Informationsveranstaltung Qualitätsmanagement für Hochdurchsatz-Genotypisierung

Statistische Qualitätssicherung von Affymetrix- Daten

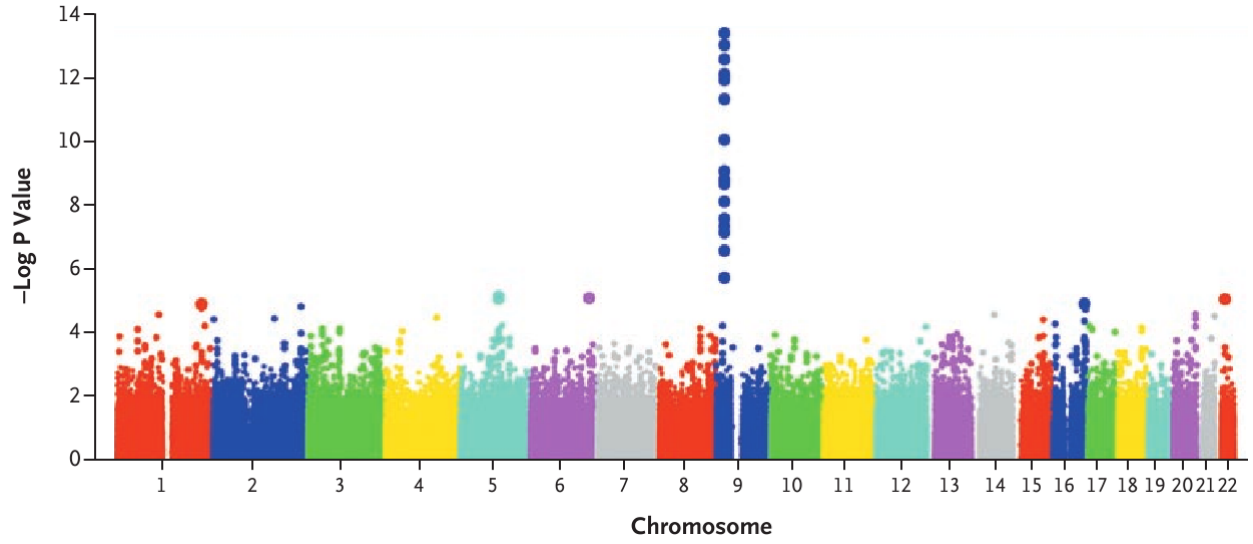
Berlin, 21.06.2010
Arne Schillert & Andreas Ziegler
Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck

SPONSORED BY THE

Federal Ministry
of Education
and Research

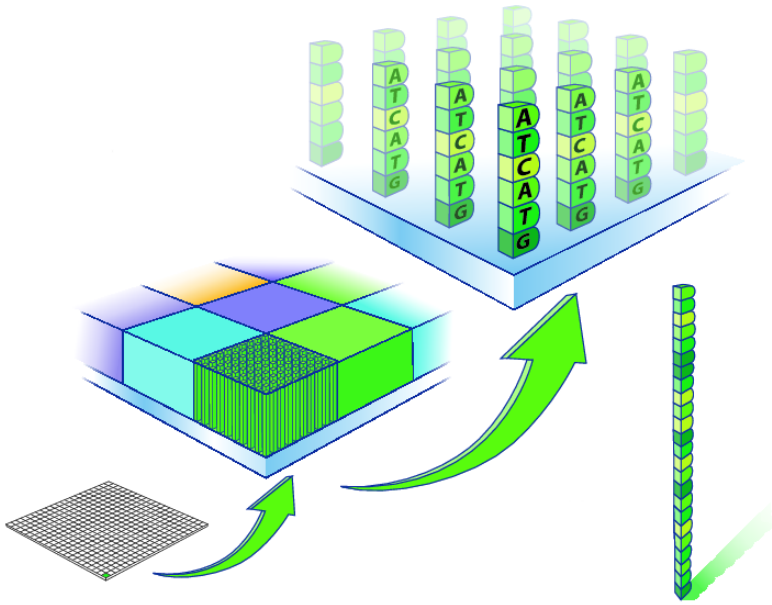
Hochdurchsatz-Genotypisierung

- Ergebnis einer genomweiten Assoziationsstudie:



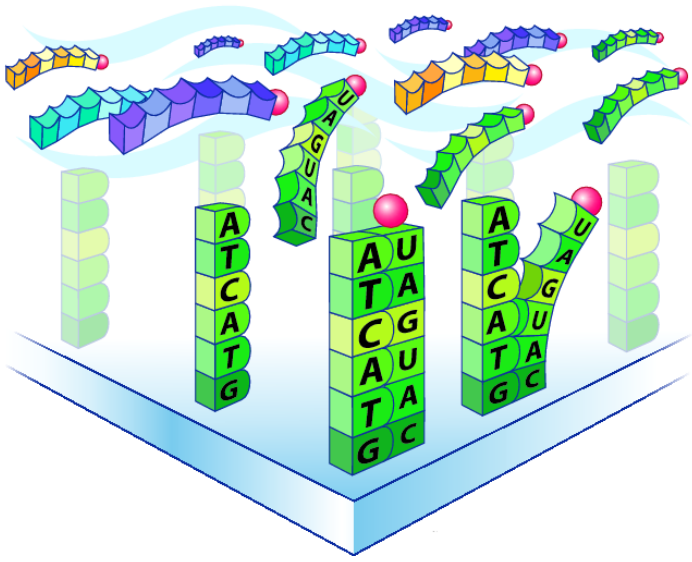
Samani, NEJM 2007, 357:443-453

Schematischer Aufbau eines Microarrays



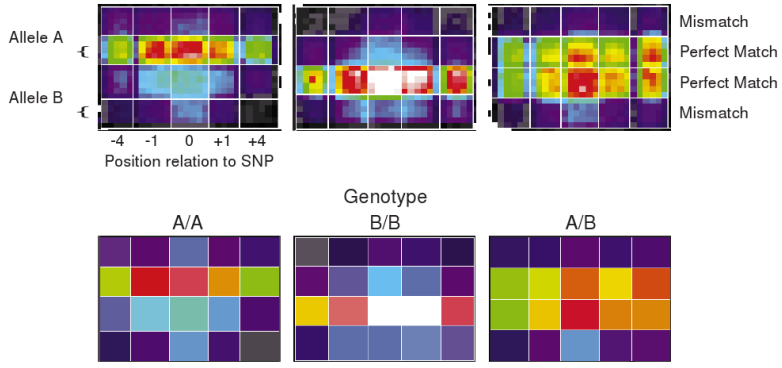
- Oligonukleotide auf Glasträger fixiert
- komplementär zur Sequenz des SNPs
- vollständig komplementär: *perfect match (PM)*
  Base des SNPs nicht komplementär: *mismatch (MM)*
- pro SNP mehrere Oligonukleotide

Ablauf eines Microarray-Experimentes



- DNA fragmentieren und markieren
- DNA-Fragmente auf Array hybridisieren lassen
- Abwaschen nicht gebundener Fragmente
- Farbstoff anhängen und anregen
- Scannen und Bild an Raster ausrichten $\rightarrow$ CEL-Datei

