

Entwicklung und Implementierung eines Auswertungswerkzeuges für Matrix-CGH-Daten:

*Markus Kreuz, Hilmar Berger, Maciej Rosolowski,
Swen Wessendorf, Carsten Schwaenen & Dirk Hasenclever*

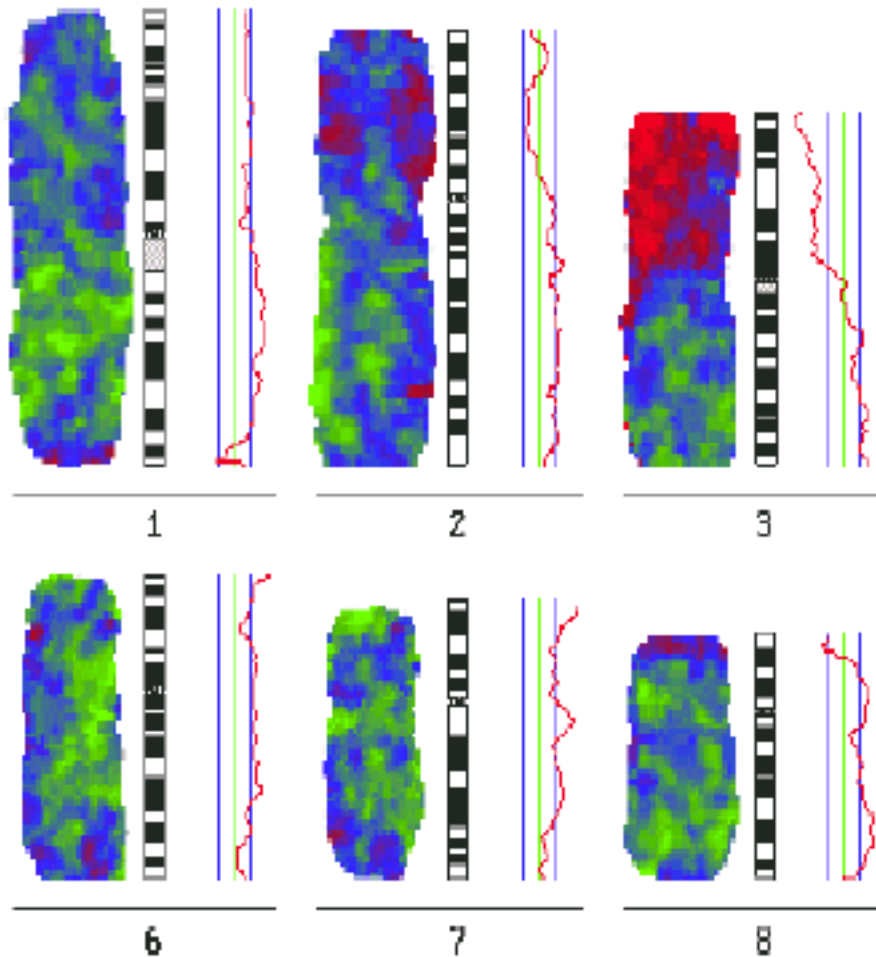
Inhalt:

- Was ist CGH/aCGH
- Analysewerkzeug für aCGH Daten:
 - (A) Implementierung
 - (B) Funktionen
 - Qualitätskontrolle
 - Normalisierung
 - Segmentierung
 - Klassifikation
 - Multi-Chip Analyse
 - Integrierte Analyse von aCGH- and Genexpressionsdaten

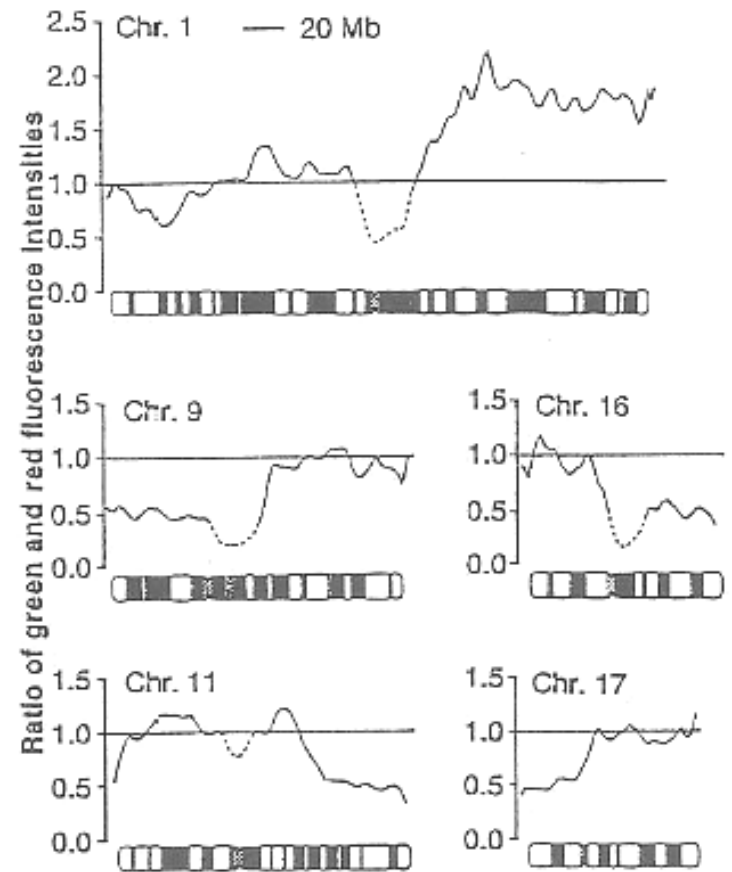
Klassische Comparative Genomic Hybridisation (CGH):

