Genomprojekte und die Kartierung des menschlichen Genoms

Peter N. Robinson

Institut für medizinische Genetik
Charité Universitätsmedizin Berlin

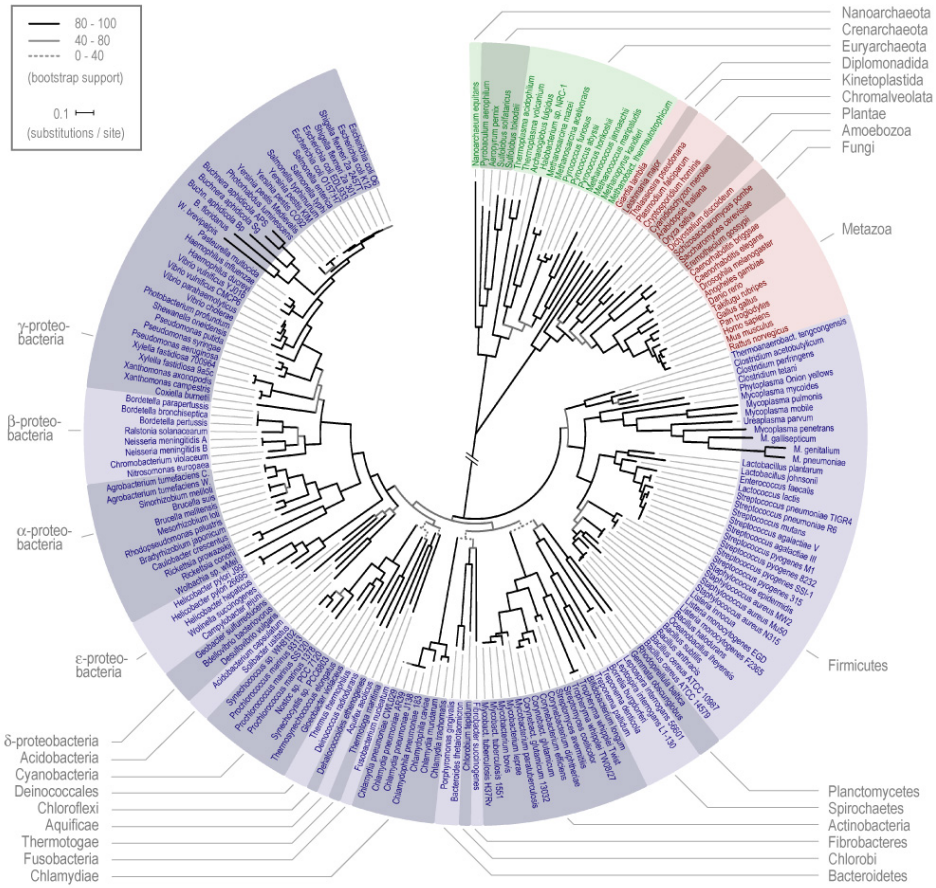
12. November 2014

Outline

1. Genomprojekte und Modellorganismen
2. Das Human-Genome Project
3. Kartierung

Vergleichende Genomics

Derzeit[1] $>$ 7012 Genome in NCBI-Datenbanken

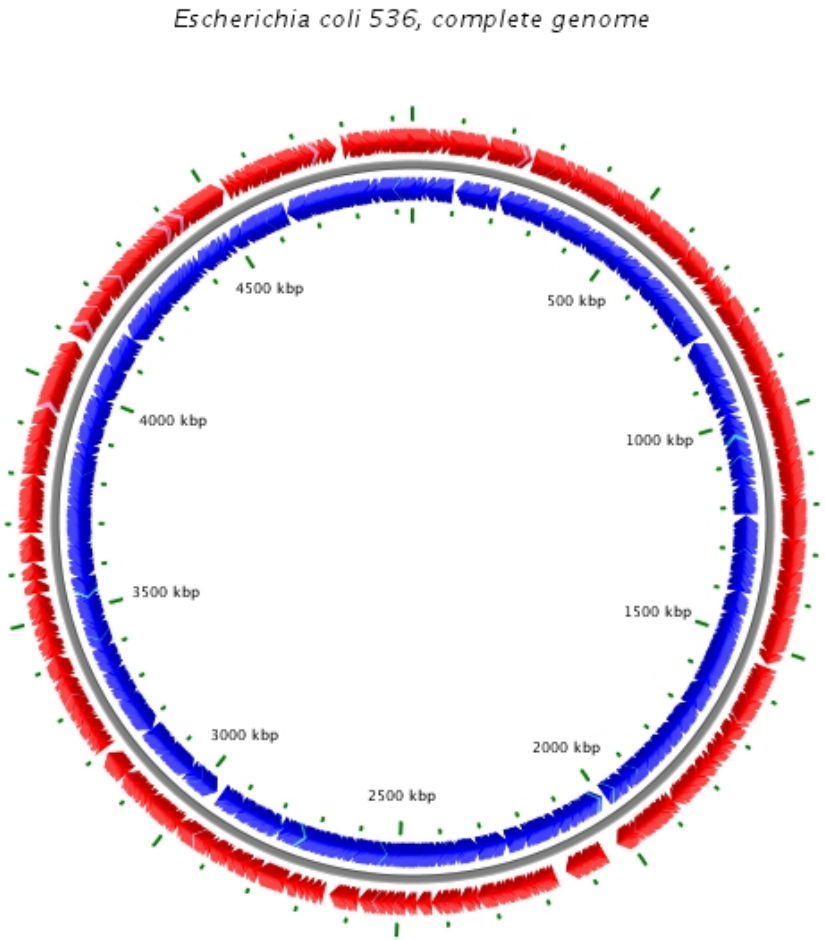


Bildquelle: `http://www.bork.embl.de/tree_of_life/`

[1] 1.04.2012. Im Jahr 2009 waren es “nur” 3021 Genome.

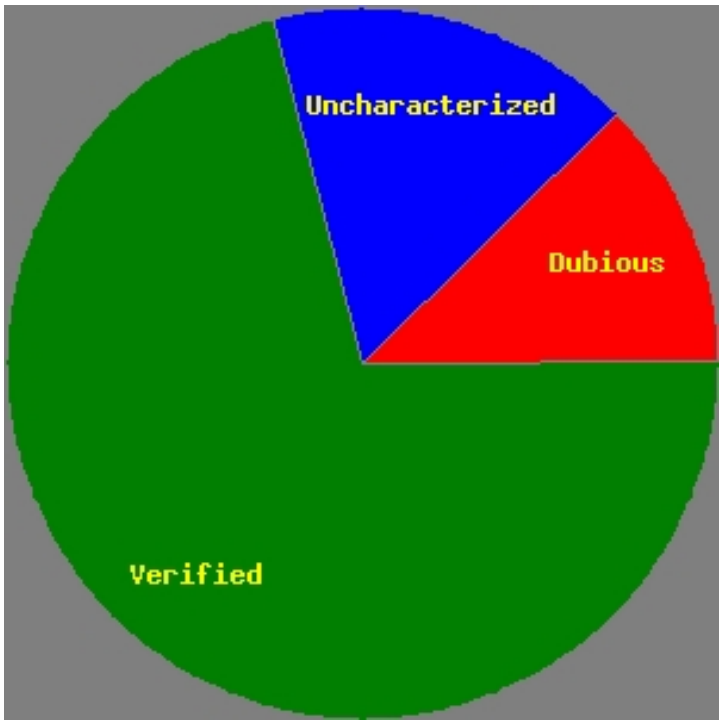
E. coli

- 4,9 Mb
- Ein Chromosom



Saccharomyces cerevisiae-Genom

- Knospungshefe
- 16 Chromosomen, Mitochondrion



- 4691 ORFs, 70.99%
- 1104 ORFs, 16.71%
- 813 ORFs, 12.30%

Bildquelle: `http://www.yeastgenome.org/`

Homologe Gene

Organismus	*H.sapiens*: homologe Gene (n)	% Gene im Organismus
E. coli	412	9%
Hefe	1785	30%
C. elegans	5019	25%
Drosophila	6057	44%
Dog	27761	81%
Chimp	29529	98%

Quelle: euGenes (Vorsicht: Zahlen nur ungefähr!)

Outline

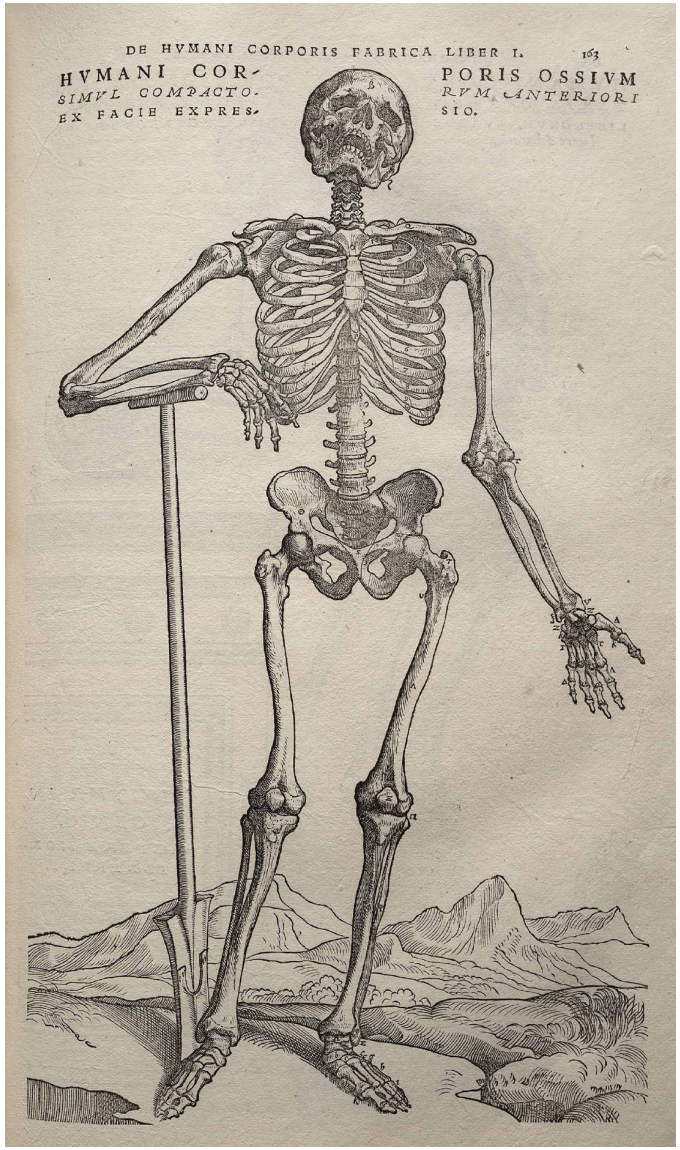
1 Genomprojekte und Modellorganismen

2 Das Human-Genome Project

3 Kartierung

Vesalius

- Andreas Vesalius
- De Humanis Corporis Fabrica (1543)
- Anfang der modernen Anatomie
- Grundlage für zahlreiche Entdeckungen

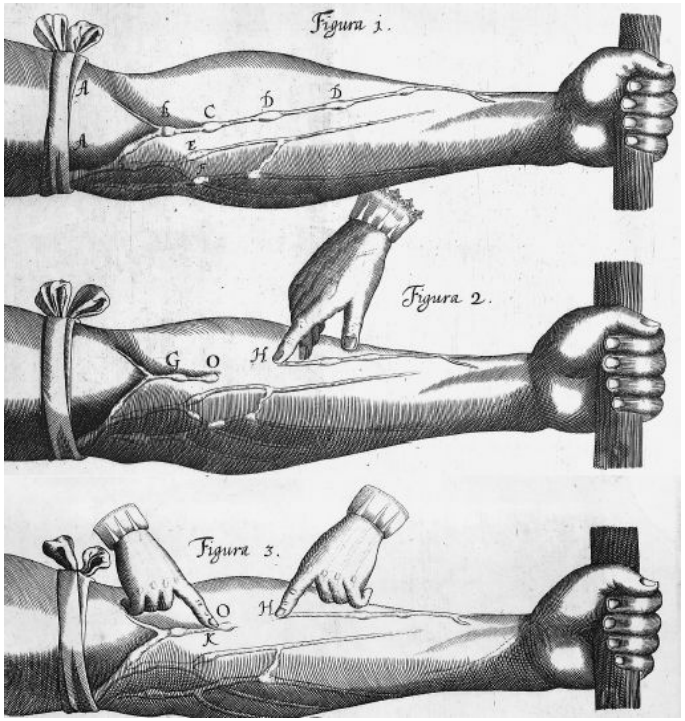


Von Galen zu Harvey

Galenos von Pergamon, 129-216 n. Chr.