Falschfarbenbild der Intensitätswerte

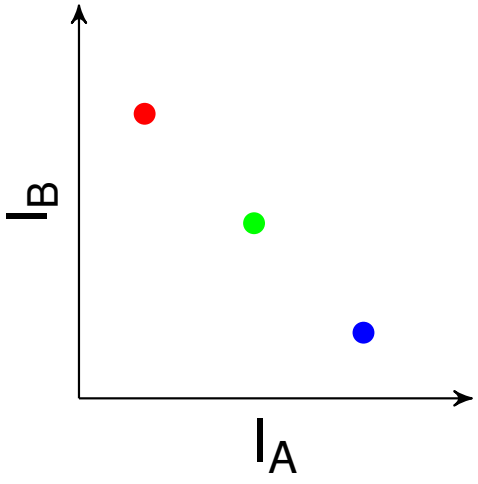


Lipshutz, Nat Genet 1999, 21:20-24

- Intensitätswerte eines Quadranten zusammengefasst
- Verhältnis von Intensität des A-Allels (I_A) zu Intensität des B-Allels (I_B) bestimmt Genotyp:

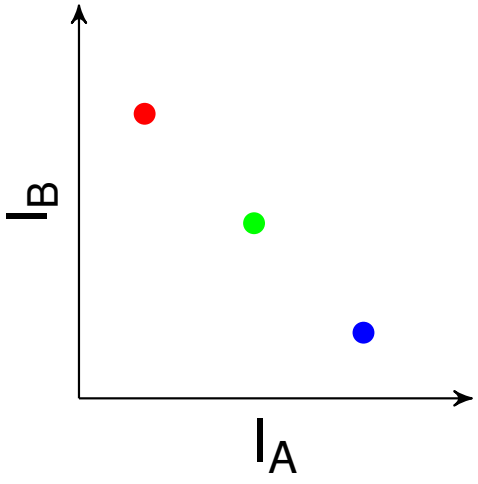
$$I_A \gg I_B \rightarrow A/A$$
$$I_A \approx I_B \rightarrow A/B$$
$$I_A \ll I_B \rightarrow B/B$$

