

Molekularbiologische Datenbanken

Kartierung von Chromsomen I
Biotechnische Grundlagen



Ulf Leser

Wissensmanagement in der
Bioinformatik



Ankündigung

- Am 2.6.2004 (nächste Woche) beginnt die Vorlesung um 9.00 s.t. Wohin: Max-Planck-Institut für Molekulare Genetik
- Am 16.6.2004 Exkursion
 - Um 10.00 Uhr
 - Max-Planck-Institut für Molekulare Genetik
 - Ihnestr. 73, Haupteingang
 - U-Bahn Station Thielplatz (U1)

Modelle von Identifikatoren

- Semantikfrei
 - Object-Ids
 - „Surrogate“ Keys
- Semantikbehaftet
 - „Sprechende Schlüssel“
 - Beispiel: „CYC_BOVIN“ = „Protein_Species“ (SP)
 - **Problem: Können sich ändern**
 - Neue Erkenntnisse über Proteinfunktion
 - Neue Erkenntnisse über Vorkommen in Species
 - Heirat
 - ...

Beispiel: GenBank

- Versuch 1: „Locus name“
 - Sprechend: „HUMHBB“ – „Human Betaglobin region“
 - Definitionen verändern sich - als Ids abgelöst
- Versuch 2: Accession-Number
 - Eindeutige, globale ID für jede Submission
 - Keine Versionierung
- Versuch 3: GID (Genbank ID)
 - Eindeutige, interne ID für jede Version einer Submission
 - Aber: Zugriff auf „aktuellsten“ Entry nicht möglich (Nur über Comment line)
- Versuch 4: NID (Nucleotide ID)
 - Eingeführt als übergreifende GID für EMBL/Genbank/DDBJ
 - Versionen von Entries – unterschiedliche NID
- Versuch 5: Accession-No.Version (Seit 1999)
 - Accession-No ist eindeutige, globale ID für jede eingesandte Sequenz
 - Bei Updates eines Eintrags – neue Version, keine neue Accession-no

2. Versionierung

- Das molekularbiologische Wissen wächst und verändert sich ständig
 - Ensembl: ca. 40% Änderungen pro Release
 - Swiss-Prot: Ca. 30% der Einträge ändern sich pro Release
- Analysen benutzen bestimmte Version (aktuellste)
- **Nachvollziehbarkeit von Analysen nicht gewährleistet**
 - wenn Daten eines bestimmten Release nicht widerherstellbar sind
- **Versionierung von Daten essentiell**
 - Version der Datenbank (Release)
 - Versionen von Objekten / Attributen

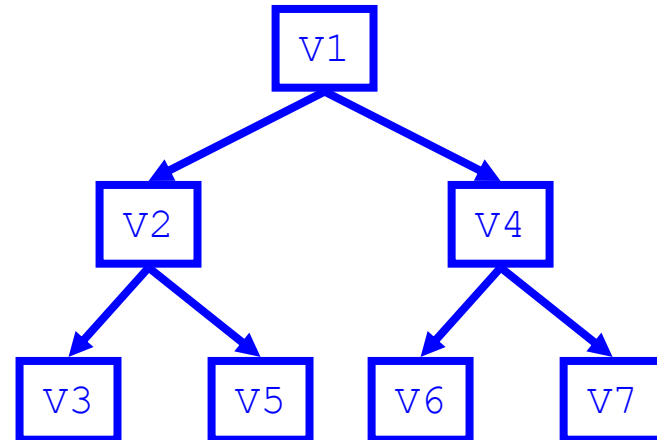
Versionierungsmodelle

Linear



- Zeitliche Reihenfolge fest
- Standardverfahren: Release / Version

Hierarchisch



- Zeitliche Reihenfolge halbgeordnet
- Verfahren von CVS
- Deutlich komplexer

Versionen im relationalen Modell

- Aufgabe
 - RDBMS soll Versionen verwalten
 - Lineare, tuple-basierte Versionierung
- Anforderungen: Zugriff muss möglich sein auf
 - Aktuelle Version
 - Zustand der Datenbank zu beliebigem Zeitpunkt d_0
- 2 Varianten
 - Single Table
 - Schattentabellen

Variante 1: Single-table

- Erweiterung jeder Tabelle T um Attribute
 - Versionsnummer V
 - ALIVE Flag A
 - VALIDFROM D
 - Schlüsselveränderung $K \rightarrow (K+V)$

Variante 2: Schattentabellen

- Pro Tabelle T anlegen einer Tabelle T^S
 - Zusätzliche Attribute
 - Versionsnummer V
 - VALIDUNTIL D
 - Schlüssel in T^S: $K \rightarrow (K+V)$
- T bleibt unverändert
- T^S speichert alte Versionen
- T speichert nur aktuellste Version

Vergleich

- Häufige Änderungen, eher wenig Lesezugriffe - Variante 1
- Seltenerer Änderungen, vor allem Zugriff auf aktuellste Version – Variante 2
- Variante 2 eher für MDB geeignet
- Außerdem zu klären
 - Referenzen / Fremdschlüssel auf versionierte Objekte
 - Entry-based Daten haben keine Referenzen
 - Identifikation von Änderungen (Delta Berechnung)

3. Widersprüchliche Daten

- Sind sehr häufig
 - Experimentelle Datenbanken: unterschiedliche Ergebnisse
 - Integrierte Datenbanken: unterschiedliche Ansichten
- Repräsentation von Widersprüchen
 - Interpretation aller Fakten als „Observations“
 - Jede Observation hat Evidenz und Quelle
- Observation versus Version
 - Versionen sind geordnet
 - Neue Versionen „besser“ als alte

Repräsentation

Datenintegration

Datenquelle
Wer, Version, Warum

Objekt

Attribut

Experimentelle
Daten

Experiment: Wer, wann, wie, wo

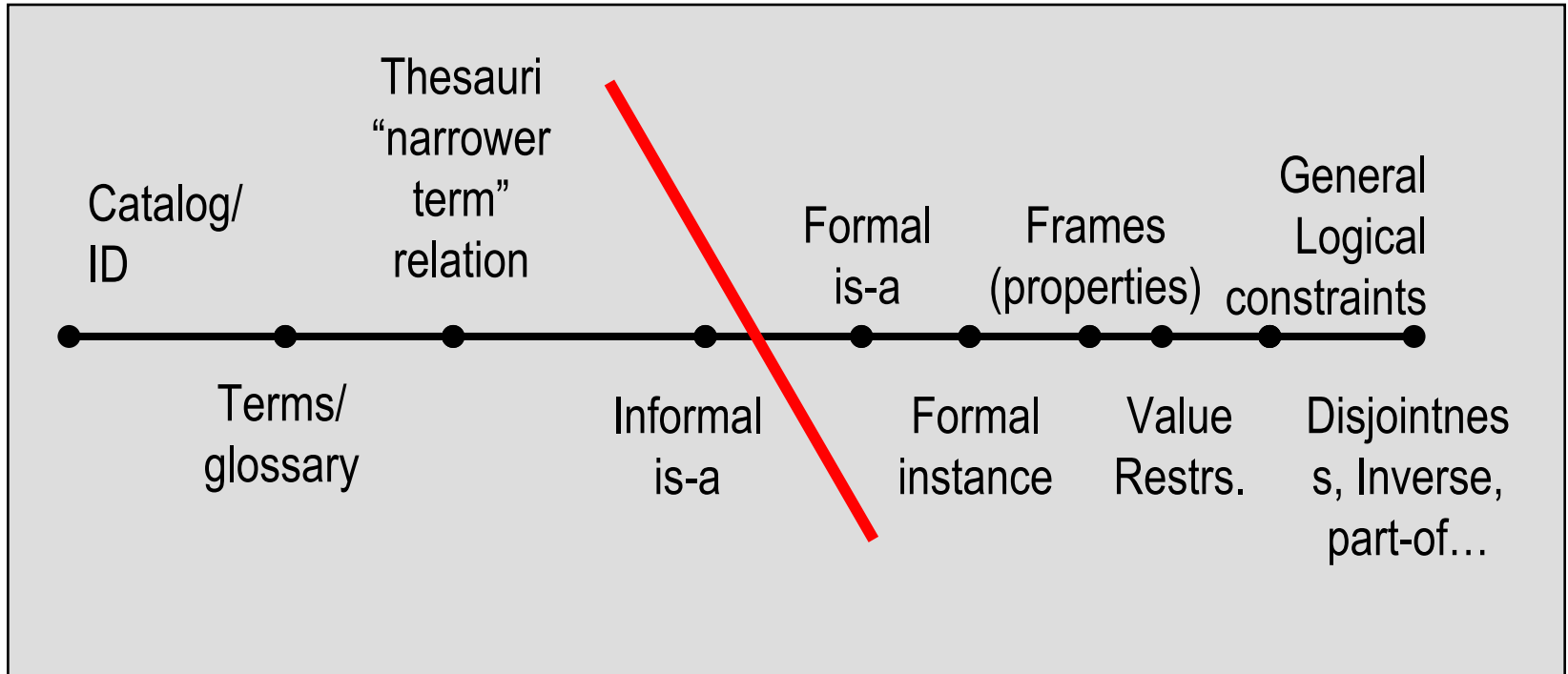
Untersuchtes
Objekt

Untersuchungs-
ergebnis

Controlled Vocabularies

- Zweck
 - Präzise Suchergebnisse
 - Validierung von Eingabemasken
 - Vergleichbarkeit von Daten: Proteinfunktion in verschiedenen Spezies, enzymatische Eigenschaften, ...
- Viel Aufmerksamkeit in den letzten Jahren (endlich)
 - Keywordlisten
 - Taxonomy Projekt
 - GeneOntology
- Vokabulare, Thesauri, Schlüsselwörter, ...
- Ontologien