Das Blut wird in der Leber erzeugt,

von hier aus fließt es in nur eine Richtung



William Harvey (1578 - 1657)

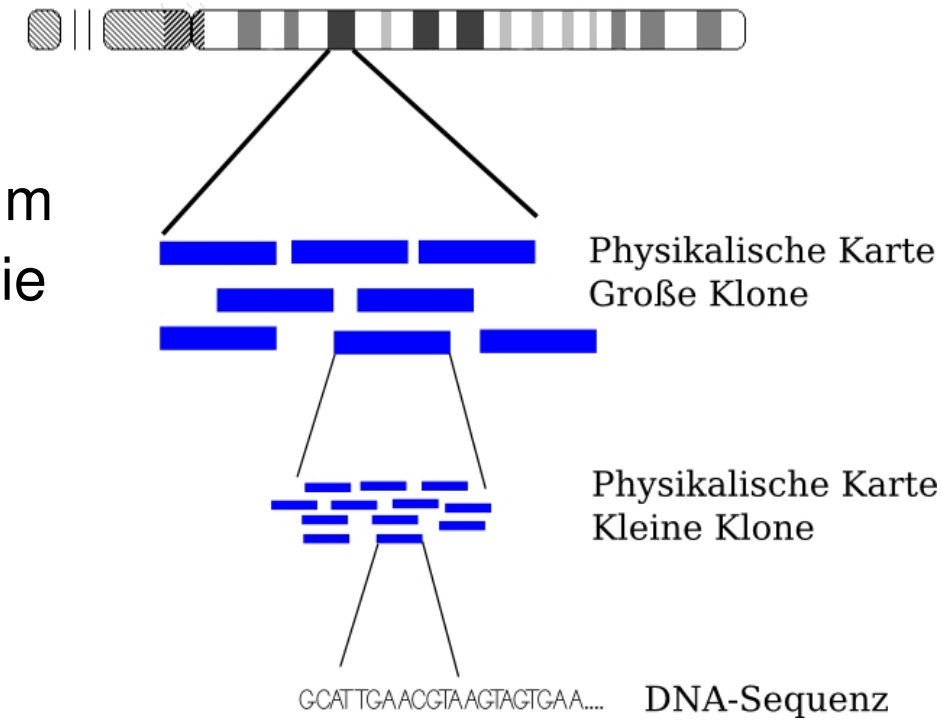
De motu cordis (1628).

Medizinische Ziele des Humangenomprojekts

- Genidentifikation bei Erbkrankheiten
- Besseres Verständnis von Genstruktur und -funktion: Grundlage einer modernen, molekularen Medizin
- Personalisiertes Medizin: Vorhersage von Krankheitsrisiken bzw. von individuellen Nebenwirkungen auf Medikamente durch Bestimmung von genetischen Varianten (Polymorphismen)
- Die Genomsequenz dürfte wie der Atlas von Vesalius Ausgangspunkt zahlreicher Entdeckungen werden

Grundlagen: Genetische Karte

- Eine genetische Karte zeigt die relativen Abstände von Markern auf einem Chromosom zueinander
- Voraussetzung/Grundlage, um eine detaillierte Karte bzw. die komplette Sequenz eines Genoms zu erstellen, d.h., ”Gerüst”



Grundlagen: Genetische Marker

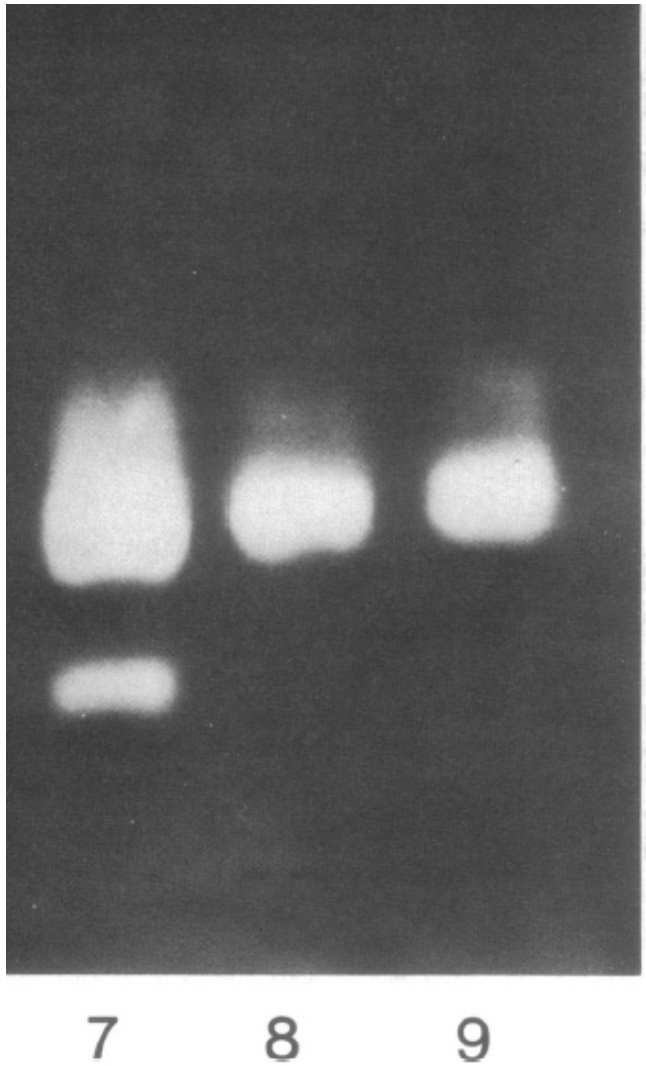
- Ein Marker ist ein beliebiges mendelndes Merkmal, mit dessen Hilfe man einen Chromosomenabschnitt in einem Stammbaum verfolgen kann.
- Weitere mathematische Einzelheiten (lod-score, θ) in späteren Vorlesungen
- Im folgenden soll gezeigt werden, welche Merkmale als Marker dienen können.

Marker im Zeitalter vor der DNA-Sequenzierung

- Schleutermann, Bias, Murdoch, McKusick (1969) Linkage of the Loci for the Nail-Patella Syndrome and Adenylate Kinase *Am J Hum Genet.* **21**: 606–630.
- Nail-Patella-Syndrome: Autosomal dominante Vererbung
- Kopplung mit ABO-Locus 1955 gezeigt
- Schleutermann et al. zeigten eine strikte Kosegregation zwischen dem klinischen Merkmal Nail-Patella-Syndrom und dem biochemischen Merkmal einer Adenylate-Kinase-Isoform

Adenylat-Kinase

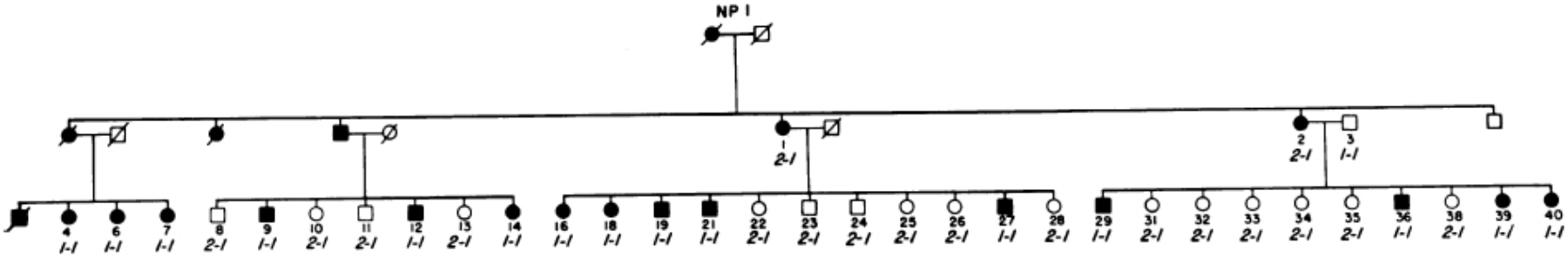
- Adenylatkinase: 2 ADP $\rightarrow$ ATP + AMP
- Unterschiedliche Isoformen, die sich elektrophoretisch trennen lassen



Nagel-Patella-Syndrom

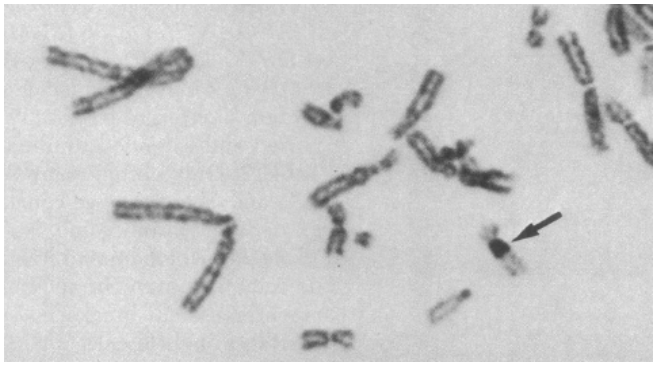
- Nail patella syndrome
- Autosomal dominant
- Klinische Diagnose einfach

Kopplung



- Die betroffenen (schwarze Symbole) erben immer die AK1-Isoform 1 vom betroffenen Elternteil
- 1-1: AK1-Isoform 1 (homozygot)
- 2-1: AK1-Isoform 1/Isoform 2
- Was fällt Ihnen auf?

AK1 und Chromosom 9



Westerveld A (1976) Assignment of the AK1:Np:ABO linkage group to human chromosome 9. Proc. Nat. Acad. Sci. USA **73**: 895–899

- Später konnte durch Hamster-Mensch Hybdridzelllinien mit jeweils einem menschlichen Chromosome gezeigt werden, dass das Gen für AK1 auf Chromosome 9 gelegen ist
- Enthielt der Hybrid ein Chromosom 9, beobachtete man AK1-Aktivität
- Enthielt der Hybrid ein anderes menschliches Chromosom, beobachtete man keine AK1-AKtivität

Frage

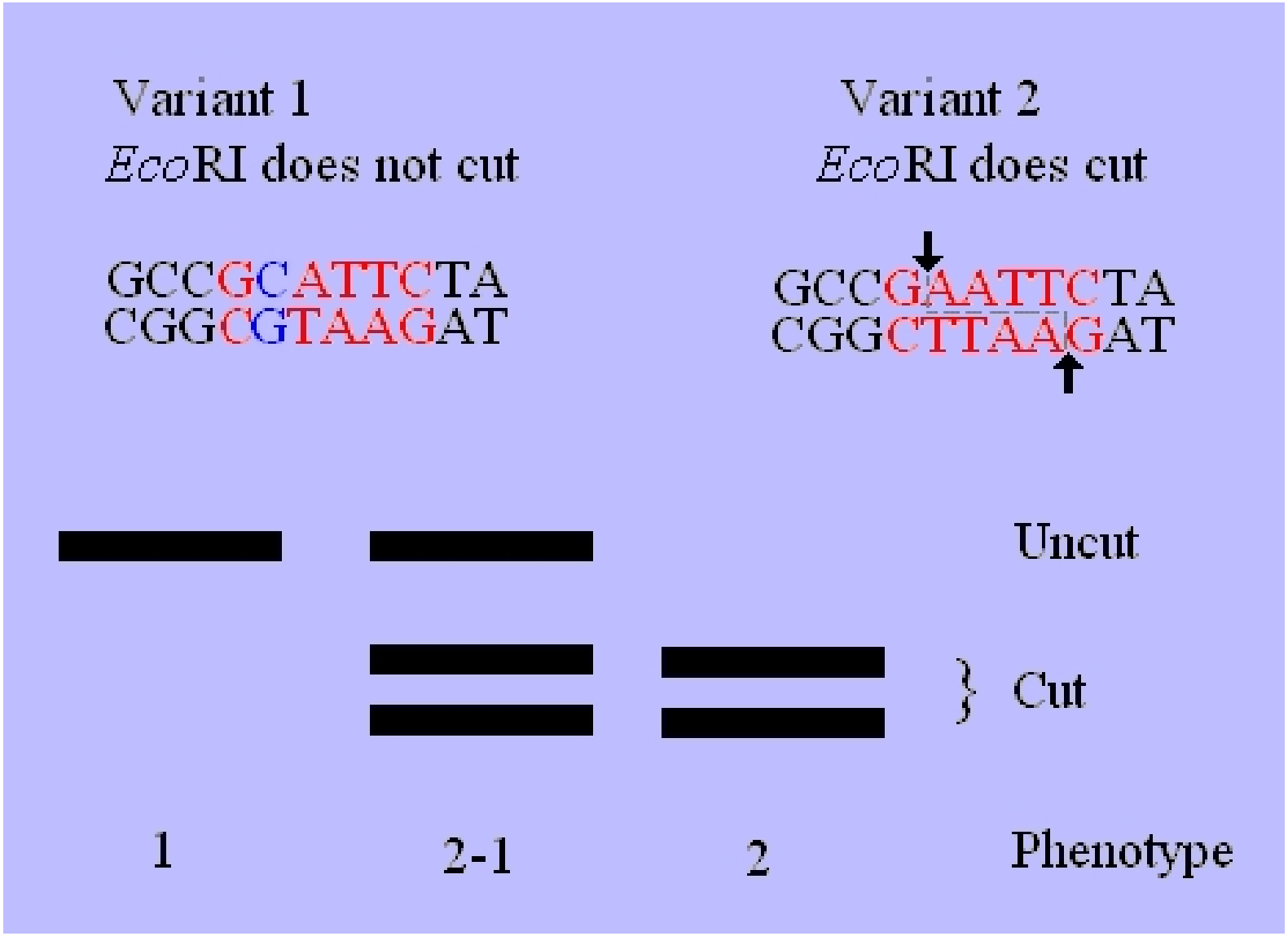
Auf welchem Chromosom ist das Gen für Nagel-Patella-Syndrom gelegen?