- Lokale Unterschiede in der DNA-Kopiezahl (Aberrationen, Amplifikationen) können detektiert werden
- DNA des Test- und Referenzgewebes wird farblich markiert
- Gemisch aus Test und Referenz DNA wird gegen Metaphase-Chromosomen hybridisiert
 - Kompetitive Reaktion
- M-Phase Chromosomen werden mit Mikroskop untersucht
- Beobachtetes Farbverhältnis in einer Genomregion spiegelt das lokale Verhältnis der DNA-Kopiezahl von Test- und Referenzgewebe wider
- Ermöglicht genomweite Analyse, Auflösung ist jedoch gering (10-20 MBp)

Farbinformation:



Analyse der Farbinformation:



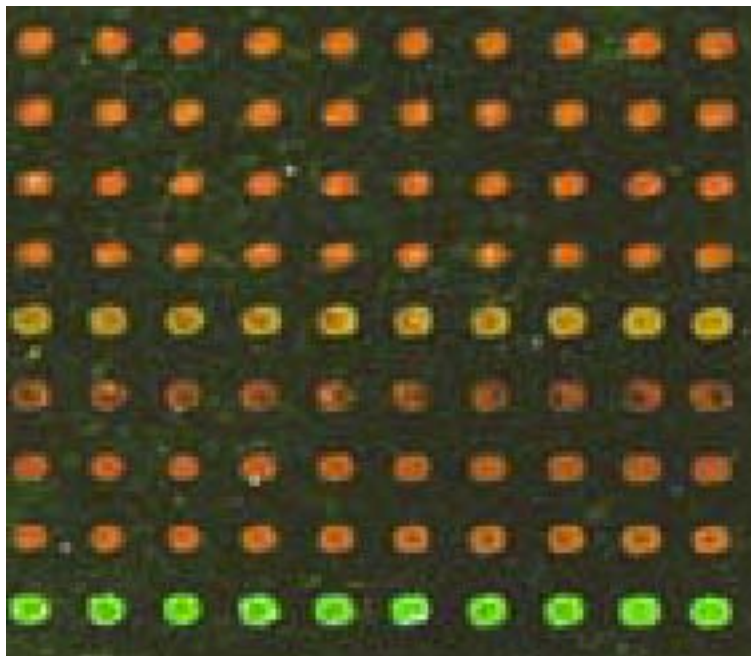
Matrix-CGH:

- Statt M-Phase Chromosomen werden Klonsonden verwendet
- Diese sind auf einem Chip platziert (vergleichbar mit Genexpressionsanalyse)
- Mehrere tausend Klone können parallel gemessen werden

- Vorteile:
 - Höhere Auflösung
 - Weniger Gewebematerial wird benötigt
 - Klondichte kann in relevanten Regionen erhöht werden

Datenverarbeitung:

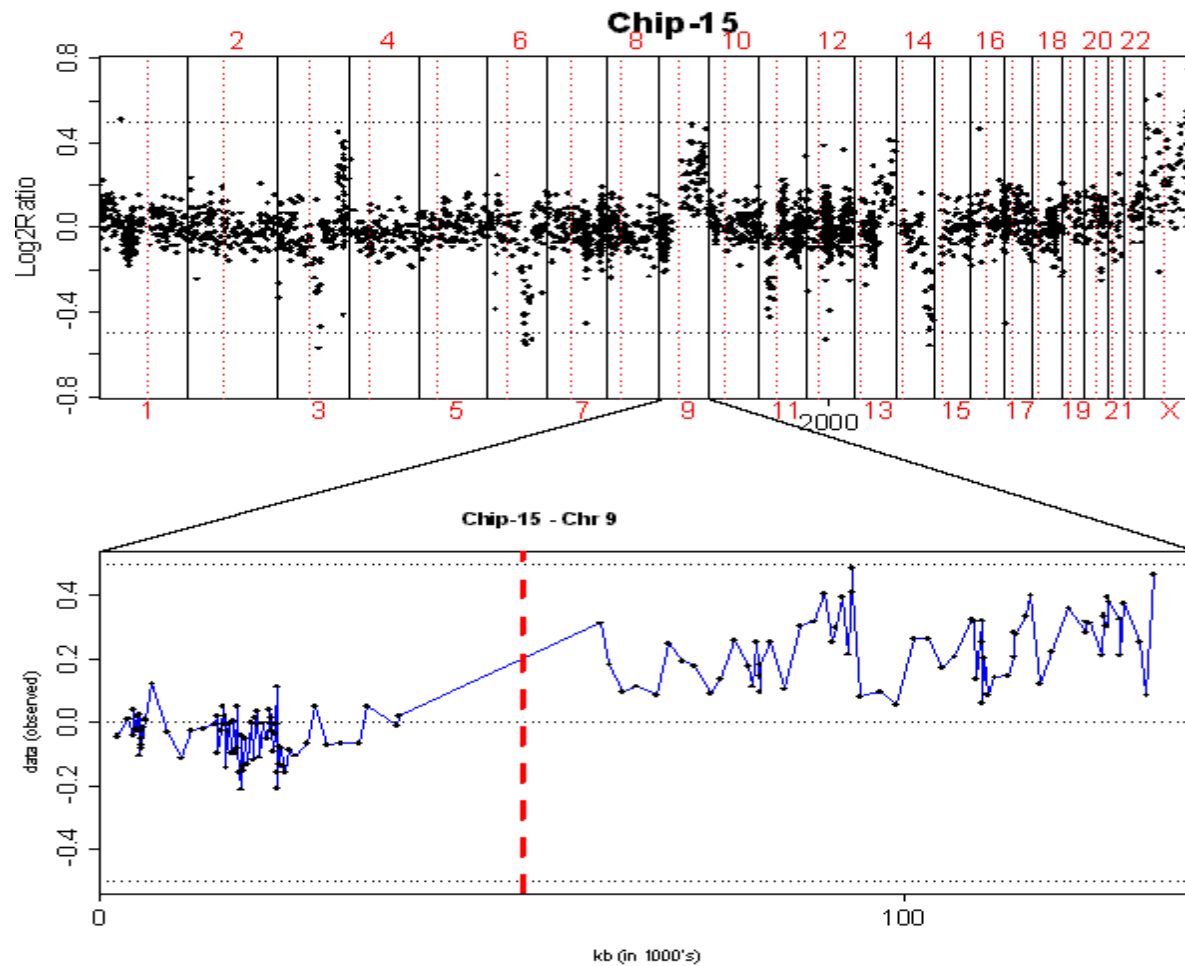
Farbinformation:



Scannen und Vorverarbeitung:

ID	Log2(Ratio)	Ratio
1p_20177817_	0,054820547	1,03872989
1p36.1_219080	-0,055906595	0,96198974
1p36.13-1p36.	0,063330476	1,04487508
1p_22908025_	-0,068225938	0,95381016
1p36.13_2333	0,147413794	1,10758222
1p_24714610_	0,182482617	1,13483505
1p36.1-1p36.2	-0,028205184	0,98063953
1p_27812052_	0,101160476	1,07263592
1p36.1_28231	0,124950981	1,09047068
1p_28828925_	-0,302820029	0,81066624
1p_30333933_	-0,273580891	0,82726366
1p_32317745_	-0,274215535	0,82689982
1p_34198917_	-0,590791656	0,66397846
1p_36414831_	-0,362109516	0,77802611
1p_37979843_	-0,340269317	0,78989384
1p34_3922872	0,097538753	1,06994657
1p34.3_39365	-0,007584759	0,99475644
1p_39452868_	-0,268726135	0,83005214

Visualisierung:



Implementierung:

- Werkzeug ist in R als eigenständiges Paket implementiert
- Analyseschritte werden durch Funktionen realisiert
- Daten sind in einem R-Objekt gespeichert

- Funktionalität umfasst:
 - Import der Daten
 - Datenanalyse
 - Präsentation der Ergebnisse im HTML-Format
 - Export der Ergebnisse in Excel oder tab-separierte txt Dateien

Qualitätskontrolle:

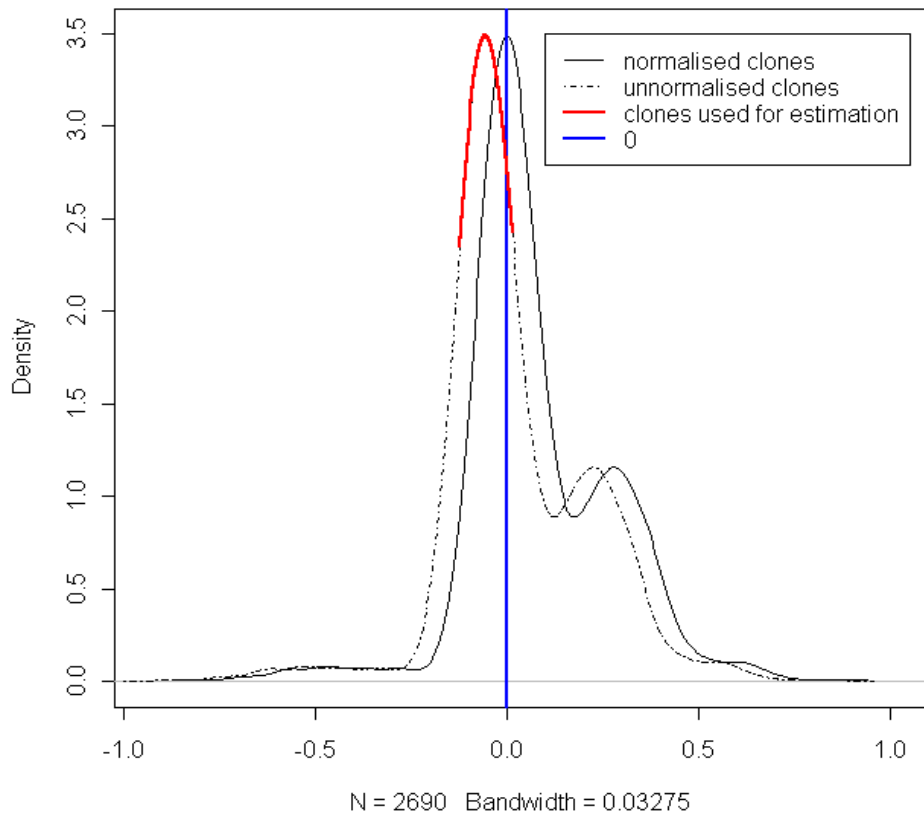
- Werkzeug markiert Klone/Chips mit:
 - Vielen Missing Values
 - Starker Streuung (Median Absolute Deviation)
 - Systematischen Verzerrungen (Bias)
- HTML Ausgabe:
 - > Benutzer kann gegebenenfalls Chips/Klone ausschließen

Vorverarbeitung – Normalisierung:

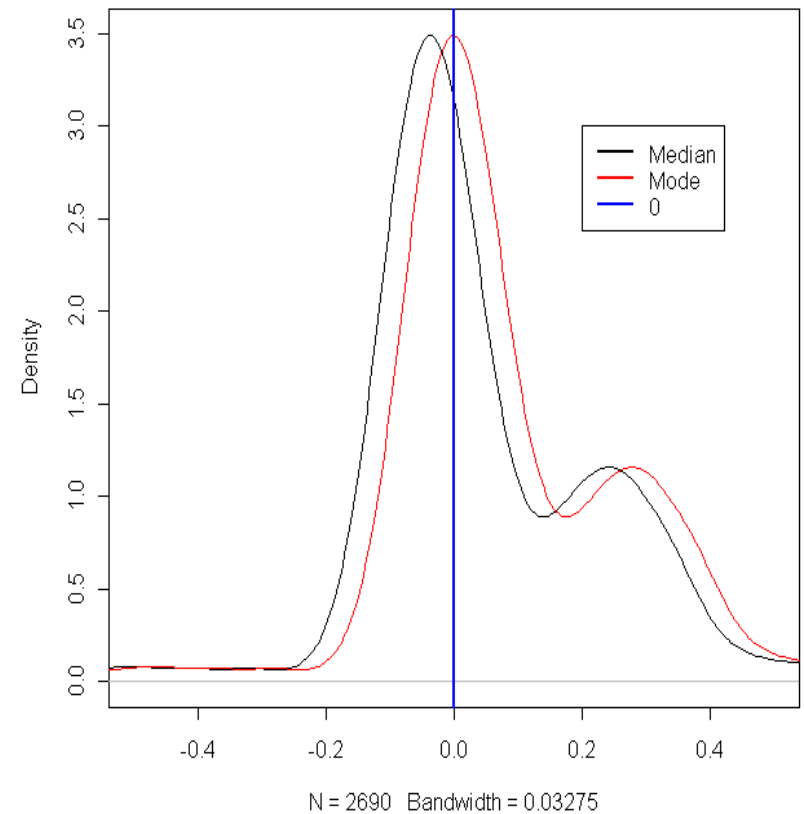
- Unterschiede in Menge/Qualität von Test und Referenz DNA unvermeidlich
 - Führt zu systematischen Abweichungen der Messwerte (Ratios zu hoch/niedrig)
- Verschiedene Ansätze um den Effekt zu schätzen/eliminieren
 - **Globale** / Intensitätsabhängige / Print-Tip Normalisierung
 - Verwende alle Klone vs. verwende Teilmenge mit geringer Aberrationswahrscheinlichkeit
 - Mittelwert vs. Median als Effektschätzer