Cluster-Plots

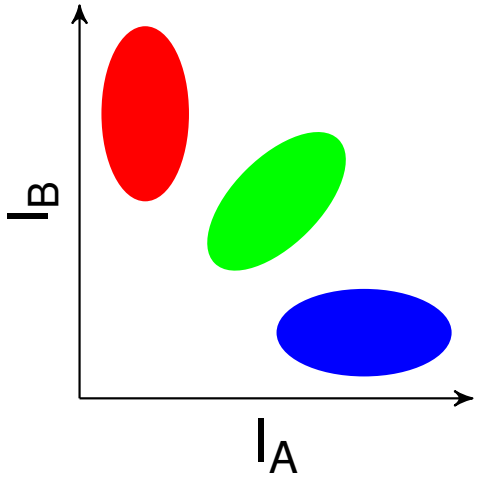


theoretisch

Cluster-Plots

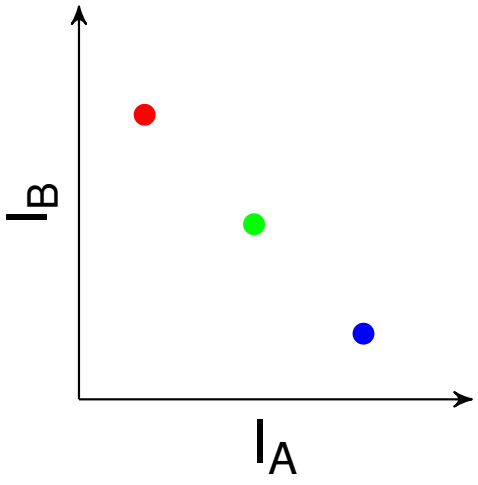


theoretisch

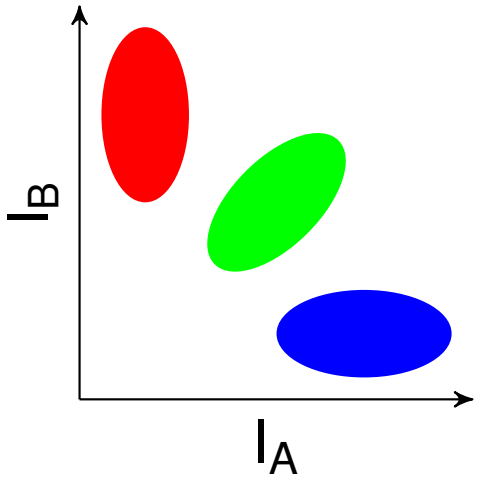


praktisch

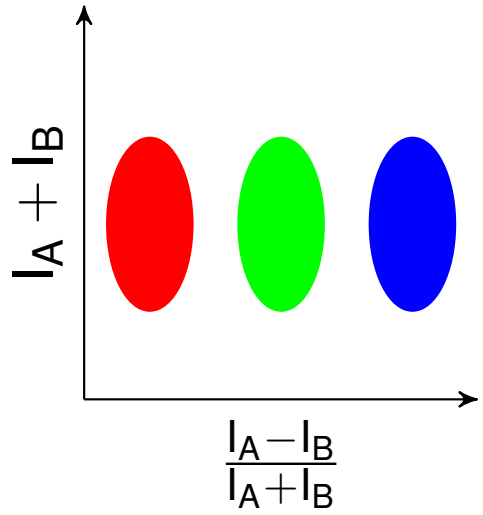
Cluster-Plots



theoretisch

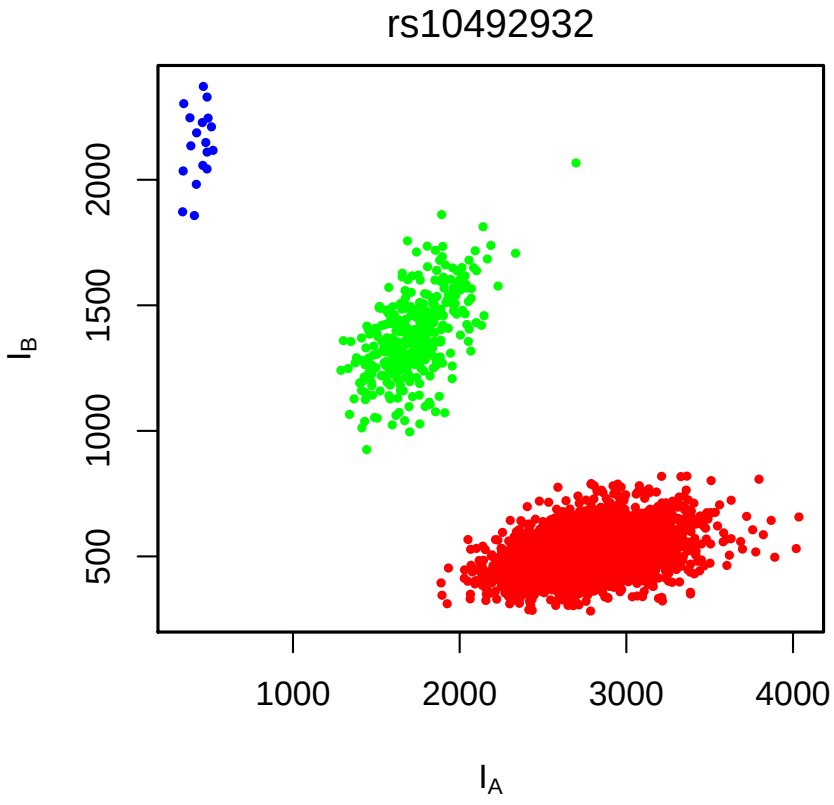


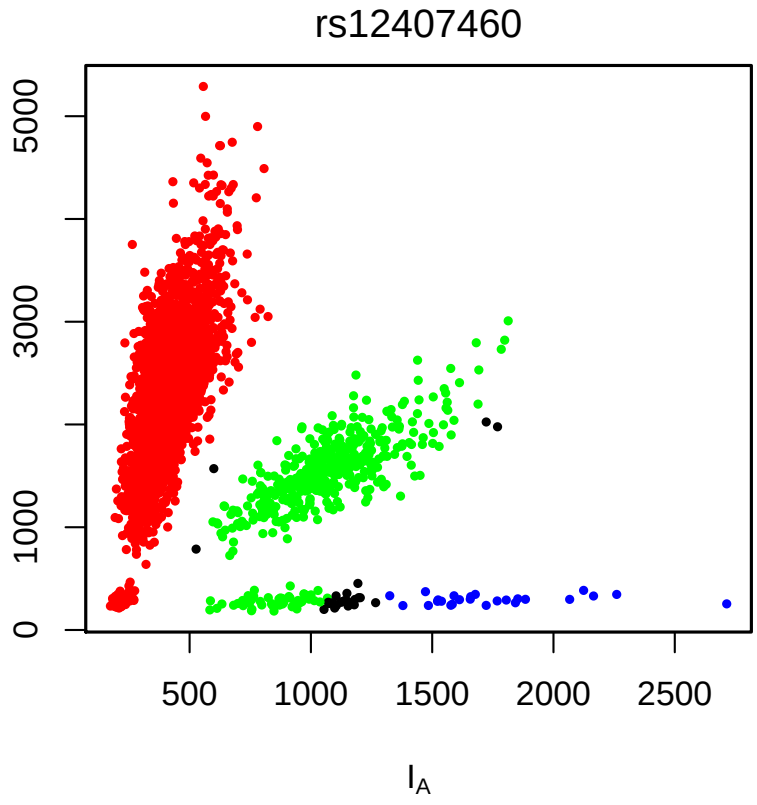
praktisch



Kontrastdarstellung

Cluster-Plots von Realdaten





Identifizierte Algorithmen

Birdseed

BRLMM **B**ayesian **r**obust **l**inear **m**odel with **M**ahalanobis distance classifier

CHIAMO ital. “Ich rufe”

CRLMM **C**orrected **r**obust **l**inear **m**odel with **M**ahalanobis distance classifier

GEL **G**enotype calling using **e**mpirical **l**ikelihood

JAPL lautmalerisch frz. “Ich rufe”

MAMS **M**ulti-**a**rray **m**ulti-**S**NP genotype calling

PLASQ **P**robe-**l**evel **a**llele-**s**pecific **q**uantification procedure

SNiPer-HD **SNiPer** **H**igh **D**ensity

Identifizierte Algorithmen

Birdseed

BRLMM **B**ayesian **r**obust **l**inear **m**odel with **M**ahalanobis distance classifier

CHIAMO ital. “Ich rufe”

CRLMM **C**orrected **r**obust **l**inear **m**odel with **M**ahalanobis distance classifier

GEL **G**enotype calling using **e**mpirical **l**ikelihood

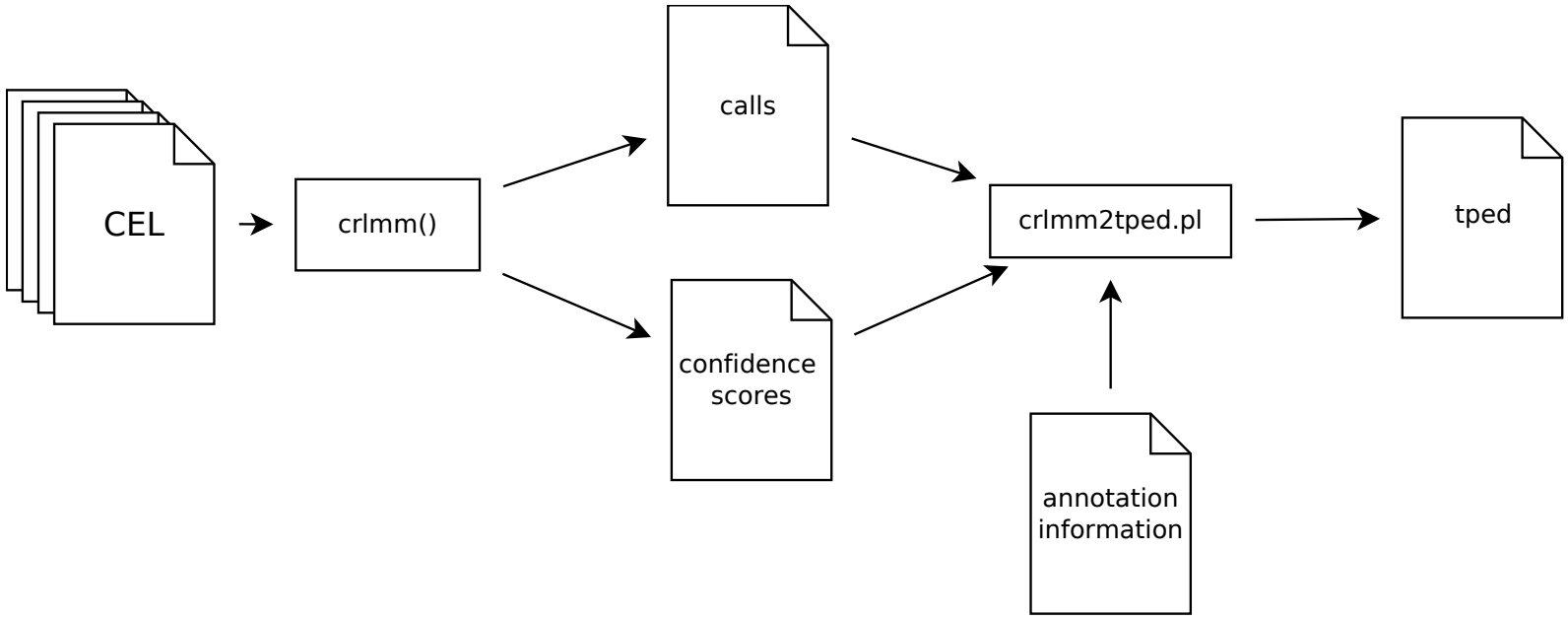
JAPL lautmalerisch frz. “Ich rufe”