DNA-Polymorphismen

- Blutgruppen, Mobilitätsvarianten von Serumenzymen u.ä. sind ab 1910 als Marker eingesetzt worden
- Preis, Aufwand jedoch sehr hoch, ein relativ kleiner Teil der genetischen Karte wurde bis 1980 erschlossen
- Durchbruch: Erkenntnis dass DNA-Polymorphismen als genetische Marker eingesetzt werden können

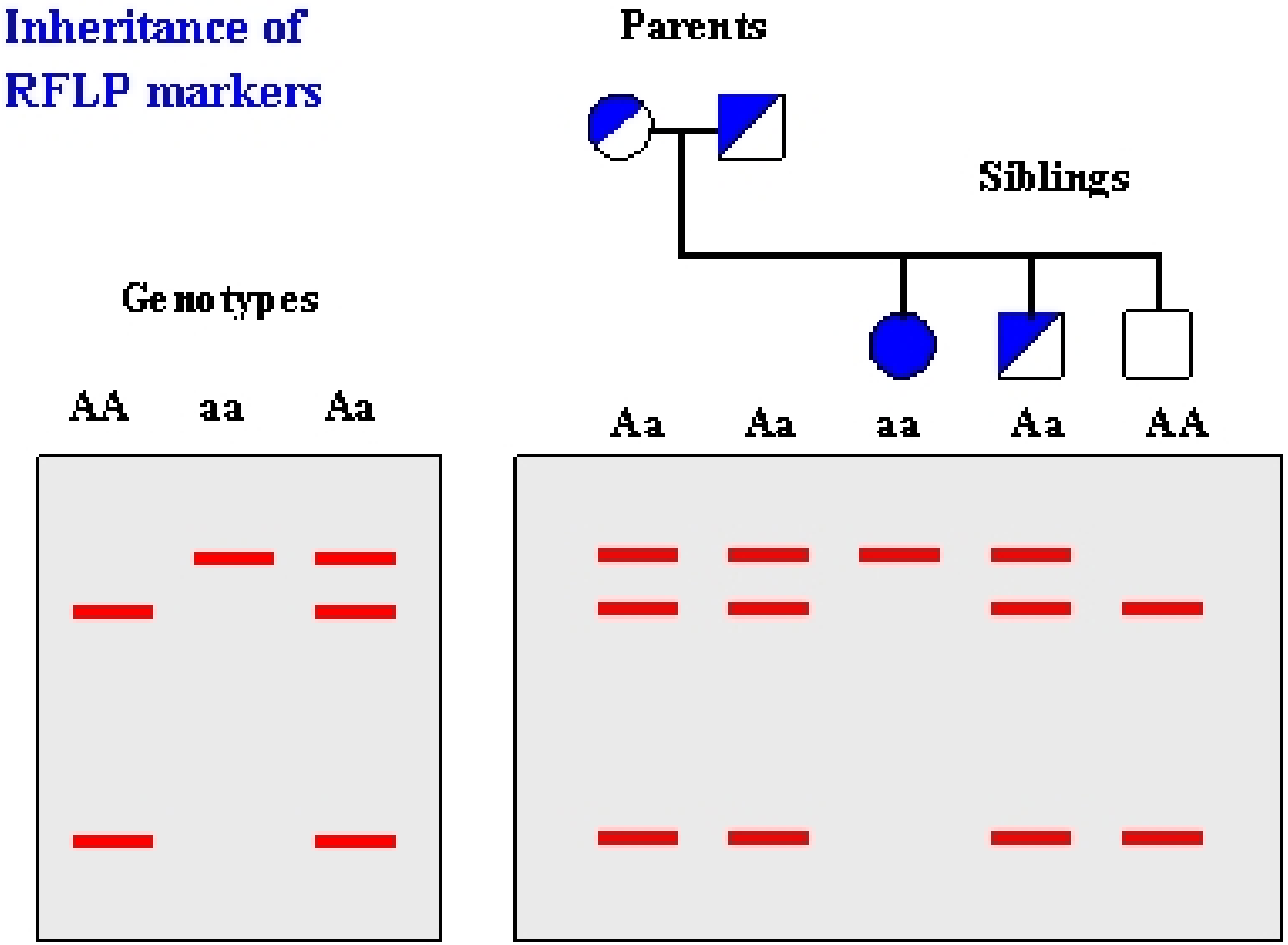
RFLP

Restriktionsfragmentlängenpolymorphismus



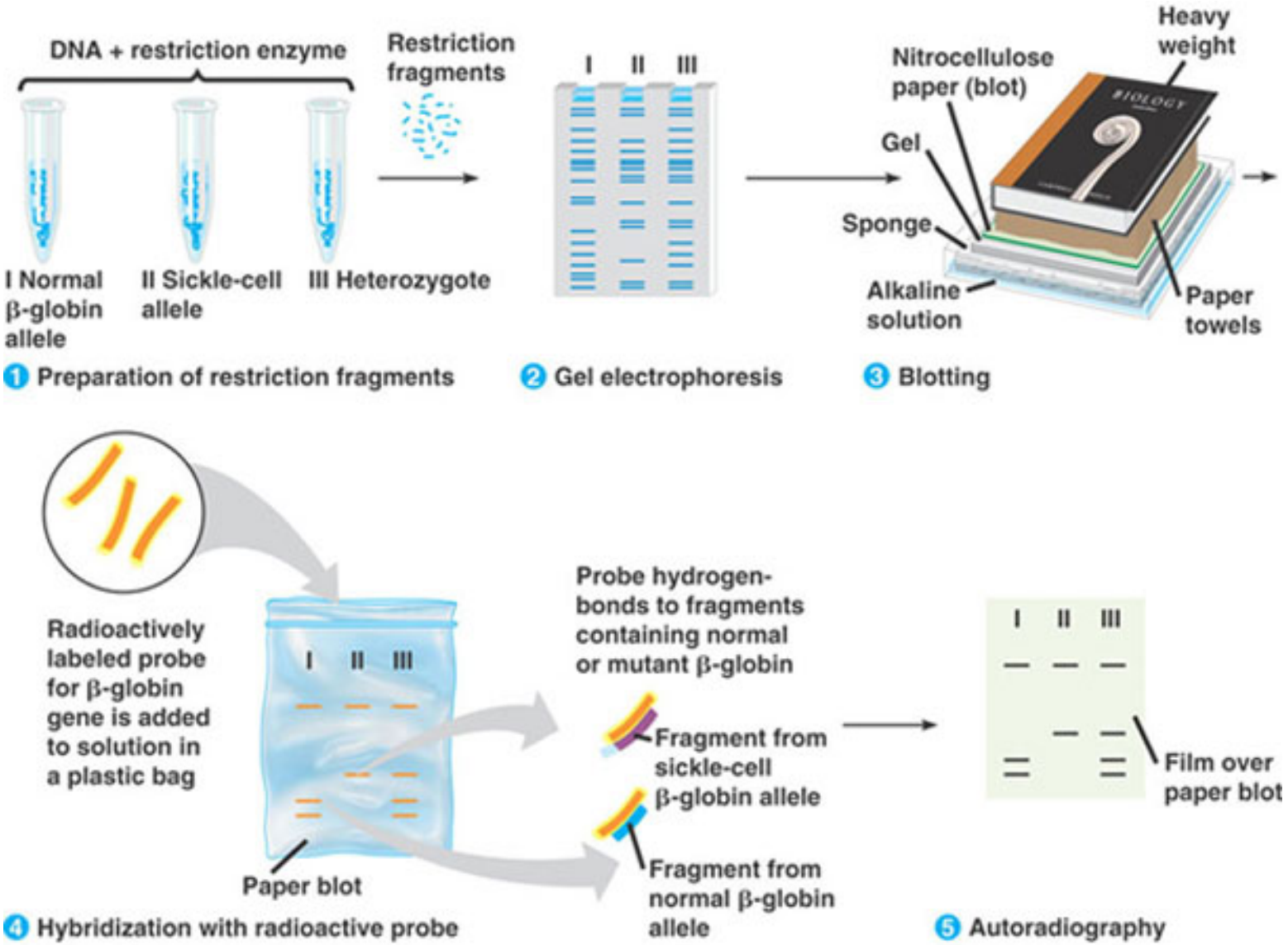
RFLP

Restriktionsfragmentlängenpolymorphismus



RFLP

- RFLP, Southern-Blot

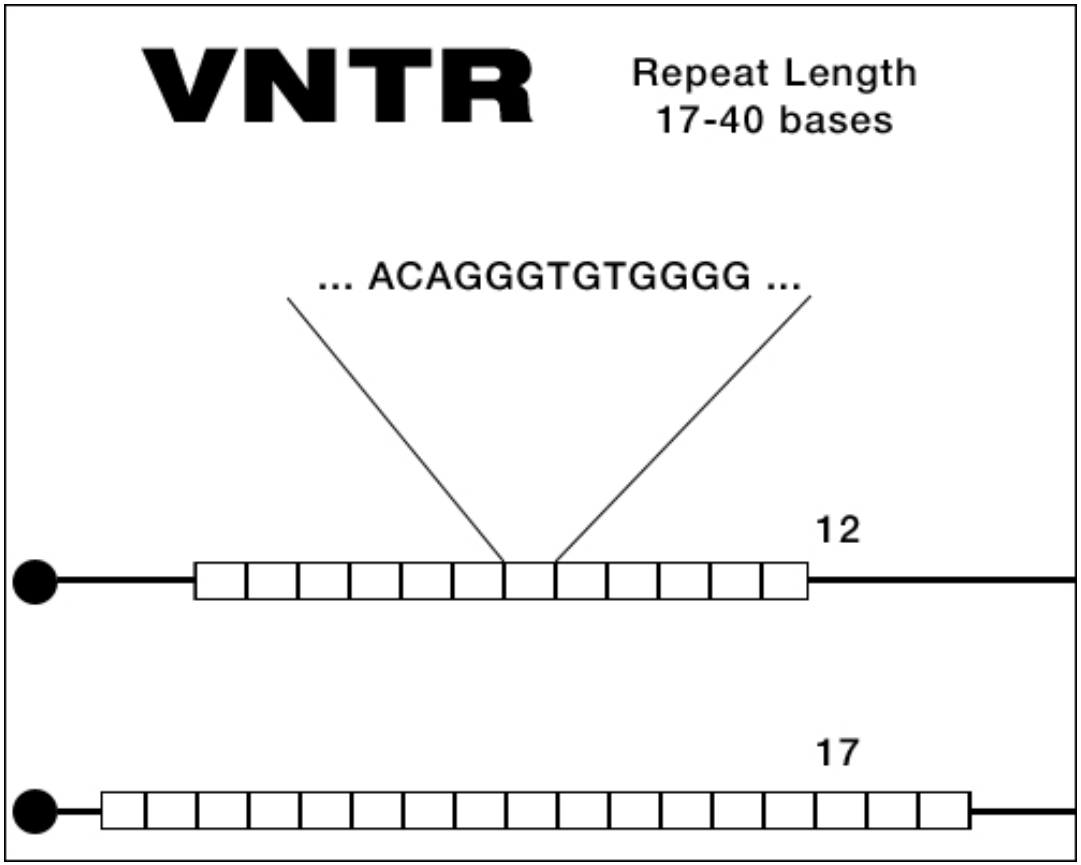


RFLP

- ca. 10^5 im menschlichen Genom
- Zwei Markerallele, maximale Heterozygotie 0,5
- Nachweis mittels Southern-Blotting, neuerdings PCR
- Relativ aufwändig