Ontology Formalisation

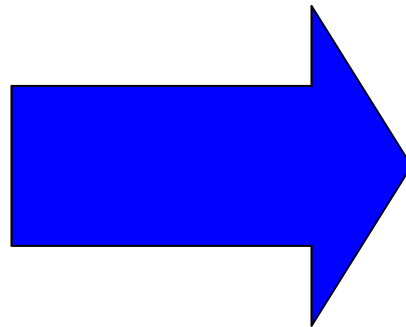


Source: Robert Stevens, From Building and Using Ontologies, U. of Manchester

Inhalt dieser Vorlesung

- Was ist Kartierung (Mapping)?
- Typen von Genomkarten
- Integrierte Karten

Teil I. Ziel des HGP



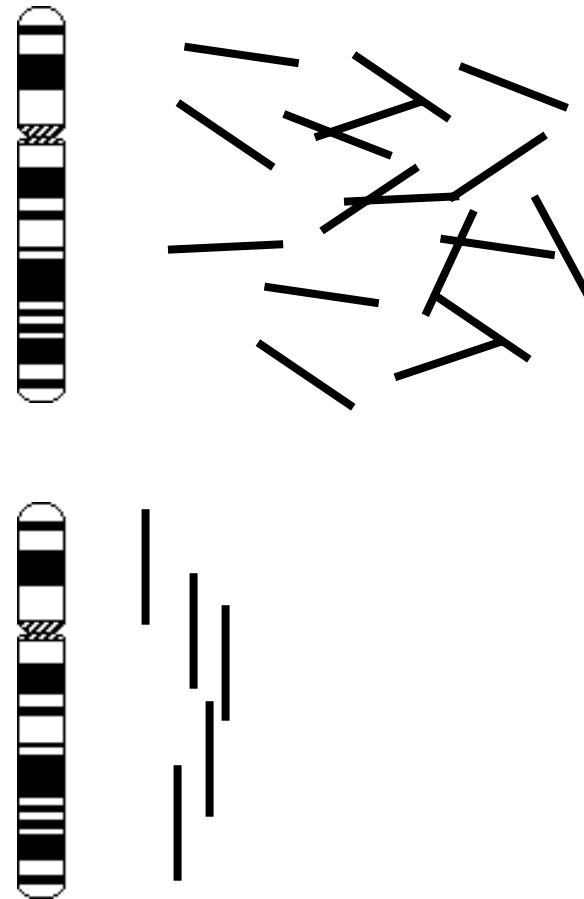
1. A
2. C
3. G
4. G
5. C
6. C
7. T
8. G
9. G
10. A
11. G
12. G
- ...
- ...
- ...
- 123242423 A
- 123242424 C
- 123242425 G
- 123242426 G
- 123242427 T

Sequenzierung

- Problem
 - Chromosome sind nicht direkt sequenzierbar
 - DNA Stücke nur bis ca. 500-1000bp sequenzierbar
- Finde eine Menge von Bruchstücken so, dass
 - Die Bruchstücke sind kurz (sequenzierbar)
 - Das Verfahren ist kostengünstig
 - Ihre Reihenfolge der Stücke ist berechenbar
- Mapping von Chromosomen
 - Bestimmung einer minimalen Menge und ihrer Anordnung von Chromosomenbruchstücken so, dass die Bruchstücke (und damit ihre Sequenz) zusammen das Chromosom möglichst komplett überdecken

Grundidee

- Chromosom isolieren
- Zufällig in Stücke brechen (Clone)
- **Minimal Tiling Path** ermitteln
- Stücke sequenzieren
- Fertig

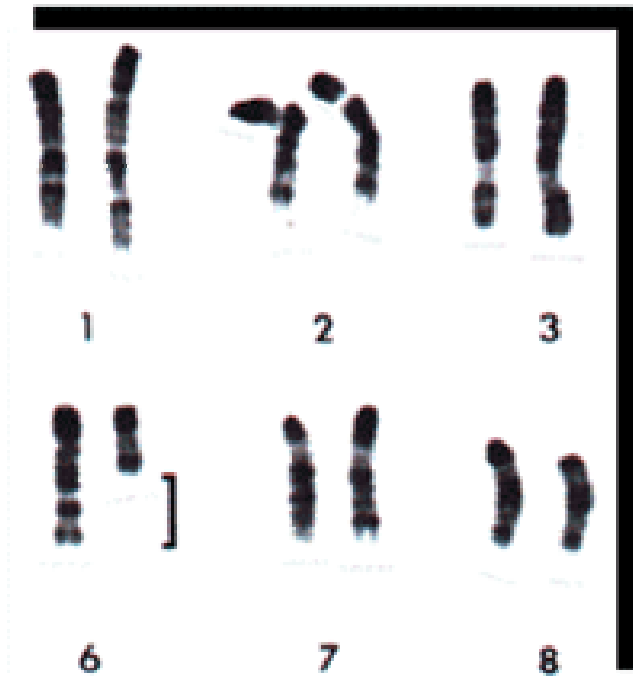


Teil II. Genomkarten und Kartierungsverfahren

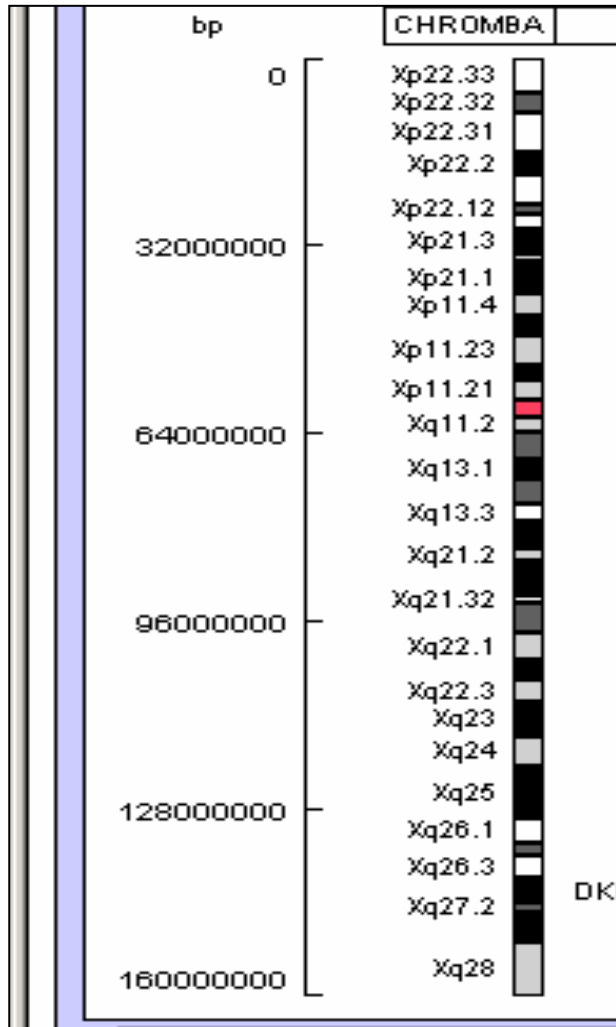
- Cytogenetische Karten
- Genetische Karten
- Physikalische Karten
 - Clone-Karten
 - Marker-Karten

II.1 Cytogenetische Karten

- Beobachtung charakteristischer Färbungen
 - „Chromosome Bands“
 - Bedingt durch Konformation der DNA
- Grobe Orientierung