MAMS **M**ulti-**a**rray **m**ulti-**S**NP genotype calling

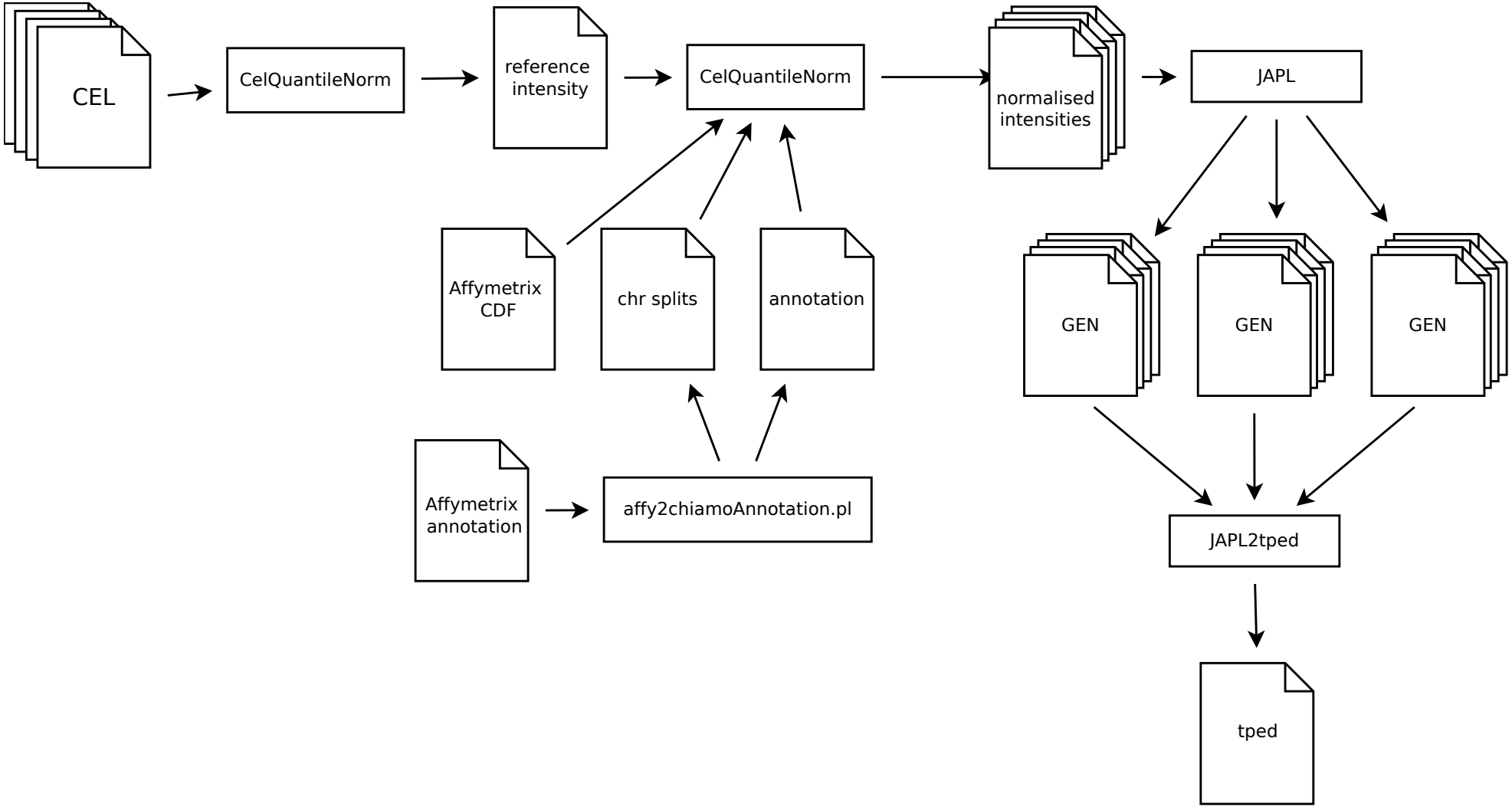
PLASQ **P**robe-**l**evel **a**llele-**s**pecific **q**uantification procedure

SNiPer-HD **SNiPer** **H**igh **D**ensity

Flow chart – CRLMM



Flow chart – JAPL



Confidence Scores – CRLMM

crlmm-calls.txt

	ID1	ID2	ID3
SNP-1	3	2	1
SNP-2	2	2	1

crlmm-conf.txt

	ID1	ID2	ID3
SNP-1	0.999	0.950	0.990
SNP-2	0.990	0.800	0.999

- Annotationsinformation:
  SNP-1: A/C; SNP-2: G/T
- Grenzwert für Confidence Scores: 0.99

tped Genotypen:

	ID1	ID2	ID3
SNP-1	C C	0 0	A A
SNP-2	G T	0 0	G G

Confidence Scores – JAPL

SNP-1.gen

SNP	chipScan	P1	P2	P3
SNP-1	ID1	5.5e-05	0	0.999
SNP-1	ID2	0.05	0.95	0
SNP-1	ID3	0	0.001	0.999

- Annotationsinformation: SNP-1: A/C
- Grenzwert für A-posteriori Wahrscheinlichkeit: 0.99

tped Genotypen:

	ID1	ID2	ID3
SNP-1	C C	0 0	A A

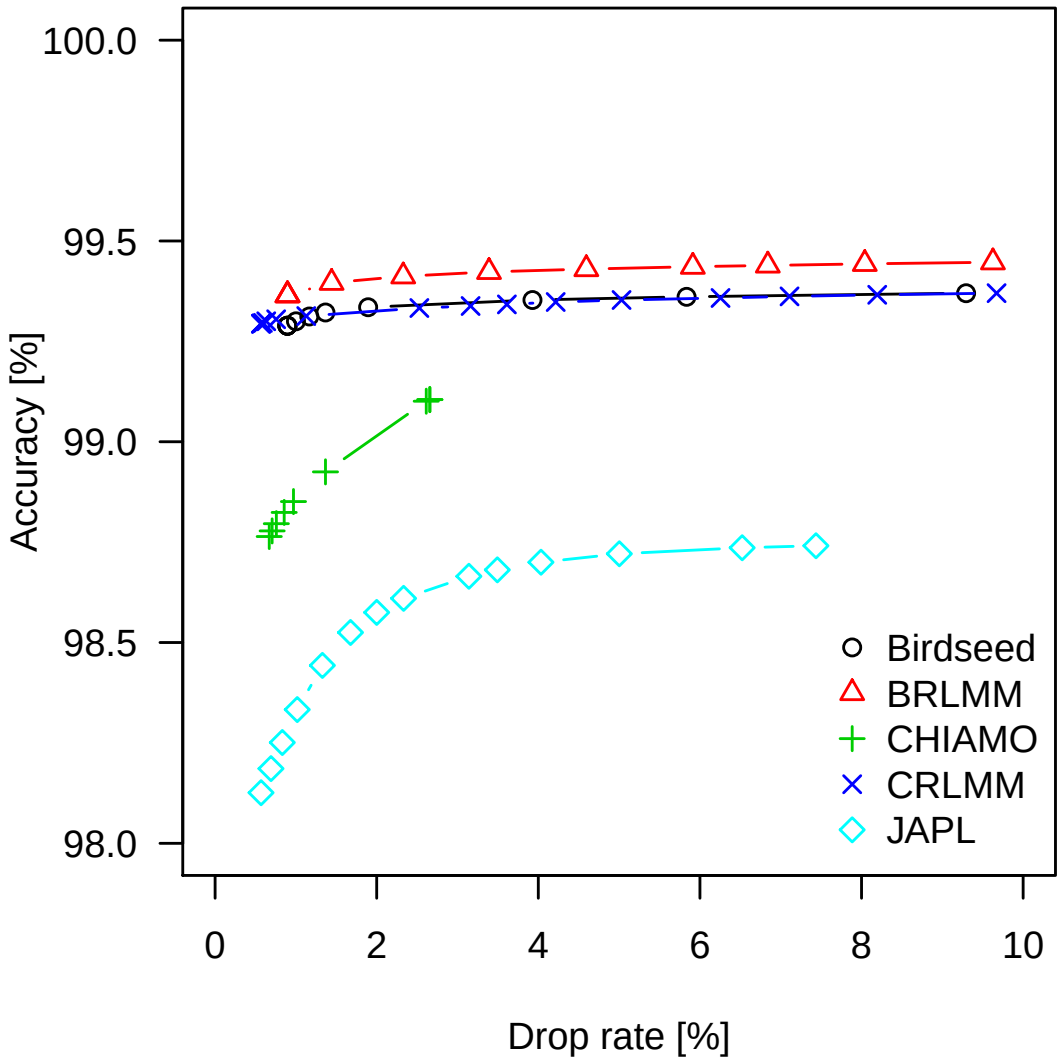
HapMap-Daten

- Genotypdaten und CEL-Dateien für das Affymetrix Human Mapping 500k Array Set und den Genome-Wide Human SNP Array 6.0
- *Goldstandard*
- Daten von 270 Individuen
- nur die der 30 CEU trios verwendet
- CEL-Dateien direkt (`www.hapmap.org`) oder als Bioconductor Paket verfügbar
- Genotypen von der PLINK Website verwendet

