BMBF-Projekt
„Qualitätsmanagement für Hochdurchsatz-Genotypisierung“

Subproject TP 2:

(E. Vicedo Jover, B. Müller-Myhsok, Th. Bettecken
Max-Planck-Institut für Psychiatrie)

Info-Veranstaltung
TMF/Charité Berlin
June 21 2010

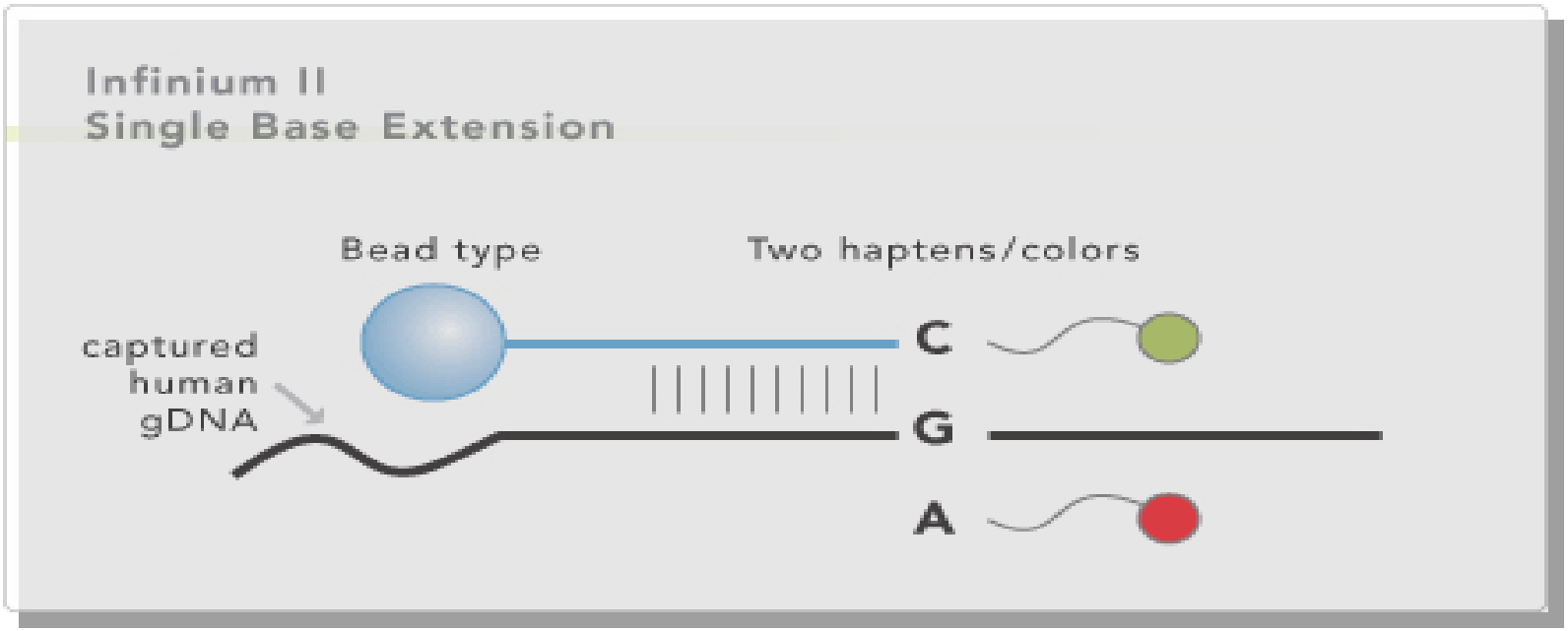
GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