Elemente eines Chromosoms



Telomer

p – der kurze Arm

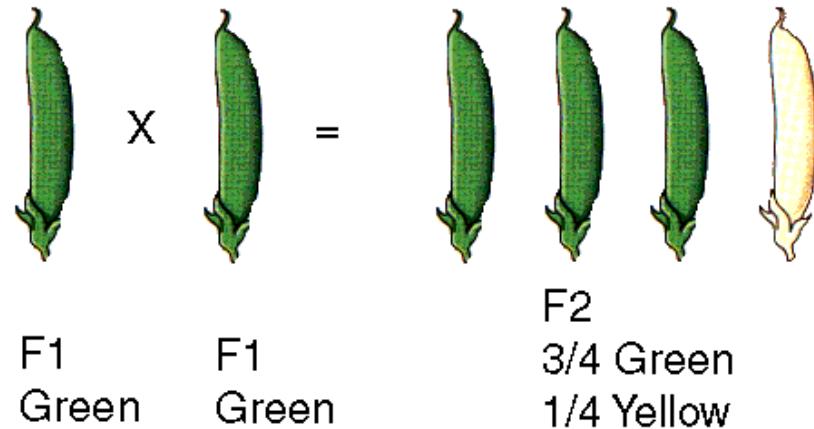
Centromer

q – der lange Arm

Telomer

II.2 Genetische Karten

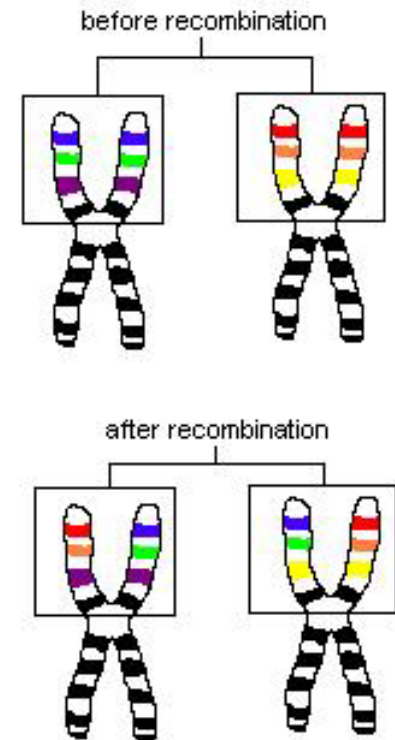
- Mendelsche Vererbungsgesetze



- Folgerung
 - Anlagen für Eigenschaften liegen in Kopien vor
 - Im konkreten Individuum dominiert eine bestimmte Ausprägung
 - Die andere ist aber verborgen vorhanden und wird vererbt
 - Ist in Nachkommen die dominierende Ausprägung nicht vorhanden, kommt die andere Ausprägung zu Tage
 - Aber: Keine Kenntnis der „Orte“ der Eigenschaften / Ausprägungen

Rekombination

- Vererben sich bestimmte Ausprägungen häufig zusammen?
 - Haben Menschen mit blaue Augen oftmals blonde Haare?
 - Haben Pflanzen mit großen Blüten auch lange Stängel?
- **Beobachtung: Ja**
 - Bestimmte Phänotypen vererben sich öfter gemeinsam als andere



Genetische Karten

- Rekombination ist Regel, nicht die Ausnahme
- Kann an allen Stellen eines Chromosoms passieren
- **Wahrscheinlichkeit der gemeinsamen Vererbung zweier Ausprägungen hängt vom Abstand ab**
 - Wahrscheinlichkeit für Rekombinationsereignis ist an jeder Stelle (nur theoretisch!) gleich
 - Je mehr „Platz“ zwischen zwei Ausprägungen, desto höher die Wahrscheinlichkeit eines Rekombinationsereignis
 - Kleiner Abstand: kaum Platz für Rekombination, häufige gemeinsame Vererbung der Ausprägungen
 - Auf verschiedenen Chromosomen: **statistische Unabhängigkeit**
- **Einheit: CentiMorgan**
 - Wahrscheinlichkeit 1% für Trennung von M_1 und M_2 : 1 cM
 - Wahrscheinlichkeit 2% für Trennung von M_1 und M_2 : 2 cM
 - ...

Erstellung genetischer Karten

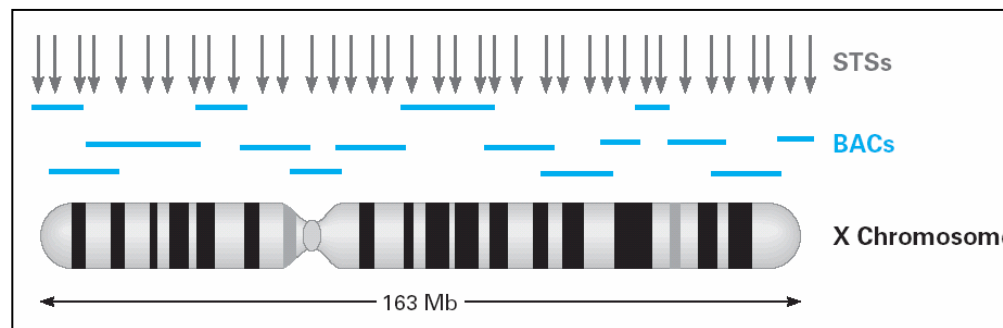
- Tiere: Züchtung
- Menschen: Grosse Stammbäume (CEPH)
- Verfahren
 - Bestimmung der Merkmale bei allen Individuen
 - Beobachtung der Wahrscheinlichkeiten der gemeinsamen Vererbung
 - Berechnung der Ordnung und Abstände der Marker
- Algorithmus
 - Prinzipiell: $O(n!)$ mögliche Ordnungen
 - **Linkage Analysis** (Siehe Literatur)

Fazit

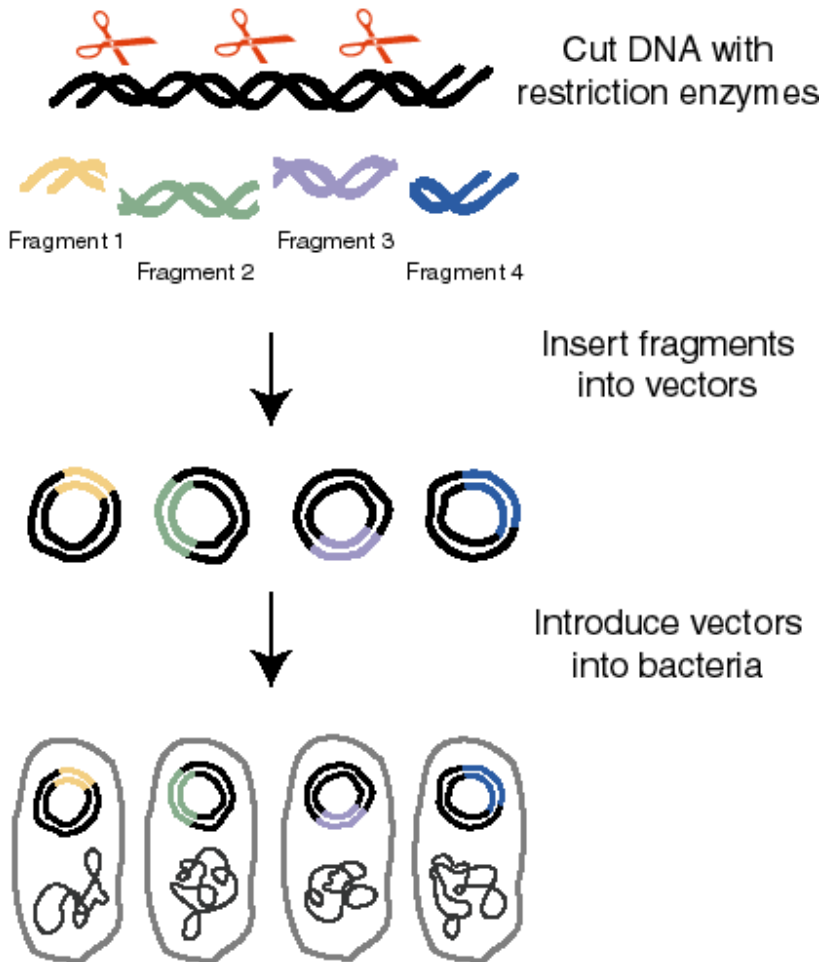
- Historisch ältestes Verfahren
- Bei beobachtbaren Ausprägungen ohne Biotechnologie möglich
- Zuverlässig
- Aber
 - Rekombination ist nicht gleichverteilt
 - Maximale Auflösung ca. 1 cM
 - Bei Menschen fehlen große Stammbäume
 - Entsprechung zur physikalischen Realität unklar
 - Keine Hilfe zur Sequenzierung

II.3 Physikalische Karten

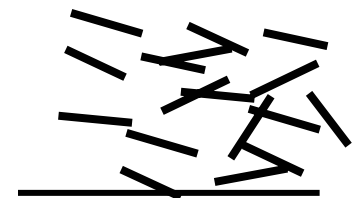
- Bestehen aus zwei Arten von Objekten
 - Clone: YAC, PAC, BAC, Cosmids, ... mit **Ausdehnung**
 - Marker: Loci, STS, EST, ... „ohne **Ausdehnung**“
- Wichtigste biotechnische Verfahren zur Erstellung
 - Hybridisierung und PCR
- Kartierungsverfahren
 - Clone Hybridisierung
 - STS Content Mapping



Gesamtübersicht

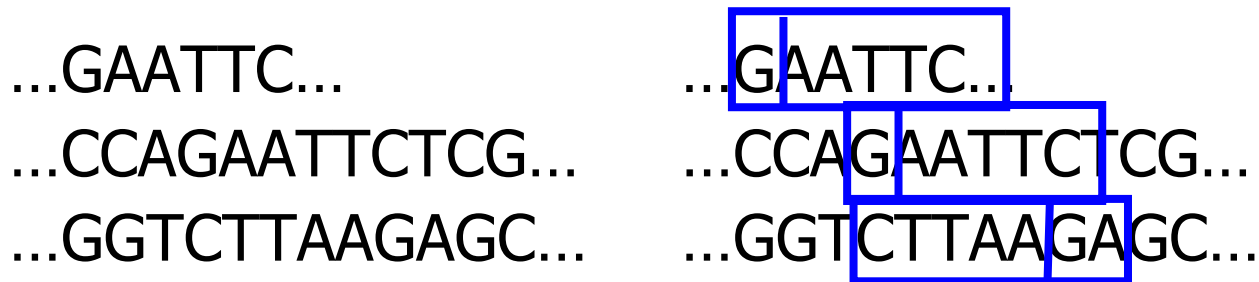


- Schneiden mit Restriktionsenzymen
 - Länge hängt ab von
 - Spezifität des Restriktionsenzym
 - Länge der Behandlung (partieller Verdau)
- Bruchstücke unterschiedlicher Länge
- Auftrennen nach Länge
 - Gelelektrophorese
- Clonierung in Bakterien
 - Vervielfältigung
- Ergebnis:

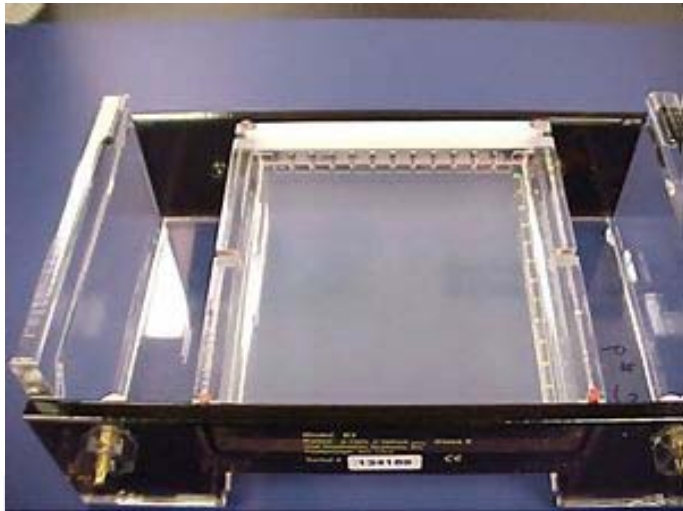
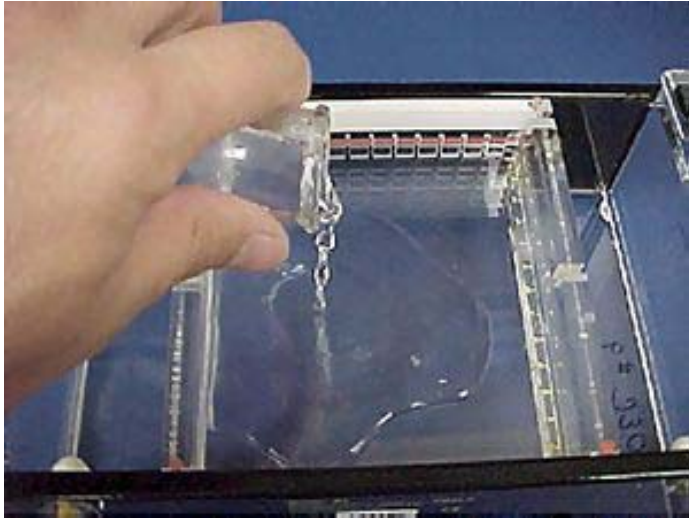


Restriktionsenzyme

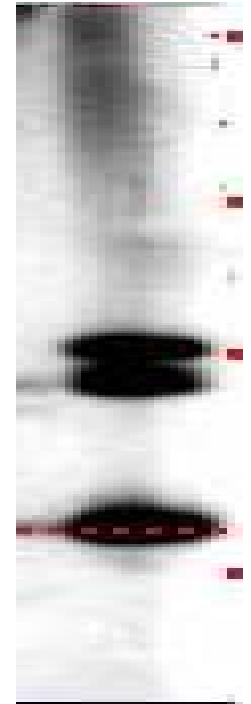
- Proteine, die DNA an **spezifischen Stellen** schneiden
 - Notwendig in Zelle als „Müllabfuhr“
 - Viele bekannt (kommerziell)
 - Partieller Verdau möglich
- Beispiel: Enzym EcoRI schneidet GAATTC nach dem führenden “G”



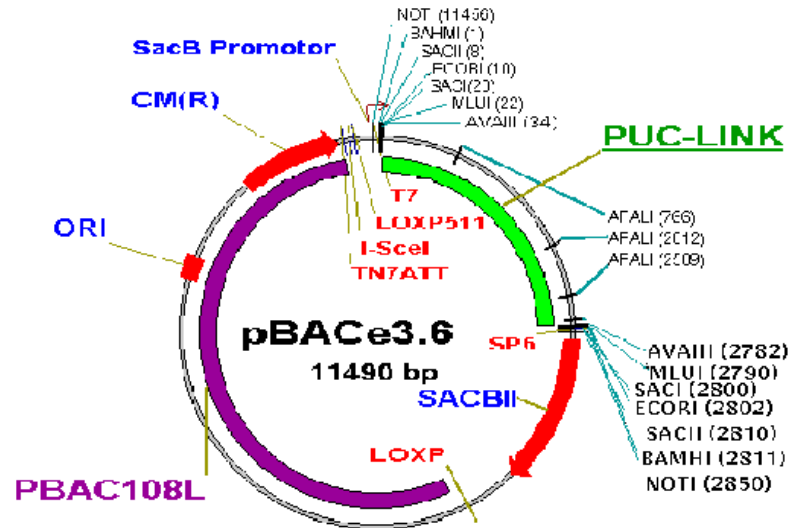
Gelelektrophorese



- Zähflüssiges Agarose-Gel
- DNA Bruchstücke werden negativ geladen
- Anlegen einer elektrischen Spannung
- DNA Bruchstücke bewegen sich im Feld richtig Anode
- **Geschwindigkeit hängt im wesentlichen von Größe des Bruchstücks ab**
 - Beweglichkeit im Gel
 - Ladung



Clonierung



- Isoliertes Bruchstück wird in einem „Vektor“ eingebracht
 - DNA replication initiation site
 - Antibiotika-Resistenzen
 - Schnittstellen zur späteren Extraktion
 - ...
- Dieser wird von Bakterie als Chromosom betrachtet

Clonetypen

- YAC: Yeast Artificial Chromosome
 - Größe zw. 150KB und 1.8MB
 - Schnell zur großen Abdeckung
 - Hohen Anteil an Chimerismen (40-60%)
- BAC: Bacterial Artificial Chromosome
 - Größe zwischen 150 und 300 KB
 - Sehr robust
 - „Arbeitspferd“ des Human Genome Projects
- Cosmids
 - Größe zwischen 35 und 45 KB
 - Kann „direkt“ sequenziert werden

Minimal benötigte Menge von Clonen

		Cosmids (40KB)	BAC (250 KB)	YAC (1000 KB)
Drosophila	180 MB	4500	720	180
Homo Sapiens	2.800 MB	70000	11200	2800

- Theoretischer Wert
 - Clone sind gleichverteilt
 - Hohe Überdeckung notwendig (6-10)

Clonebibliotheken

- Mapping erfordert viele Versuche
- Identität der Clone muss gleich bleiben
- Bibliotheken
 - Mengen von Bakterien mit Clonen
 - Lassen sich einfach vervielfältigen
 - Internationale Verwendung
 - Identifikation: Koordinaten
 - ICRFy900F18, CEPH44G, 128_H_44
- Zur Weiterverarbeitung: Übertragung auf **Filter**
 - Extraktion der "Nutzlast" aus Bakterien
 - Fixierung auf Material (Nylon, Glas, Silikon)



Clonebibliotheken kann man kaufen

The screenshot shows the RZPD Colony Filters web interface. At the top, there is a navigation bar with the RZPD logo, the title "RZPD Colony Filters", a "Help" link, and a "Complete Library List" button. Below the navigation bar is a search filter section with the following settings:

- Product Type = Colony Filters
- Organism = Homo sapiens (Human)

There are "AND" buttons and a "Go On" button to the right of the filter section. Below the filter section is a table of available libraries for purchase.

