

# Aufbau des menschlichen Genoms

Peter N. Robinson

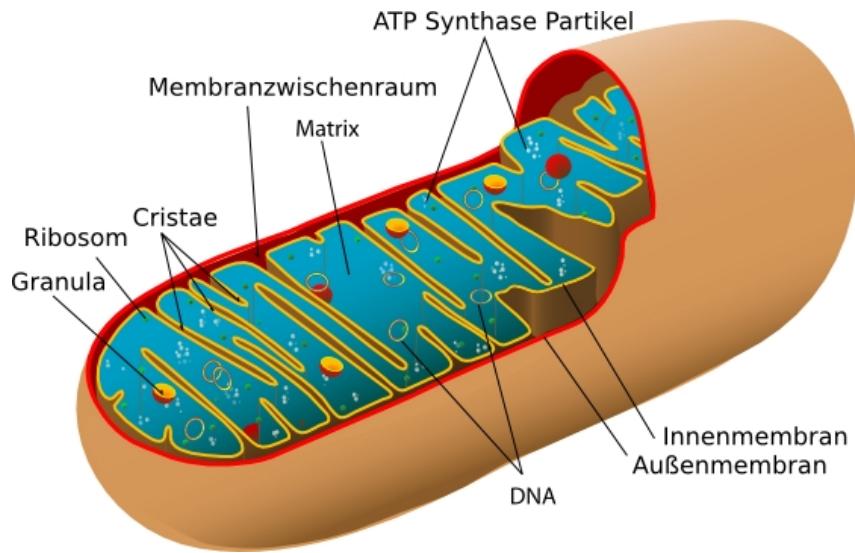
Institut für medizinische Genetik  
Charité Universitätsmedizin Berlin

9. Juni 2008

## Heute

- Einige Aspekte des menschlichen Genoms:
  - ▶ Kerngenom vs. Mitochondriengenom (S&R Kapitel 9.1)
  - ▶ CpG-Inseln
  - ▶ Repetitive, nichtkodierende DNA (S&R Kapitel 9.5)
  - ▶ Proteinkodierende Gene und Genfamilien (S&R Kapitel 9.3)
  - ▶ Expression proteinkodierender Gene (S&R Kapitel 10.1, 10.2)
- In der letzten Vorlesung: RNA: Spleißen, miRNA und andere RNA-Familien

# Das Mitochondrion



Wikipedia commons

- Kraftwerk eukaryontischer Zellen
- Besonders wichtig für Zellen, die viel Energie verbrauchen: Muskel, Nerven, Herzmuskel

## Kerngenom und Mitochondriengenom

Urheberrechtlich geschütztes Bild entfernt

# Mitochondriopathien

- 37 vom mitochondrialen Genom kodierte Gene: oxidative Phosphorylierung (13), rRNA (2), tRNA (22)
- Weitere vom Kerngenom kodierte Proteine werden aus dem Zytosol 'importiert'
- Mitochondropathien: Symptome vor allem in Organen, die viel Energie verbrauchen: Myopathie, Kardiomyopathie, Taubheit, Retinopathie, Diabetes mellitus, verschiedene ZNS-Symptome wie Epilepsie, Schlaganfall
- ca. 1:10.000 Menschen betroffen.

## Das Mitochondriengenom

- 93% der DNA: kodierend
- keine Introns
- 16,6 kb

Urheberrechtlich geschütztes Bild entfernt

# Der mitochondriale genetische Code

- Etwas anderer Code als im Kerngenom
- Das mitochondriale Genom kodiert alle tRNA-Moleküle für die eigene Proteinsynthese ( $n=22$ )
- 8 tRNAs: erkennen Familien von vier Codons, die sich an der 3. Base unterscheiden
- 14 tRNAs: erkennen Paare von Codons: XYPu oder XYPy
- $8 \times 4 + 14 \times 2 = 60$
- Die übrigen 4 Codons sind Stopp-Codons

## Die fünfte Base: 5-Methylcytosin

Urheberrechtlich geschütztes Bild entfernt

- Bei vielen mehrzelligen Tieren wird ein Teil der Cytosinreste methyliert
- DNA-Methyltransferase
- Bei Vertebraten ist das CpG-Dinukleotid Ziel für eine Cytosinmethylierung
- Beim Menschen wird ein Anteil von ca. 3% der Cytosinreste methyliert, vor allem in CpG