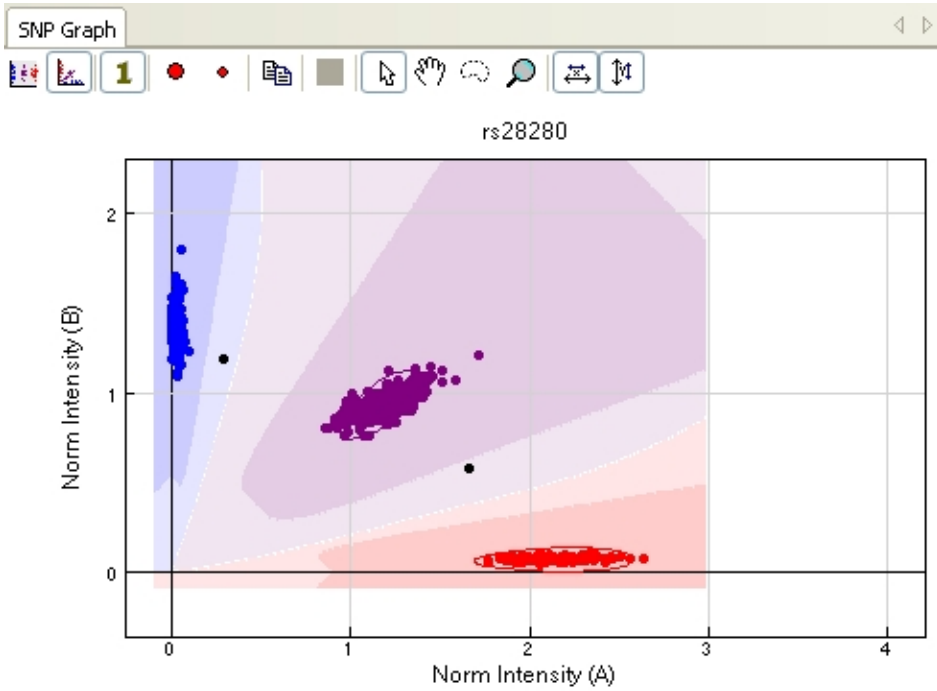
Ziele

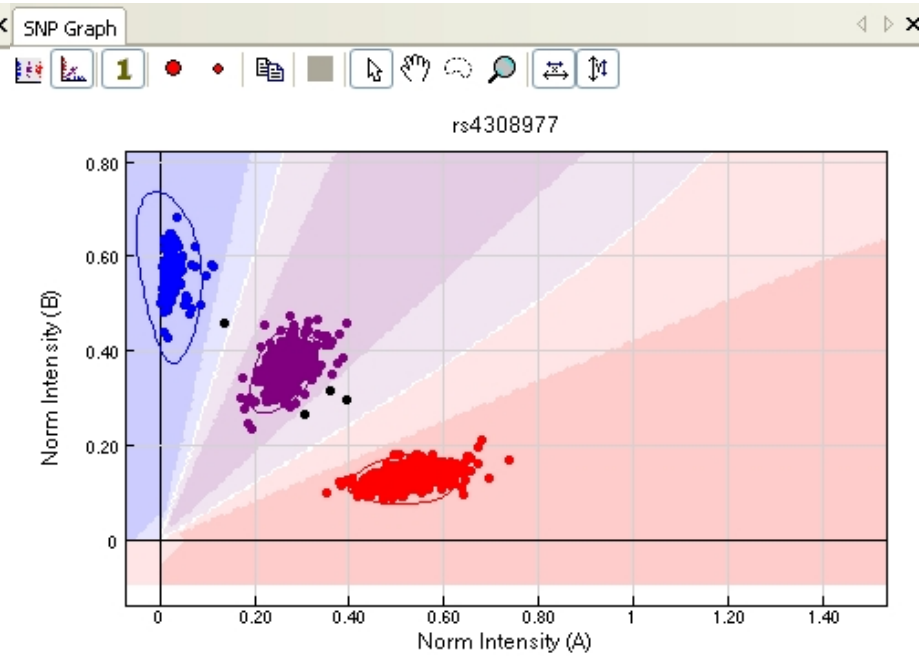
1. Interpretation von Allelen nicht-kanonischer Cluster

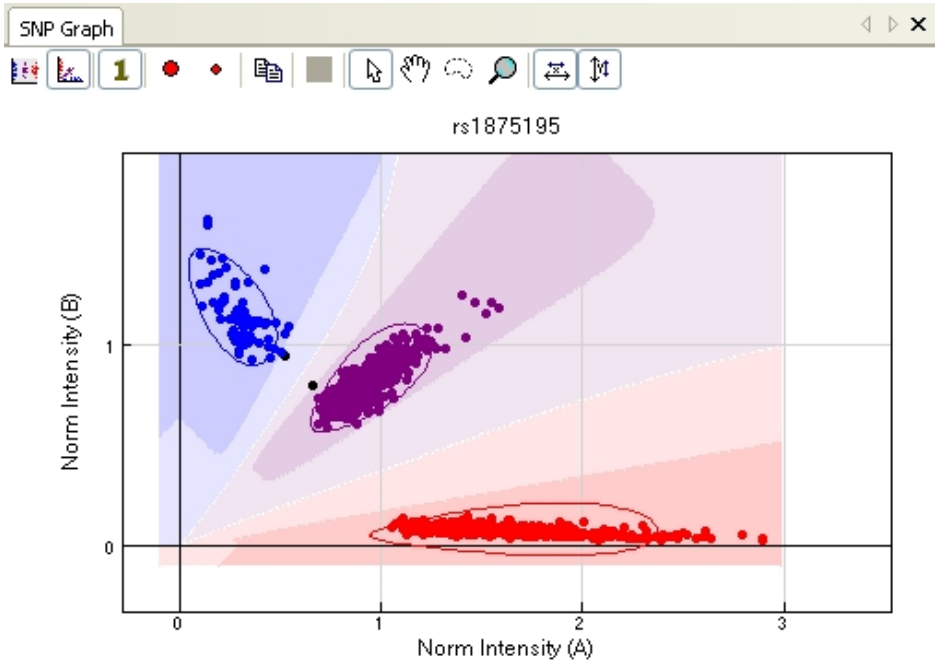
(SNPs in direkter Nachbarschaft, CNVs, SVs, andere)