Berechnung der Konkordanz

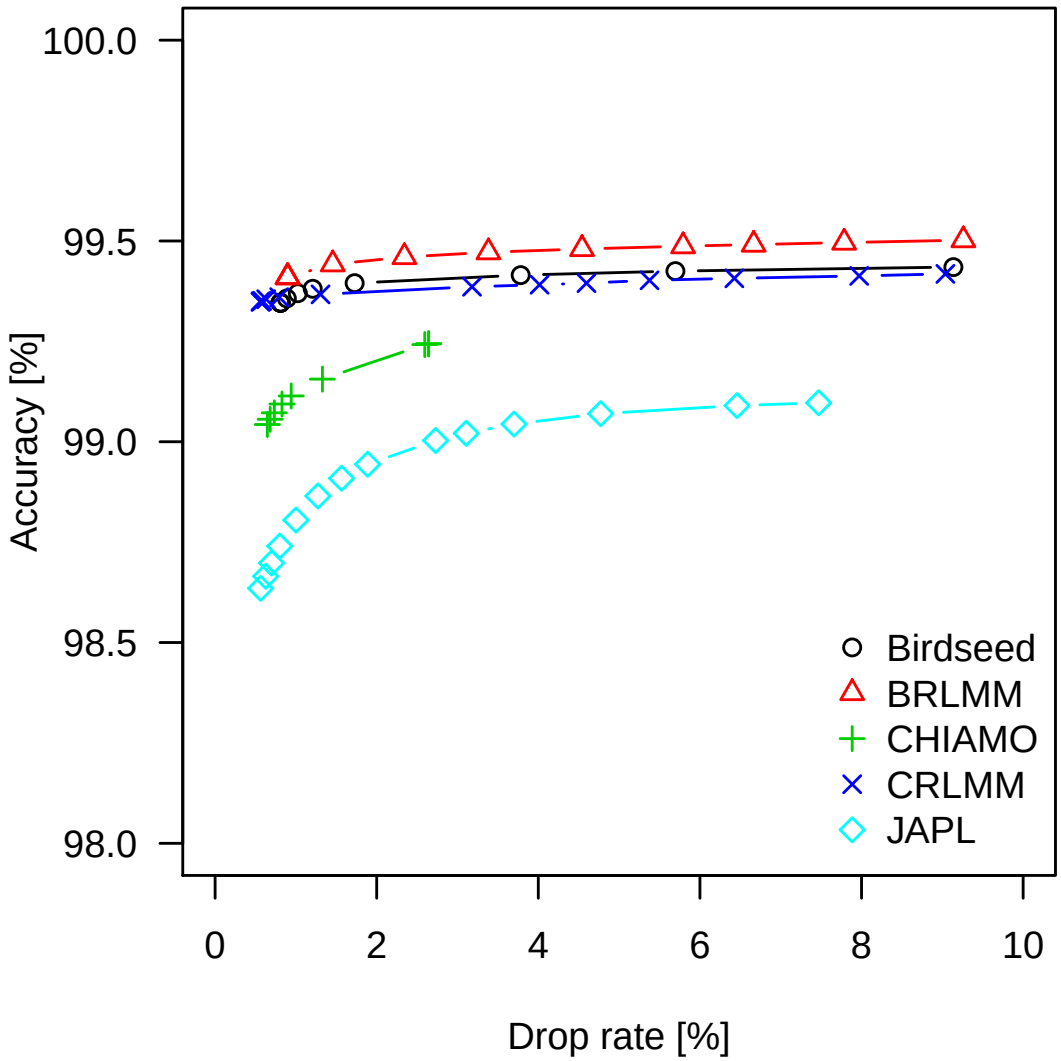
- PLINKs `merge-mode 7` verwendet, d.h. fehlende Genotypen nicht berücksichtigt
- mit HapMap-Daten verglichen
- Bewertung der Güte mittels ADPs:
  - *Accuracy vs. Drop rate Plots* (Lin, Genome Biol 2008, 9:R63)
  - für verschiedene Grenzwerte des Confidence Scores Konkordanz und Anteil fehlender Werte bestimmt