Schätzer für Modalwert:

Normalisation using mode-estimator



Median vs. Mode

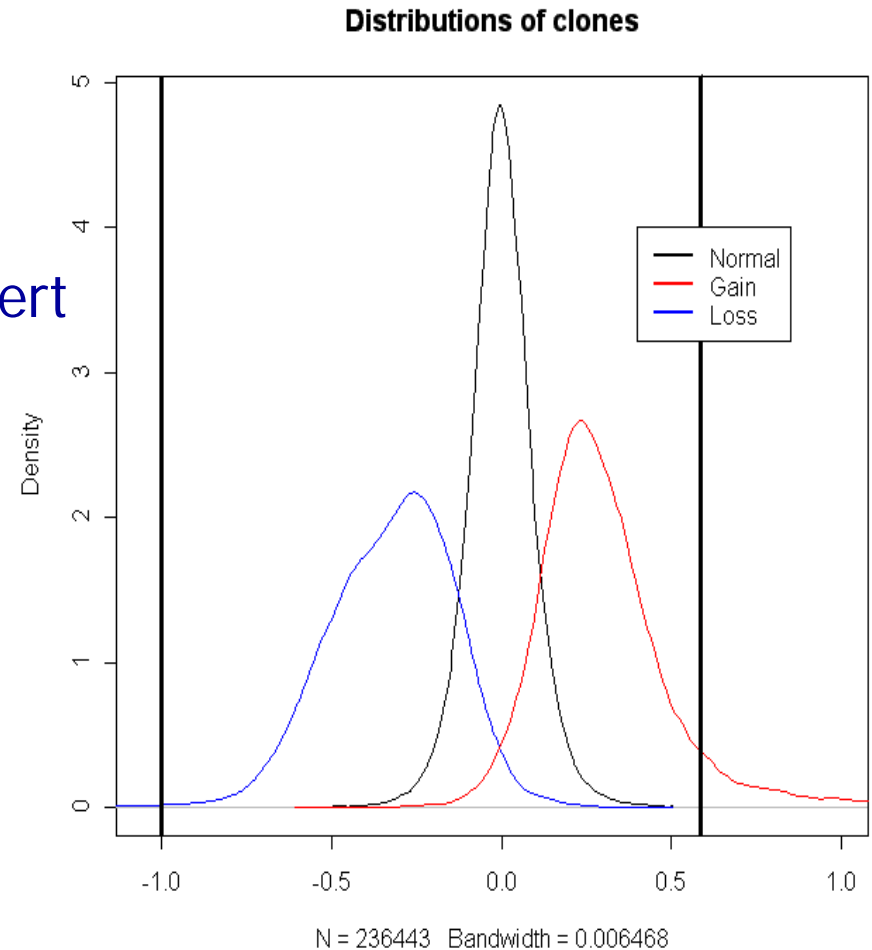
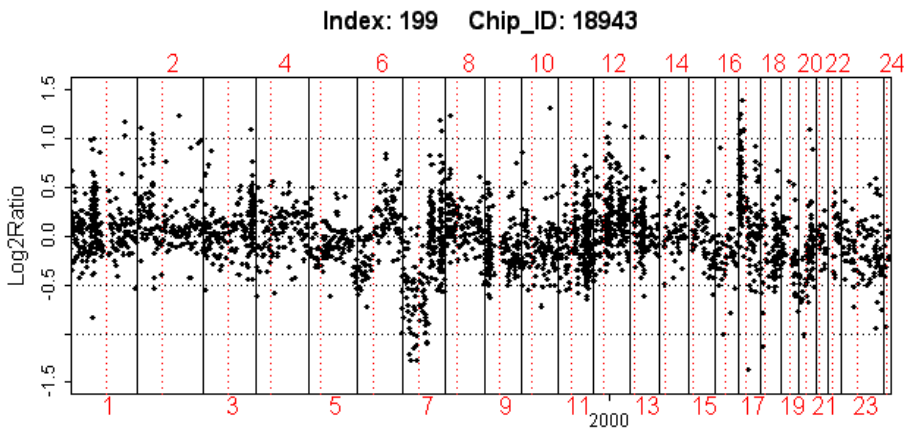


Segmentierung:

- Jedem Klon soll ein Status, Zugewinn/Normal/Verlust, zugewiesen werden
 - Verwendung von Schwellwerten
 - Niedrige signal to noise ratio
 - > Schwache Aberrationen bei Betrachtung einzelner Klone schwierig zu finden
 - Aberrationen umfassen meist längere Segmente
 - Information von benachbarten Klonmesswerten wird genutzt um Streuung der Messwerte zu reduzieren