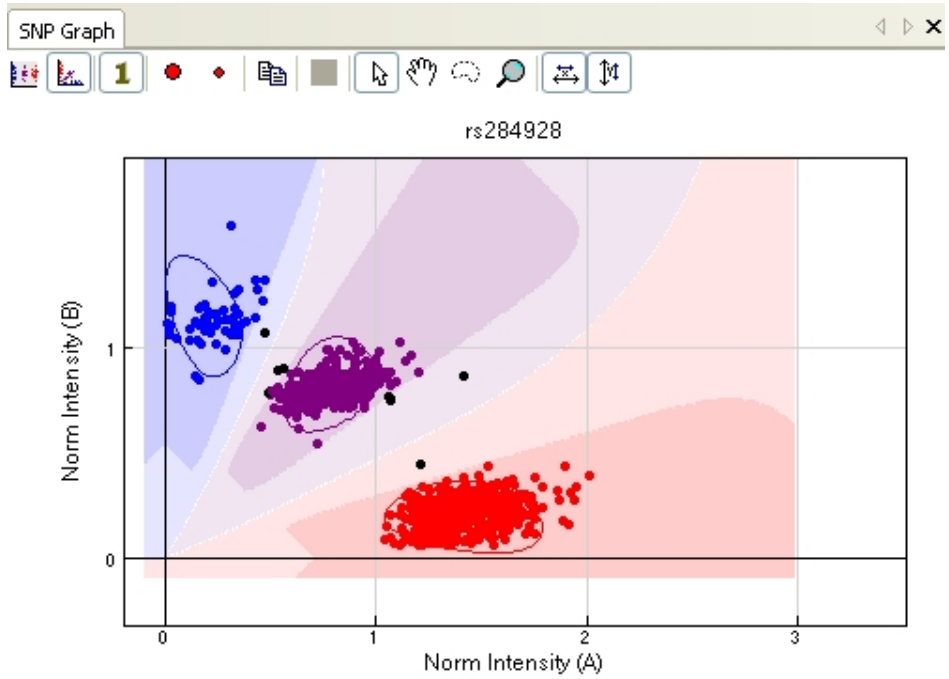
2. Dokumentation von Genotypisierungs Projekten

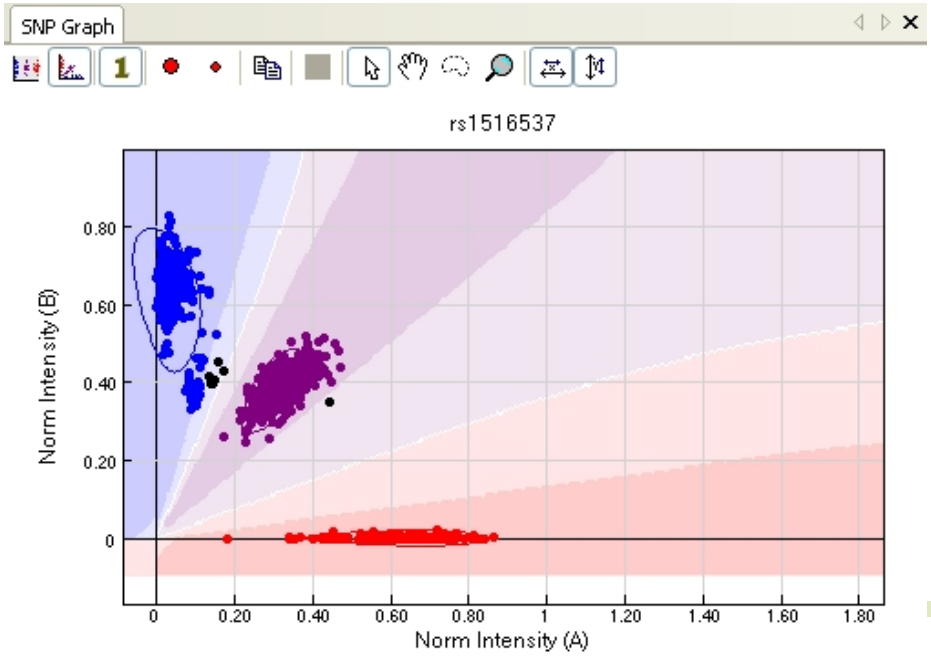

Infinium II
Single Base Extension
Bead type
Two haptens/colors
captured
human
gDNA
C
G
A