# Desaminierung von 5-Methylcytosin

Urheberrechtlich geschütztes Bild entfernt

- 5-Methylcytosin ist chemisch instabil
- Desaminierung:  $C \Rightarrow T$

## Depletion von CpG in der Evolution

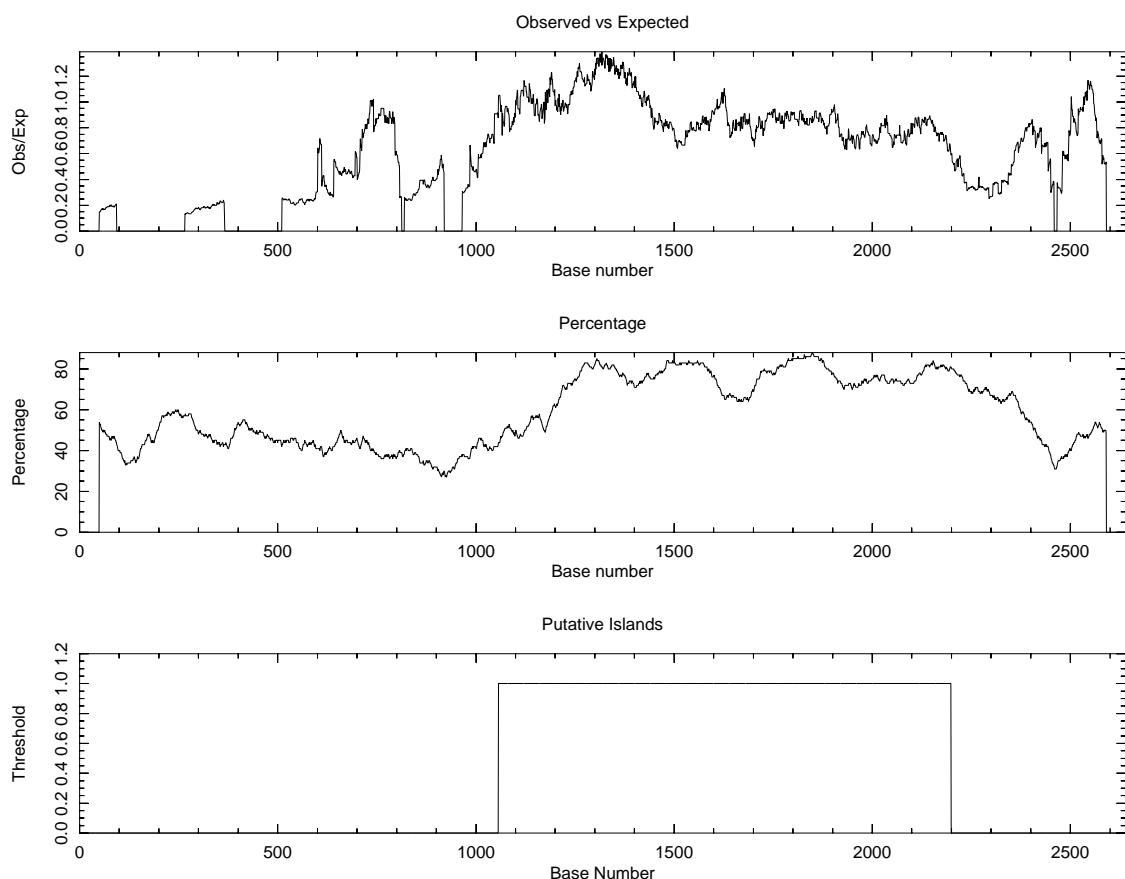
Urheberrechtlich geschütztes Bild entfernt

- CpG  $\Rightarrow$  TpG (und im komplementären Strang zu CpA)
- Das CpG-Dinukleotid: ca. 20% der auf Grund von  $p_C$  und  $p_G$  erwarteten Frequenz im Genom

# CpG-Insel

- Nichtmethylierte Regionen im Genom, die häufig am 5'-Enden von Hauskeeping-Genen und Entwicklungs-assoziierten Genen gelegen sind
- Der nichtmethylierte Status verlangsamt den Abbau der CpG-Dinukleotide während der Evolution
- Ca. 56% der menschlichen Gene sind mit CpG-Inseln assoziiert.
- CpG-Inseln überlappen häufig mit dem Promoter
- Definition: G+C-Gehalt über 50%, Verhältnis: beobachtete vs. erwartete CpGs von  $\geq 0,6$

# CpG-Insel



# Exon-Intron-Struktur

Urheberrechtlich geschütztes Bild entfernt

- Wenig Gene ohne Introns
- Durchschnittliche Exonlänge ca. 200 bp
- Intronlänge variiert stark

## Das menschliche Genom und seine Statistik (1)

Genomgröße	~ 3 200 Mb
Euchromatin	~ 2 900 – 3 000 Mb
Konstitutives Heterochromatin	> 200 Mb
Stark konservierter Anteil	> 100 Mb (> 3%)
kodierend	~ 50 Mb (~ 1,5%)
duplizierte DNA-Segmente	> 150 Mb (> 5%)
Nicht kodierende, repetitive DNA	> 50% des Genoms