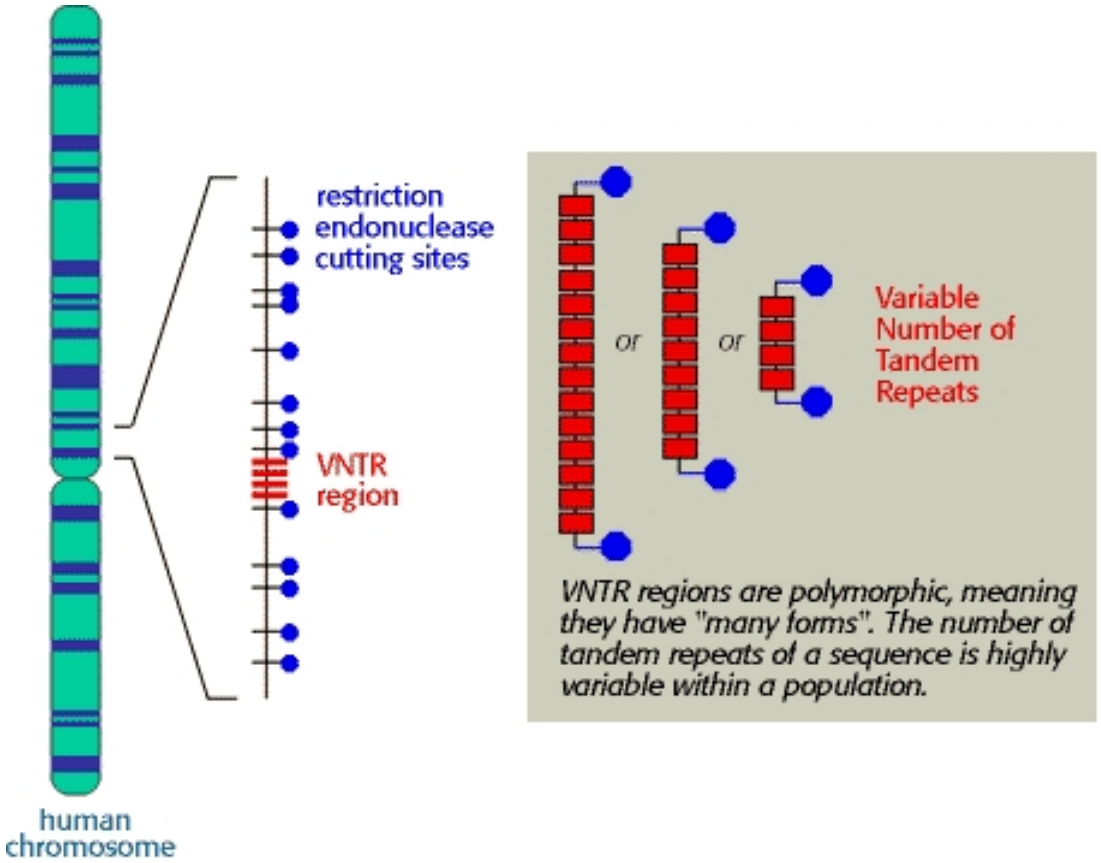
VNTR

- Variable Number of Tandem Repeats
- 10–100b
- Synonym: Minisatellit



VNTR

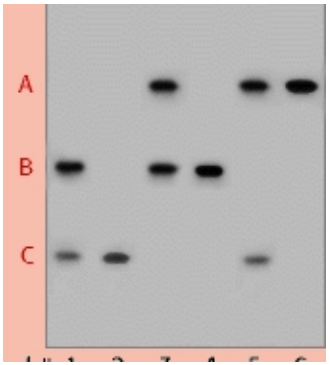
- Hochinformativ, da viele Allele

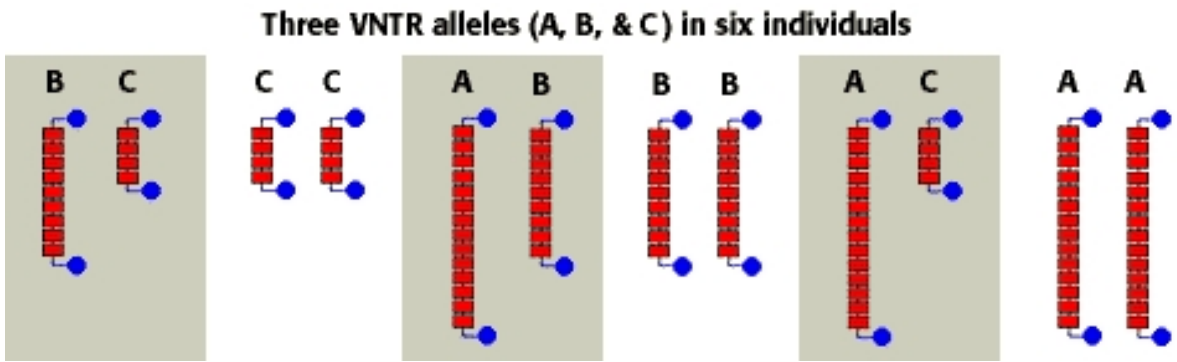


Bildquelle: www.biology.arizona.edu

VNTR

- Aufwändiger Nachweisverfahren (Southern-Blotting nach Restriktionsanalyse)





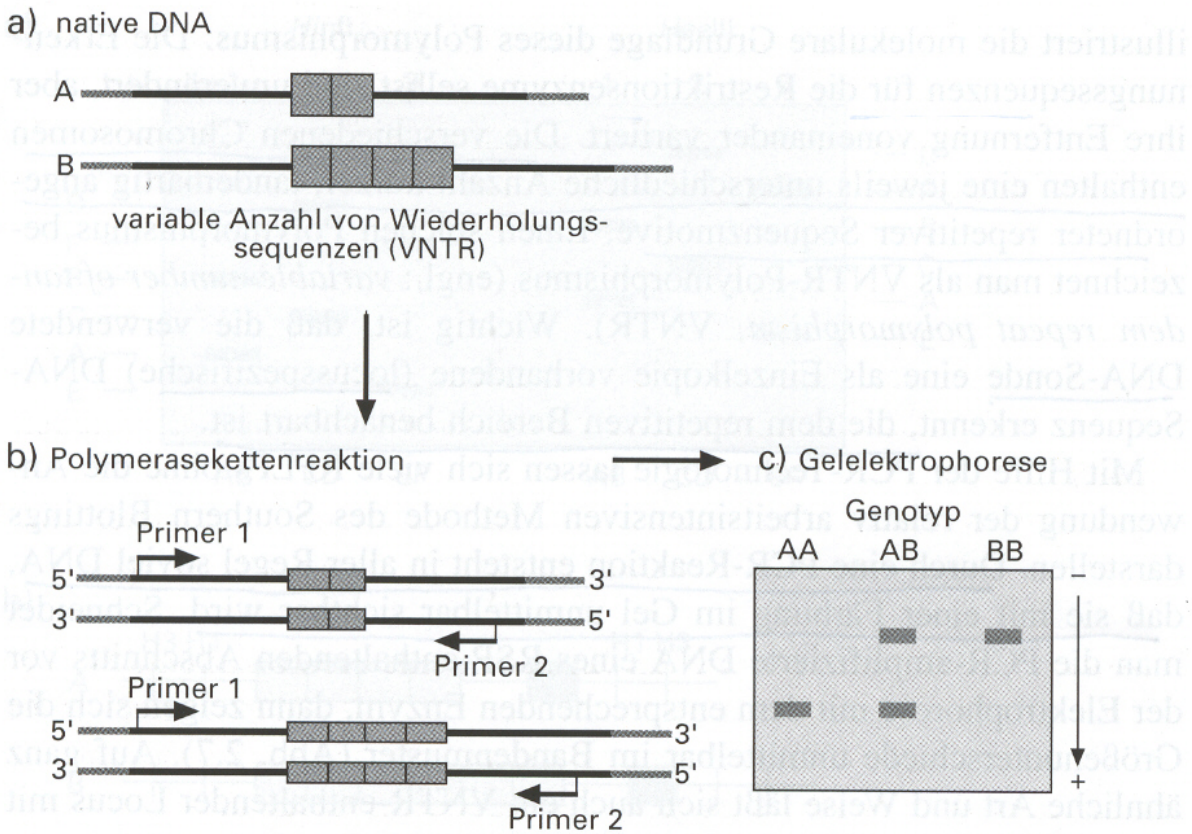
Bildquelle: www.biology.arizona.edu

VNTR

- VNTR (auch Minisatellit genannt)
- Viele Allele, hoch informativ
- geschätzt 10^4 im menschlichen Genom
- Nachweis: Southern-Blot, radioaktive Sonden (aufwändig)
- Ungleiche Verteilung im Genom
- Daher sind Minisatelliten nicht gut geeignet für Hochdurchsatzkartierung

Mikrosatelliten

- VNTR-Mikrosatelliten (2-10nt) sind mittels PCR leicht nachzuweisen
- Meist $(CA)_n$ repeats



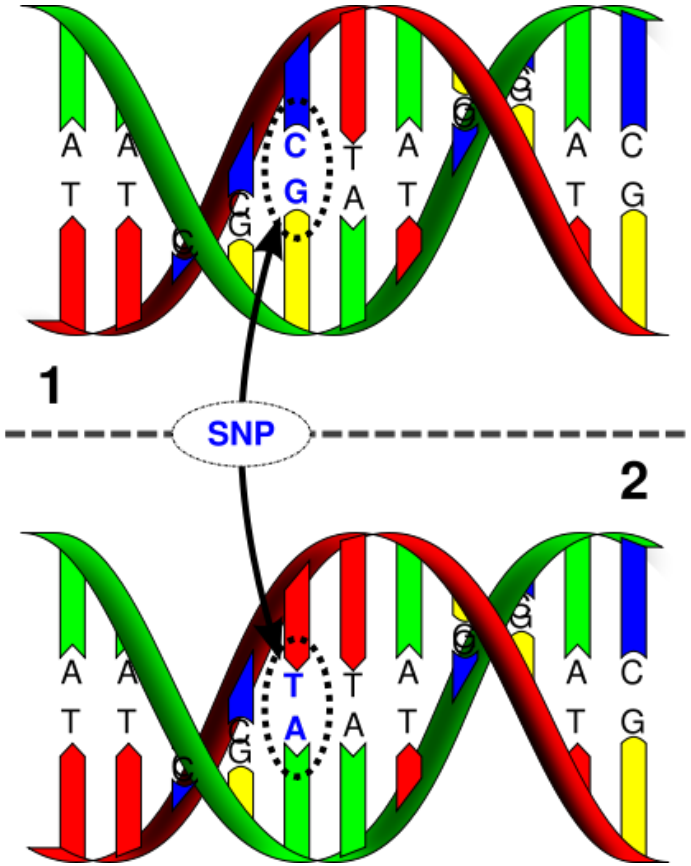
Bildquelle: Zentrale für Unterrichtsmedien im Internet e.V.