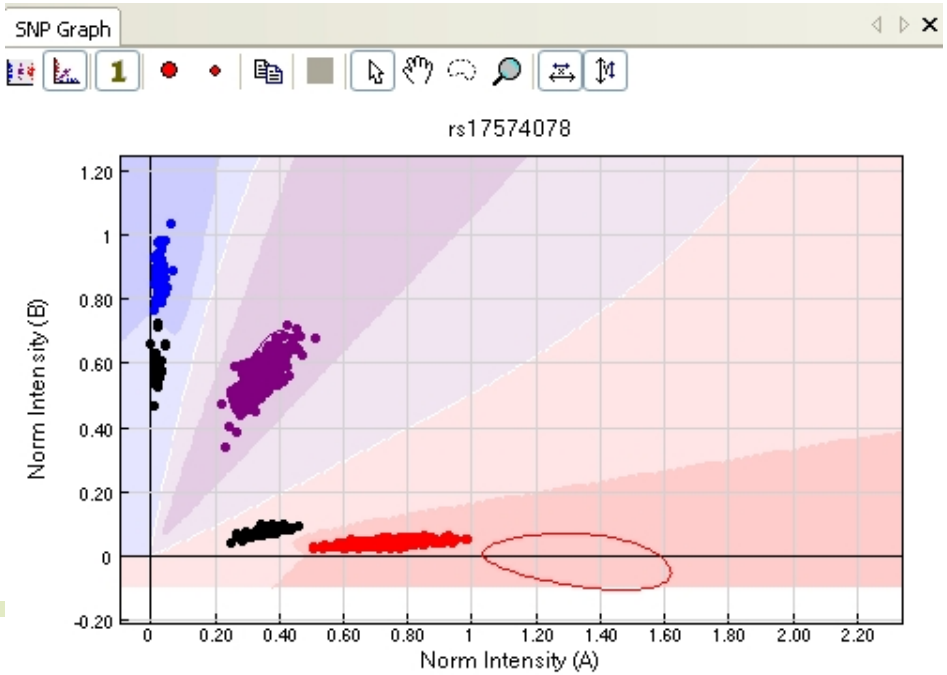

SNP Graph
rs28280
Norm Intensity (B)
Norm Intensity (A)
0
1
2
3
4

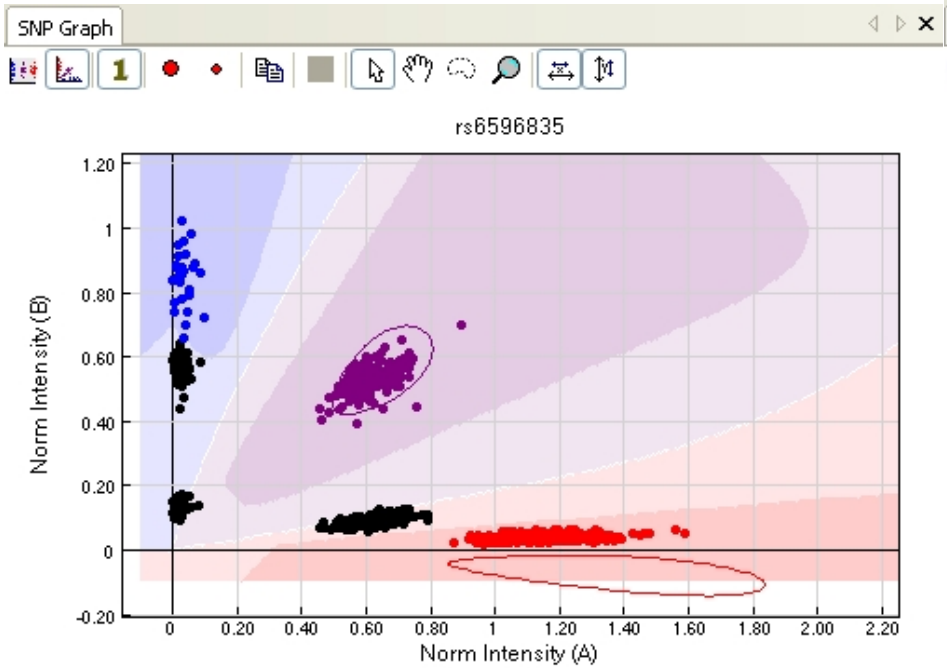

SNP Graph
rs4308977
Norm Intensity (B)
Norm Intensity (A)
0
0.20
0.40
0.60
0.80
1
1.20
1.40

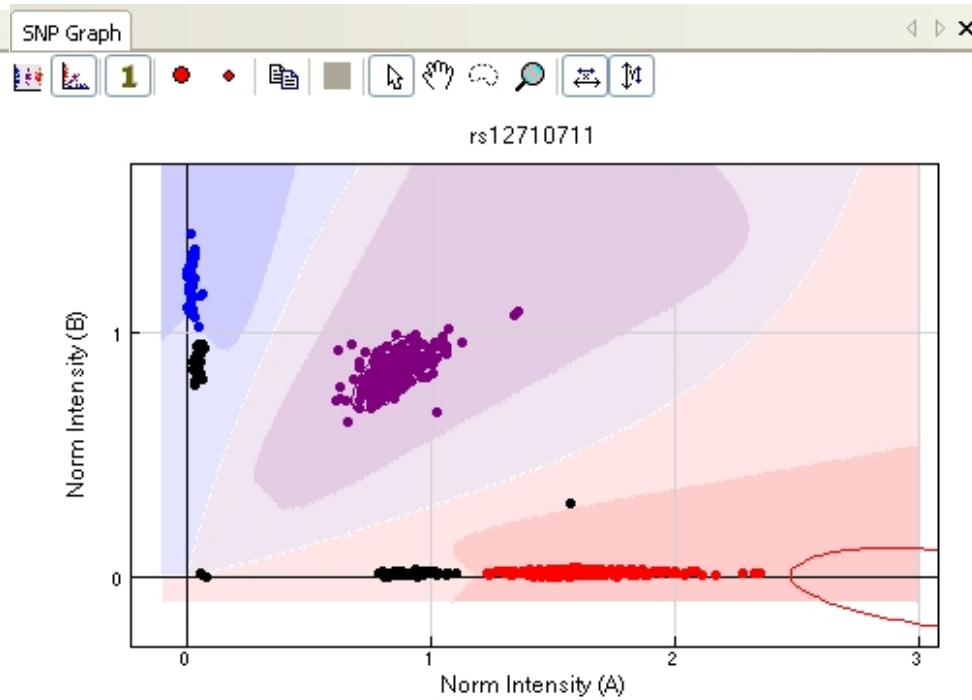

SNP Graph
rs1875195
Norm Intensity (B)
Norm Intensity (A)
0
1
2
3


SNP Graph
rs284928
Norm Intensity (B)
Norm Intensity (A)
0
1
2
3


SNP Graph
rs1516537
Norm Intensity (B)
Norm Intensity (A)