Library Name	Library No.	Additional Conditions	Filter per Set	Set Price [EUR]	Select to Order	Quantity
Human 4X YAC (ICRF)	900		3	375.00	<input type="checkbox"/>	<input type="checkbox"/>
Human 4Y YAC (ICRF)	901		-	upon request		
Human Adipocyte cDNA	452		1	125.00	<input type="checkbox"/>	<input type="checkbox"/>
Human Adrenal Cortical Carcinoma cDNA	681		1	125.00	<input type="checkbox"/>	<input type="checkbox"/>
Human Adrenal Gland cDNA	592		1	125.00	<input type="checkbox"/>	<input type="checkbox"/>
Human BAC (RPCI-11)	753		16	1820.00	<input type="checkbox"/>	<input type="checkbox"/>
Human Bone Marrow cDNA	413		1	125.00	<input type="checkbox"/>	<input type="checkbox"/>

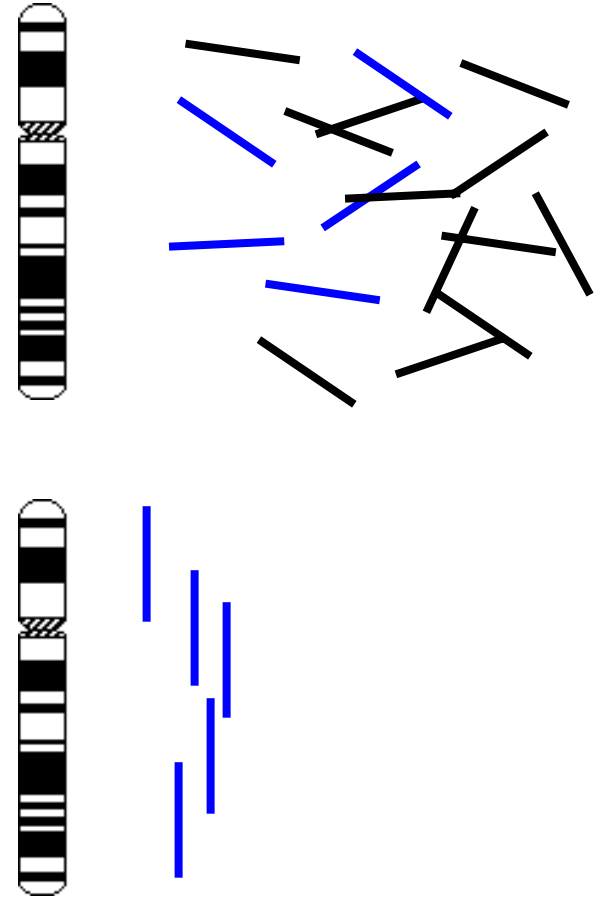
At the bottom of the browser window, the status bar shows "Document: Done (8.863 secs)".

DNA Marker

- Eindeutig im Genom lokalisierbare „Punkte“
- Bei allen Individuen einer Spezies gleich
- STS
 - Sequence-tagged Sites
 - Eindeutige Sequenzen
 - Länge 300-500 Basenpaare
- EST
 - Expressed Sequence Tags
 - STS aus kodierenden Abschnitten
 - ESTs machen heute über 80% aller bekannten Sequenzen aus!
- Nachweis: PCR oder Hybridisierung

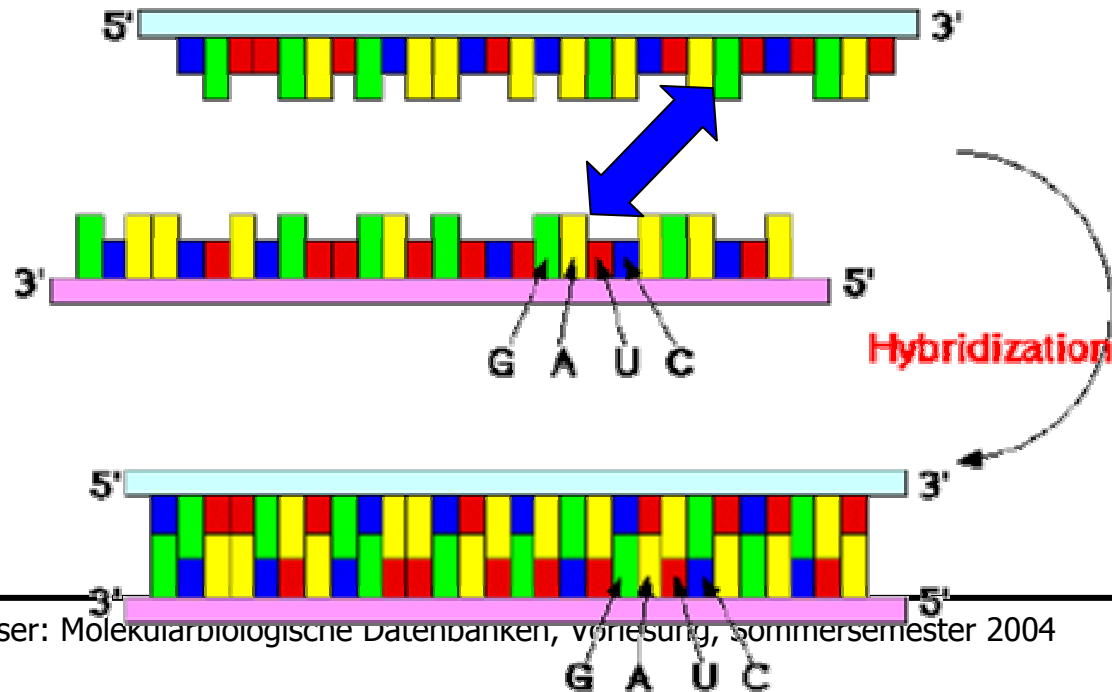
Kartierungstechniken

- Ziel ist der Minimum Tiling Path
- Was kann man über Clone/Marker herausbekommen ?
 - Clone überlappen
 - Hybridisierung
 - Clone enthalten Marker
 - PCR



Hybridisierung (Sketch)

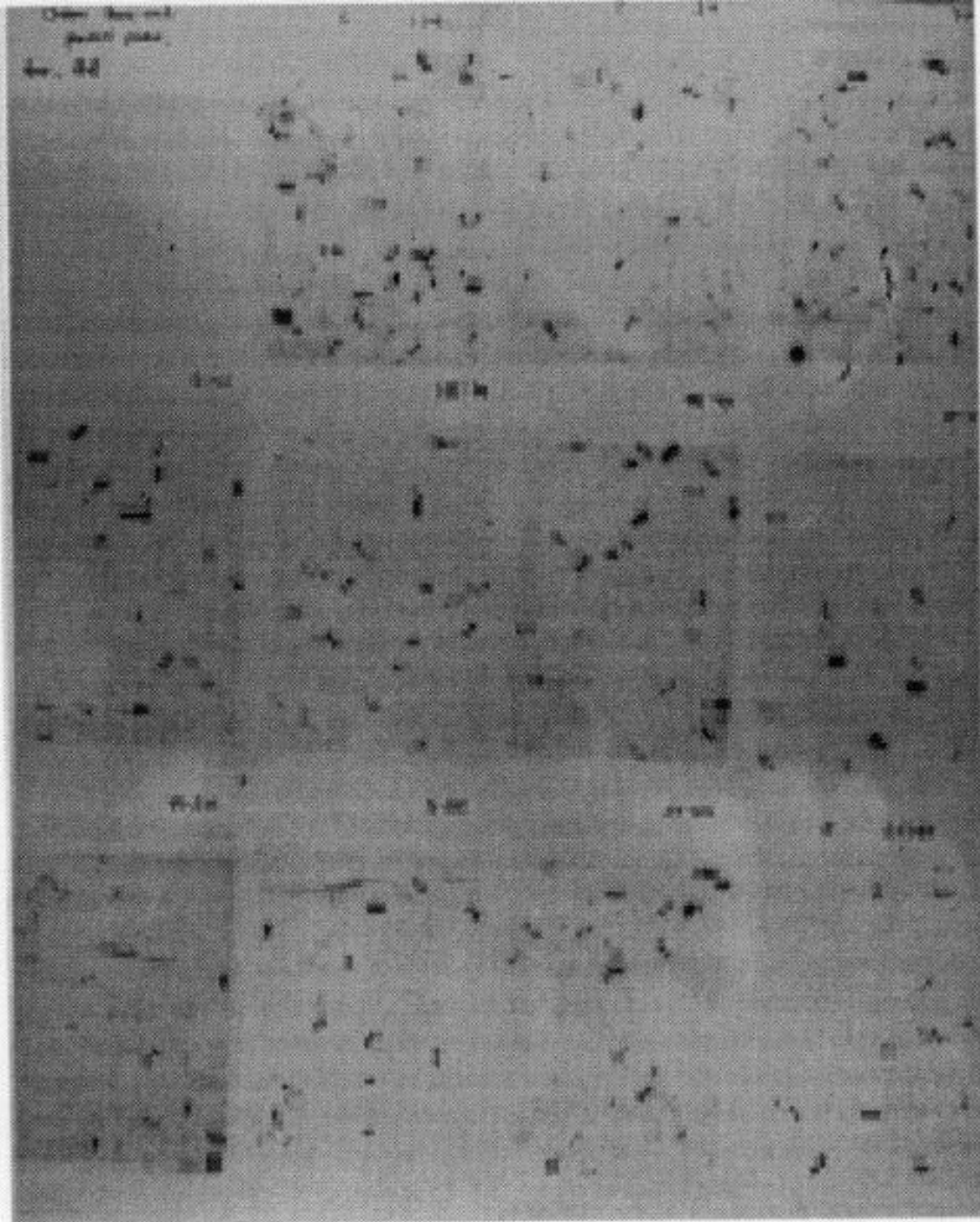
- Ein Clone (Probe, Sonde) wird markiert (radioaktiv, fluoreszierend)
- Probe wird mit Bibliothek auf Filter zusammengebracht
- Clone in Probe hybridisiert mit den Clonen in der Bibliothek, die komplementäre DNA tragen
- Nicht hybridisierte Clone werden abgewaschen