Segmentierung (2):

- Einsatz von Glättungs- oder Segmentierungsverfahren
 - Streuung wird reduziert
 - Vielzahl von Methoden existiert



Segmentierung (3):

- Einige Segmentierungsmethoden im Überblick:
 - **Olshen (2004)** -> **Circular Binary Segmentation (CBS)**
 - **Fridlyand (2004)** -> **Hidden Markov Model (HMM)**
 - Wang (2004) -> Cluster Along Chromosomes (CLAC)
 - Bilke (2004) -> Topological Statistics
 - Jong (2003) -> aCGH Smooth
 - Autio (2003) -> CGH-Plotter
 - Picard (2005) -> CGHseg

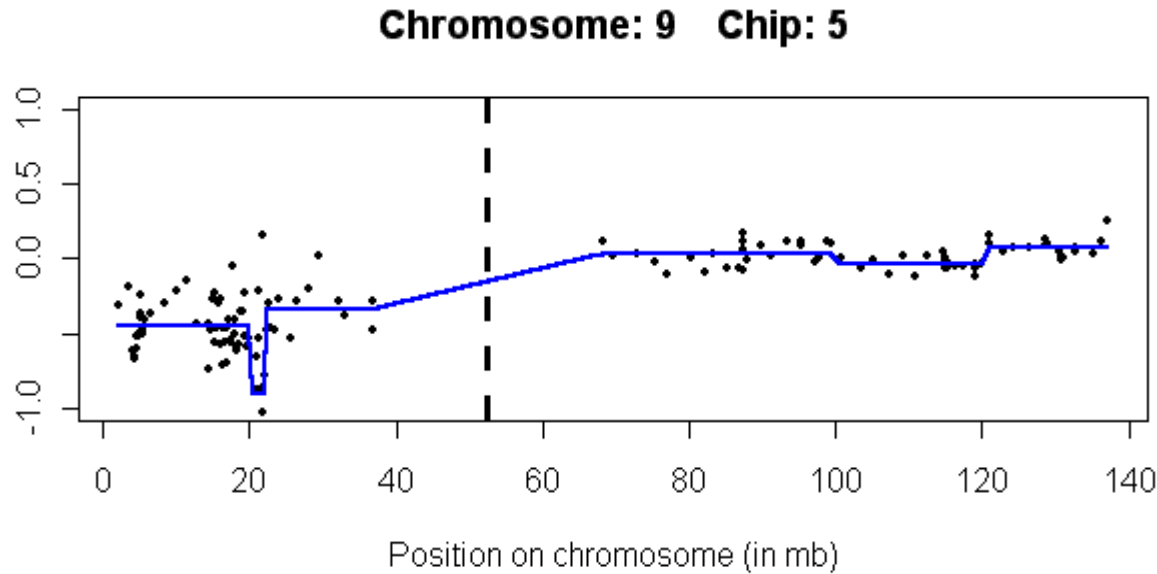
Segmentierung (4):

- Methode von Fridlyand et al.:
 - Basiert auf Hidden-Markov-Modellen (HMM)
 - Klone werden verschiedenen Zuständen zugeordnet
 - Jedem Zustand ist eine unterschiedliche Ausgabeverteilung zugeordnet
 - Transitionsmatrix initialisiert, dass Zustandswechsel unwahrscheinlich
 - Für jedes Chromosom wird HMM mit 1 bis N Zuständen angepasst
 - Auswahl des besten HMM mittels AIC
 - Weitere Analyse des HMM (Ausreißer, Transitionen...)

Segmentierung (5):

- Methode von Olshen et al.:
 - Change-point Problem
 - Circular binary segmentation
(Erweiterte Version der Binären Segmentierung)
 - Jedes mögliche Segment wird mittels Permutationstest auf signifikante Abweichung des Mittelwerts geprüft

Segmentierung (6):



- **Vergleich Olshen vs. Fridlyand:**
 - Olshen zeigt bessere Ergebnisse
 - Fridlyand erzeugt keine ausreichende Glättung
-> Hohe False Discovery Rate

