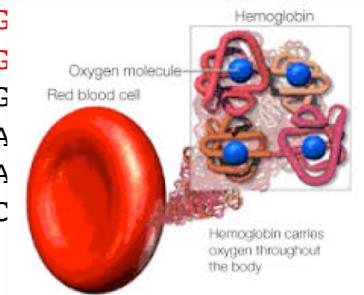


Evolutionäre Geschichte

Struktur



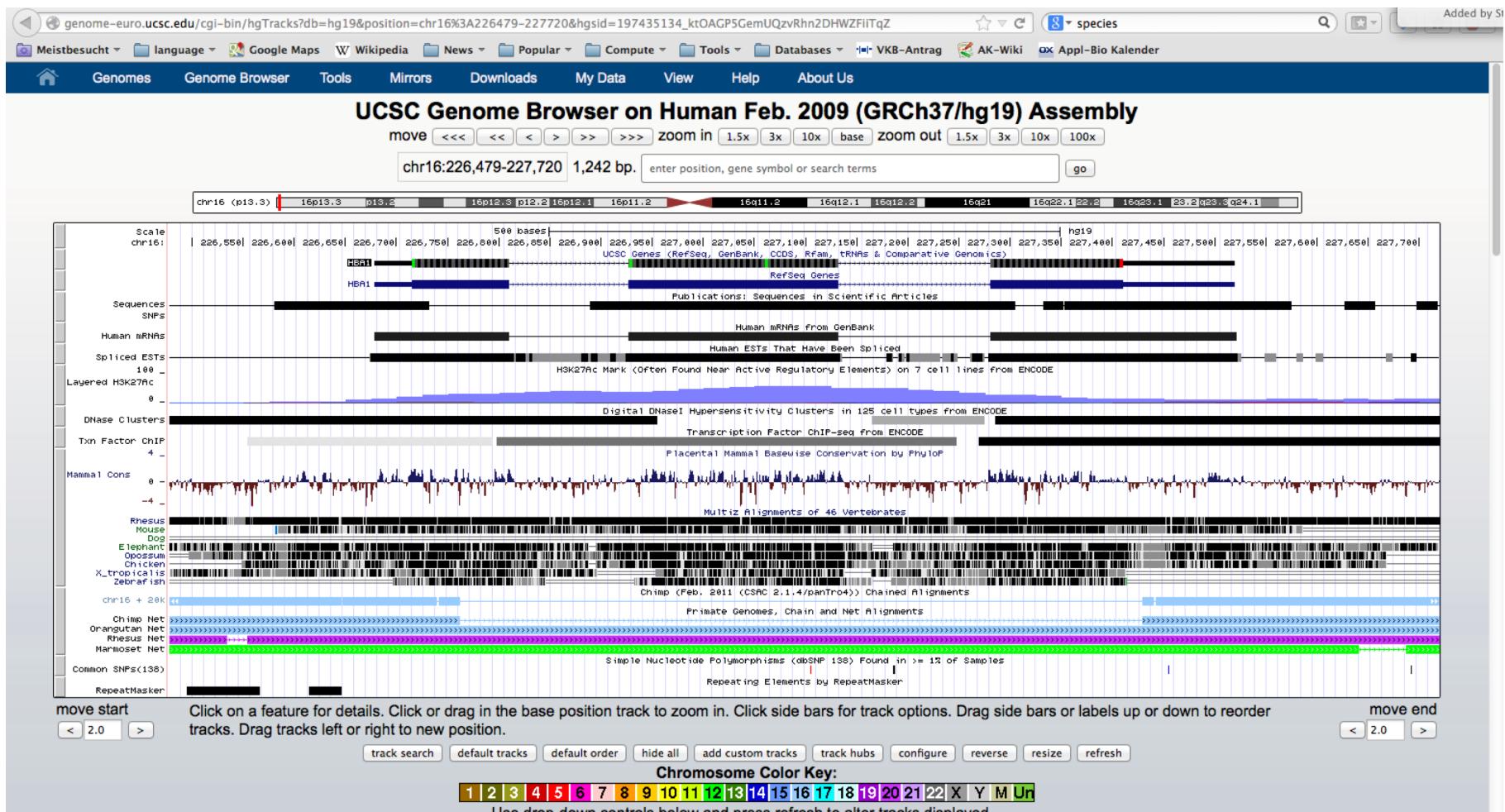
Genomic DNA
shear
= = = = selected ~200-300 bp fragments
attach adapters to create sequencing library
apply to flowcell
cluster generation by solid phase PCR (bridge amplification)
sequencing by synthesis with reversible terminators



Daten-Historie

Funktion

Am Ende steht die Annotation



Biologische Grundlagen

Wörterbuch

grundlage



Grund·la·ge

Substantiv [die]

etwas, das die unerlässliche Voraussetzung für etwas ist.

"Lebenslanges Lernen ist eine Grundlage für den Erfolg."



Übersetzungen, Wortherkunft und weitere Definitionen

Grundlage 0: Characteristics of a eukaryotic cell

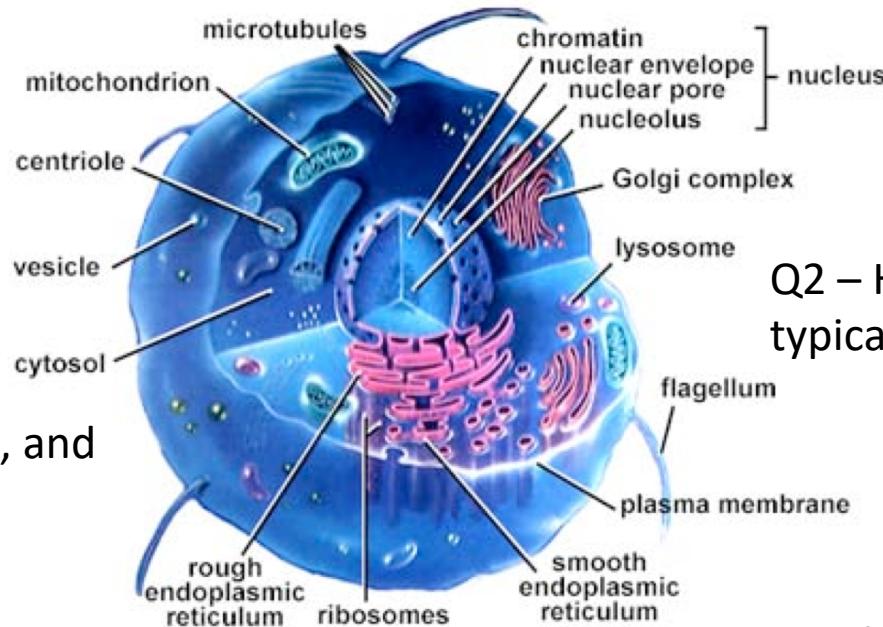
Q1 - How many different genomes do we have in a eukaryotic cell?

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

Q6 - What does the endosymbiont hypothesis postulate?

Q9 - What is 'alternative splicing'?

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



Q2 – How many genes does a typical eukaryote have?

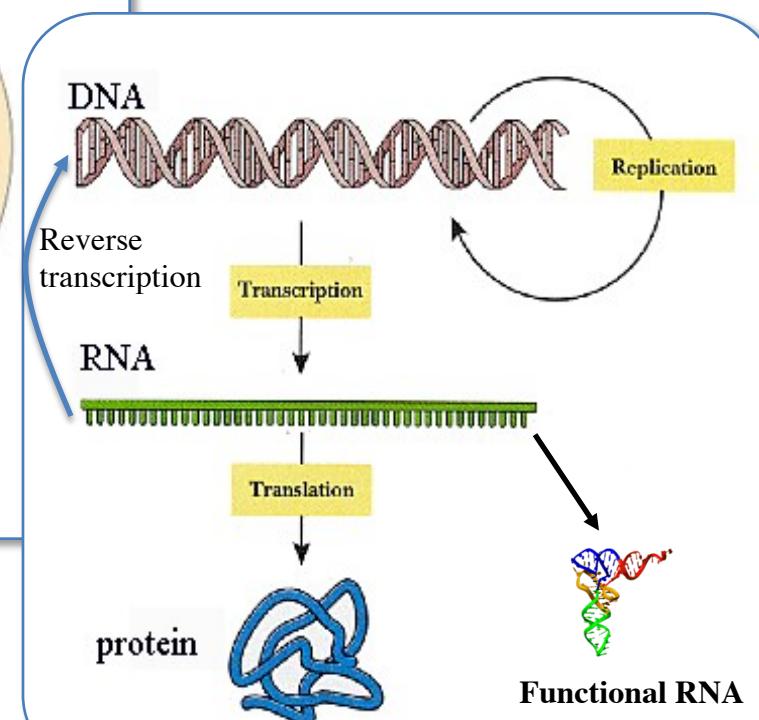
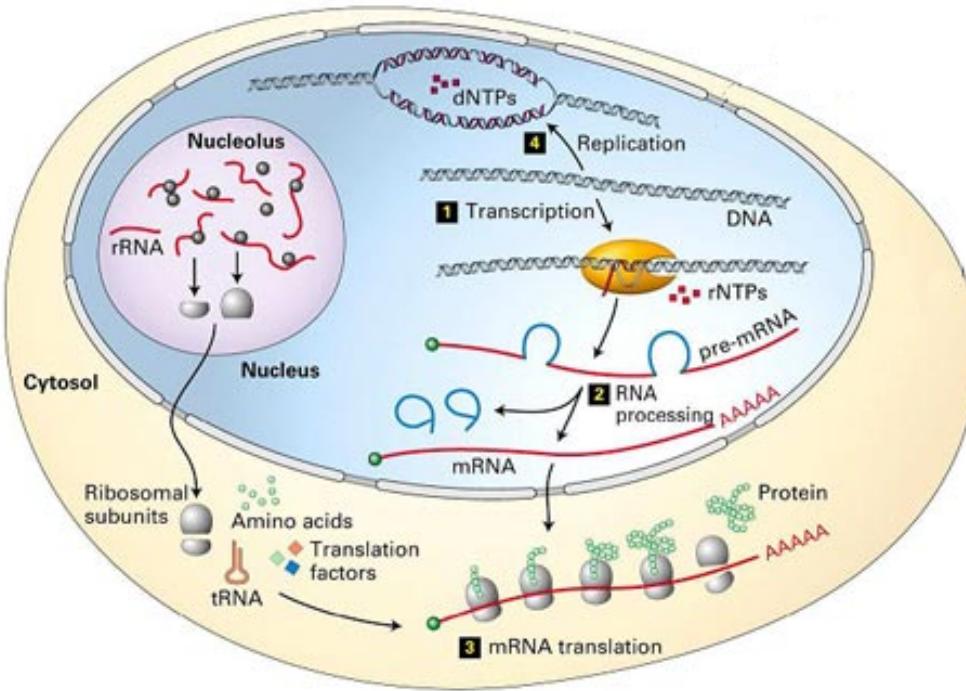
Q8 – What are exons, what are introns, and what characterizes the process 'splicing'?