Ergebnis

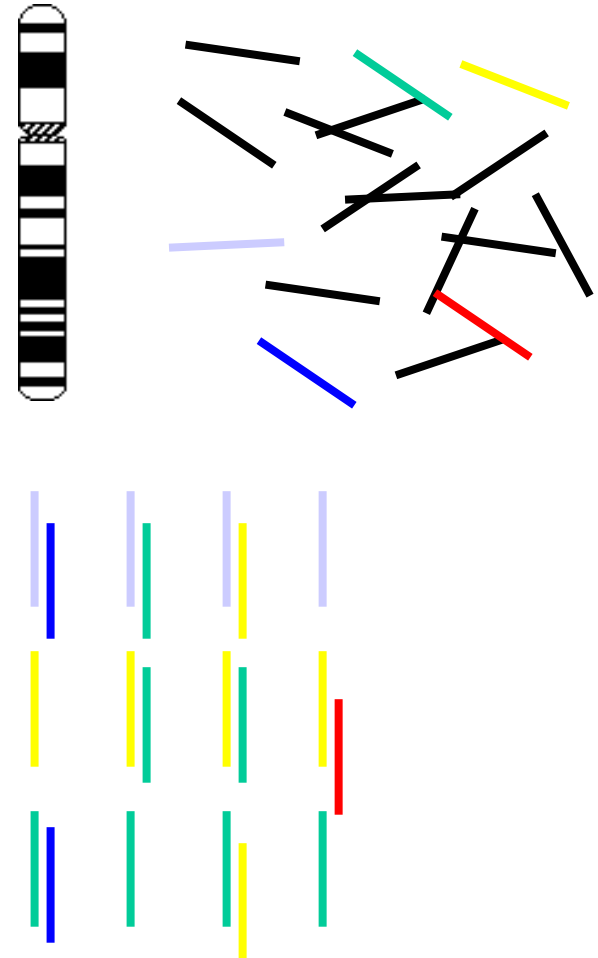
- Manuelle Auswertung
- Imageanalyse
- Spot-Strategien
- Clone-überlappungen

Quelle: Choi, Wing, Caltech



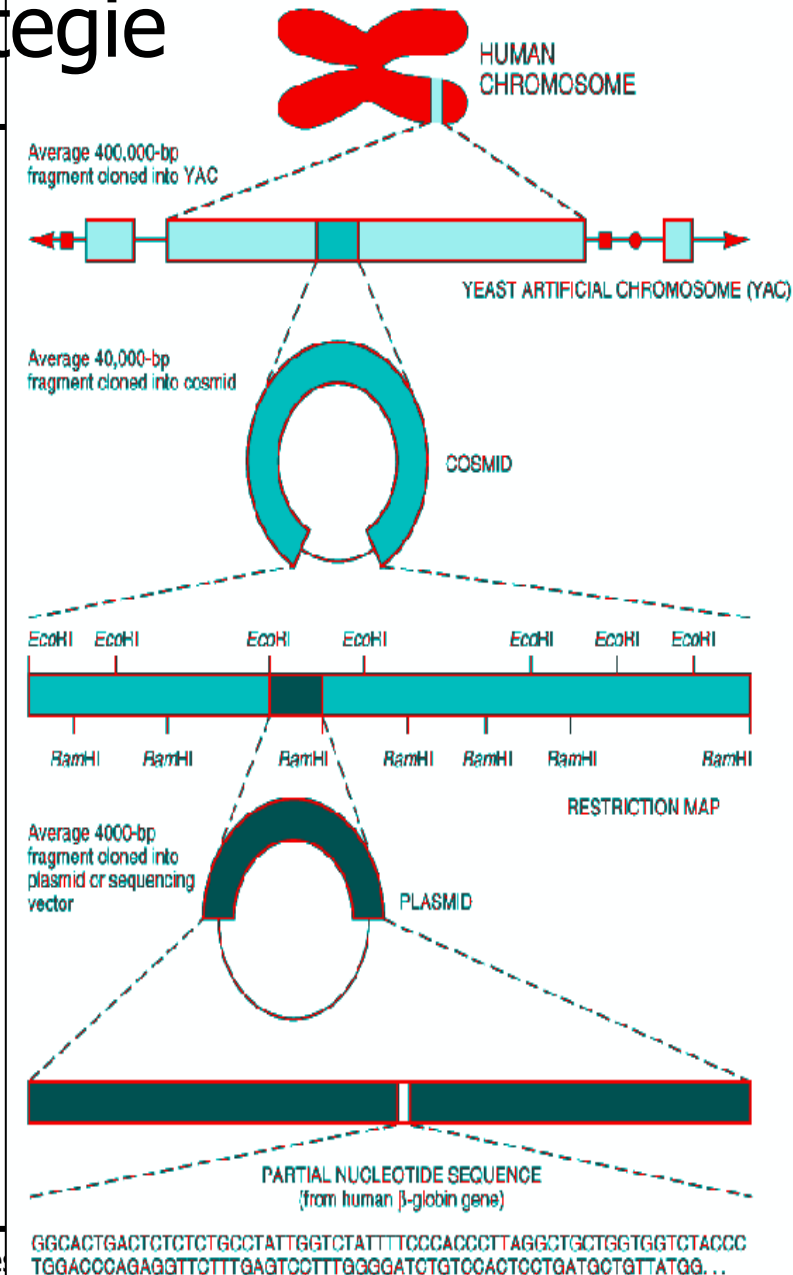
Abstraktes Ergebnis

- Gegeben
 - Bibliothek
 - Menge von Proben
 - Hybridisierungsergebnisse
- Folgerungen
 - Clone-Überlappungen
 - Clone-Nicht-Überlappungen
- Mapping - Aufgabe
 - Konsistente Reihenfolge der Clone finden
 - Algorithmen später

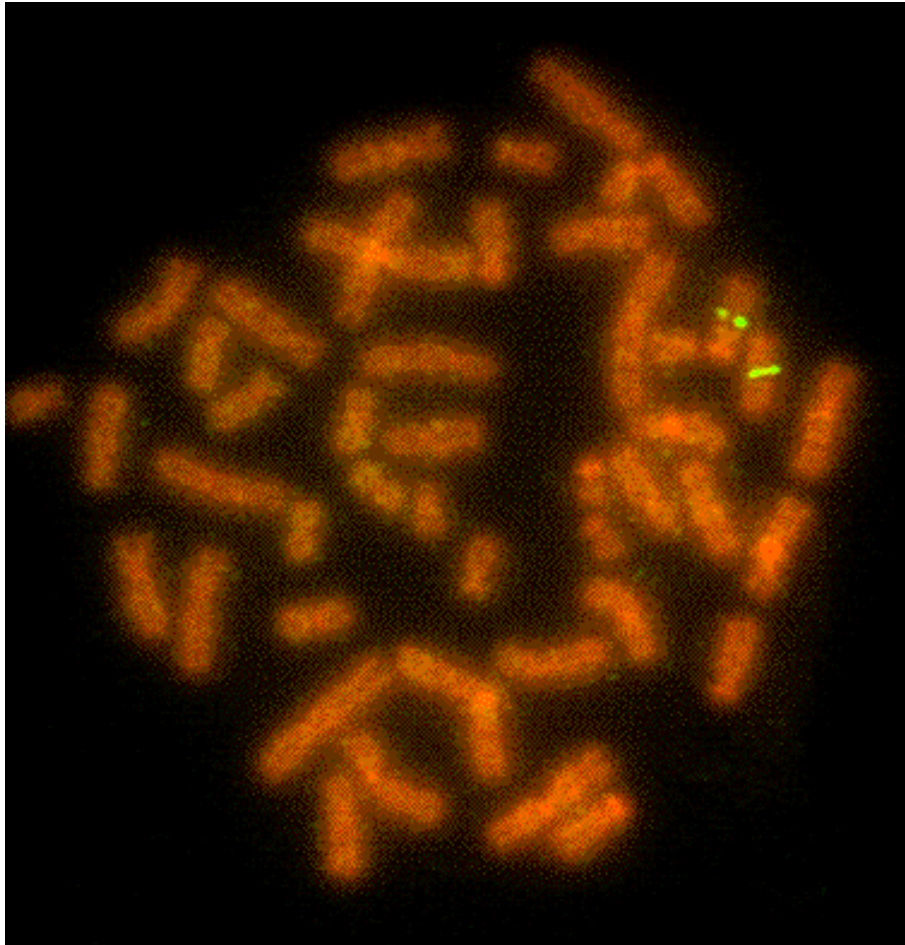


Beispiel für Mapping-Strategie

- Kurze Clone – viele Hybridisierungen notwendig
- Lange Clone – nicht sequenzierbar
- **Mehrstufige Verfahren**
- Beispiel
 1. Stufe: YAC – YAC Hybridisierung
 - YAC Minimum Tiling Path, Y_1, \dots, Y_n
 2. Stufe: YAC – Cosmid Hybridisierung
 - Alle Cosmids zu Y_i , picken und neue Library machen
 3. Stufe: Cosmid – Cosmid Hybridisierung
 - Cosmid Minimum Tiling Path pro Yac
 4. Stufe: YAC Übergangszonen klären
 - Cosmid Minimum Tiling Path



Einschub: In Situ Hybridisation



Clone kann man direkt mit Chromosomen hybridisieren

- Zuordnung von Clonen zu Chromosomen
- Zuordnung von Clonen zu Chromosomenbändern