# Das menschliche Genom und seine Statistik (2)

- Wie viele Gene hat das menschliche Genom?
- Noch unbekannt. Laut <http://www.ensembl.org> Version 49 (2008)

Bekannte proteinkodierende Gene	21.541
Vorhergesagte proteinkodierende Gene	1.119
Pseudogene	2.081
RNA-Gene	4.421

## Gendichte Regionen & Überlappende Gene

Urheberrechtlich geschütztes Bild entfernt

## "Genwüsten"

- Beispiel: Genwüsten auf Chromosom 7
- Insgesamt 20 solche Genwüsten mit ca. 20460 kb

Größe	Ort	Gene innerhalb der Region	Repetitive DNA (%LINEs/ SINEs)
1850 kb	7q11.22	2 vorhergesagte Gene	> 35%
1740 kb	7q31	3 vorhergesagte Gene; 1 Pseudogen	> 37%
1700 kb	7p12.2	3 vorhergesagte Gene; 1 Pseudogen	> 34%

## "Genwüsten"

- Einige Genwüsten zeigen eine auffällige Konservierung und scheinen mehrere regulatorische Elemente, die für die benachbarten Gene bedeutsam sind †.
- Benachbarte Gene oft mit Funktion in der Entwicklung

Urheberrechtlich geschütztes Bild entfernt

† Ovcharenko I (2005) Evolution and functional classification of vertebrate gene deserts  
Genome Res. 15:137–145.

# Gibt es Junk DNA?

## Onion Test

The onion test is a simple reality check for anyone who thinks they have come up with a universal function for non-coding DNA. Whatever your proposed function, ask yourself this question: Can I explain why an onion needs about five times more non-coding DNA for this function than a human?

— T. Ryan Gregory

Zwiebelgenom: ~ 15 Gb

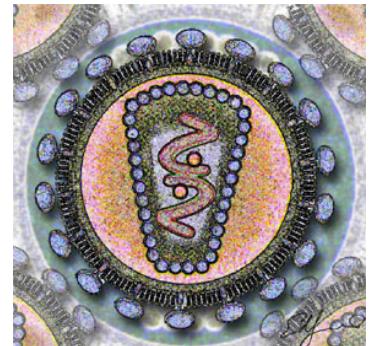


## Transposons

- Aus Transposons abgeleitete Repeats ⇒ ca. 40% des menschlichen Genoms
- "springendes Gen"
- Bewegliche DNA-Sequenzen, die zu verschiedenen Stellen im Genom wandern können
- Vermutlich aus Retroviren hervorgegangen

# Retroviren (1)

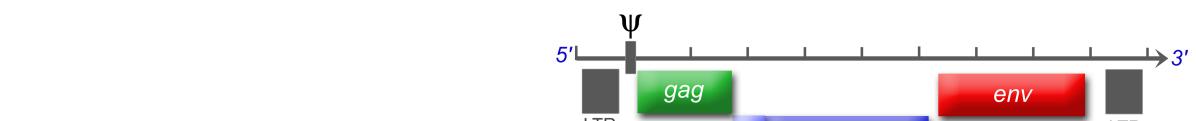
- ssRNA Viren
- Unterschiedliche Retroviren befallen Säugetiere, Vögel, Fische,...
- Beim Menschen HIV und HTLV-I



Wikipedia commons

## Retroviren (2)

- Einzelsträngiges RNA-Genom von 7–12 kb
- Drei Gene und zwei *Long Terminal Repeats* (LTR)
- Nach Eintritt in die Wirtszelle wird die RNA mittels Reverser Transkriptase in DNA umgeschrieben und mittels Integrase in das Genom der Wirtszelle eingefügt.
- Die Reverse Transkriptase (Virusenzym) ist relativ ungenau (hohe Mutationsrate)

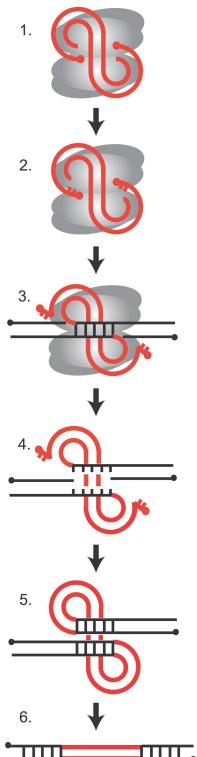


- *gag*: Nukleokapsidproteine
- *pol*: Protease, Reverse Transkriptase und Integrase.
- *env*: die Proteine der Hülle

Wikipedia commons

# Retroviren (3): Integrations ins Wirtsgenom

- ① Präintegrationskomplex (virale DNA, Integrase).
- ② Die Integrase entfernt zwei Nukleotide von den 3'-Enden der viralen DNA, so dass 5'-Überhänge entstehen.
- ③ Die Integrase schneidet die Wirts-DNA an einer zufälligen Stelle
- ④ Die Paarung der Wirts-DNA wird aufgelöst.
- ⑤ DNA-Reparaturenzyme, Ligasen
- ⑥ Virusgenom nun vollständig ins Wirtsgenom integriert