Q4 - What does a ribosome do?

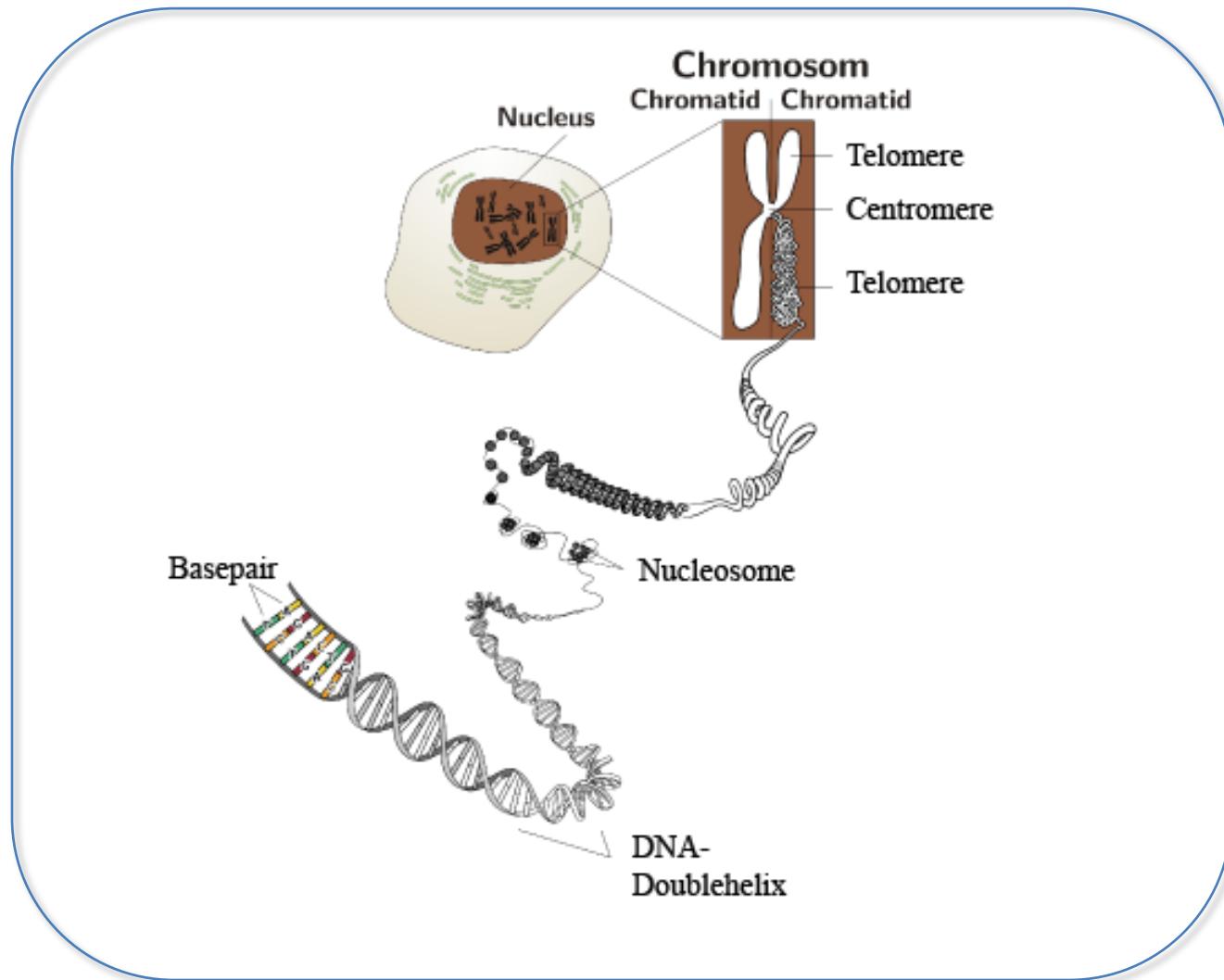
Q5 - What does a mitochondrion do?

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

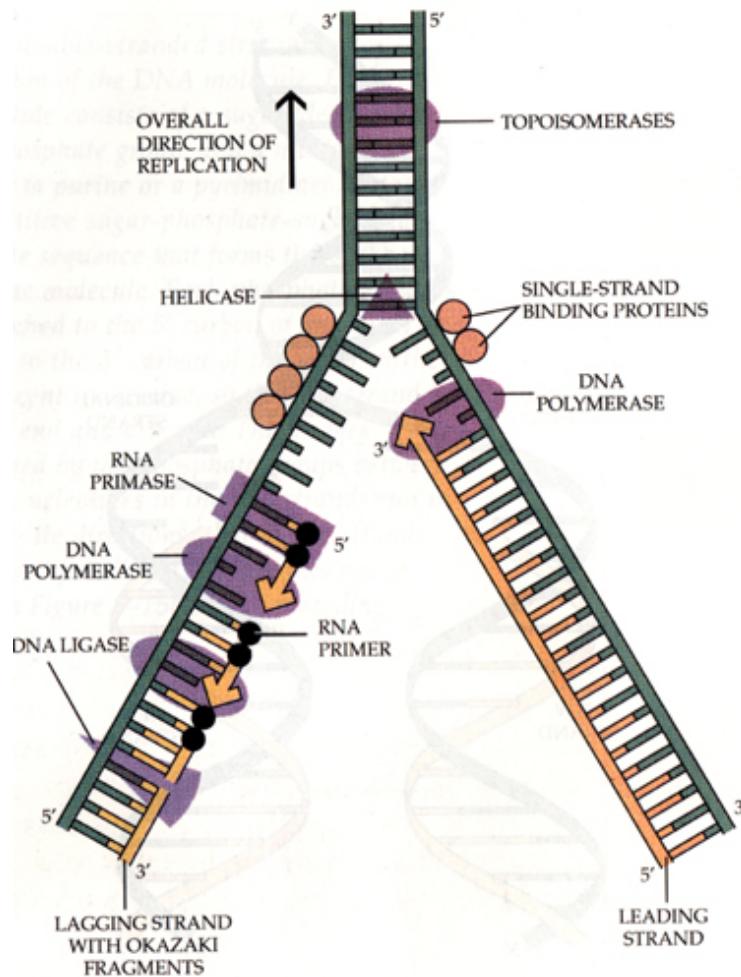
Grundlage 1: Informationsfluss in einer Zelle (Das ‚zentrale Dogma‘ der Molekularbiologie)