Mikrosatelliten

- Hochinformativ, da viele Allele
- ca. 10^5 im menschlichen Genom
- Bestimmung durch PCR, auch automatisierte Muiltiplex-PCR

SNPs

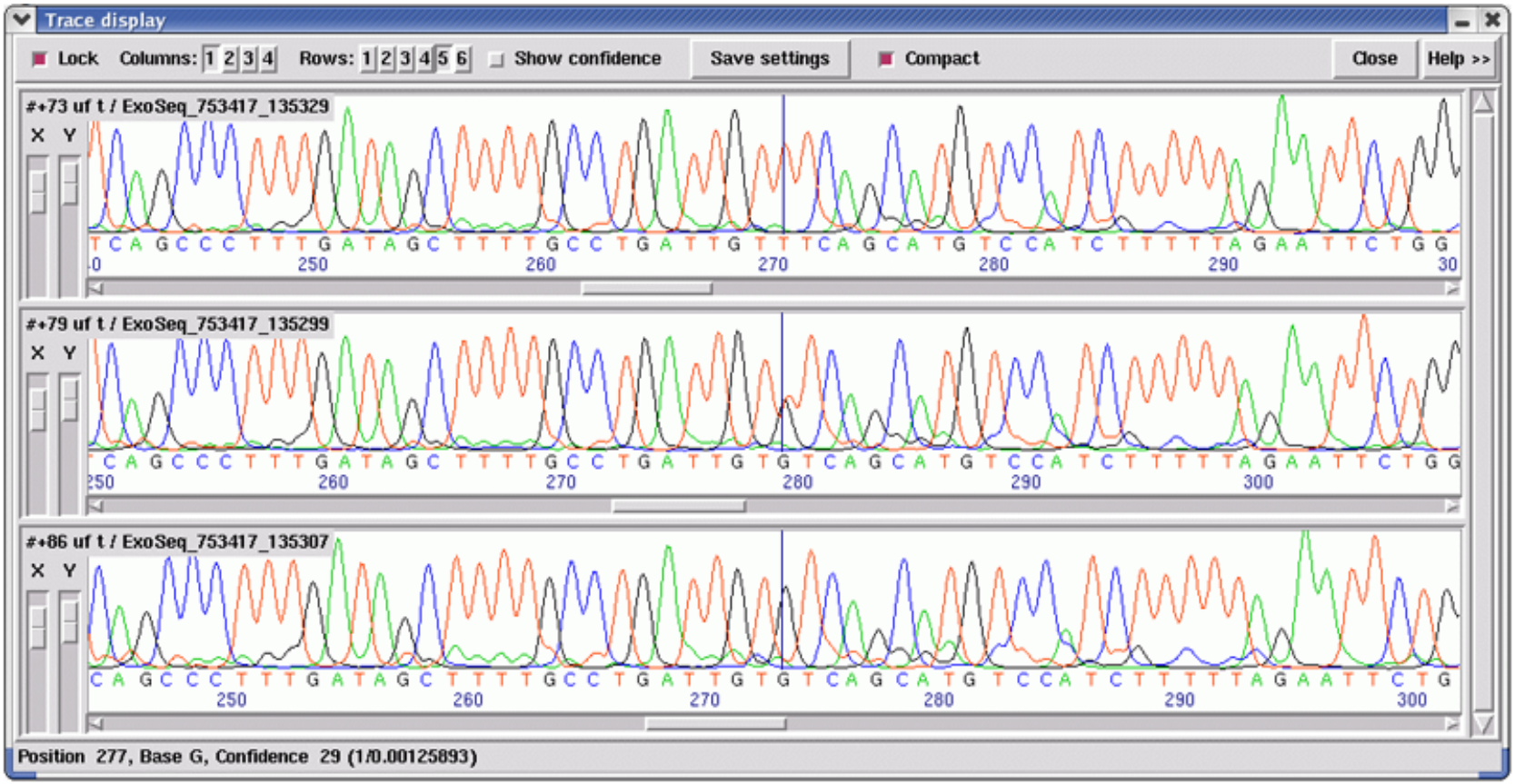
- Single Nucleotide Polymorphism
- Variationen von einzelnen Basenpaaren



Bildquelle: Wikipedia commons

SNPs

- G/T-SNP von oben nach unten T/T, G/T und G/G.



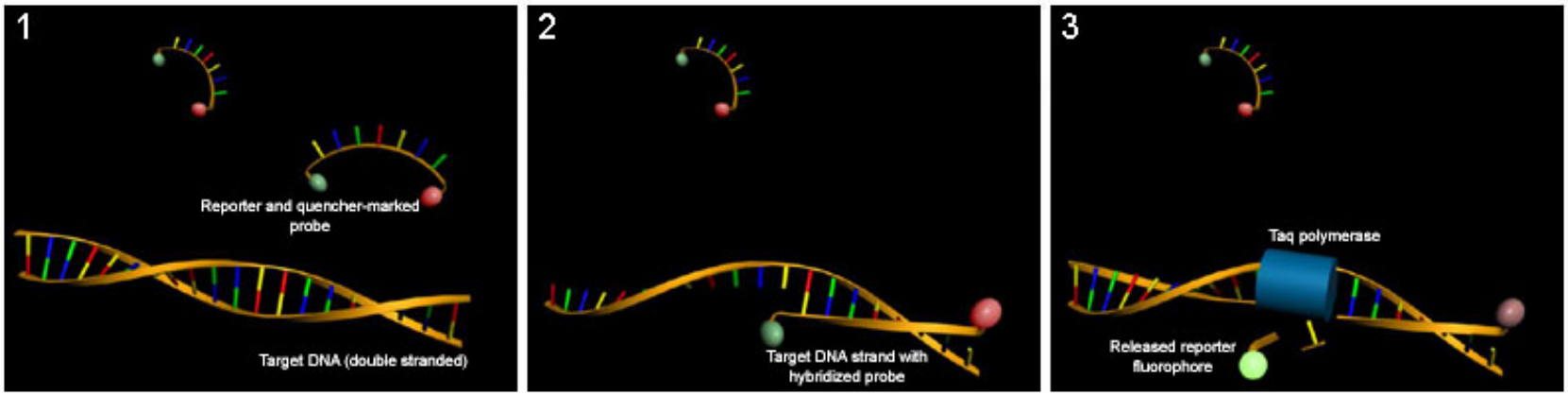
SNPs

- Bestimmung durch automatisierte Verfahren in großem Maßstab möglich

Bildquelle: www.ikmb.uni-kiel.de

SNPs

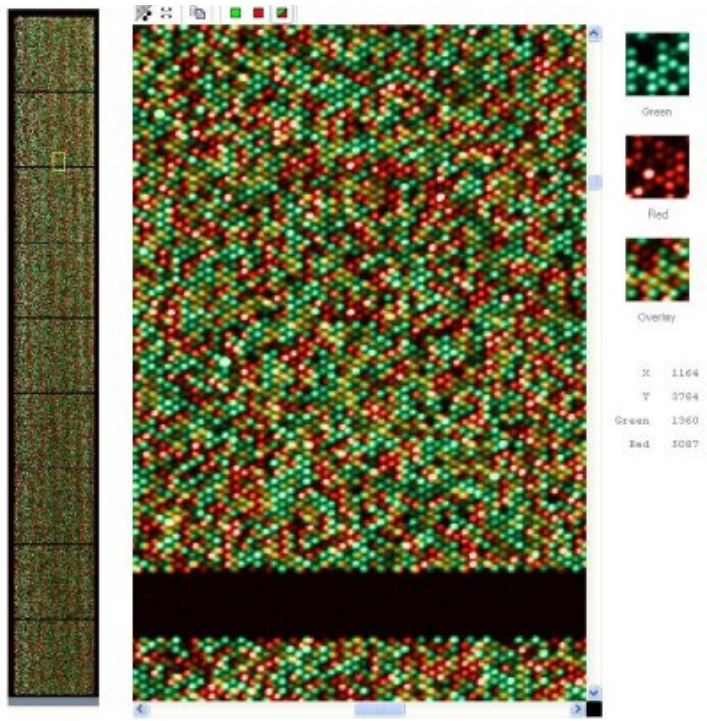
- Nachweis der beiden Allele z.B. durch Taqman-Sonden



Bildquelle: Wikipedia Commons

SNPs

- Neuerdings Nachweis auch mittels Mikroarray
- Hier genomweiter Satz von 550000 SNP Markern (Illumina)



www.sanger.ac.uk

SNPs

Bemerke...

- RFLPs $\subset$ SNPs
- Die meisten SNPs erzeugen/zerstören jedoch keine Restriktionsschnittstelle und sind daher keine RFLPs
- Der große Vorteil der SNPs besteht in der Möglichkeit, mit den o.g. Hochdurchsatzverfahren eine hohe Dichte an Markern effizient und günstig bestimmen zu können.

Outline

1 Genomprojekte und Modellorganismen

2 Das Human-Genome Project

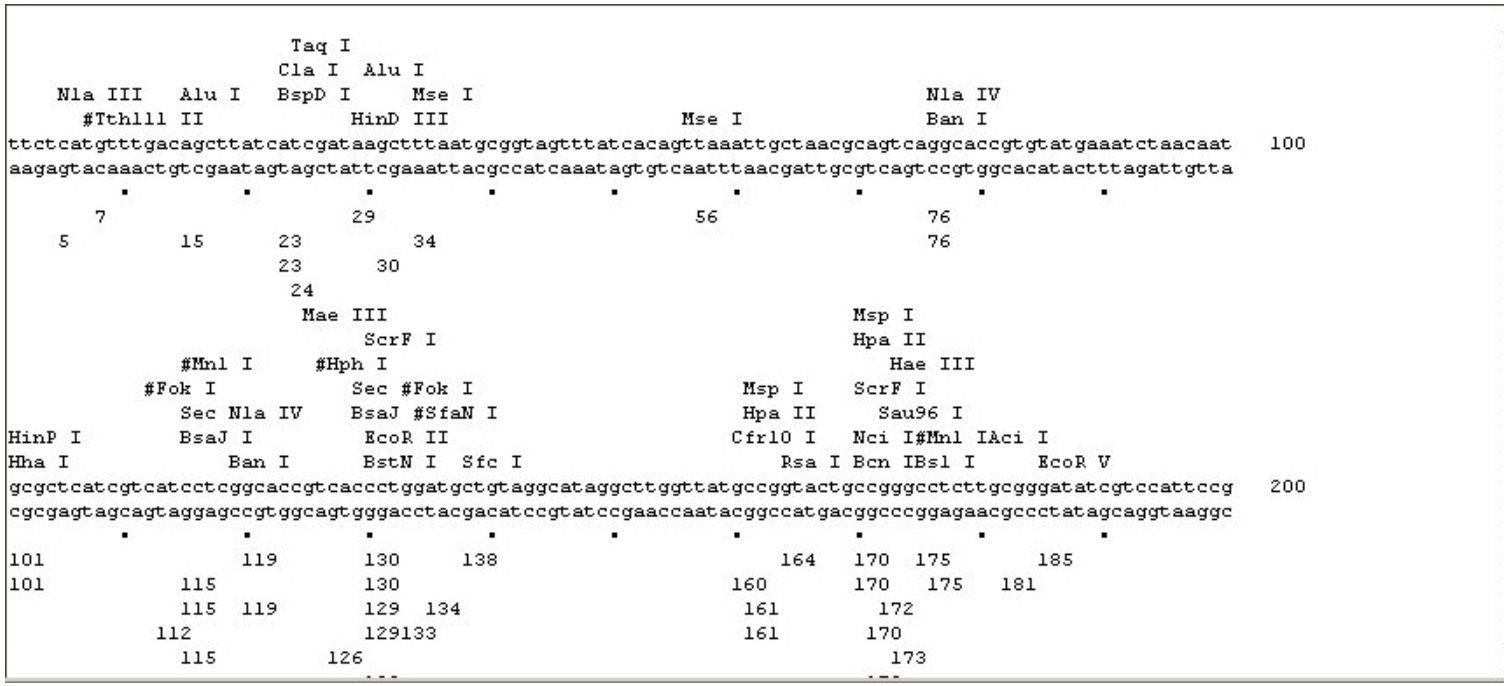
3 Kartierung

Kartierung

Im folgenden soll nun gezeigt werden, wie man mittels der o.g. Marker das menschliche Genom kartiert hat