Wikipedia commons

# Retroviren (4): LTR

- Long Terminal Repeat
- Integration ins Wirtsgenom (zusammen mit Integrase)
- Promoter/Enhancer-Sequenzen: Bringt Transkriptionsmaschinerie der Wirtszelle dazu, die zwischen den LTRs gelegenen viralen Gene zu transkribieren
- Auch TATA-Box, und Bindungssequenzen für Transkriptionsfaktoren der Wirtszelle
- Expression durch Wirtszellenenzyme RNA pol II, ...

Kingsman SM, Kingsman AJ (1996)

Urheberrechtlich geschütztes Bild entfernt

Eur. J. Biochem. 240:491–507.

# Retrotransposons

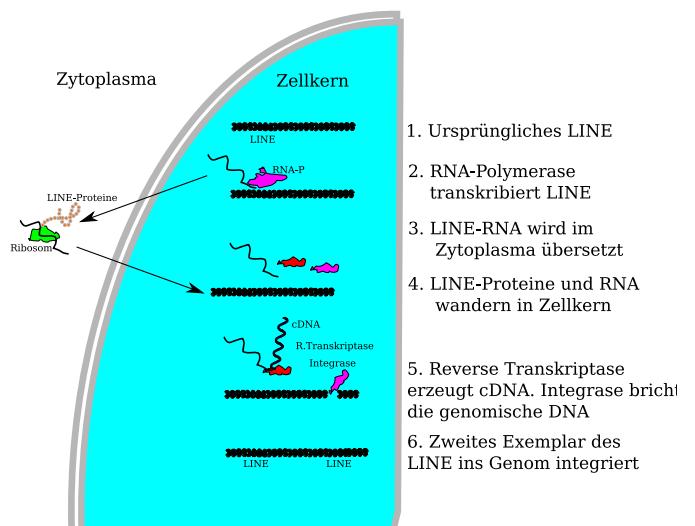
- Transponieren ("Springen") mittels Reverser Transkriptase
- Kopiertransposition: Eine Kopie wird hergestellt, welche sich an anderer Stelle im Genom integriert
- Drei Transposonklassen der Säuger:
  - ▶ LINEs
  - ▶ SINEs
  - ▶ Retrovirusähnliche Elemente mit LTR-Sequenzen

## LINEs

- Long interspersed nuclear elements ("Lange, eingestreute Kernelemente")
- Autonom transponierbare Elemente, d.h., LINEs kodieren alle für die Retrotransposition notwendigen Produkte, etwa Reverse Transkriptase
- Drei Familien, LINE-1, LINE-2, LINE-3, die zusammen etwa 20% (!) des Genoms ausmachen
- LINE-1: Größte Familie (17% des Genoms), das einzige LINE, bei dem noch aktive Transpositionen stattfinden

# LINE: Lebenszyklus

- Vollständiges LINE: 6,1 kb
- Kodiert zwei Proteine:  
RNA-bindendes Protein und  
Integrase/Reverse-Transkriptase  
(RT)
- Bei der Integration setzt die RT  
nicht immer bis zum 5'-Ende fort,  
sodass viele eingefügte LINEs  
verkürzt/funktionslos sind



## SINE-Sequenzen

- Short interspersed nuclear elements
- 100–400 bp lang
- Wichtigstes Familienmitglied: *Alu*-Sequenzen (häufigste Sequenz des menschlichen Genoms, etwa alle 3 kb)
- Codieren keine Proteine, können sich nicht autonom vermehren

# Alu-Lebenszyklus

- a) ~ 300 bp mit 3' oligo(dA) Schwanz, flankiert von kurzen direkten repeats (schwarze Pfeile). A und B: RNA-Polymerase III Promotoren<sup>b</sup>
- b) Retrotransposition durch RNA-P III über das Ende des Alu-Elements hinaus (bis nächstes Stoppsignal, TTTT)
- c) Insertion
- d) Neues Alu-Element

Urheberrechtlich geschütztes Bild entfernt

Batzer MA, Deininger PL (2002) Alu repeats and human genomic diversity. Nature Genetics Reviews 3:370–380

<sup>b</sup> RNA-P III katalysiert die Bildung von tRNA, 7SL-RNA und 5S rRNA. Die 7SL-RNA ist für den intrazellulären Transport von Proteinen zuständig. Alu-Elemente sind aus einem 7SL-RNA-Gen hervorgegangen