Accuracy vs drop rate plot – alle SNPs



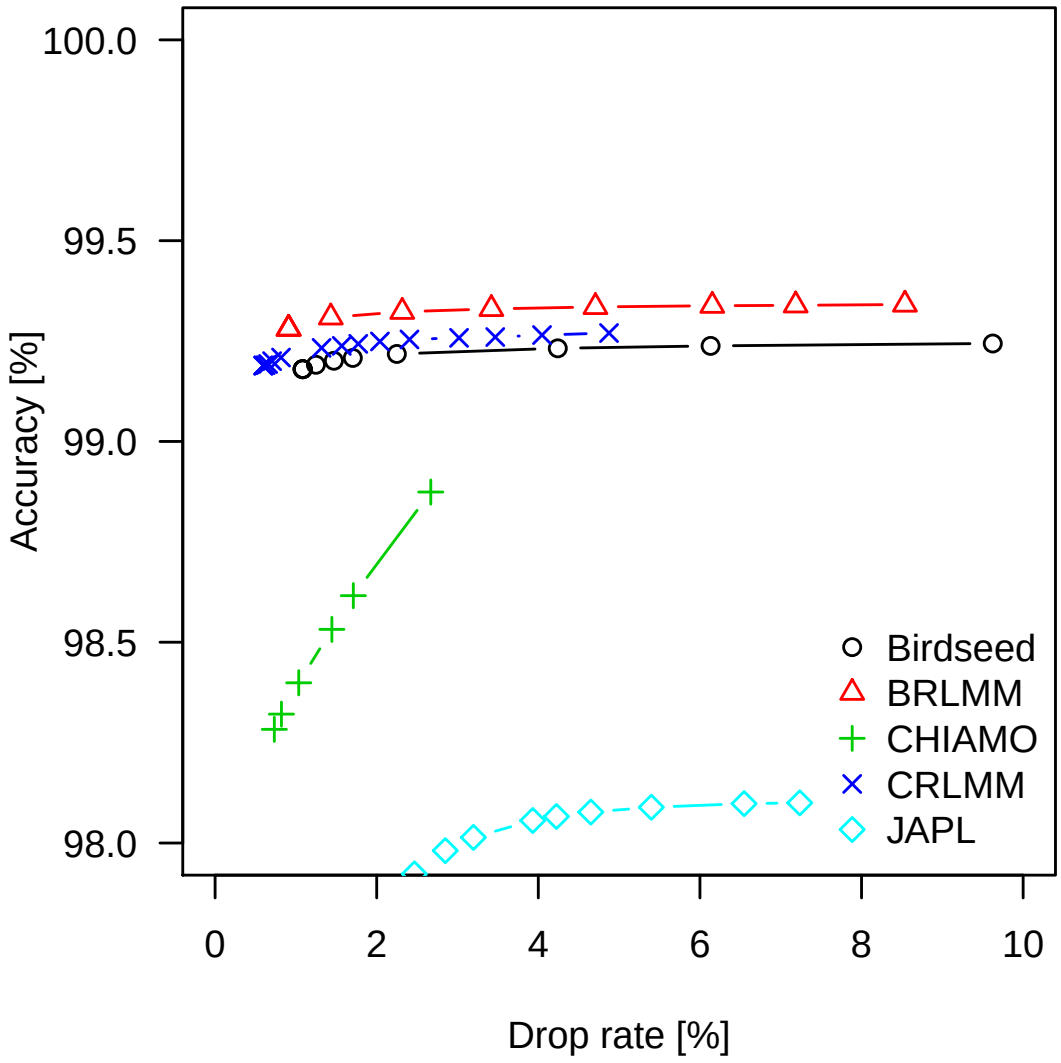
- 482,203 SNPs

ADP – Häufige SNPs



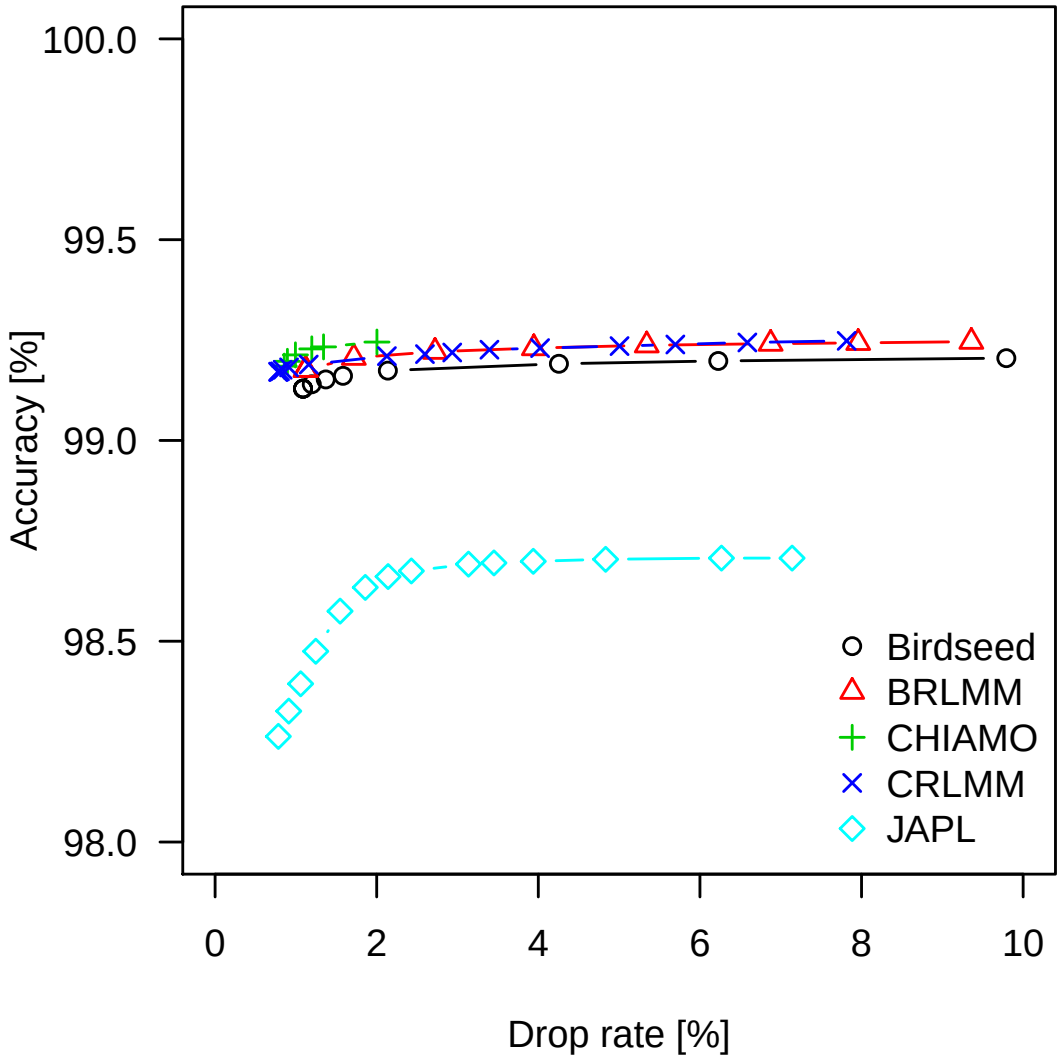
- MAF $\geq$ 10%
- 321,883 SNPs

ADP – Seltene SNPs

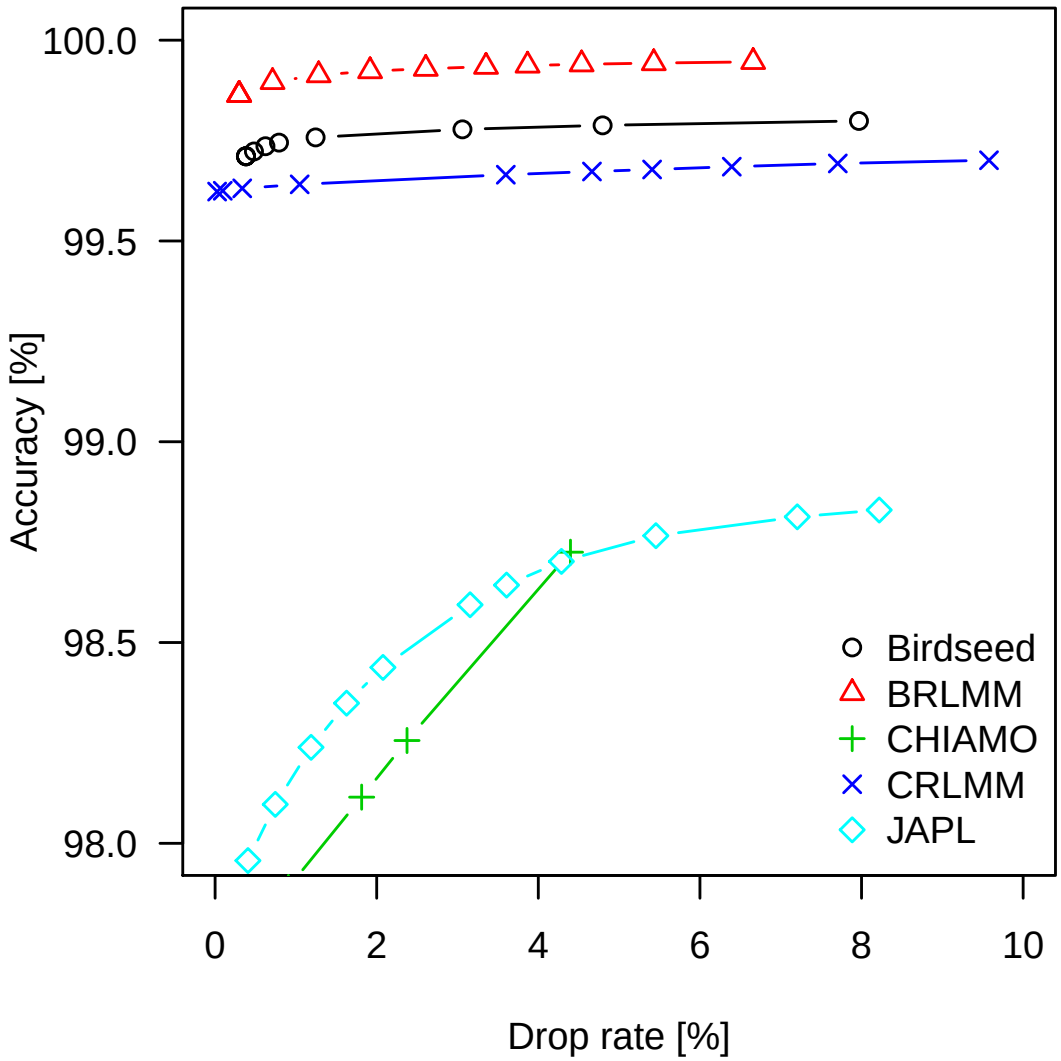


- MAF < 10%
- 160,320 SNPs

ADP – Homozygote Genotypen



ADP – Heterozygote Genotypen



Ergebnisse der Qualitätskontrolle

Kriterien:

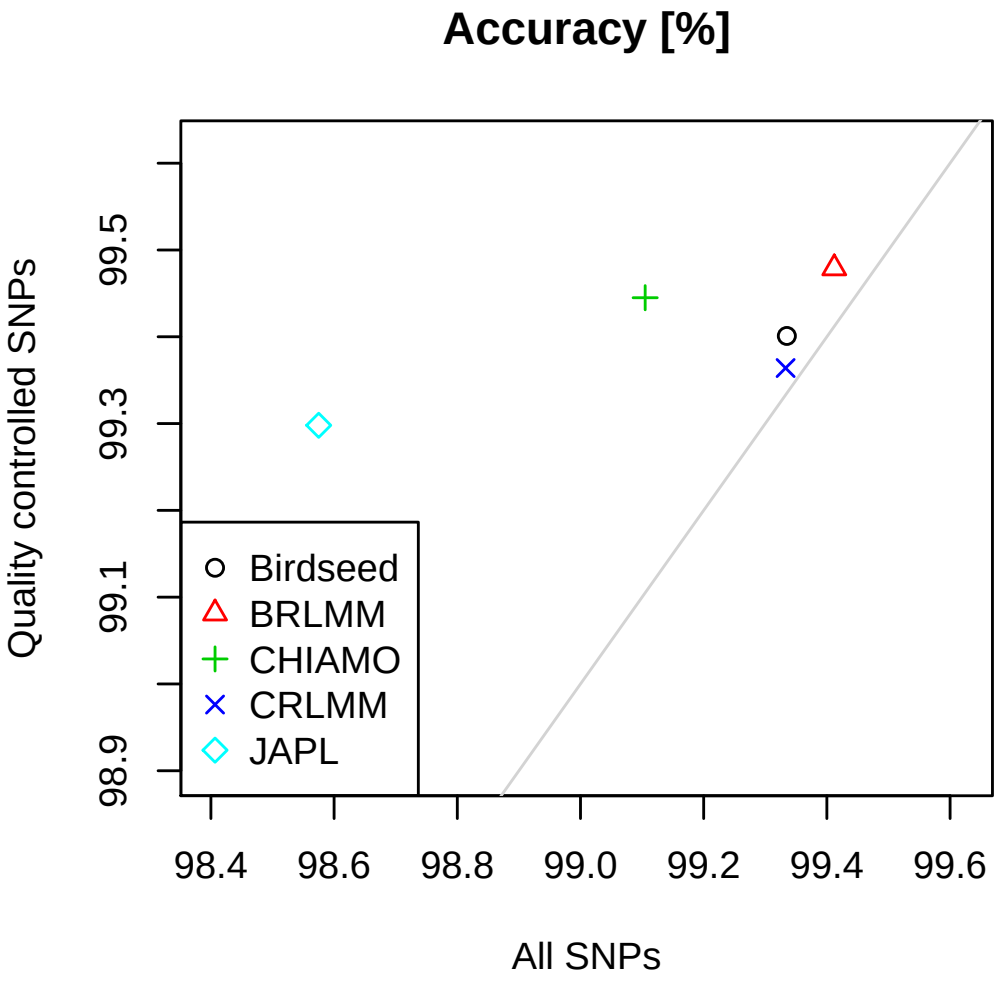
- Anteil fehlender Werte (MiF) < 2%
- Häufigkeit des seltenen Allels (MAF) > 1%
- Abweichungen vom HWE (HWE) p > 0.0001

Anzahl der ausgeschlossenen SNPs:

Algorithmus	MiF	MAF	HWE	Summe
BRLMM	153656	77349	67727	213373
Birdseed	111111	77026	65970	166750
CHIAMO	166683	87709	91074	233785
CRLMM	21794	82027	80411	98013
JAPL	66046	69697	70107	136456

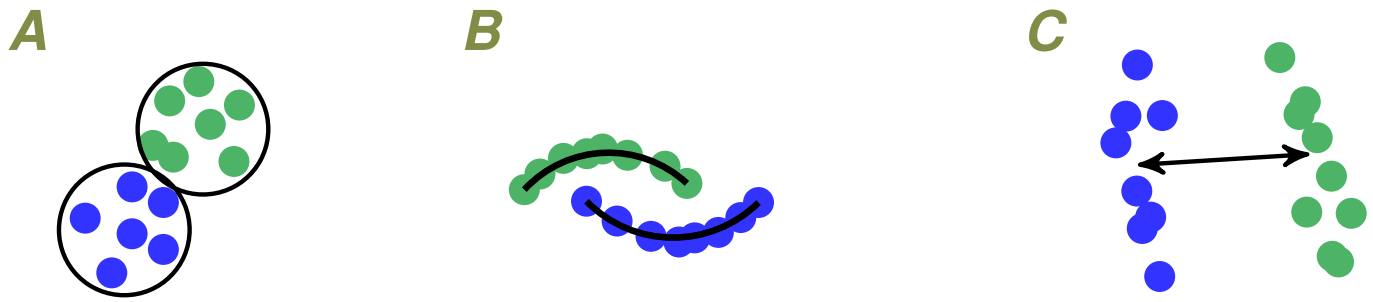
SNPs: 482,203

Konkordanz vor/nach der QC



Clustervaliditätsmaße

- Beurteilung von Cluster-Plots = interne Validität eines Clusterings in Cluster-Analysen
- Kriterien für die Validität:
  - **A** *Kompaktheit*
  - **B** *Verbundenheit*
  - **C** *Trennbarkeit*
  - **D** Kombinationen von A–C



- Alternative Idee: Pertubations- Analyse (Teo, Ann Hum Genet 2008, 72: 368-374)

A: Kompaktheit

- gemessen durch *Cluster-spezifische Intra-Cluster-Varianz* $\mathbb{V}ar(\mathrm{IC}_k)$ und die *Gesamt-Intra-Cluster-Varianz* $\mathbb{V}ar(\mathrm{IC})$

$$\mathbb{V}ar(\mathrm{IC}_k) = \frac{1}{n_k} \sum_{i=1}^{n_k} d_k^2(i, \mu_k)$$

$$\mathbb{V}ar(\mathrm{IC}) = \frac{1}{n} \sum_{k=1}^{3} \sum_{i=1}^{n_k} d_k^2(i, \mu_k)$$

- In der Praxis jedoch *root mean square distance (RMSD)*:

$$\mathrm{RMSD} = \sqrt{\mathbb{V}ar(\mathrm{IC})}$$

B: Verbundenheit – *Connectivity*

- Bestimmung mittels „Nächster Nachbar“-Methoden
- Für Probe i des Clusters k wird der j-te nächste Nachbar $nn_{i(j)}$ bestimmt

$$C_{i,\mathrm{nn}_{i(j)}} = \begin{cases} 0 & \text{falls } i \text{ und } \mathrm{nn}_{i(j)} \text{ denselben Genotyp,} \\ \frac{1}{j} & \text{falls } i \text{ und } \mathrm{nn}_{i(j)} \text{ verschiedene Genotypen} \end{cases}$$

- Kennzahl für die Verbundenheit für die J nächsten Nachbarn:

$$Conn = \sum_{i=1}^{n} \sum_{j=1}^{J} C_{i,\mathrm{nn}_{i(j)}}$$

- *Conn* groß $\Rightarrow$ Cluster zweier Genotypgruppen liegen sehr dicht beieinander