DNA-Restriktionskarte

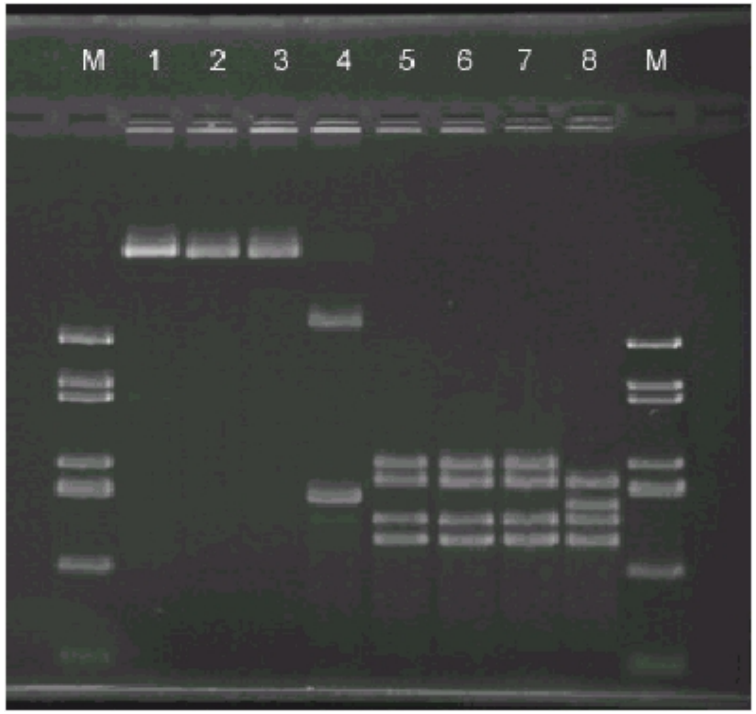
- Eine Karte der Restriktionsstellen in einer DNA-Sequenz
- Falls die DNA-Sequenz bekannt ist, so ist die Karte trivial zu erstellen



z.B. Webcutter http://rna.lundberg.gu.se/cutter2/

Restriktionskartierung einer unbekannten DNA-Sequenz

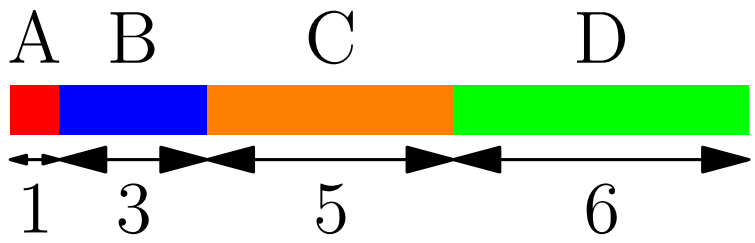
- Ein wesentlich schwierigeres Problem
- DNA-Klon wird mit Enzym verdaut $\rightarrow$ N Fragmente
- Fragmente werden durch Elektrophorese nach Fragmentlänge getrennt



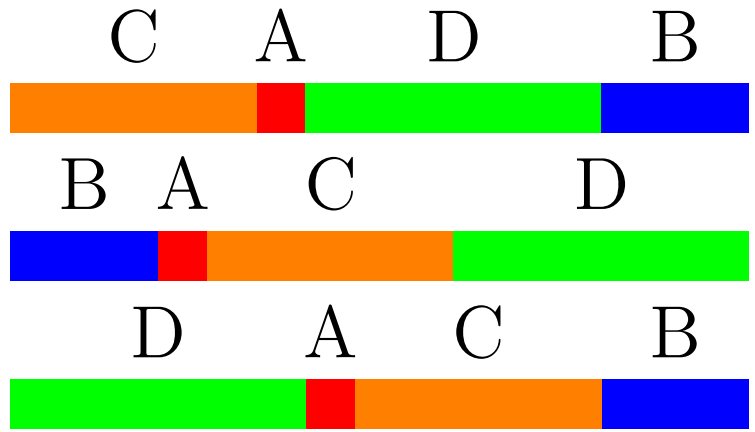
Bildquelle: Edelstein et al 2000

Restriktionskartierung

- Was lernen wir von einem einfachen Restriktionsverdau?



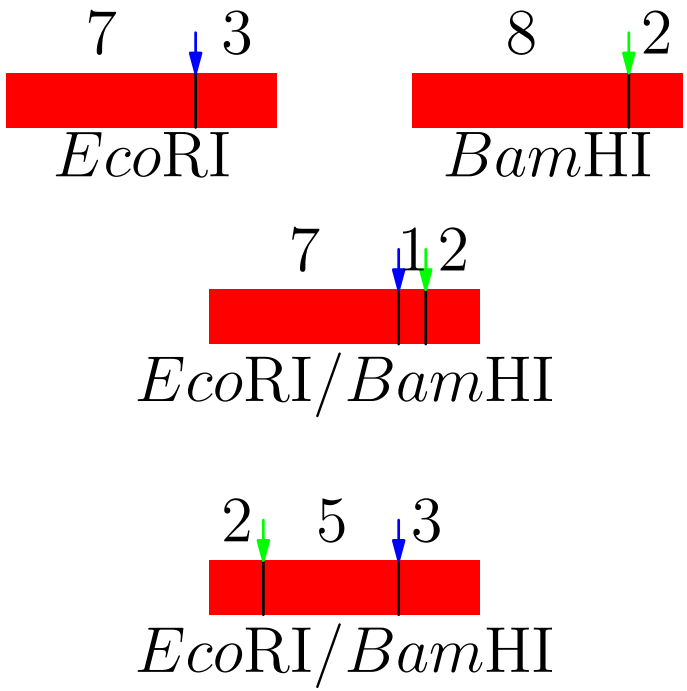
N! Permutationen:



...

Restriktionskartierung

- Doppelverdau (Beispiel: Fragment der Länge 10 kb)
  1. Verdau mit *Eco*RI
  2. Verdau mit *Bam*HI
  3. Verdau mit beiden Enzymen
- Reihenfolge der Schnittstellen anhand des Musters bestimmen



Doppelverdau-Problem

- Gegeben sei ein DNA-Segment, das jeweils mit Enzym A, Enzym B bzw. beiden Enzymen verdaut wird:
  - dA: Fragmentlängen nach Verdau durch Enzym A
  - dB: Fragmentlängen nach Verdau durch Enzym B
  - dX: Fragmentlängen nach Doppelverdau A/B
- Ziel:
  - Positionen der Schnittstellen für Enzym A
  - Positionen der Schnittstellen für Enzym B

Doppelverdau-Problem

Permutationen einer Menge $[A, B, C, D]$

$[A, B, C, D]$	$[A, B, D, C]$	$[A, C, B, D]$
$[A, C, D, B]$	$[A, D, B, C]$	$[A, D, C, B]$
$[B, A, C, D]$	$[B, A, D, C]$	$[B, C, A, D]$
$[B, C, D, A]$	$[B, D, A, C]$	$[B, D, C, A]$
$[C, A, B, D]$	$[C, A, D, B]$	$[C, B, A, D]$
$[C, B, D, A]$	$[C, D, A, B]$	$[C, D, B, A]$
$[D, A, B, C]$	$[D, A, C, B]$	$[D, B, A, C]$
$[D, B, C, A]$	$[D, C, A, B]$	$[D, C, B, A]$

- 1. Wahl: n, 2. Wahl $n-1$, 3. Wahl $n-2, \ldots :\longrightarrow n!$

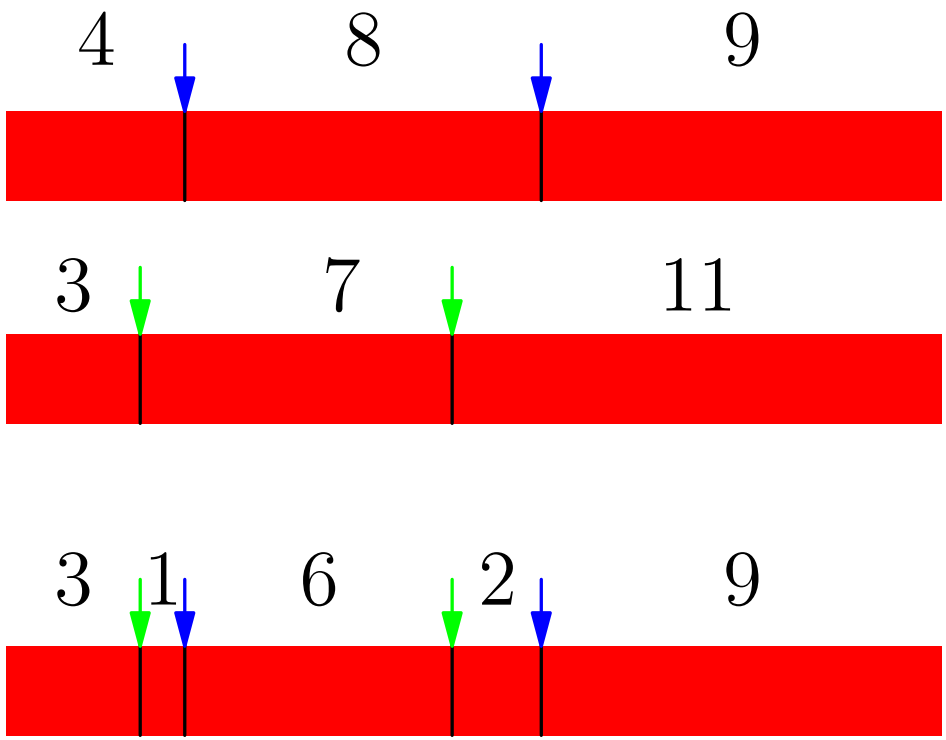
Doppelverdau-Problem

Eine Lösung durch ”Brachialgewalt”

- $\mathscr{A}$: Menge aller Permutationen nach Verdau durch A
- $\mathscr{B}$: Menge aller Permutationen nach Verdau durch B
- $\mathscr{A}\mathscr{B}$: Menge aller Permutationen nach Doppelverdau

Doppelverdau-Problem: matlab/octave

- Wir wollen die folgende Karte berechnen
- Unsere Beobachtung:
  - Verdau mit A: 9,8,4
  - Verdau mit B: 11,7,3
  - Doppelverdau: 9,6,3,2,1



Doppelverdau-Problem: matlab/octave

```
A=[9 8 4 ];
B=[11 7 3];
AB=[9 6 3 2 1];

[a,b,ab]=doubledigest(A,B,AB);
```

- Vektoren A,B und AB definieren
- Funktionsaufruf liefert a,b und ab (korrekte Reihenfolge der Fragmente) zurück

Doppelverdau-Problem: matlab/octave

```
 function [a,b,ab] = doubledigest(A,B,AB)
%function doubledigest(A,B,AB)
```

- matlab/octave-Funktionen werden in m-Datei gespeichert
- Name der Funktion stimmt mit dem Dateinamen überein
- Erste Zeile der Datei enthält Funktionssignatur
- %: Kommentar

Doppelverdau-Problem: matlab/octave

```
pA=perms(A);
pB=perms(B);
pAB=perms(AB);
```

- `perms`: Eingebaute Funktion
- `perms(x)`: liefert alle Permutationen des Vektors x zurück.

```
octave:2>A= [ 1 2 3];
octave:2> perms(A)
ans =

   1   2   3
   2   1   3
   1   3   2
   2   3   1
   3   1   2
   3   2   1
```

Doppelverdau-Problem: matlab/octave

```
for i=1:length(pA)
    for j=1:length(pAB)
        if compatible(pA(i,:),pAB(j,:))
            for k=1:length(pB)
                if compatible(pB(k,:),pAB(j,:))
                    a=pA(i,:);
                    b=pB(k,:);
                    ab=pAB(j,:);
                    return;
                end
            end
        end
    end
end
```

Doppelverdau-Problem: matlab/octave

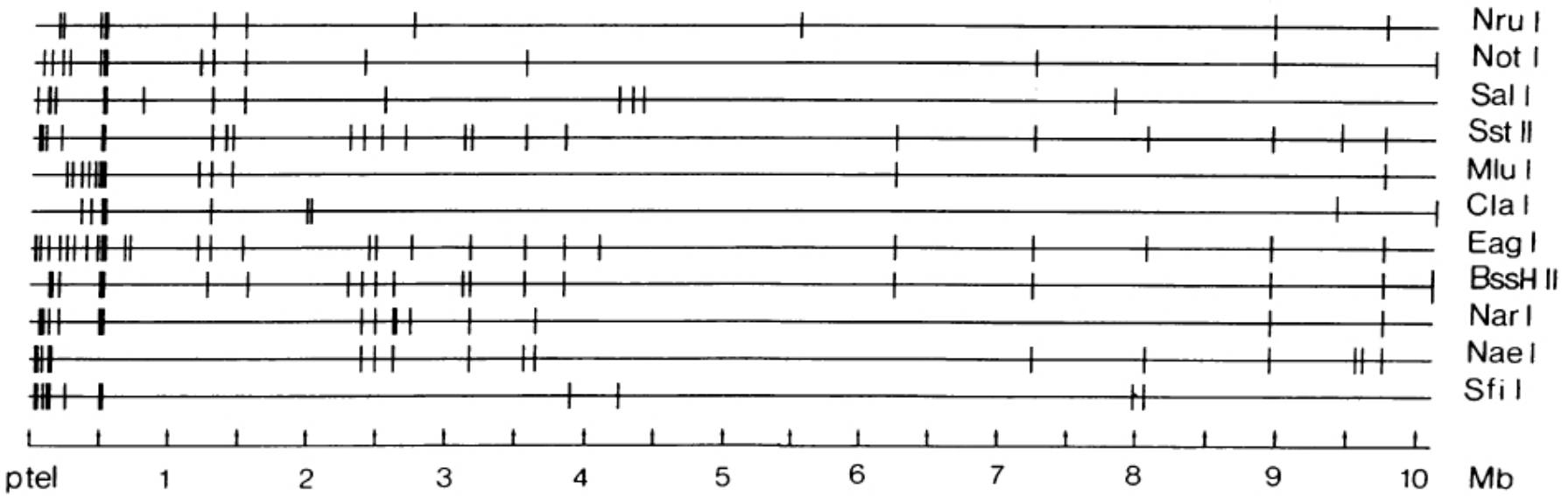
```
function c = compatible(x,ab)
cAB=cumsum(ab);
cX=cumsum(x);
mem=ismember(cX,cAB);
c = sum(mem)==length(mem);
return;
```