Grundlage 2: Organisation der genetischen Information in einer Zelle



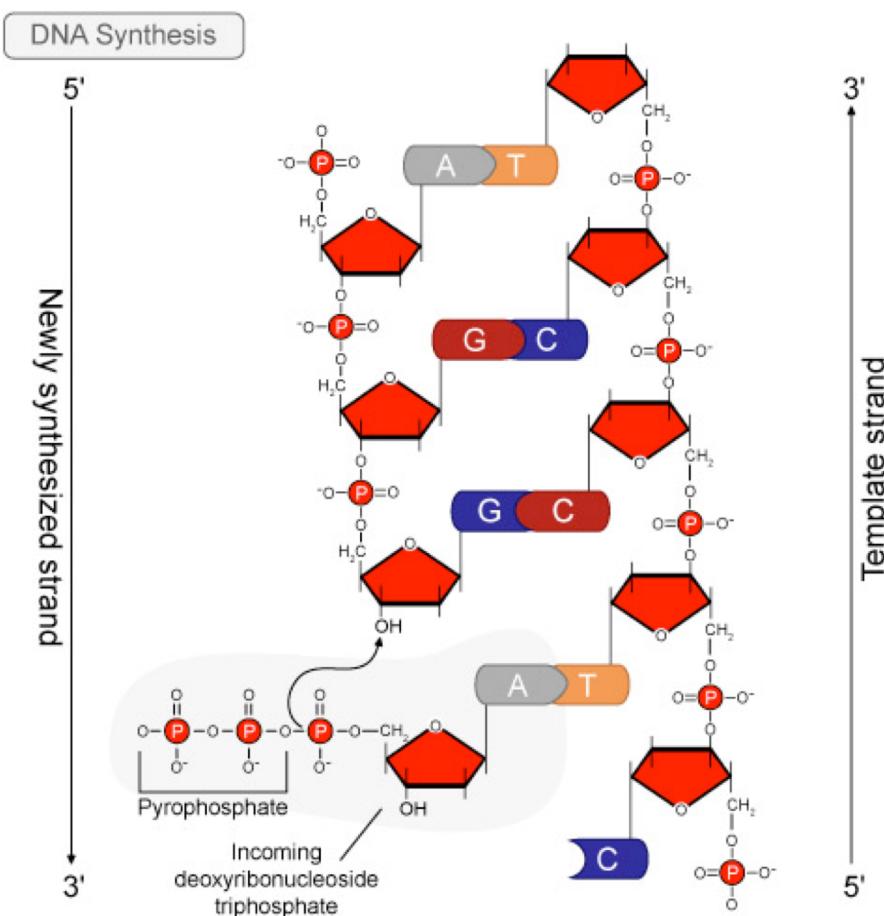
Grundlage 3: DNA-Replikation ist semi-konservativ



Die 3 Grundprinzipien
enzymatischer DNA
Synthese:

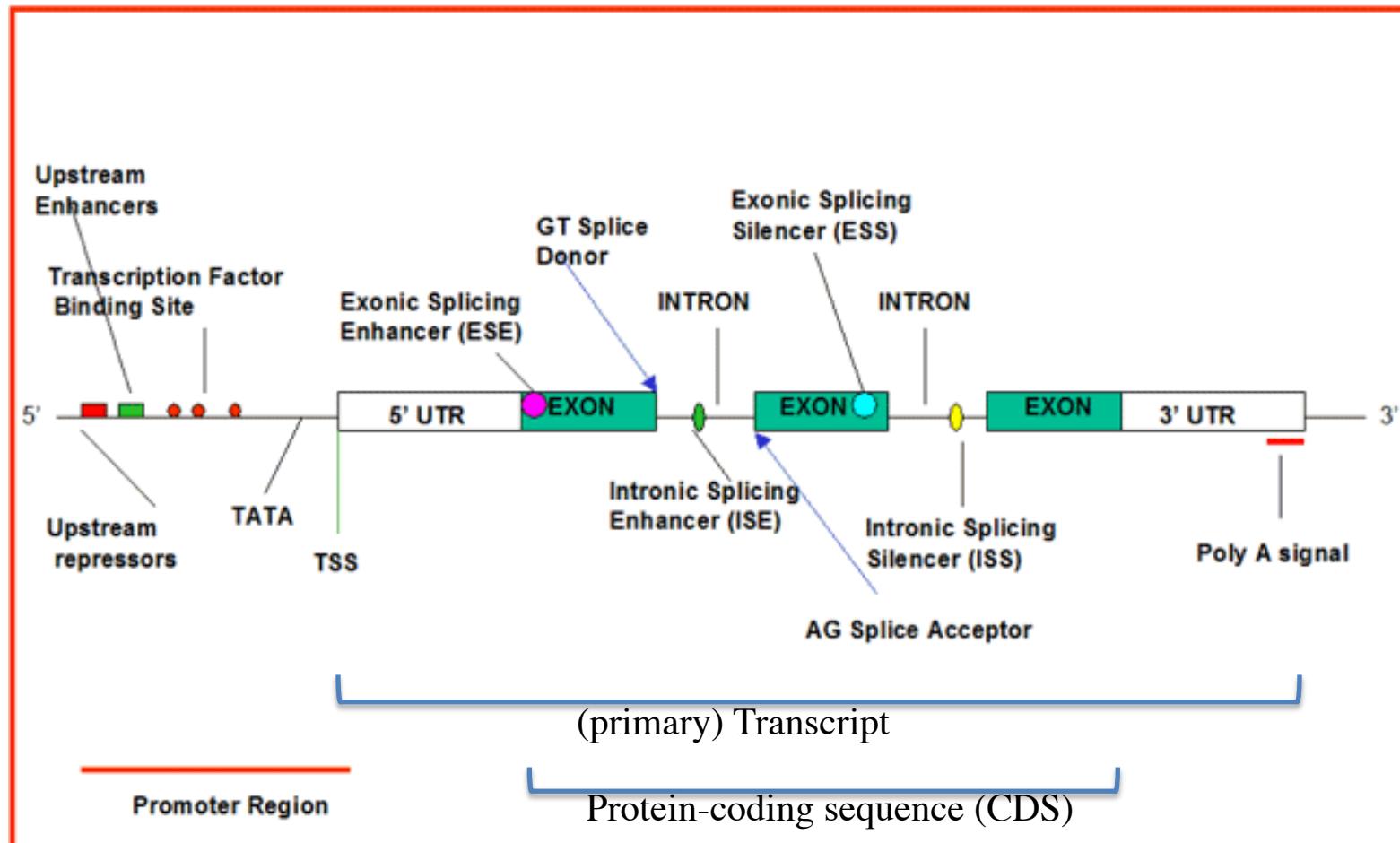
- 5' → 3'
- template guided
- primed

Grundlage 4: Die molekularen Prinzipien der enzymatischen DNA Synthese



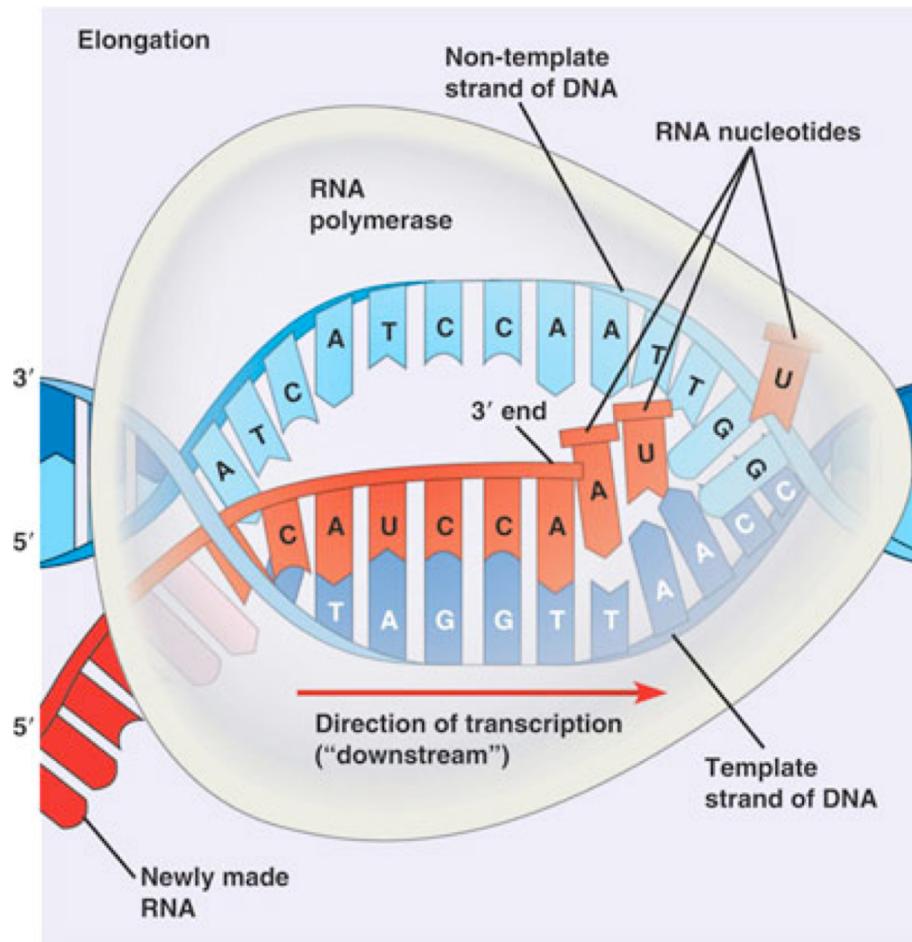
Dept. Biol., Penn State ©2004

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



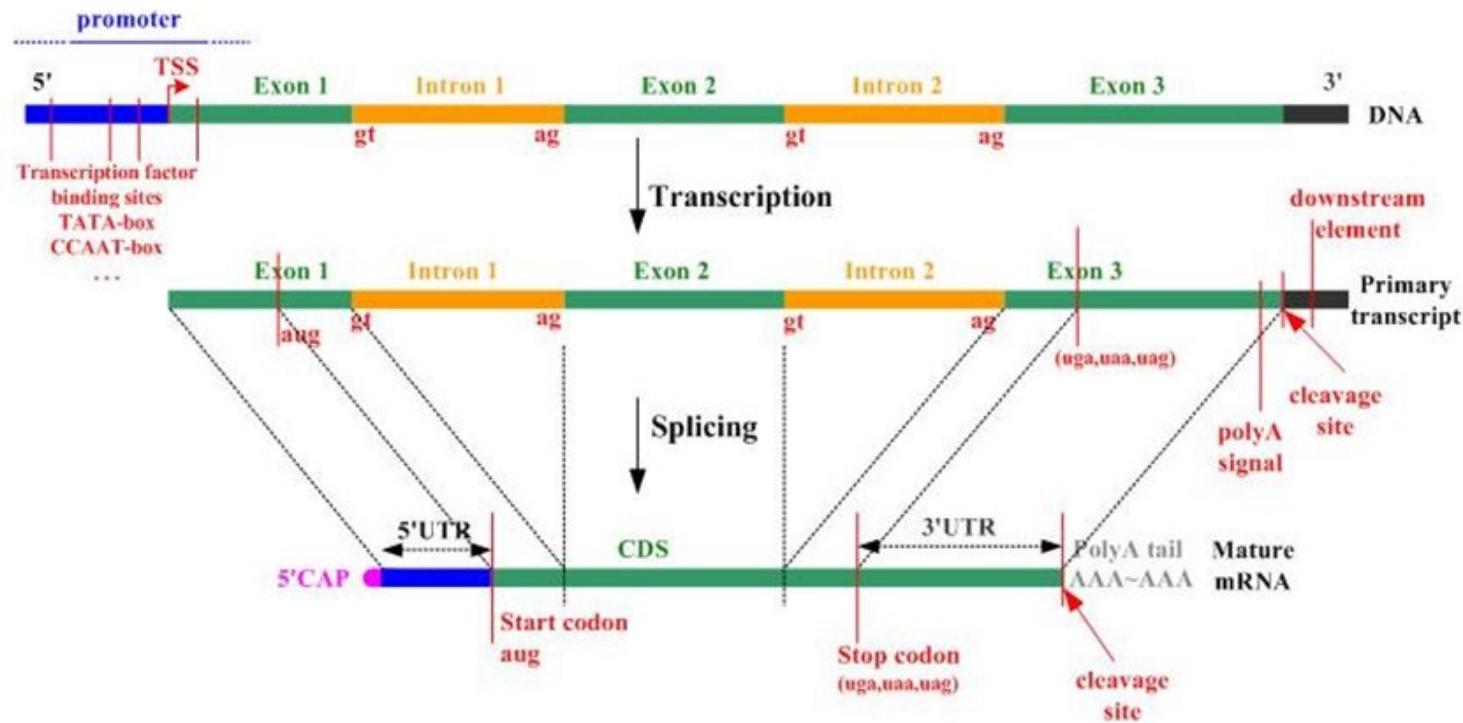
TSS: Transcription start site

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



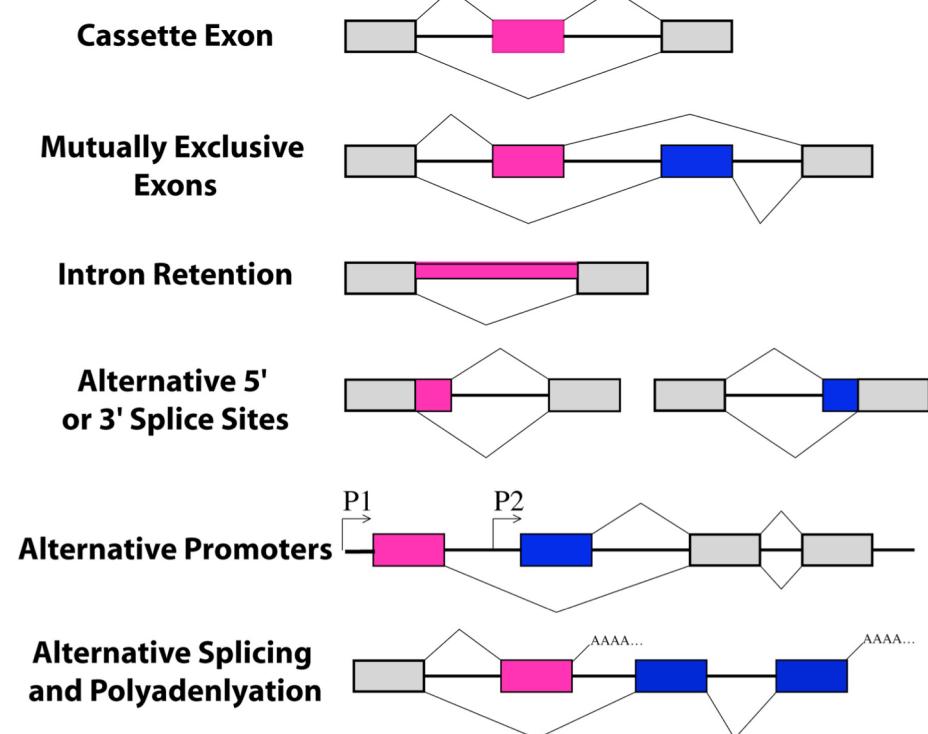
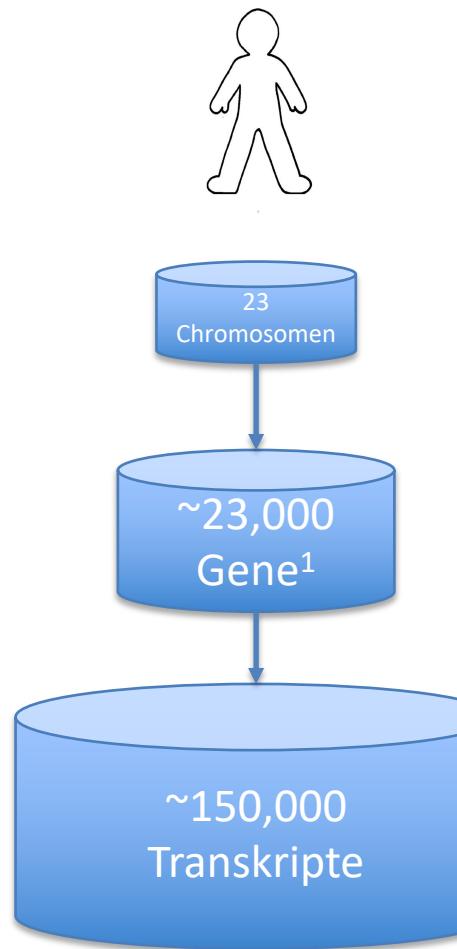
*make sure to get the orientation and direction of synthesis right

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



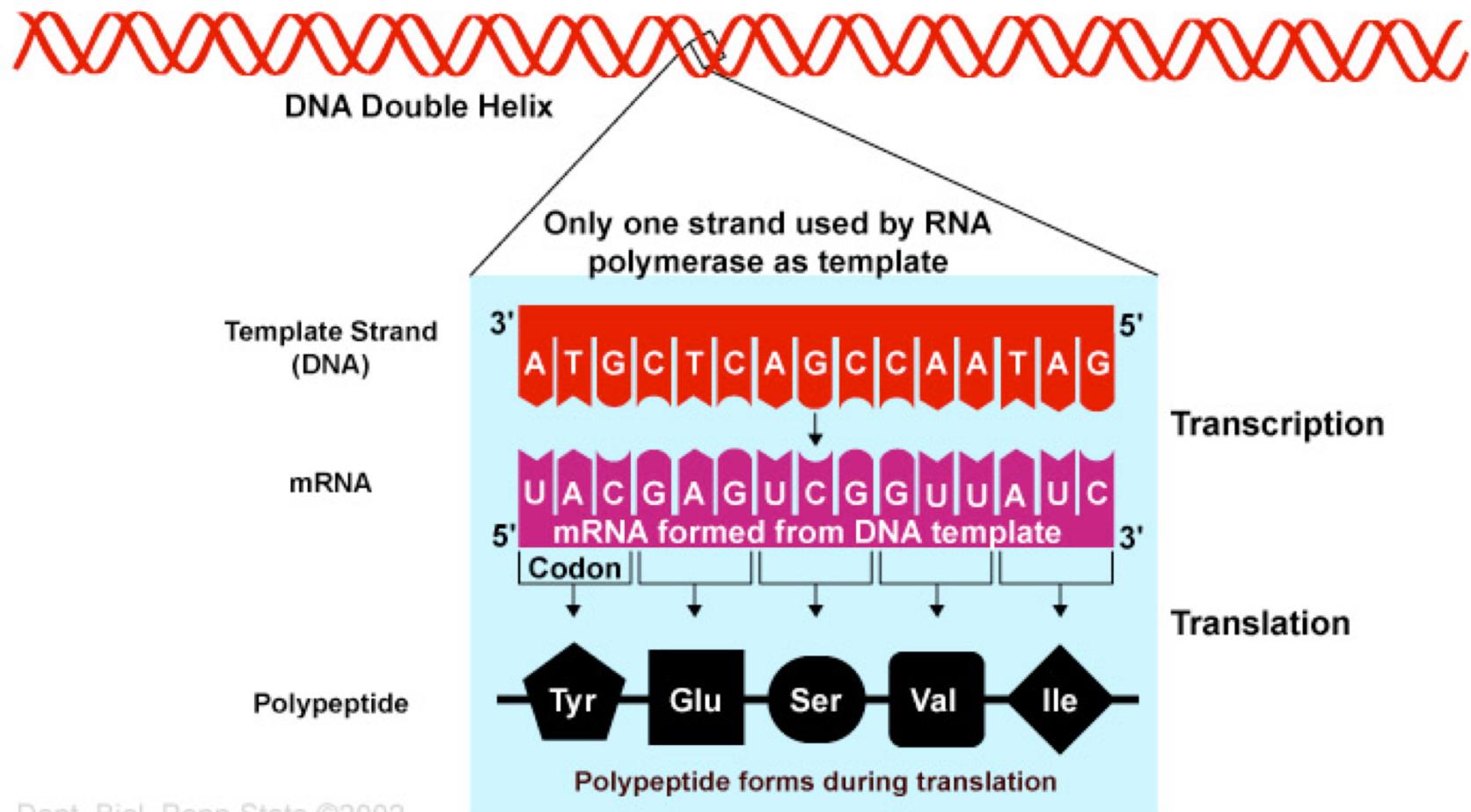
Adapted in part from
<http://online.itp.ucsb.edu/online/infobio01/burge/>