SNP Graph
rs17574078
Norm Intensity (B)
Norm Intensity (A)


SNP Graph
rs6596835
Norm Intensity (B)
Norm Intensity (A)


SNP Graph
rs12710711
Norm Intensity (B)
Norm Intensity (A)

Approach

1. Classification
2. Identification of non canonical clusters
   - 2.1 Manually
   - 2.2 By software
3. Interpretation
   - 3.1 Sequencing of selected samples
     - 3.1.1 Classical sequencing
     - 3.1.2 2nd Gen Sequencing around the SNPs loci
   - 3.2 In silico analyses
     - 3.2.1 Searching for other SNP variants in direct neighbourhood of SNPs with non canonical clusters
     - 3.2.2 Searching for known CNVs and SV variants in direct neighbourhood

Scheduled Deliverables

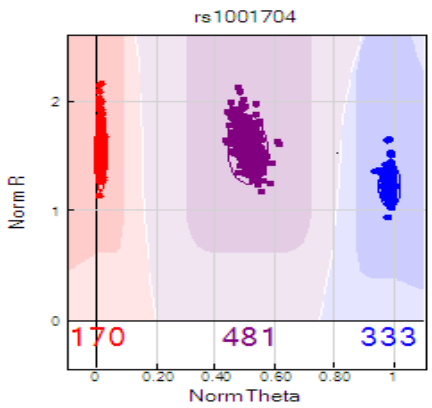
1. Catalogue of canonical and non canonical SNP clusters
2. Classification of SNPs, percentages
3. Script/Software for algorithmic classification of SNPs
4. Catalogue for Interpretation
5. Form for Documentation for Genotyping Projects

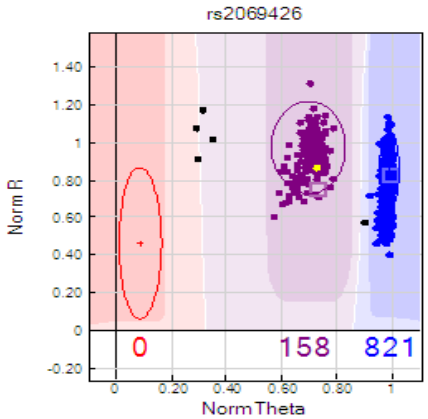
Data used

- **Illumina Genotyping BeadChips Human Hap 550 (“550k”)**
  - Samples:
    - All with a Call Rate > 0.995 => 986 – 48 = **938 Samples**
  - SNPs:
    - All, except: X,Y, XY and mitochondrial SNPs
      i.e. 561.466 – 14.008 = **547.458 SNPs**

Cluster Type Definition

Based on clustering of fluorescence intensity points





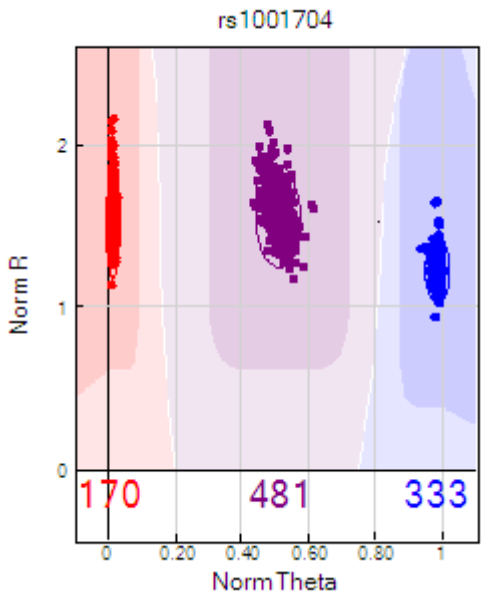
Minimum number of samples for a new cluster = 1 % of total

When no cluster for this SNP, then **MZP** =1

Catalogue of Cluster Types

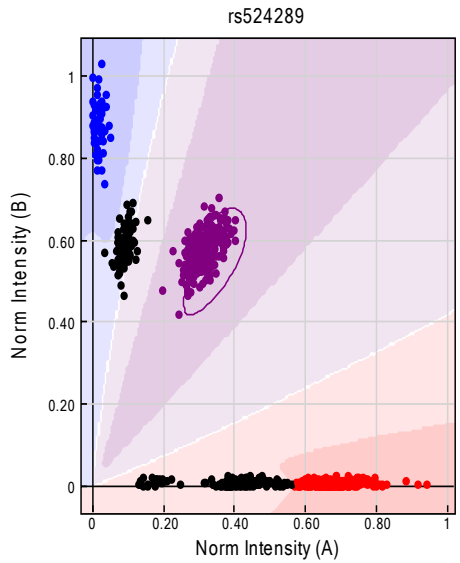
Cluster Type	Description
0	Broad clusters
1	One cluster
2	Two clusters
3	Three clusters
4	Four clusters
5	Five clusters
6	Six clusters
11	Bad separation
99	Others
21	Zero cluster + 1 cluster
22	Zero cluster + 2 clusters
23	Zero cluster + 3 clusters
29	Zero clusters + undefined