## Mit Retroelementen assoziierte Krankheiten

- Die Insertion einer neuen Kopie eines Retroelements kann z.B. zur Unterbrechung der protein-kodierenden Sequenz eines Gens führen.
- Zum Beispiel gibt es Berichte über die Entstehung einer Hämophilie A durch Insertion eines LINEs in ein Exon des Faktor-VIII-Gens
- Jedoch eher selten verglichen mit Punktmutationen

# Alu-Elemente und Translokationen

- Beispiel: t(11;22) (q23;q11) ist eine relativ häufig beobachtete balanzierte Translokation
- Nachkommen von Trägern dieser balancierten Translokation können auf Grund einer 3:1 meiotischen Nondisjunction ein der[22]-Syndrom, d.h., sie haben ein zusätzliches Chromosom 22pter-22q11::11q23-11qter
- 47,XX,+der(22)t(11;22)(q23;q11) bzw.  
47,XY,+der(22)t(11;22)(q23;q11)
- Häufige Symptome: Geistige Retardierung, Fehlbildungen der äußeren Ohren, Gaumenspalte, Herzdefekte

Urheberrechtlich geschütztes Bild entfernt

Medne et al. (2008), Eintrag in Gene Reviews (NIH)

## t(11;22)

- Die Bruchpunkte für die t(11;22) Translokation liegen in spezifischen Alu-Elementen auf Chromosom 11 und 22<sup>†</sup>
- Ursache der Translokation: Alu-Alu-Rekombination

Urheberrechtlich geschütztes Bild entfernt

<sup>†</sup> Hill AS et al. (2000) The most frequent constitutional translocation in humans, the t(11;22)(q23;q11) is due to a highly specific alu-mediated recombination. Hum Mol Genet.9:1525-32.

- Beide Alu-Elemente dieser Translokation enthalten eine Kern-Alusequenz (5' –CCTGTAATCCCAGCACTTGGGAGGC–3')<sup>†</sup>
- Molekulare Mechanismen der Translokationsentstehung noch umstritten (interchromosomal homologe Rekombination, nonhomologous end-joining,?), aber die Sequenzhomologie der Alu-Elemente spielt vermutlich eine wichtige Rolle

Urheberrechtlich geschütztes Bild entfernt

---

<sup>†</sup> Hill AS et al. (2000) The most frequent constitutional translocation in humans, the t(11;22)(q23;q11) is due to a highly specific alu-mediated recombination. Hum Mol Genet.9:1525-32.

## Nicht nur junk?

- Homologe Rekombination zwischen LINEs bzw. Alu-Sequenzen hat durch die Bildung von großen DNA-Umlagerungen wesentlich zur Evolution des menschlichen Genoms beigetragen.
- Retroelemente haben auch eine Rolle bei Genduplikation gespielt und somit die Bildung von Genfamilien begünstigt
- Regulatorische Sequenzen von Retroelementen sind im Verlauf der Evolution auch für die Funktionen "normaler" Gene rekrutiert worden
- Bei vielen Spezies werden SINEs unter Stressbedingungen vermehrt transkribiert. Die Transkripte binden an die Proteinkinase PKR und blockieren diese. Da die PKR normalerweise die Proteinsynthese hemmt, stimulieren die SINEs somit die Proteintranslation
- Da es Abertausende von SINEs gibt, kann somit eine sehr schnelle Regulation erreicht werden.

# Pseudogene

- Schadhafte Kopien von Genen
- Verkürzte oder mutierte Fragmente
- Ca. 21.000 proteinkodierende Gene
- ca. 11.000 nichtprozessierte Pseudogene
- ca. 8.000 prozessierte Pseudogene

## Pseudogene (2)

Urheberrechtlich geschütztes Bild entfernt

# Nichtprozessierte Pseudogene

- Duplikation eines Gens mit Exons, Introns, Promoter
- Auf Grund von falschen Stoppcodons in Exonsequenzen leicht zu erkennen

Urheberrechtlich geschütztes Bild entfernt

Beispiel: Humanes *RPL21*-Gen und eines seiner Pseudogene  $\Psi RPL21$

Gerstein & Zheng, The Real Life of Pseudogenes, Scientific American, 2006

## NF1-Pseudogene

- NF1-Gen: 17q11.2
- 11 nichtprozessierte Pseudogene bzw. Kopien von Genfragmenten auf 7 verschiedenen Chromosomen

Urheberrechtlich geschütztes Bild entfernt

# Prozessierte Pseudogene

- Schadhafte Kopien von Genen, die die Exonsequenzen eines funktionellen Gens enthalten sowie am ENde eine Oligo(dA)-Sequenz
- Solche prozessierte Pseudogene sind auf der Ebene der cDNA durch Retrotransposition kopiert worden (wahrscheinlich unter Beteiligung des LINE-1-Transpositionssystems)
- Prozessierte Pseudogene werden normalerweise nicht exprimiert