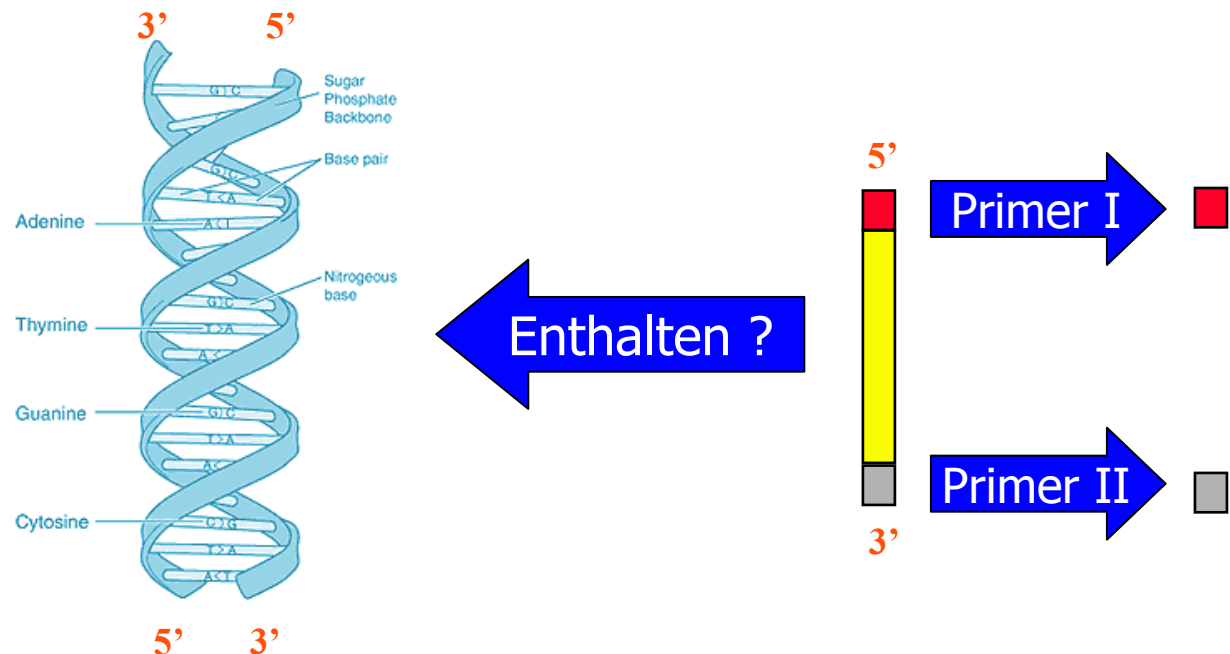
PCR (Sketch)

- **Polymerase Chain Reaction**

- Methode zur Vervielfältigung von DNA Bruchstücken
- Zusammen mit Gelelektrophorese hochempfindliches Nachweisverfahren
- Nachzuweisende Sequenz muss bekannt sein
- Vielfältige Verwendung
 - „Genetischer Fingerabdruck“
 - Forensik
 - Vaterschaftstests

PCR Illustration

- Gegeben:
 - Doppelsträngige DNA D
 - Probe S mit bekannter Sequenz
- Frage:
 - $S \in D$?



PCR Idee

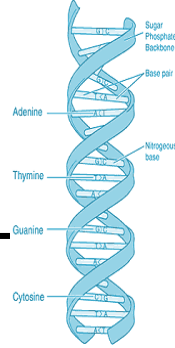
- Zutaten

- Doppelsträngige DNA kann durch Erhitzen in einzelsträngige aufgebrochen werden
- Polymerase ist ein Enzym, dass an einzelsträngiger DNA mit doppelsträngigem Teilstück bindet und das doppelsträngige Teilstück verlängert
 - Vorhandensein der notwendigen Nukleinsäuren vorausgesetzt

- PCR Anfang

- In Lösung: zu untersuchenden DNA, Primer, Polymerase, Nukleinsäuren

PCR Zyklen



Heat
→

Cool
↙

Heat
→

Cool
↙

Heat
→

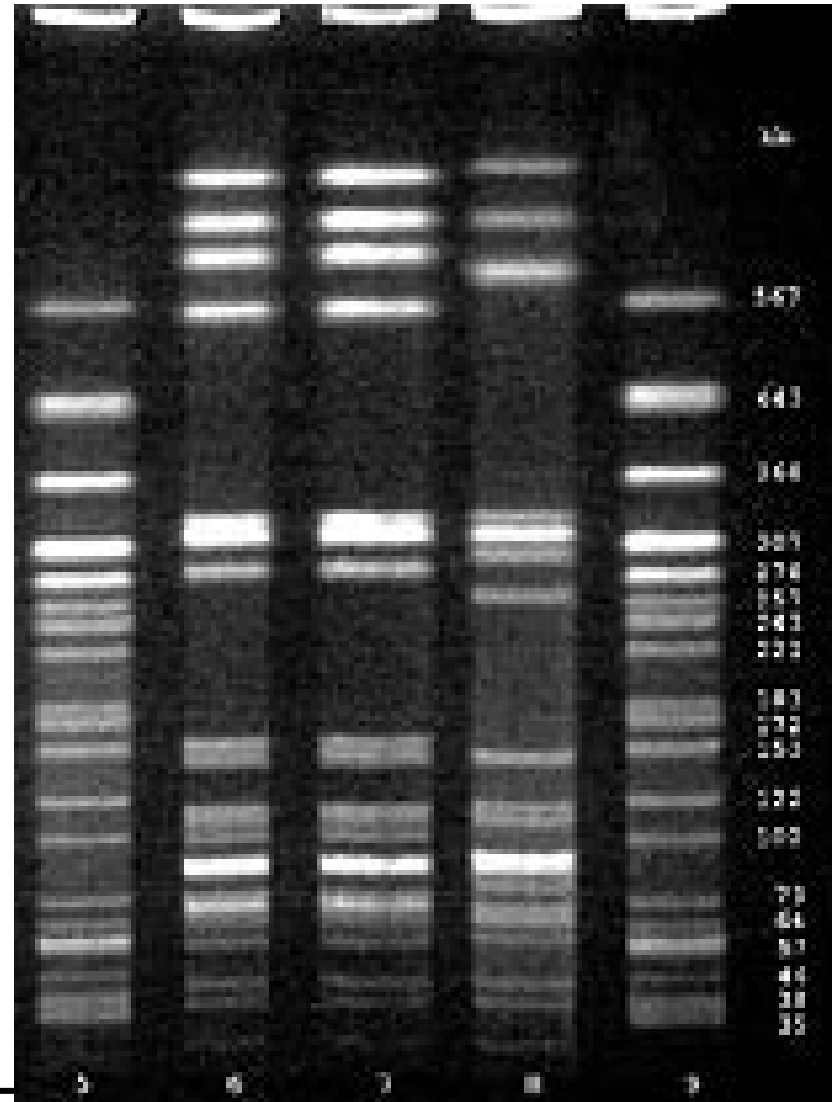
Cool
↙

Heat
→

etc.

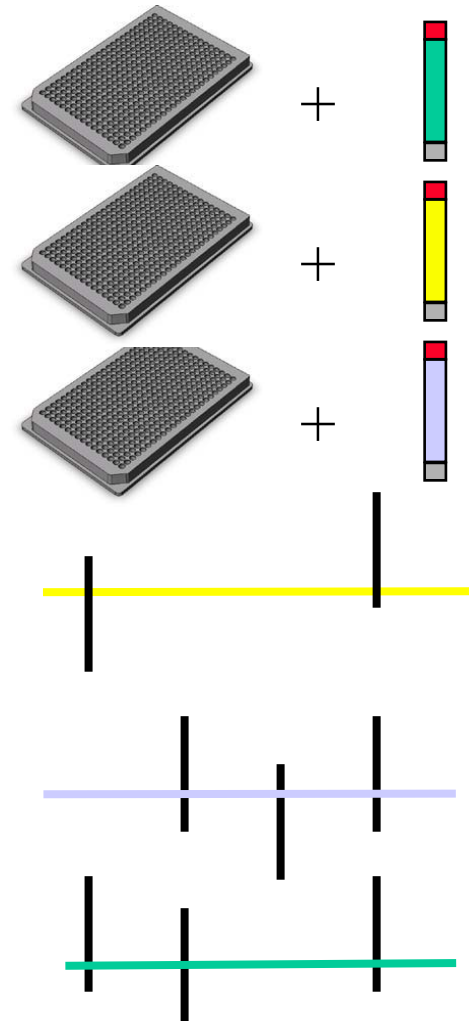
Nachweis

- Ergebnis:
- $S \in D$: Lösung enthält sehr viel S
- $S \notin D$: Lösung enthält fast kein S
- Nachweis:
Auftrennung im Gel



STS Content Mapping

- Gegeben
 - Menge von Clonen
 - Menge von STS
- Gesucht
 - Welcher der Clone enthält welche STS ?
- Verfahren
 - PCR mit jedem STS
- Ergebnis
 - „Anchoring“ der Clone
 - Algorithmen später