Cluster Definitions, Examples



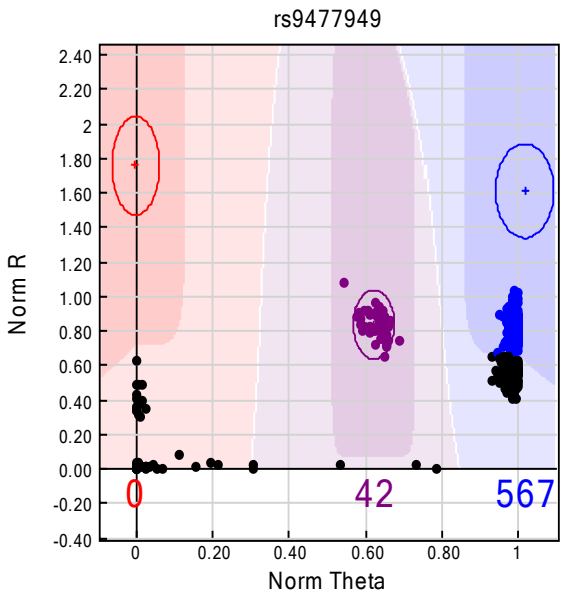
Cluster Type 3

(Canonical Cluster)

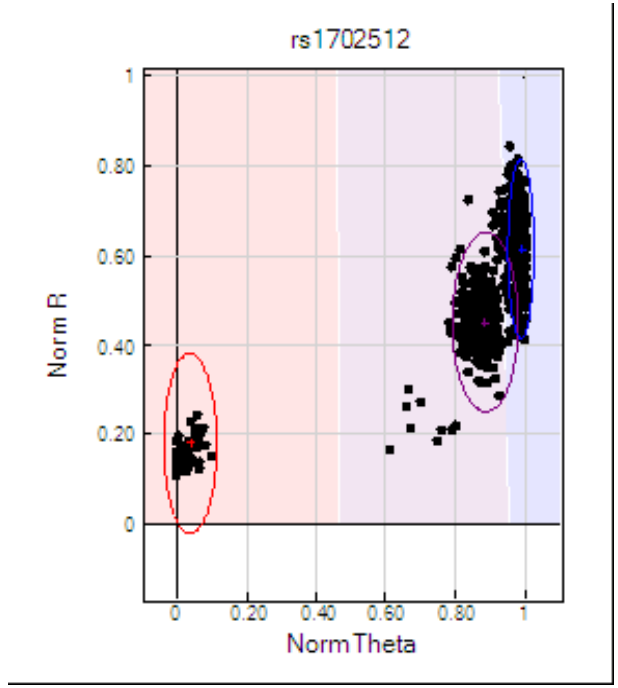


Cluster Type 5

Cluster Definitions, Examples 2



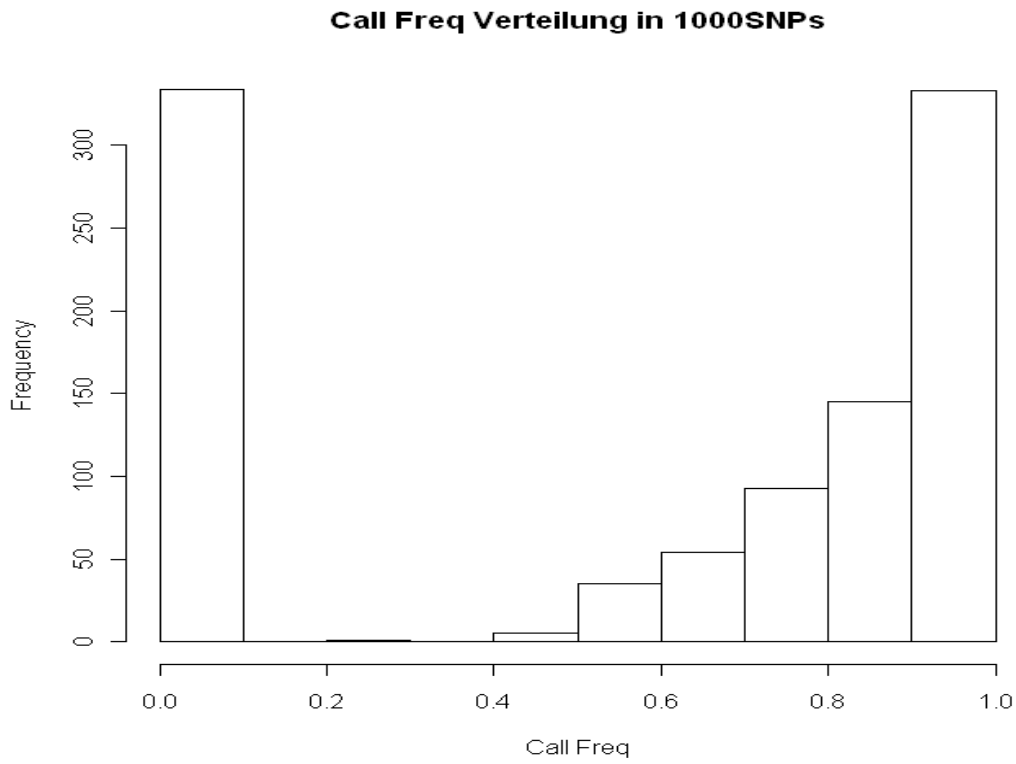
Cluster Type 23
(Zero cluster/Null alleles plus 3)