Unterschiedliche Exon-Usage kann aus einem Gen viele Genprodukte generieren

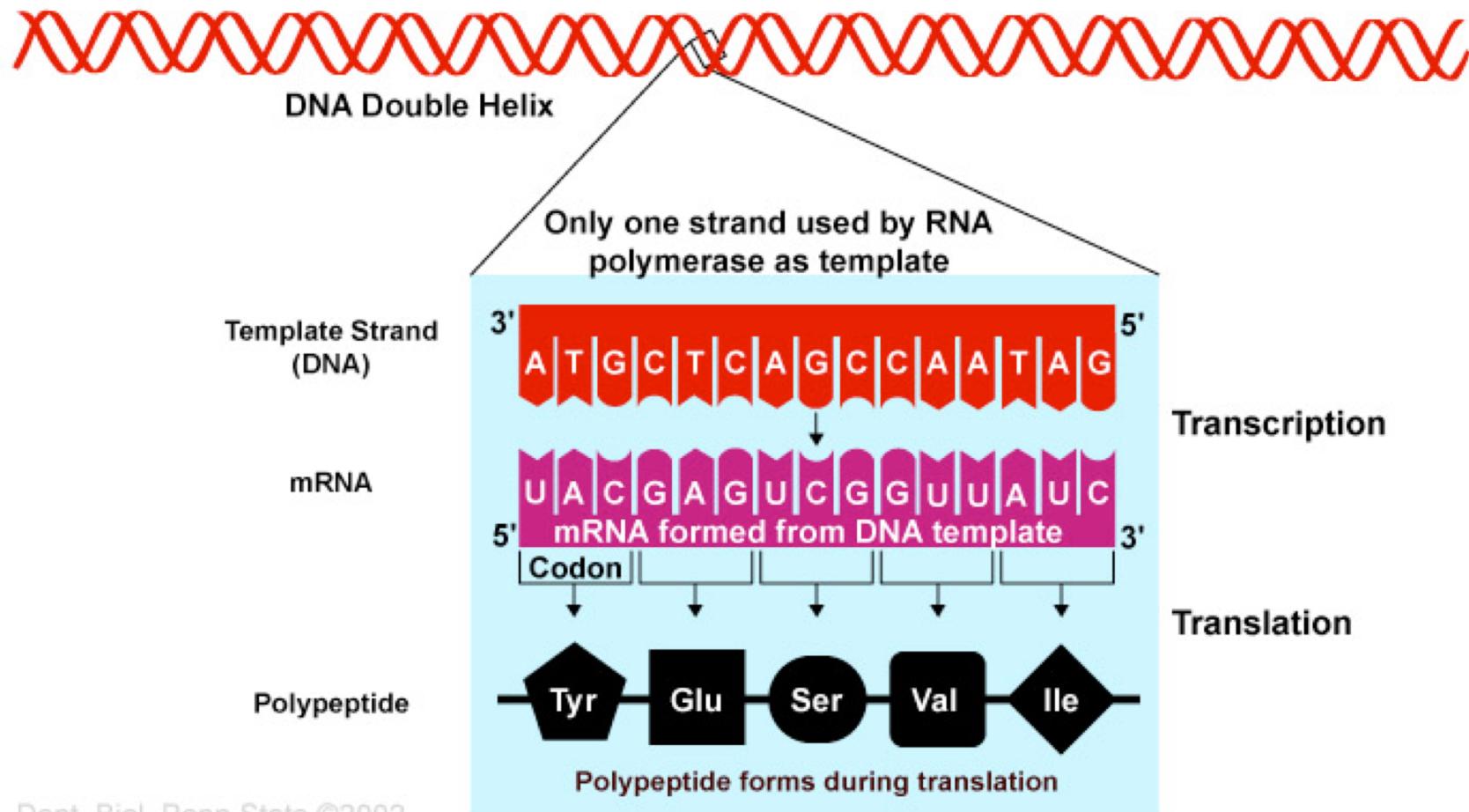


1. Protein-kodierende „known genes“

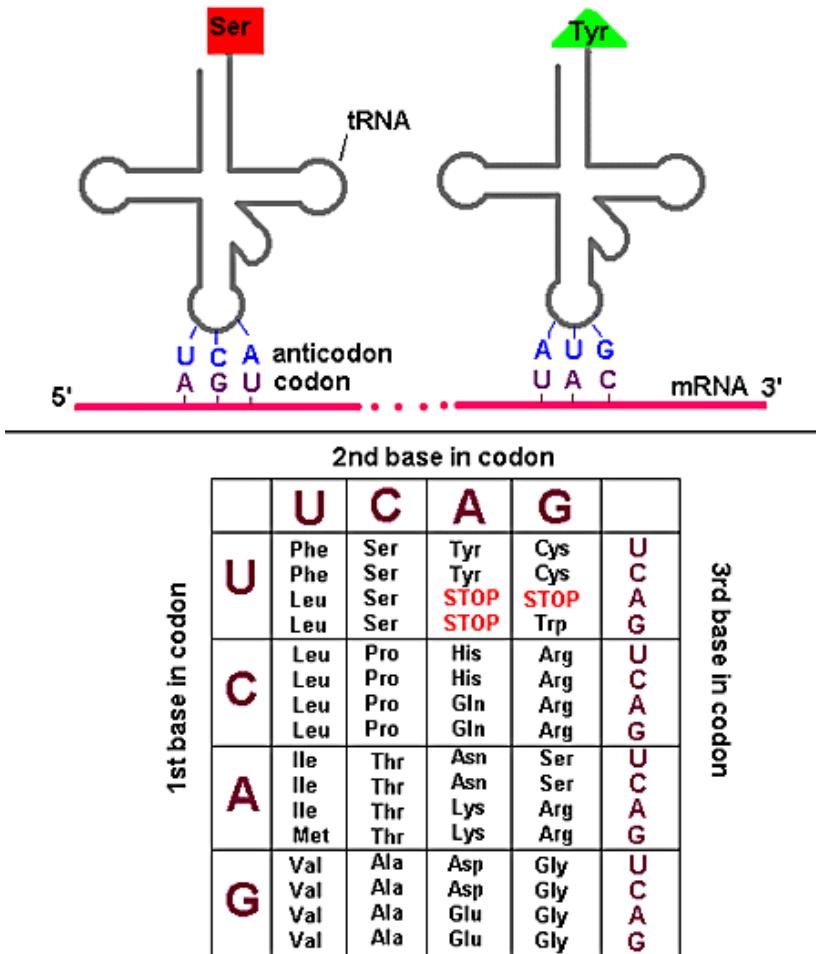
Grundlage 7: Die Information ist in Form von Basentriplets (Codons) abgelegt