- AB, korrekte Reihenfolge: 3 – 1 – 6 – 2 – 9
- Kumulative Summe : 3 – 4 – 10 – 12 – 21
- A korrekte Reihenfolge: 4 – 8 – 9
- A, kumulative Summe: 4 – 12 – 21
- `ismem(A,AB)` liefert Vektor zurück dessen Einträge angeben, ob A(i) Mitglied von AB ist
- Falls Reihenfolge von A mit der von AB übereinstimmt, enthält mem nur '1'
- `c` wird dann auf 1 (wahr) gesetzt

Doppelverdau-Problem

- Nicht geeignet für realistische Probleme
- Das Doppelverdau-Problem ist NP-schwierig

Restriktionskartierung: Beispiel



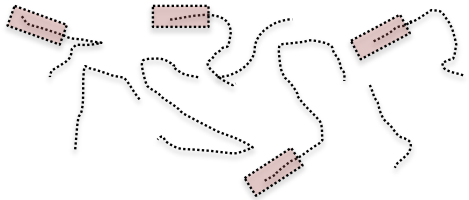
Petit et al. (1990) Long-range restriction map of the terminal part of the short arm of the human X chromosome.

Proc. Nati. Acad. Sci. USA Vol. 87, pp. 3680–3684.

Genomsequenzierung

- Die ersten Schritte der Genomsequenzierung und -assemblierung

Genome: 3.2 Gb
Many copies of genome



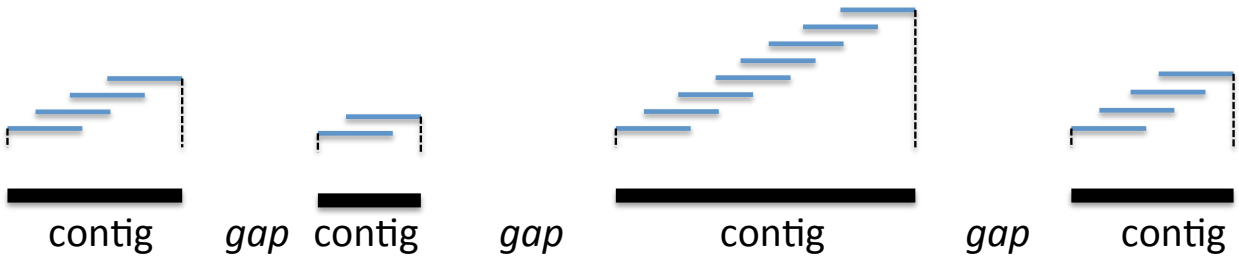
Reads: 500bp
Only one end sequenced
Not all fragments sequenced

```
tgctgctcctacaacatcggccgtgcctg
              atcggccgtgcctggaataagccct
```

Find overlapping reads

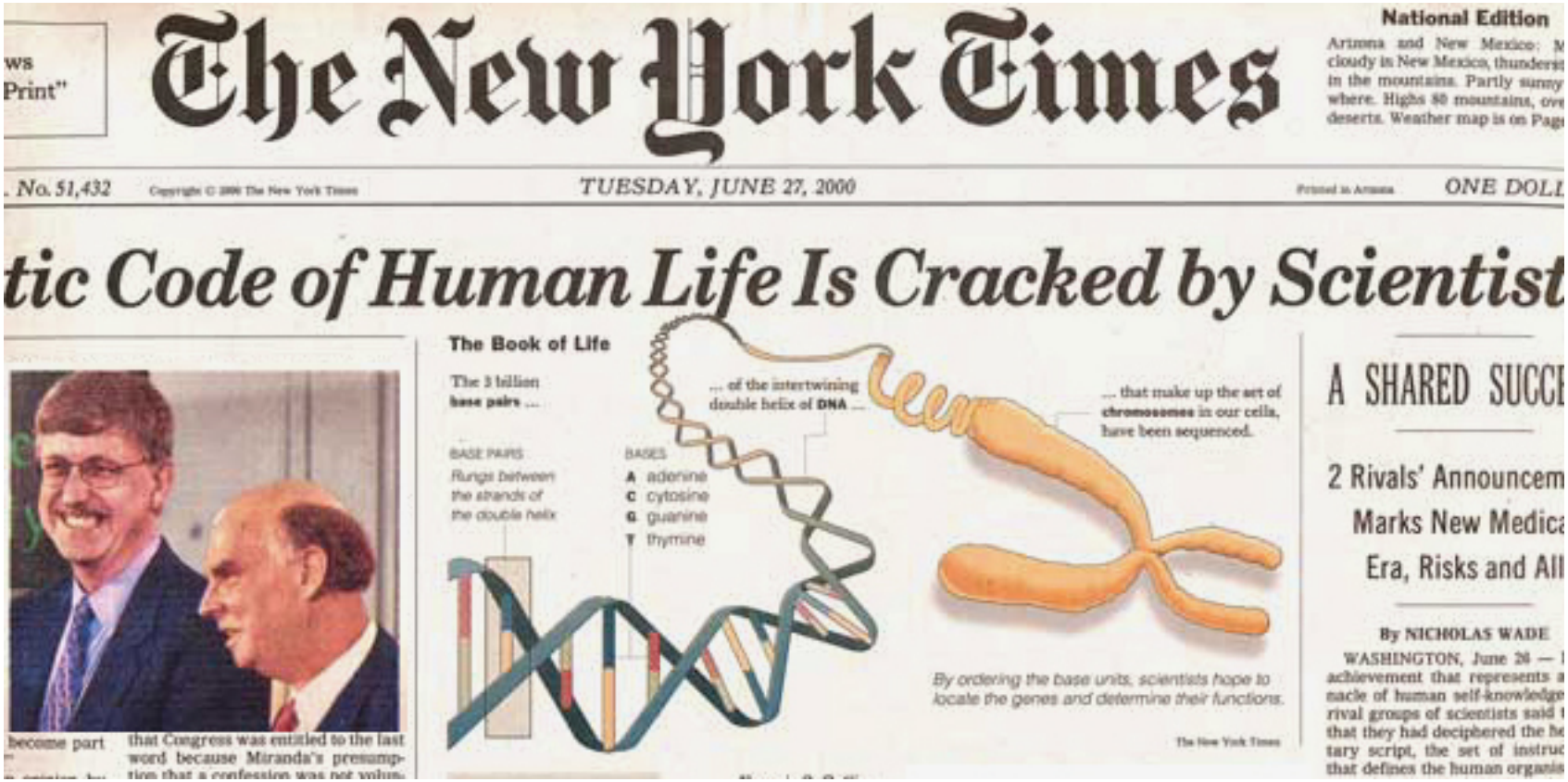
...tgctgctcctacaacatcggccgtgcctggaataagccct...

Merge overlapping reads into contigs



Result of assembly is set of contigs with gaps

Genomsequenzierung



The New York Times

National Edition

No. 51,432 TUESDAY, JUNE 27, 2000 ONE DOLL

tic Code of Human Life Is Cracked by Scientist

The Book of Life

The 3 billion base pairs ... of the intertwining double helix of DNA ... that make up the set of chromosomes in our cells, have been sequenced.

BASE PAIRS Rungs between the strands of the double helix

BASES A adenine C cytosine G guanine T thymine

By ordering the base units, scientists hope to locate the genes and determine their functions.

The New York Times

A SHARED SUCCE

2 Rivals' Announcem Marks New Medica Era, Risks and All

By NICHOLAS WADE

WASHINGTON, June 26 — achievement that represents a nacle of human self-knowledge rival groups of scientists said that they had deciphered the he tary script, the set of instruc that defines the human organis

become part that Congress was entitled to the last word because Miranda's presump-

- Ein Wettbewerb ums humane Genom: whole-genome shotgun vs. BAC by BAC

Shotgun-Sequenzierung

Volume 6 Number 7 1979 Nucleic Acids Research

A strategy of DNA sequencing employing computer programs

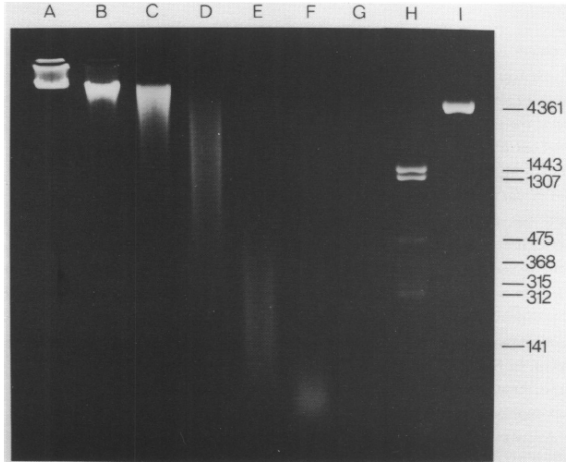
R.Staden

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

Received 23 March 1979

Schrottflinte

Shotgun-Sequenzierung



1. DNAse I Verdau der genomischen DNA (schneidet an zufälligen Stellen)
2. Klonierung kleinerer Fragmente in einen Vektor, um eine Bibliothek mit (nahezu) allen Sequenzen des Genoms zu erzeugen
3. Neben der Klonierungsstelle im Vektor befindet sich eine Universalseqeunz, so dass ein Universal PCR-Primer verwendet werden kann, um alle Sequenzen zu amplifizieren/sequenzieren
4. Im Anschluss wird per Hand (frühe 1970er Jahre) oder mittels Computers nach überlappenden Sequenzen gesucht.

Shotgun-Sequenzierung

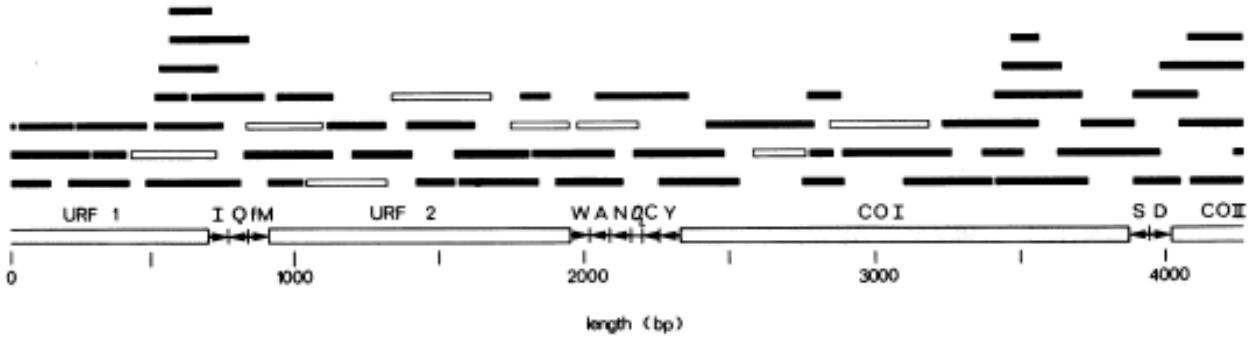


Figure 2. Sequence distribution of cloned DNase I fragments. The sequences of the cloned inserts that were overlapped to produce the complete sequence of the 4257 bp EcoRI fragment are depicted as bars and are positioned to show the contribution of each clone to the final sequence. Solid bars represent sequences from the initial random selection of 48 clones; open bars represent the eight confirming sequences that were done after T-track screening of 36 additional clones (see text). Shown schematically are several bovine mitochondrial genes identified from the DNA sequence of the EcoRI fragment [31]. These include the coding regions for cytochrome c oxidase subunits I and II (COI and COII) and two other large unidentified reading frames (URF1 and URF2) that presumably also code for proteins. Genes for mitochondrial tRNAs are labelled according to the one-letter amino acid code and are depicted as → or ← depending on whether the tRNAs have the sense of the L- or the H-strand, respectively. O_L is the presumptive origin of L-strand synthesis during mtDNA replication.