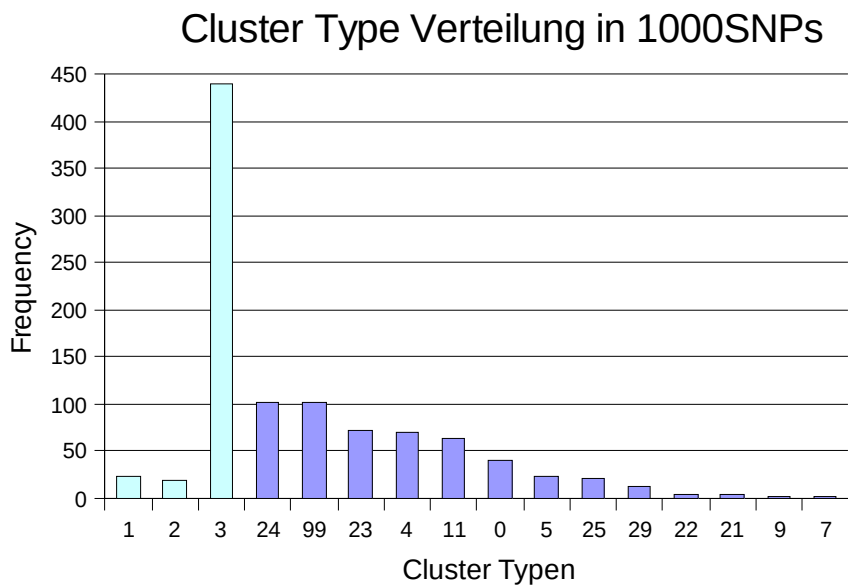
Cluster Type 11

Manual/Visual Screening of Fluorescence Intensities

1000 SNPs witch call frequency 0.0 to 1.0, enriched in call frequencies < 99%



Manual/Visual Screening of Fluorescence Intensities



Cluster Typ	Frequency
1	23
2	18
3	440
24	102
99	101
23	73
4	69
11	63
0	41
5	23
25	21
29	13
22	5
21	4
9	3
7	1

Algorithm “TypeCluster”

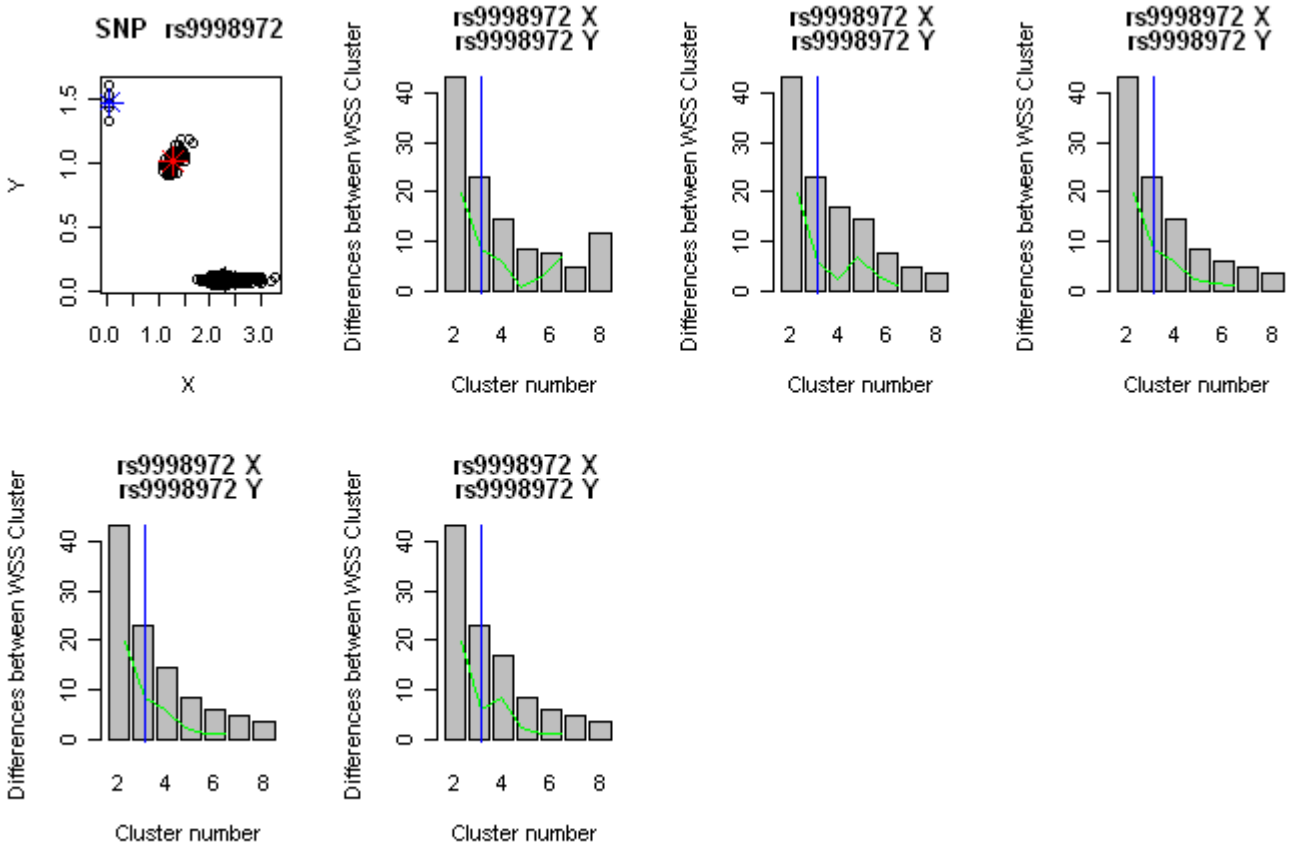
Criteria :