Die Information ist in Form von Basentriplets (Codons) abgelegt



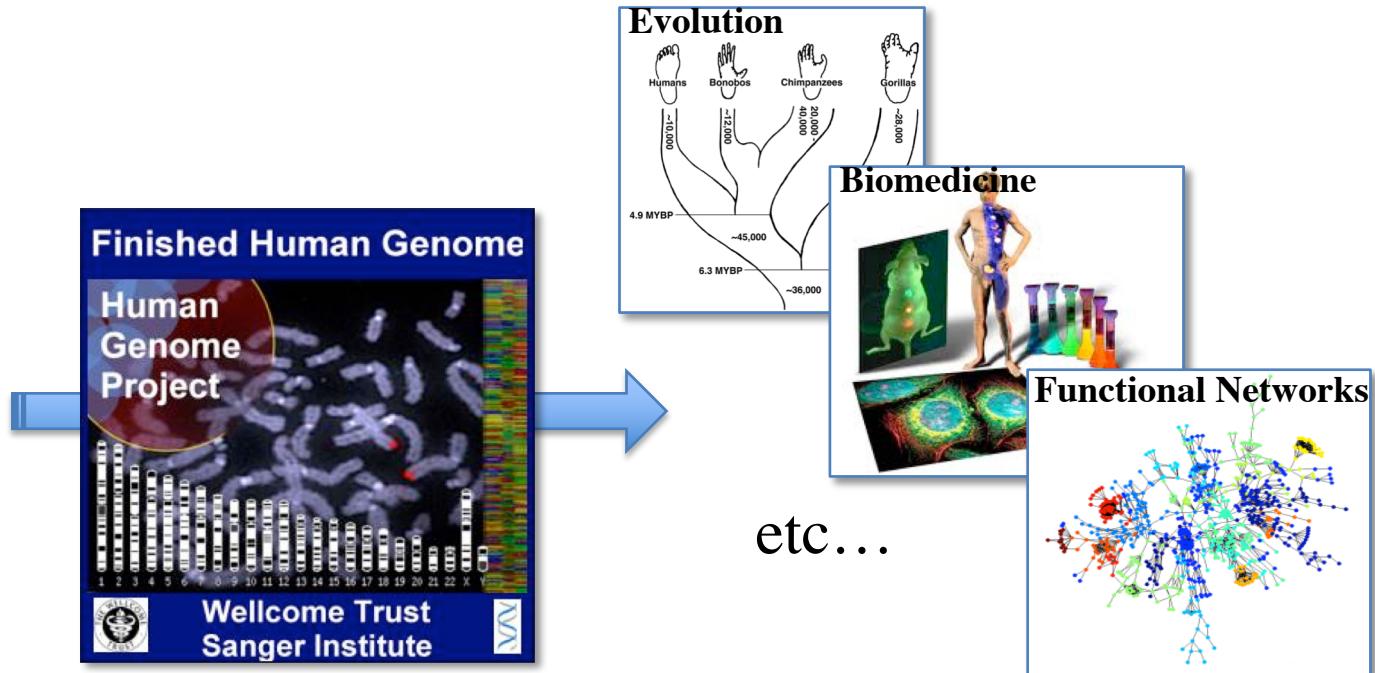
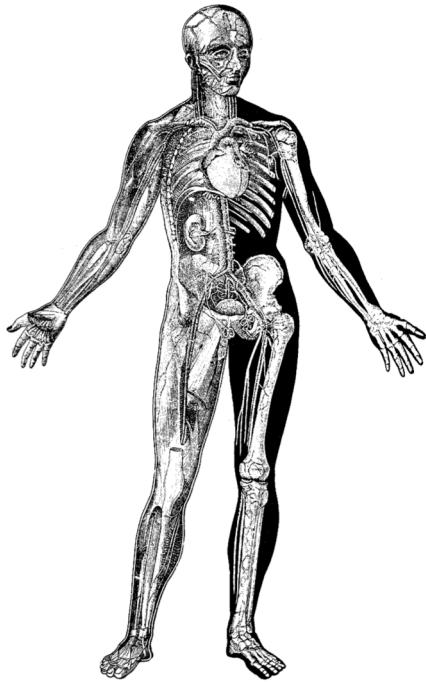
Der genetische Code... ist degeneriert* (und es gibt mehr als nur den Standard-Code!!)



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



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



Rhinolophus ferrumequinum
1,929,400,000



Amoeba proteus
290,000,000,000



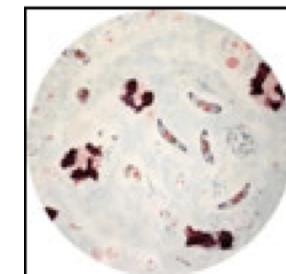
Amoeba dubia
670,000,000,000



Human immunodeficiency
virus type 1
19,750



Bufo bufo
6,900,000,000



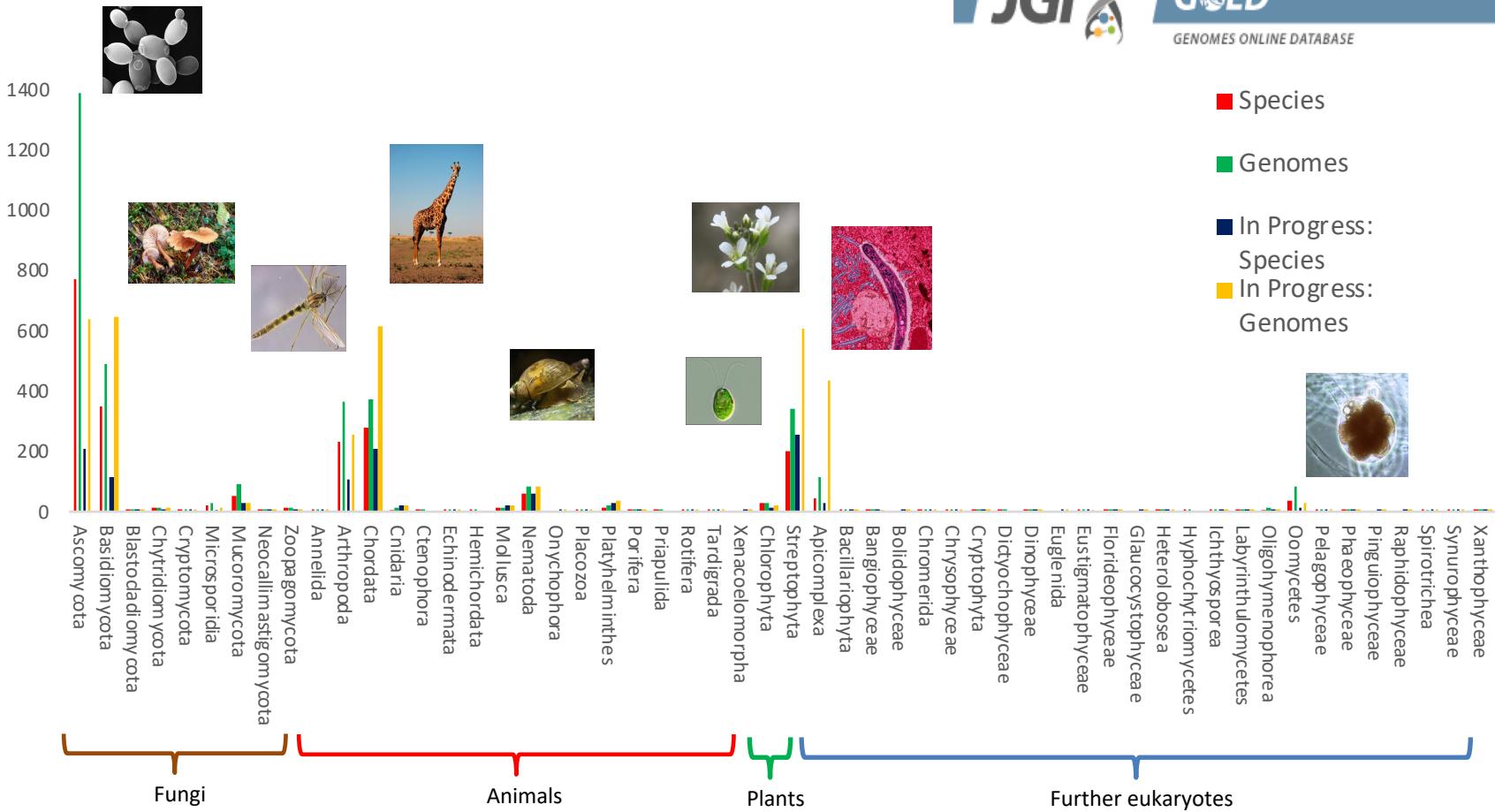
Plasmodium falciparum
25,000,000

Was wurde bis 2018 sequenziert?



GOLD

GENOMES ONLINE DATABASE



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)

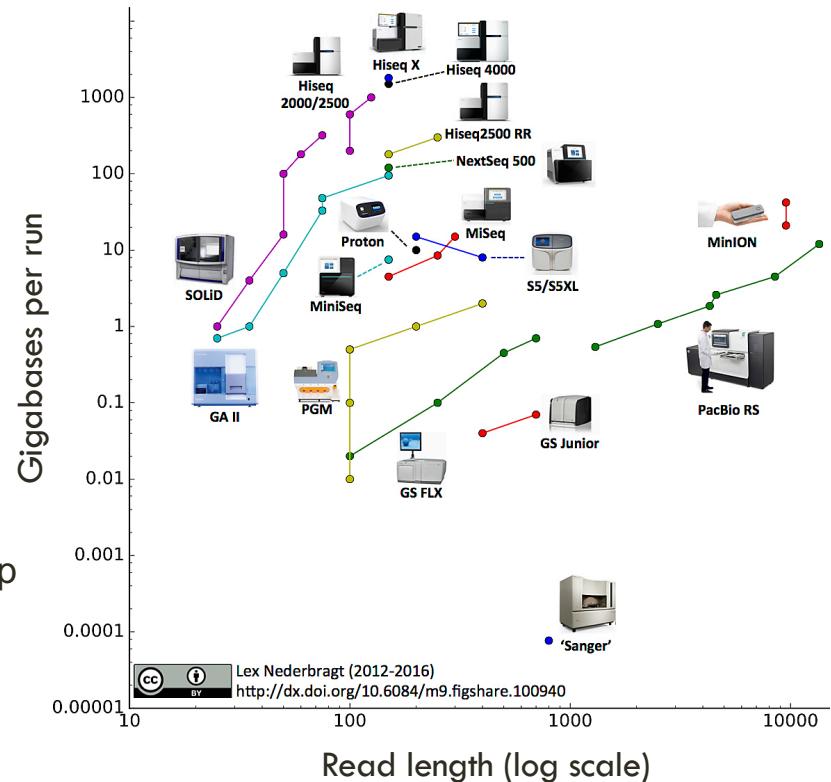
- **Sanger** method: Sequencing by synthesis
- **Maxam-Gilbert** method: chemical sequencing