1. Überlappende Sequenzen werden zu “Contigs” zusammengefügt

Shotgun-Sequenzierung

The continuing rapid fall in the cost of computer components is making it possible for most DNA sequencing laboratories to have their own small computer. The fact that DNA sequencing is now a fast procedure, and the availability of computers gives the possibility of more efficient overall strategies for sequence determination. Outlined below is one such strategy which takes into account cloning technology, the speed of DNA sequencing and the ability of computers to handle and compare data. This is followed by a

Staden R. (1979) A strategy of DNA sequencing employing computer programs. *Nucleic Acids Res* **6**:2601-10.

1. Shotgun-Strategien waren wichtig bei der Erschließung des humanen und anderer Genome
2. In angewandter Form bis heute wichtig (für Next-Generation Sequencing)

BAC by BAC

- BAC: bacterial artificial chromosome
- BACs: inserts von 100,000–300,000 nt
- BACs sind um einiges kleiner als das humane Genom und umso einfacher zu assemblieren
- Hierarchische Assemblierung: Erst die einzelnen BACs, dann Assemblieruing der überlappenden BACs

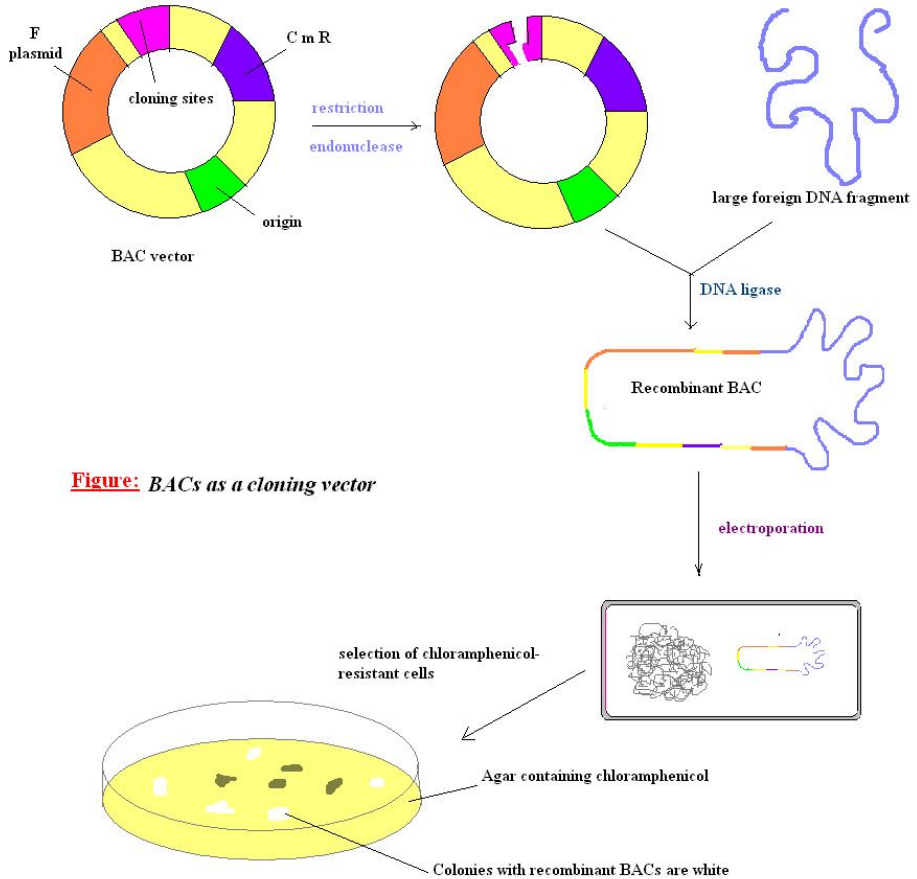
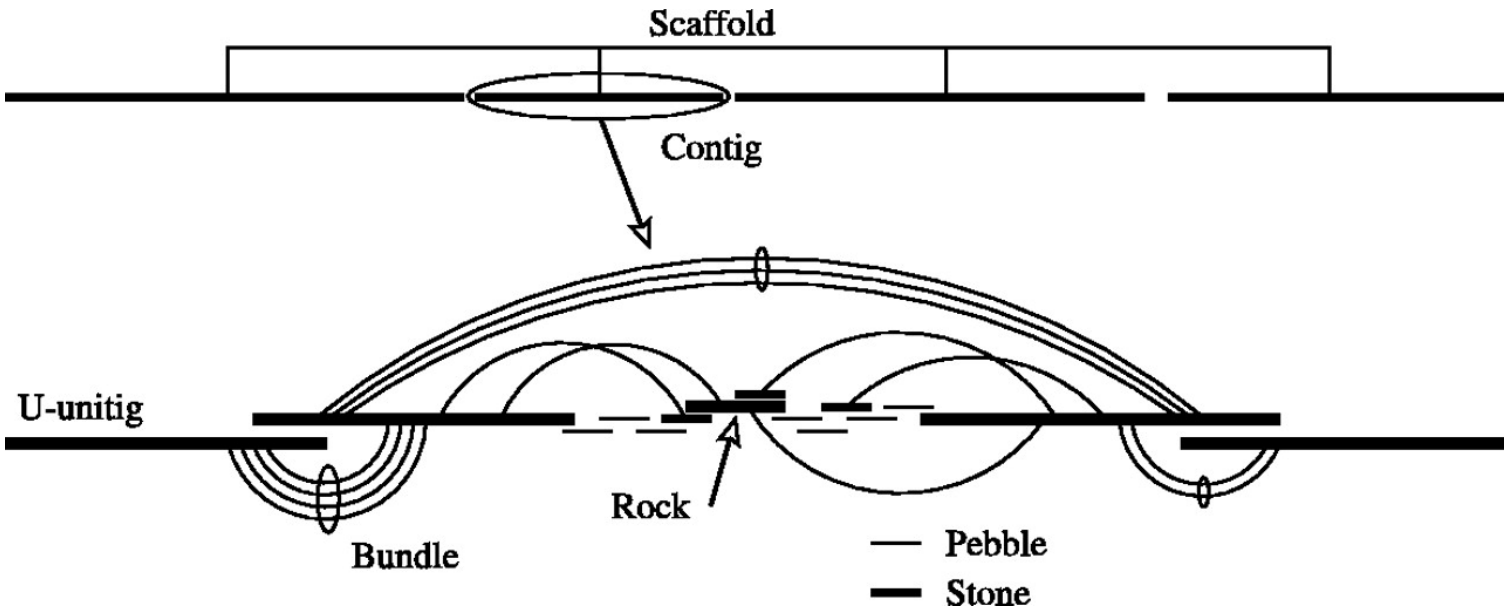


Bild: wikipedia

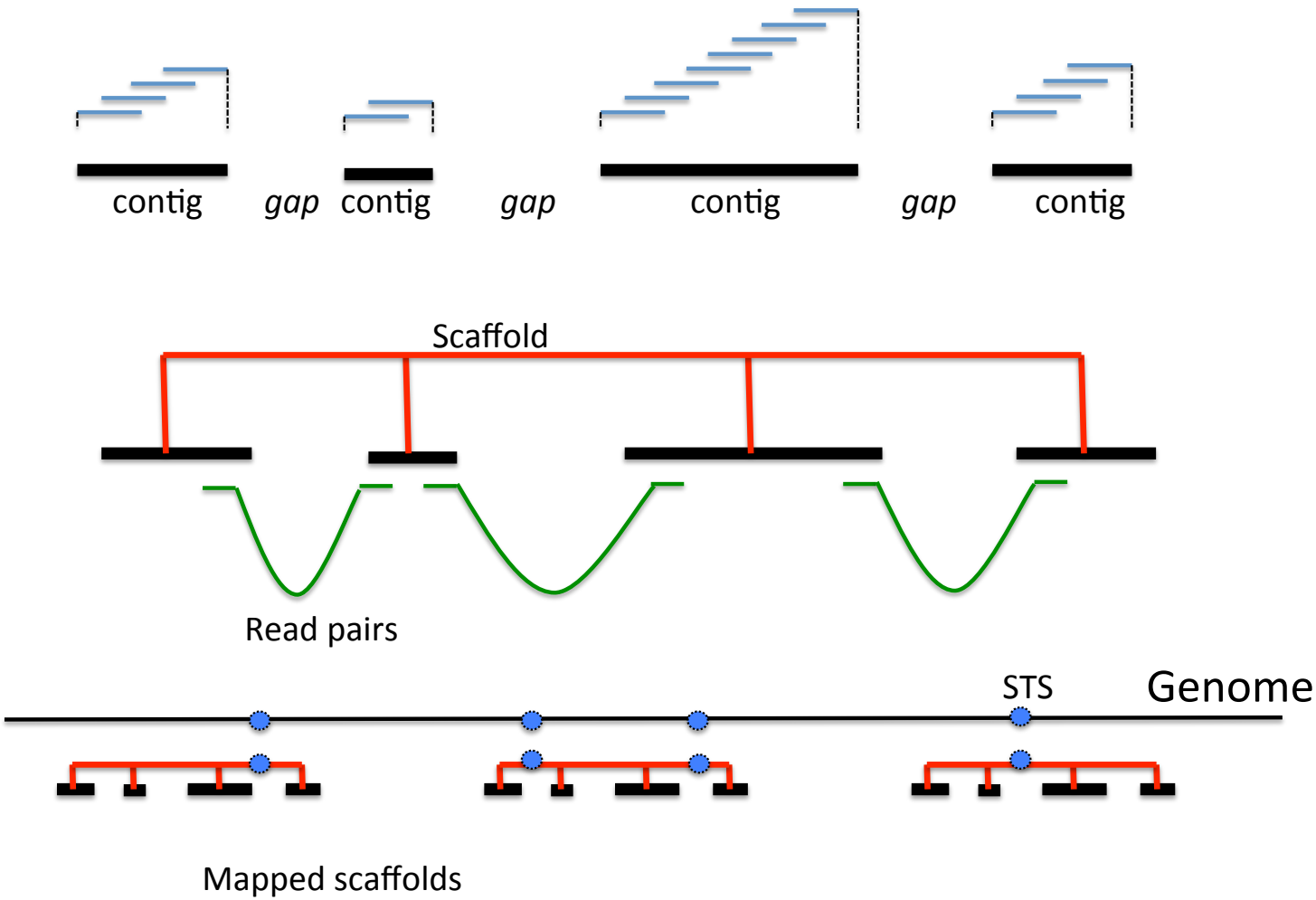
Whole genome shotgun



Myers EW (2000) A Whole-Genome Assembly of Drosophila *Science* **287**:2196–2204

- jeder gegen jeden paarweises Alignment
- Merge to contigs if overlap big enough
- contigs: klein = rock, kleiner = stone, noch kleiner = pebble.

Whole genome shotgun



- Zusammenpuzzeln der contigs

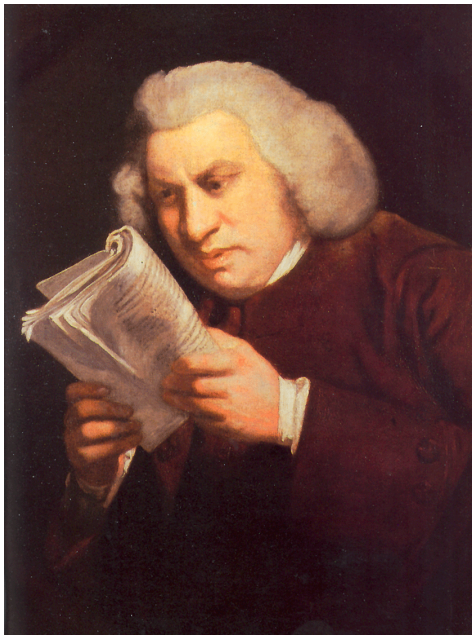
Hausaufgabe #1b

Das Doppelverdauproblem

- In dieser Aufgabe wollen wir die Skriptsprache R verwenden, um das Doppelverdauproblem wie beim matlab-Skript von Vorlesung 2 zu lösen.
- Reichen Sie Ihr Skript und Ihre Lösung ein.
- Fragen zu R?
- install.packages (gtools Bibliothek)

The End of the Lecture as We Know It

- Kontakt:
  peter.robinson@charite.de
- Strachan und Read Kapitel
  Kapitel 8.1, 8.2, 8.3, 13.2



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).