C: Trennbarkeit

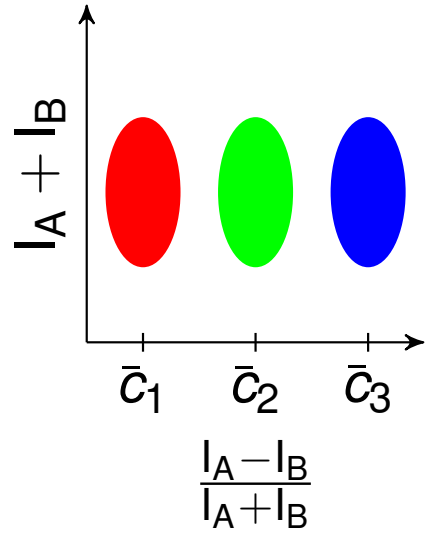
- Vielzahl von Maßen vorgeschlagen
- für Intensitätsdaten minimaler Abstand zwischen Clustern sinnvoll
- alternativ durchschnittlicher Abstand zum Heterozygoten-Cluster
- werden jeweils Abstände der Clusterzentren betrachtet: $d(\mu_k, \mu_{k'})$
- minimaler Inter-Cluster-Abstand (*minD*)

$$minD = \min_{k \neq k'} d(\mu_k, \mu_{k'})$$

- durchschnittlicher Inter-Cluster-Abstand (*meanD*)

$$meanD = \frac{d(\mu_1, \mu_2) + d(\mu_2, \mu_3)}{2}$$

D: Cluster-Separation-Criterion



- nur Cluster-spezifische Kontrast-Mittelwerte $\bar{c}_k$ und -Streuungen berücksichtigt

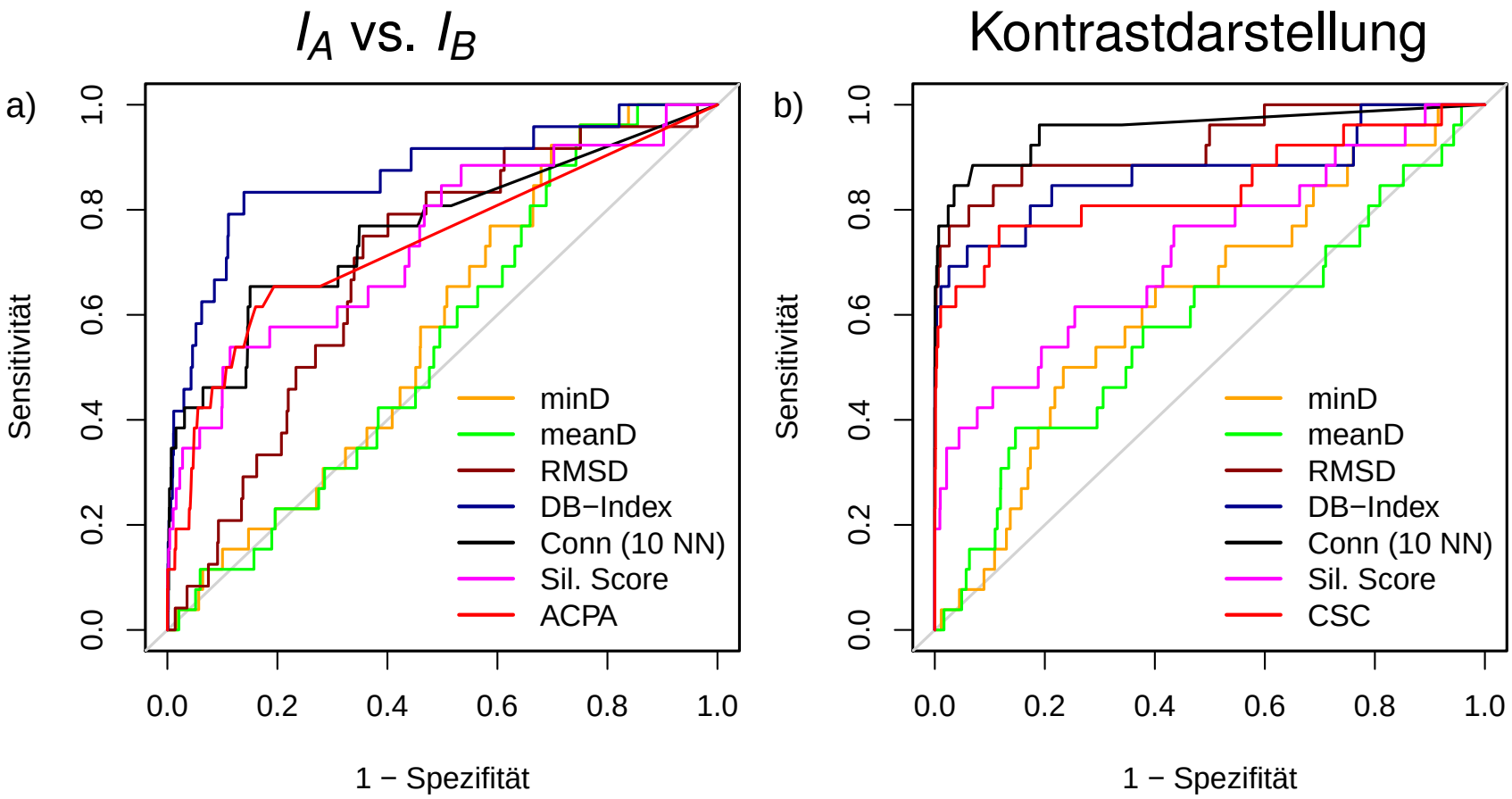
$$CSC = \min \left\{ \frac{\bar{c}_2 - \bar{c}_1}{\sigma_1 + \sigma_2} , \frac{\bar{c}_3 - \bar{c}_2}{\sigma_2 + \sigma_3} \right\} .$$

Praktische Evaluation

- Daten der Gutenberg-Herz-Studie Mainz
  (3194 Individuen, 649.491 qualitätskontrollierte SNPs)
- 5000 SNPs zufällig ausgewählt
- Bewertung der Güte durch zwei erfahrene Beurteiler $\Rightarrow$ Goldstandard
- Vergleich mit ausgewählten Clustermaßen

ROC-Kurven

- Vergleich der Clustermaße mit Goldstandard:



Zusammenfassung

- Unsere Empfehlung: CRLMM
  - Hohe Konkordanz bei geringem Anteil fehlender Werte
  - Einfach zu benutzen (Bioconductor-Pakete)
  - Schnell
- Beschränkungen dieser Analyse:
  - HapMap-Daten für Trainung und Evaluation der Modelle verwendet
  - Stichprobengröße sehr gering, aktuelle GWA-Studien $>$ 2000 Individuen
- Betrachtung der Cluster-Plots notwendig
- Clustervaliditätsmaße ermöglichen objektive Bewertung
- Automatisierung wird als R-Paket implementiert

Danke für die Aufmerksamkeit!

Affymetrix Genotyping Microarrays

Name	Kurzbeschreibung
Human Mapping 10K 2.0 Array	10.204 SNPs, PM+MM
Human Mapping 100K Set	116.204 SNPs, 2 Arrays, PM+MM
Human Mapping 500K Array Set	500.568 SNPs, 2 Arrays, PM+MM
Genome-Wide Human SNP Array 5.0	500.568 SNPs und 420.000 CNV-Sonden, nur PM
Genome-Wide Human SNP Array 6.0	906.600 SNPs und 946.00 CNV-Sonden, nur PM
Axiom Genotyping Solution	567.096 SNPs, komplettes Neudesign der Plattform