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

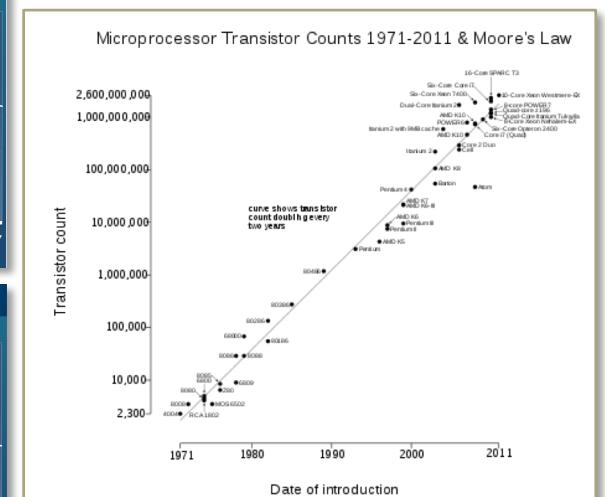
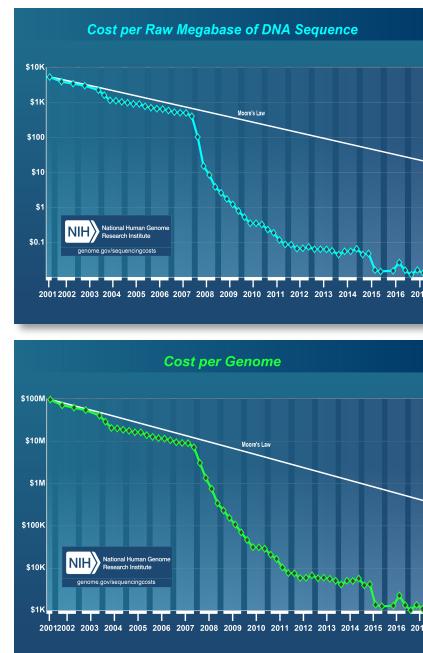
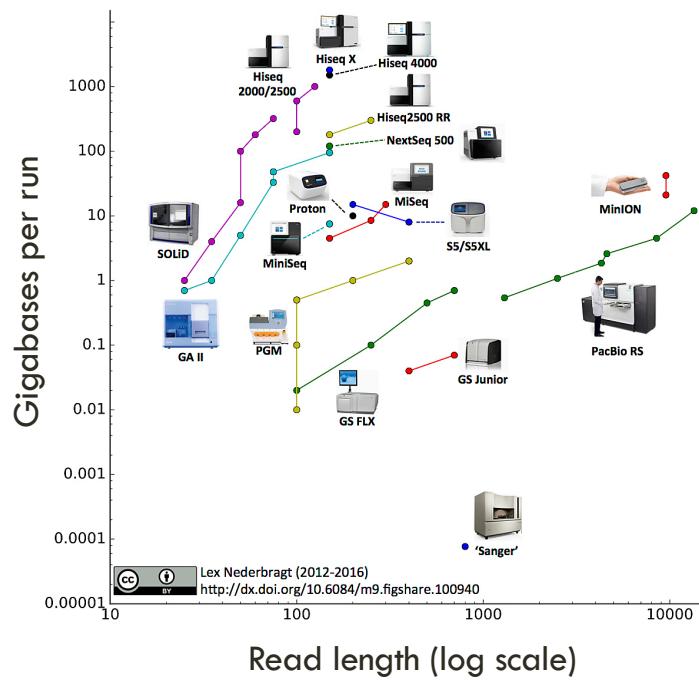
3rd generation (2015)

- **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: wikipedia

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

Die Sequenzierung gesamter Genome – Wo findet man die Daten?

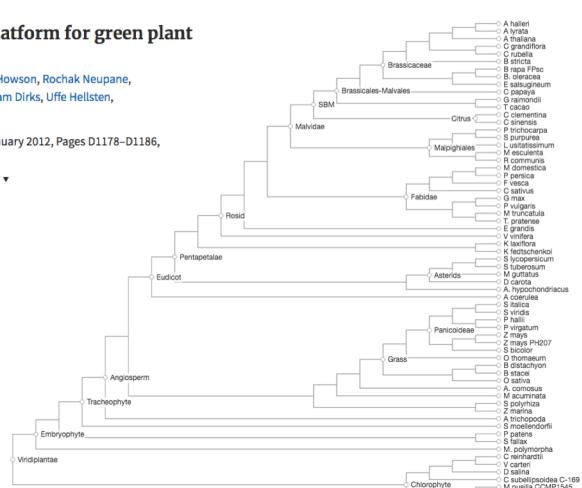
Phytozome: a comparative platform for green plant genomics

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](#)

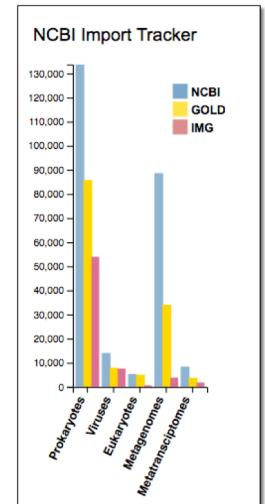


[NCBI Genome](#)

The NCBI Genome search interface. It features a search bar with dropdown menus for 'Genome' (selected), 'Limits', and 'Advanced'. Below the search bar is a blue banner with a stylized chromosome image and a pencil icon. The main content area is titled 'Genome' and contains the text: 'This resource organizes information on genomes including sequences, maps, chromosomes, assemblies, and annotations.' There are also links for 'Sign in to NCBI' and 'Help'.



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



[Flybase](#)



About Directory Tools Downloads

[JGI GOLD](#)

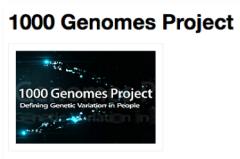
[EnsemblGenomes](#)



About us | Genomes | Data types | Data access | FAQs

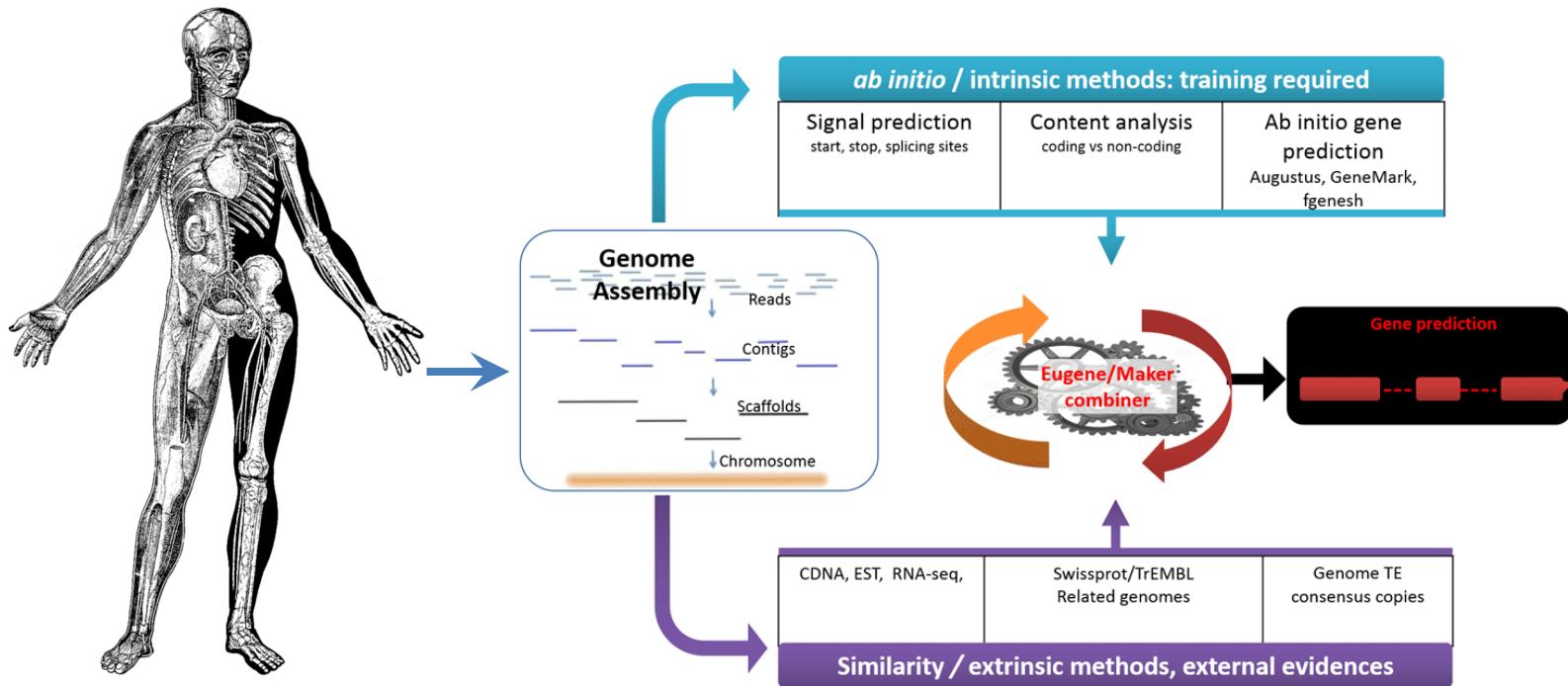
Ensembl Genomes: Extending Ensembl across the taxonomic space.

Large scale genome sequencing projects – Why?