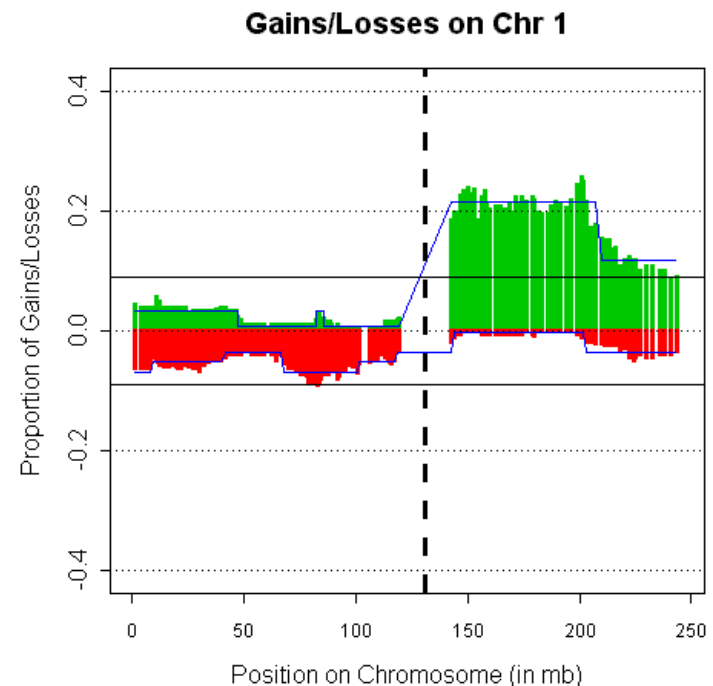
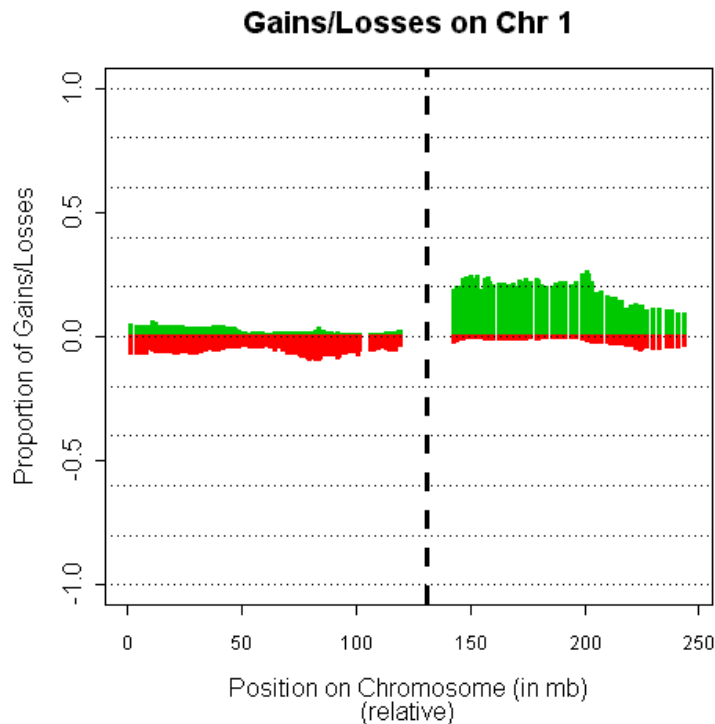
Klassifikation:

- Für jedes Segment wird entschieden welchen Status es erhält: Zugewinn/Normal/Verlust
 - Verwendung von Schwellwerten
- 2 verschiedene Ansätze
 - i) Konstanter Schwellwert
 - ii) Streuungsabhängiger Schwellwert
- Auswahl von Verfahren abhängig von Heterogenität der Experimente/Chips
- Auswahl der Schwelle erfolgt zur Zeit manuell
 - Großer Einfluss von DNA-Qualität und Tumorzellanteil

Multi-Chip Analyse:

- Automatische Detektion von rekurrenten chromosomalen Imbalancen (Rekurrente Regionen):
 - Basiert auf relativen Häufigkeiten von Zugewinnen/Verlusten
- Hidden-Markov-Modelle mit 1 bis 5 Zuständen werden an die relativen Häufigkeiten von Zugewinnen/Verlusten angepasst
 - Mittels AIC wird bestes HMM ausgewählt
 - > Ergebnis: Regionen mit gleicher Häufigkeit von Zugewinnen/Verlusten

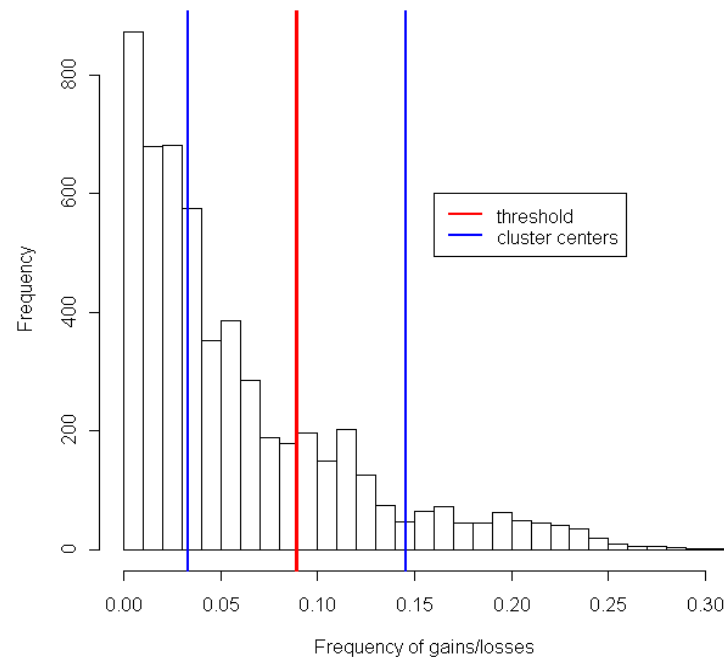
Multi-Chip Analyse (2):