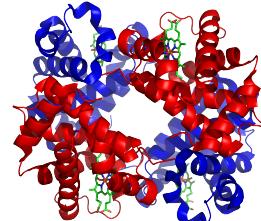
## Prozessierte Pseudogene

- Entstehung eines prozessierten Pseudogens

Urheberrechtlich geschütztes Bild entfernt

# Genfamilien

- Ein großer Teil der Gene des menschlichen Genoms gehören zu Genfamilien, die unter sich einen hohen Grad an Sequenzübereinstimmung zeigen
- Beispiel: Globingenfamilie
  - ▶  $\alpha$ -Globin
  - ▶  $\beta$ -Globin
  - ▶ Myoglobin
  - ▶ Weitere Familienmitglieder in Bakterien und Pflanzen
- Vgl. Strachan & Read für weitere Beispiele

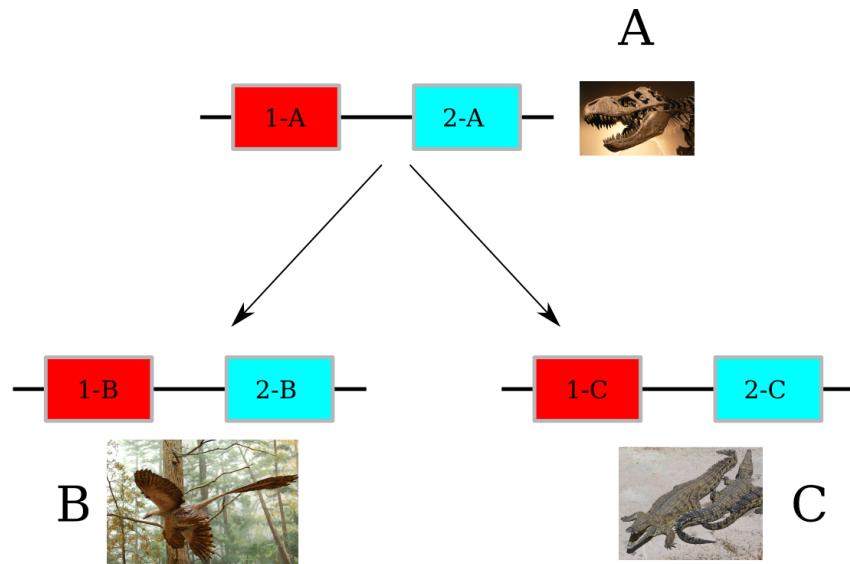


## Homologie, Paralogie und Orthologie

- **Homologie:** Beschreibt eine Beziehung zwischen zwei Sequenzen, die einen gemeinsamen evolutionären Ursprung haben
  - ▶ Ein allgemeiner Begriff
  - ▶ Umfasst Orthologe und Paraloge
  - ▶ Häufiger Fehler: "Gen X und Gen Y zeigen eine Homologie von 90%"  $\Rightarrow$  Entweder sind Sequenzen homolog oder nicht! Besser: "Gen X und Gen Y sind homolog und zeigen eine Sequenzidentität von 90%"

# Orthologe

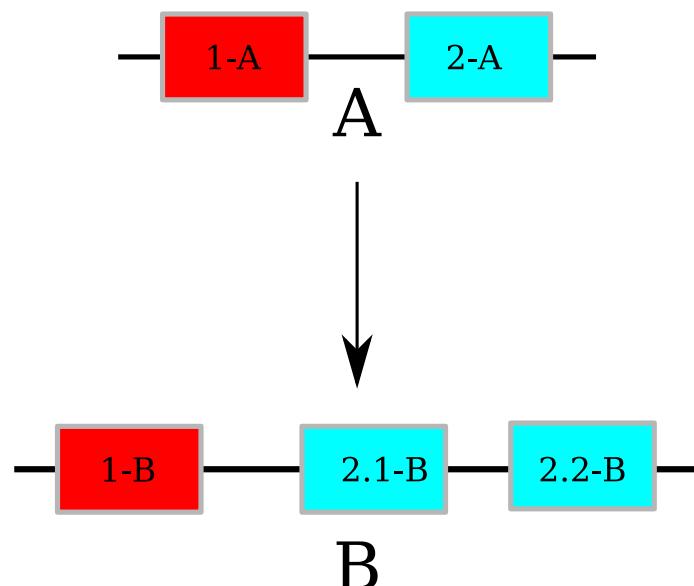
- Ortholog sind Gene, die durch Speziation aus einem einzigen Gen vom gemeinsamen Vorfahren entstanden sind
- Hier Gene 1 und 2 stammen von Organismus A (Vorfahr von Organismen B und C); die von diesen Genen entstandene Gene in B und C sind **ortholog**