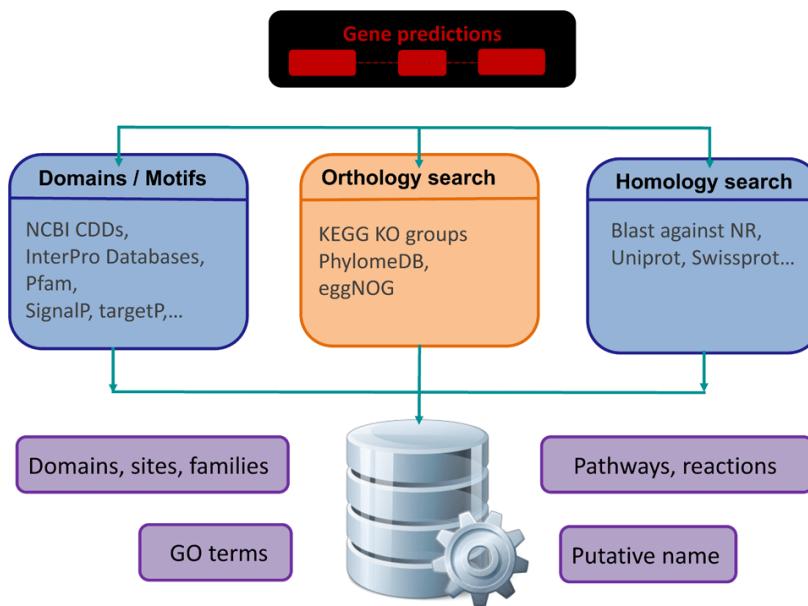


- Access to the genome sequence of related organisms / species
- Generate catalogues of
 - the encoded protein-coding and RNA genes
 - transposable elements
 - Structural variation
 - Genetic diversity within populations
 - Metabolic pathways -> natural compounds
- Assessing the metabolic capacities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

The prediction of genes from genome assemblies



Functional annotation of Genes *in silico*



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)

UNIVERSITY OF CALIFORNIA
SANTA CRUZ Genomics Institute

UCSC Genome Browser Gateway

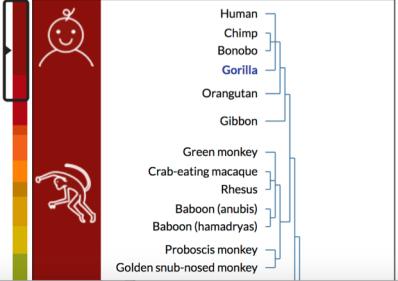
Browse>Select Species

POPULAR SPECIES

Human Mouse Rat Fruity Worm Yeast

Enter species or common name

REPRESENTED SPECIES



Human
Chimp
Bonobo
Gorilla
Orangutan
Gibbon
Green monkey
Crab-eating macaque
Rhesus
Baboon (anubis)
Baboon (hamadryas)
Proboscis monkey
Golden snub-nosed monkey

Find Position

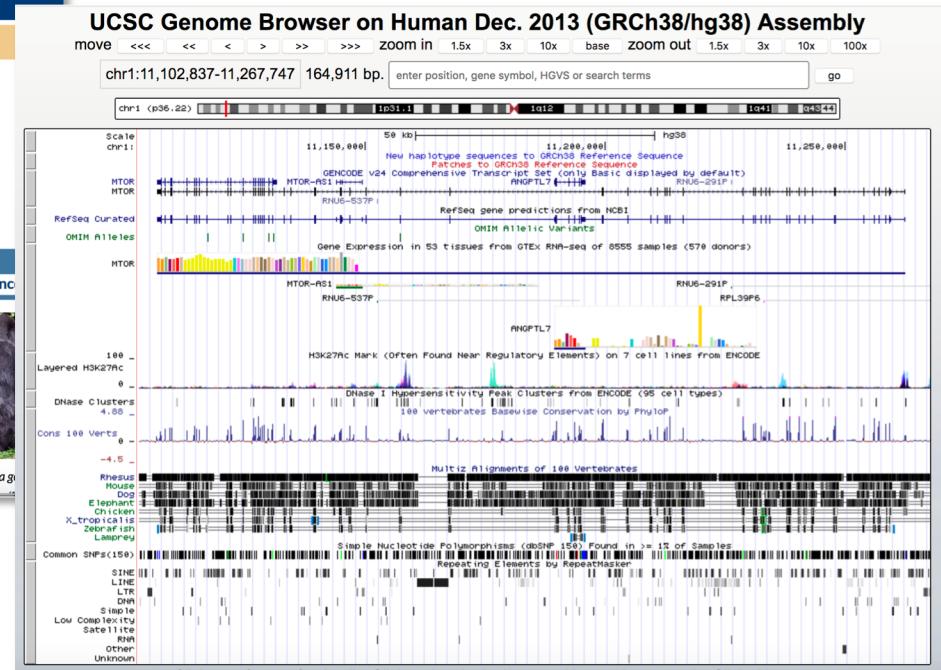
Gorilla Assembly
Dec 2014 (gorGor4.1/gorGor4)

Position/Search Term
Enter position, gene symbol or search terms
Current position: chr5:23,396,981-23,454,173

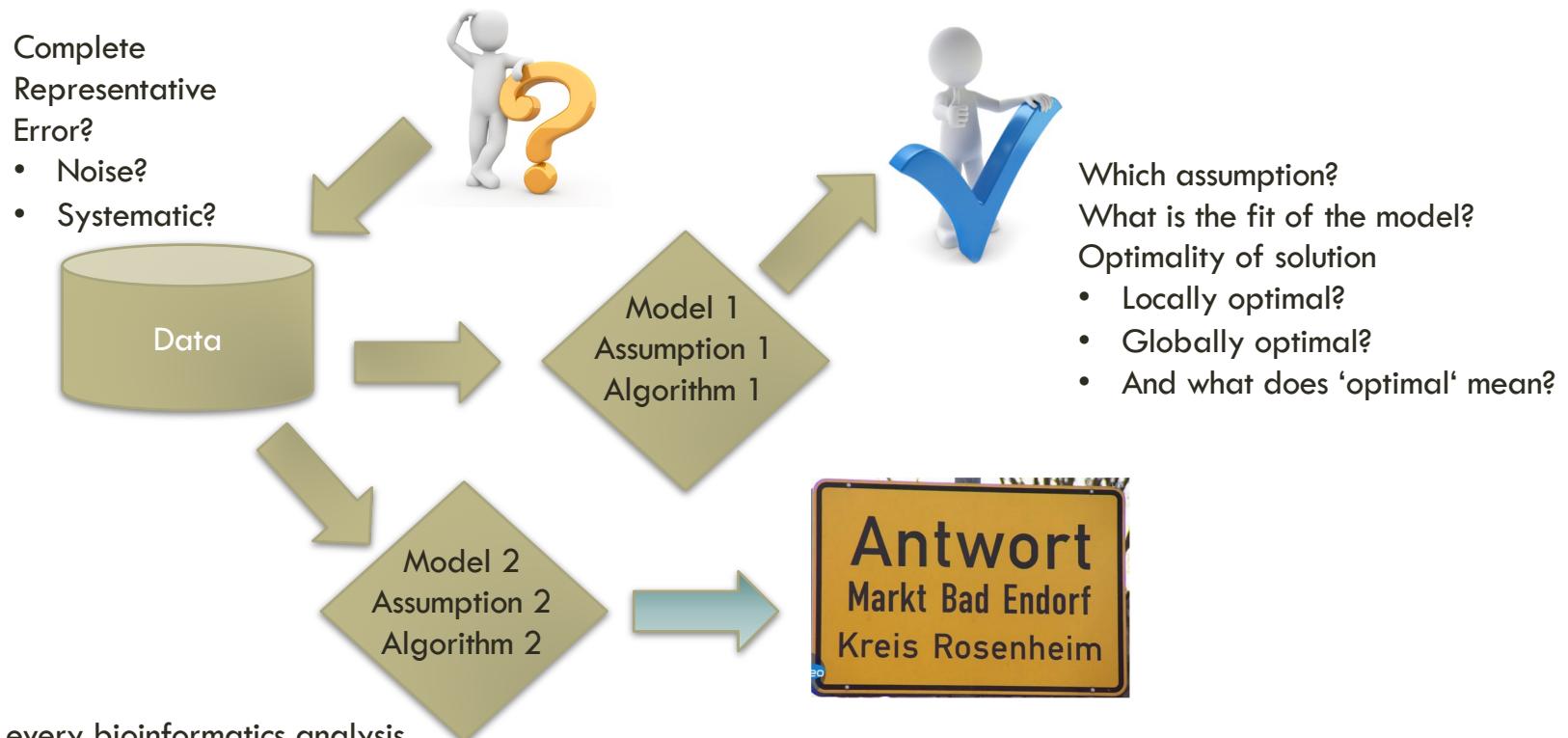
GO

Gorilla Genome Browser - gorGor4 assembly
view sequence

UCSC Genome Browser assembly
ID: gorGor4
Sequencing/Assembly provider ID: Wellcome Trust Sanger Institute
gorGor4
Assembly date: Dec. 2014
Accession ID: GCA_000151905.3
NCBI Genome ID: 2156 (Gorilla gorilla gorilla)
NCBI Assembly ID: 503571



WHAT TO CONSIDER IN A TYPICAL COMPARATIVE GENOMICS ANALYSIS¹?



¹ And actually in every bioinformatics analysis

DNA SEQUENCING TECHNOLOGIES