- Jedes Segment wird mit Schwellwert verglichen
-> Zuordnung: Sporadische/Rekurrente Region
 - Adaptiver Schwellwert wird ermittelt

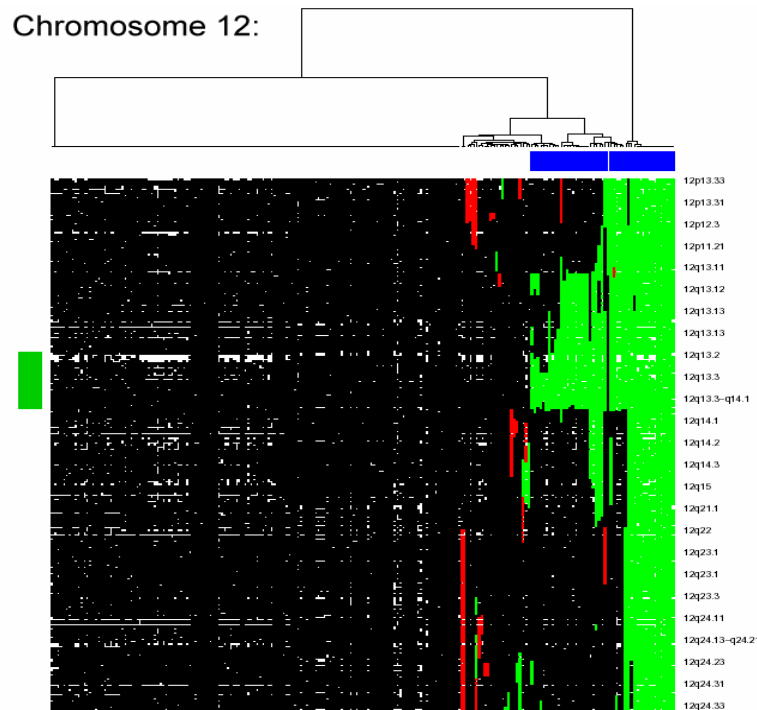
Multi-Chip Analyse (3):

- Zwei Populationen Modell:
 - Schätzen der Häufigkeit von sporadischen/rekurrenten Regionen
 - Verwendung von PAM Clustering
 - Mittelwert der Clusterzentren bildet den Schwellwert



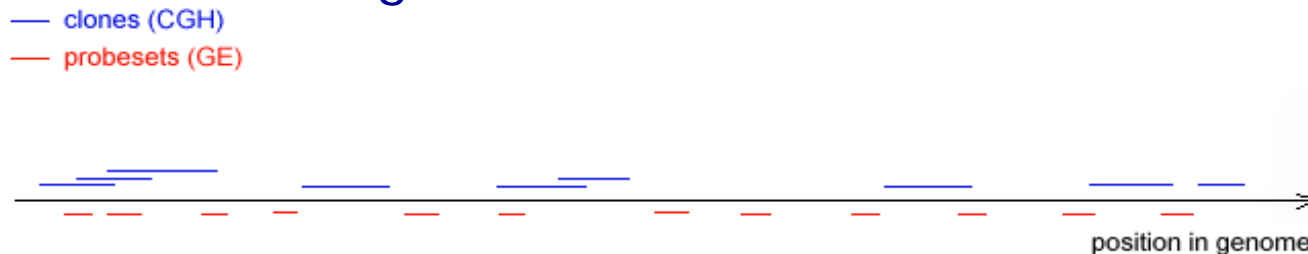
Multi-Chip Analyse (4):

- Mittels voting Algorithmus wird entschieden ob ein Chip in einer rekurrenten Region einen Zugewinn/Verlust aufweist
 - Anteil an Klonen mit Zugewinn/Verlust wird mit Schwellwert verglichen



Integrierte Analyse von aCGH und Genexpression:

- Werkzeug enthält Algorithmus um Genexpressions/CGH Datenpaar zu generieren
 - Kopiezahl der Gene wird geschätzt
 - Keine 1 zu 1 Abbildung möglich, nicht jedes Gen wird von einem Klon abgedeckt



- Kartierungsalgorithmus basiert auf CGH-Segmenten
=> Interpolation zwischen Klonmesswerten
- Ergebnis: Jedes Probeset wird durch Wertepaar repräsentiert

Vielen Dank für die
Aufmerksamkeit