Bilder: Wikipedia commons

# Paraloge

- Paraloge sind Gene, die durch Duplikation entstanden sind
- Beispiel  $\alpha$  und  $\beta$ -Globingene (Komponenten von Hämoglobin). Das Opossum (und vermutlich unsere evolutionären Vorfahren) hatten lediglich  $\beta$ -Globin



- Unter Paralogen unterscheidet man noch In– und Outparalogen
  - ▶ Outparalogs (Paraloge Gene, deren Duplikation vor einem Speziationsereignis stattgefunden hat)
  - ▶ Inparalogs (Paraloge Gene, deren Duplikation nach einem Speziationsereignis stattgefunden hat)

Urheberrechtlich geschütztes Bild entfernt

Wheeler et al. (2001) PNAS 98:1101–1106

## In– und Outparaloge

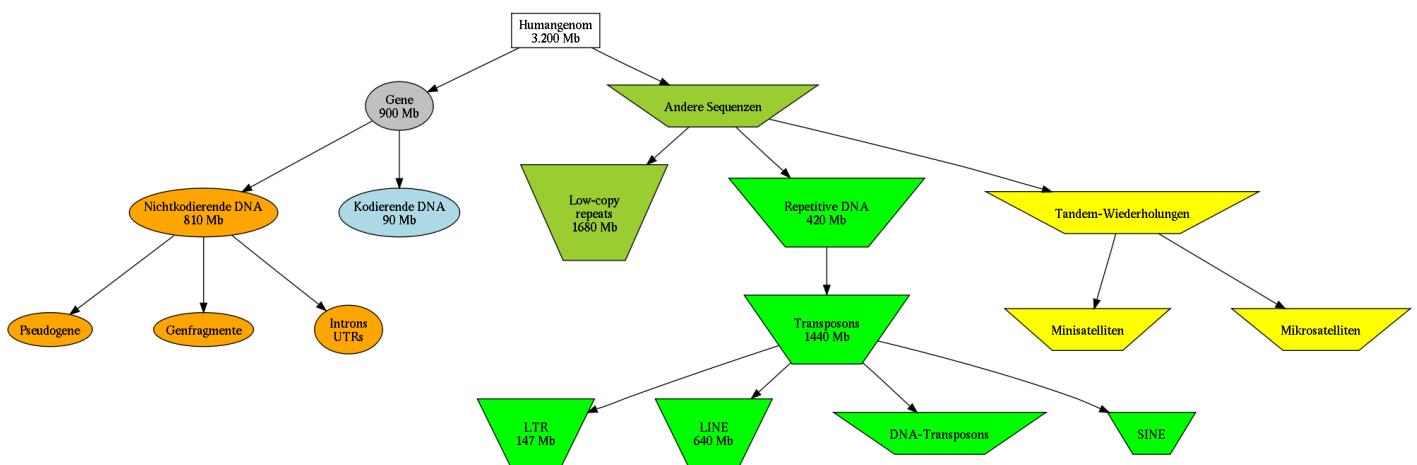
- ① Duplikation von Protein A in Spezies A ⇒ Protein A1 und A2
- ② Speziation: Entstehung von Spezies B und C
- ③ Duplikation von Protein A2 in C-Genom ⇒ Protein C2 und C3
- Innerhalb des C-Genoms sind C2 und C3 **Inparaloge**: **Nach** Speziation entstanden
- C2 und C3 sind co-ortholog mit B2
- B1 und B2 sind **Outparaloge**: Duplikation and Divergenz **vor** Speziation

Urheberrechtlich geschütztes Bild entfernt

O'Brien et al (2005) Inparanoid: a comprehensive database of eukaryotic orthologs Nucleic Acids

Research, 33:D476–D480

# Die Zusammensetzung des menschlichen Genoms



## Die Kontrolle der Genexpression

- Die Expression menschlicher Gene auf verschiedenen Ebenen gesteuert
- Eines der wichtigsten Ziele in der Bioinformatik ist das Verständnis der Netzwerke, welche die Genexpression steuern
- Regulation erfolgt auf mehreren Ebenen
  - ▶ Transkription
  - ▶ Posttranskriptional
  - ▶ Epigenetische Mechanismen (beruhen nicht auf Veränderung der Gensequenz)
- Heute: Kurze Einführung in transkriptionelle Regulation

# cis und trans

- Gallia **cisalpina**: *Gallien diesseits der Alpen*
- Gallia **transalpina**: *Gallien jenseits der Alpen*



Wikipedia commons

## cis und trans (2)

- Genregulation auf Ebene der Transkription erfolgt über die Bindung von Proteinen an regulatorische DNA-Sequenzen
- **trans**-Faktoren: Die regulatorischen Proteine (Transkriptionsfaktoren) werden von entfernt liegenden Genen kodiert und gelangen erst nach ihrer Synthese im Zytoplasma und ggf. Aktivierung wieder in den Zellkern an ihren Wirkungsort, deshalb bezeichnet man sie als **trans**-aktiv
- **cis**-Sequenzen: Die regulatorische DNA-Sequenzen, an welche die trans-Faktoren binden, liegen i.d.R. in der Nachbarschaft des regulierten Gens und werden deshalb als **cis**-aktiv bezeichnet