Teil III. Integrierte Karten

➤ Integration von Karten

- Eines Typs: Viele Labore erstellen Karten der gleichen Region
 - Vergleich zweier BAC Karten
- Unterschiedlichen Typs: Verschiedene Sichten auf die selbe Region
 - Kombination genetischer und physikalischer Karten
- Vorteile
 - Auflösen von Konflikten
 - Ergänzung fehlender Informationen
 - Verifikation durch Redundanz
 - Höhere Robustheit durch Einsatz unterschiedlicher Techniken
- Beispiele
 - Orientierung und Abstand von STS-Markern durch genetische Karten
 - Positionierung von Clone-Contigs durch genetische Marker
 - Zuordnung von Clonen zu Chromosomen durch FISH

Voraussetzungen

- Karten müssen **identische Objekte** teilen
- Identische Objekte müssen als solche erkannt werden
- Entscheidungsfindung bei widersprüchlichen Daten (oft ohne ausreichende oder vergleichbare Evidenz)
- Integration unterschiedlicher mathematischer Verfahren

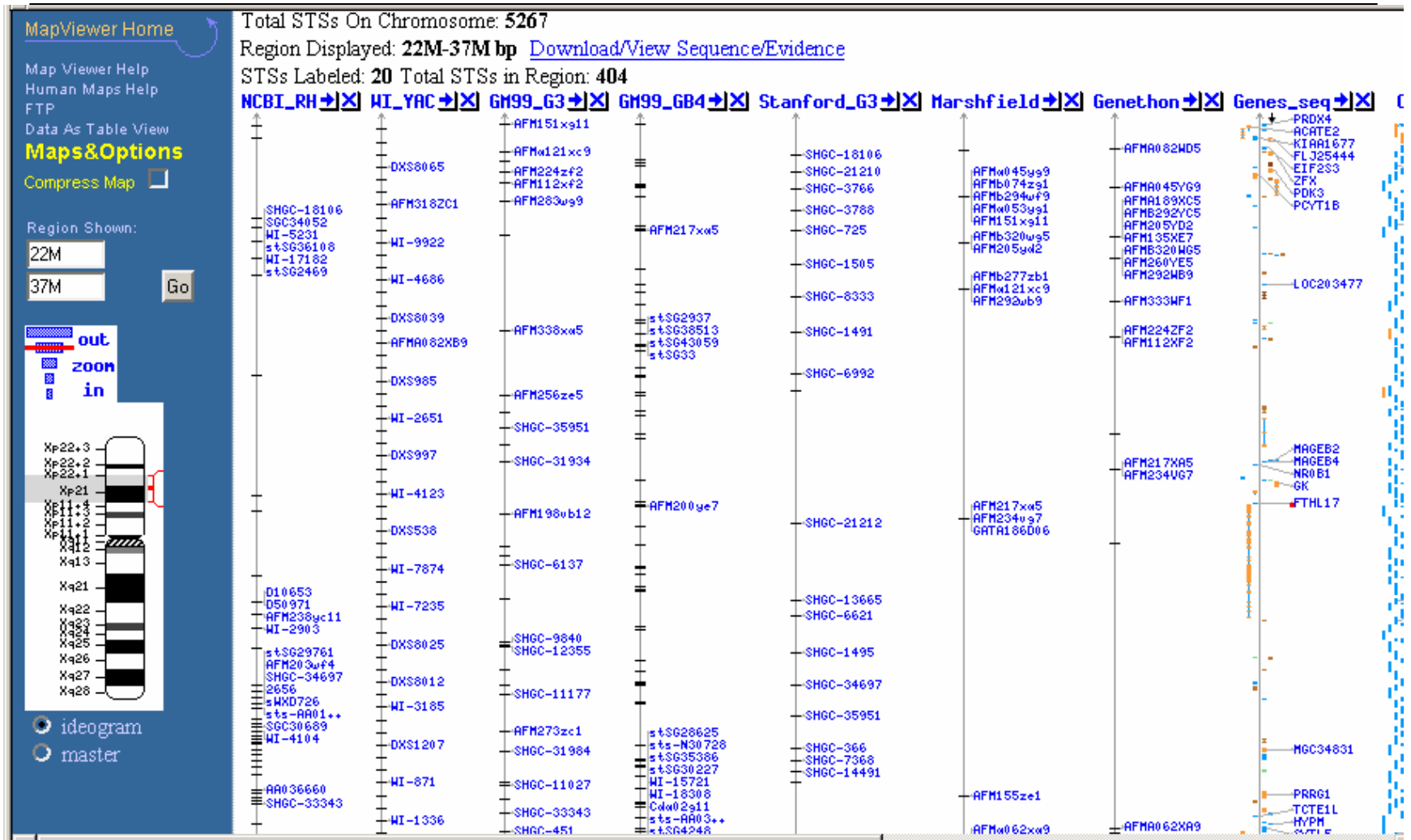
Vielfältige Probleme

- Objektnamen
 - Identifizieren gleicher Objekte sehr schwierig
 - Clone/Marker mit >20 Synonymen keine Seltenheit
- Klassennamen
 - Marker, Locus, Gene alle nicht klar definiert
 - Ein Gen auf Karten kann alles mögliche sein:
 - Gen selber (aber: Gennamen nicht standardisiert)
 - ESTs und cDNAs
 - SNPs, Mutationen, Polymorphismen
 - Splicevarianten
 - Proteinname
 - ...
 - Bereich zwischen 40-200 Basen bis 2 MB

Probleme –2–

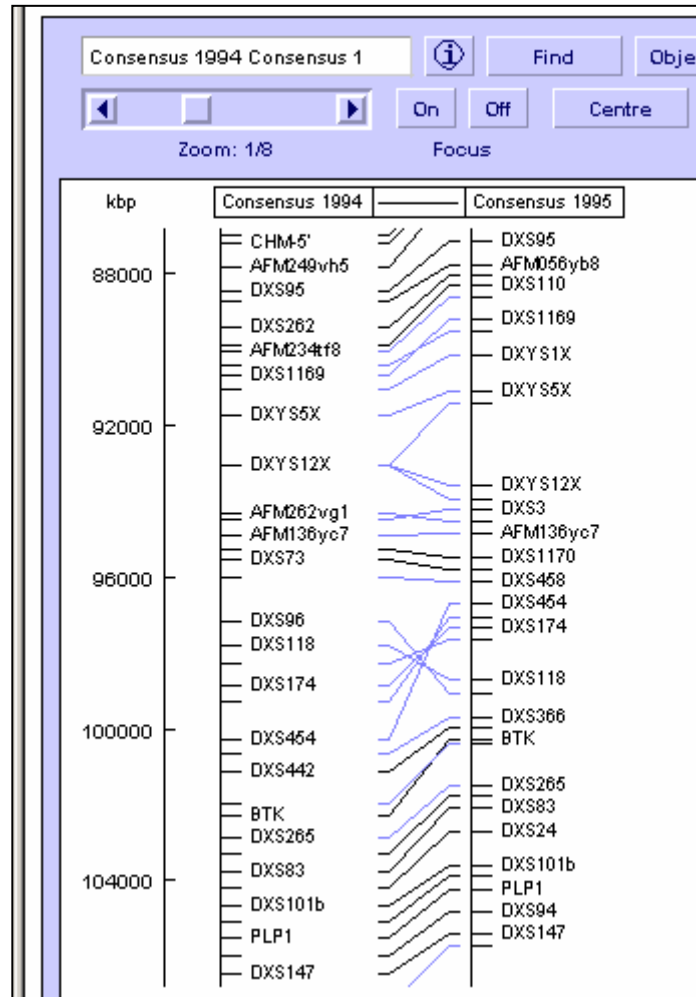
- Typische Integrationsmethode sehr anfällig
 - Eigentlich:
 - Rohdaten (Überlappungen) integrieren
 - Karte neu berechnen
 - Aber:
 - Rohdaten nicht verfügbar
 - Unterschiedliche Protokolle
 - Komplexität der Algorithmen zu groß
 - Also
 - Manuelle Integration der (veröffentlichten) Ergebniskarten
- Keine Einschätzung der Güte mehr möglich
- Umrechnung zwischen Koordinatensystemen schwierig
 - 1cM \approx 1MB nur Durchschnittswert
 - Konfidenzintervalle der Koordinaten beachten

Beispiel



Quelle: <http://www.ncbi.nlm.nih.gov/mapview>

Divergenzen



Quelle: <http://ixdb.mpimg-berlin-dahlem.mpg.de>

Fazit

- Physikalische Kartierung traditionell Voraussetzung für Sequenzierung
- Vielzahl von etablierten Methoden verfügbar
- Aber
 - Aufwändige Verfahren, manuelle Schritte
 - Keine Bevorzugung „interessanter Regionen“
 - Fehleranfällige Daten und komplexe Algorithmen
- Berechnung von Karten
 - Rohdaten: Hybridisierungen oder STS Content Mapping
 - Algorithmen müssen Fehler verkraften können
- Vollständigkeit, Robustheit und Qualität lässt sich durch Integration unterschiedlicher Karten steigern
 - Aber mit vielen Problemen behaftet
- (Nahezu) alle bekannten Karten sind integrierte Karten