# Kernpromotoren

Urheberrechtlich geschütztes Bild entfernt

- Der Core-Promoter bindet an mehr oder weniger ubiquitäre Transkriptionsfaktoren und ermöglicht so eine basale Transkriptionsrate, die durch Aktivität des weiter upstream gelegenen Promoters bzw. von Enhancern oder Silencern modifiziert werden kann
- Elemente eines Core-Promoters: TATA-Box, INR, DPE, BRE

## Insulinpromoter

Urheberrechtlich geschütztes Bild entfernt

- Ubiquitäre oder allgemein exprimierte Transkriptionsfaktoren: schwarz
- für  $\beta$ -Zellen des Pankreas spezifische Faktoren: rot
- CRE, cAMP-Response-Element; NRE, negativ regulatorisches Element.
- PDX1 bindet an vier Sequenzmotive mit der Struktur C(C/T)TAATG, die im Insulinpromotor vorkommen (A1, A2, A3, A5)

# Promoter des $\alpha$ -Globingens

Urheberrechtlich geschütztes Bild entfernt

- Die regulatorische HS-40-Sequenz des  $\alpha$ -Globingens enthält viele Erkennungselemente für erythroidspezifische Transkriptionsfaktoren.

## Strukturmotive

Urheberrechtlich geschütztes Bild entfernt

- Strukturmotive, die häufig in Transkriptionsfaktoren vorkommen
  - ▶ HTH, Helix-Turn-Helix; HLH: Helix-Loop-Helix

# Bindung an DNA

Urheberrechtlich geschütztes Bild entfernt

- Die Bindung von konservierten Strukturmotiven in Transkriptionsfaktoren an die Doppelhelix.

## Steroidrezeptoren

Urheberrechtlich geschütztes Bild entfernt

- ER, Östrogenrezeptor; GR, Glucocorticoidrezeptor; PR, Progesteronrezeptor; RAR, Retinsäurerezeptor; TR, Thyroxinrezeptor; VDR, Vitamin D-Rezeptor.
- Kleine hydrophobe Hormone diffundieren durch die Plasmamembran, binden an und aktivieren Kellkernhormonrezeptoren, welche in den Kern wandern und an spezifische DNA-Response-Elemente binden, und somit die (50–100) Zielgene aktivieren

# Steroidrezeptoren (2)

Urheberrechtlich geschütztes Bild entfernt

- Der Glucocorticoidrezeptor wird normalerweise durch die Bindung des Inhibitorproteins Hsp90 inaktiviert. Die Bindung von Glucocorticoiden an den Rezeptor setzt Hsp90 frei, der Rezeptor dimerisiert und aktiviert dann spezifische Gene, die im Promotor ein Glucocorticoid-Response-Element enthalten.

## Signalübertragung über den cAMP-PKA-Weg

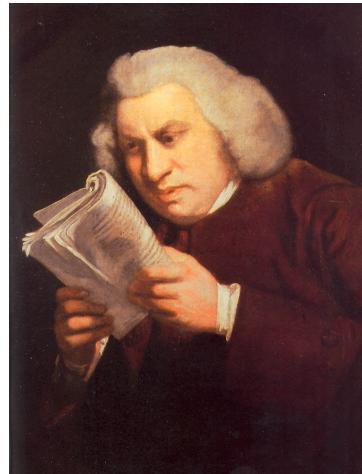
- R: regulatorische Untereinheiten der Proteinkinase A
- C: katalytische Untereinheiten
- CRE: cAMP-Response Element
- CREB: CRE-bindendes Protein

Urheberrechtlich geschütztes Bild entfernt

# The End of the Lecture as We Know It

- Diese Vorlesungsdias stehen unter der GNU-Lizenz für freie Dokumentation<sup>b</sup>
- Kontakt:  
[peter.robinson@charite.de](mailto:peter.robinson@charite.de)
- Strachan & Read Kapitel 9,  
10.1–10.2

<sup>b</sup> <http://www.gnu.org/licenses/fdl.txt>



Lectures were once useful; but now, when all can read, and books are so numerous, lectures are unnecessary. If your attention fails, and you miss a part of a lecture, it is lost; you cannot go back as you do upon a book... People have nowadays got a strange opinion that everything should be taught by lectures. Now, I cannot see that lectures can do as much good as reading the books from which the lectures are taken. I know nothing that can be best taught by lectures, except where experiments are to be shown. You may teach chymistry by lectures. You might teach making shoes by lectures!

Samuel Johnson, quoted in Boswell's Life of Johnson (1791).