- 1. Minimal Intensities (X, Y) :
  Clusters of Type 21, 22, 23 (with Null Alleles) ?
  Clusters with low intensities

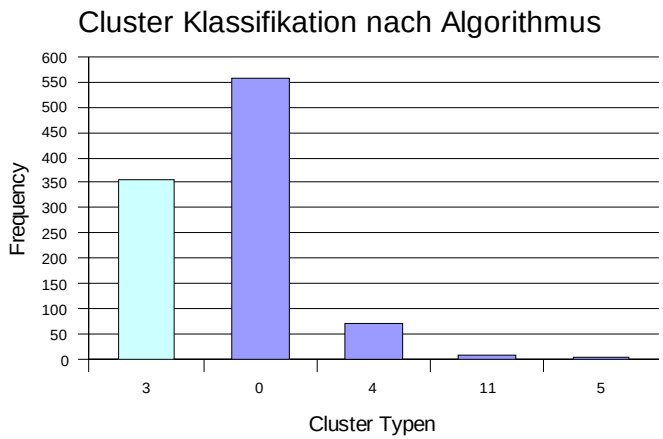
- 2. Cluster Separation (Theta):
  Clusters of types 0 and 11

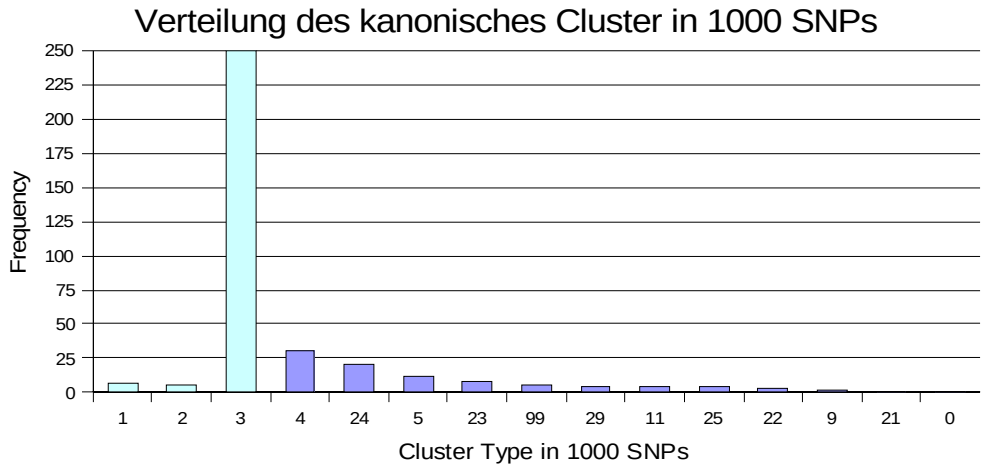
- 3. kmeans Algorithm calculates the number of remaining cluster types (1, 2, 3, 4, 5, 6, 99)

kmeans Algorithm



Results of Algorithmic Classification of Clusters





Cluster Type	Frequency
3	**357**
0	**558**
4	**73**
11	**8**
5	**4**

Cluster Type	Frequency
1	**7**
2	**6**
3	**250**
4	**31**
24	**21**
5	**12**
23	**8**
99	**5**
29	**4**
11	**4**
25	**4**
22	**3**
9	**2**
21	**0**
0	**0**

Approach

1. Classification
2. Identification of non canonical clusters
   - 2.1 Manually
   - 2.2 By software
3. Interpretation
   - 3.1 Sequencing of selected samples
     - 3.1.1 Classical sequencing
     - 3.1.2 2nd Gen Sequencing around the SNPs loci
   - 3.2 In silico analyses
     - 3.2.1 Searching for other SNP variants in direct neighbourhood of SNPs with non canonical clusters
     - 3.2.2 Searching for known CNVs and SV variants in direct neighbourhood

Attempt of NGS Sequencing around SNP loci

Fragment DNA

Ligate adaptors

Hybridize ligation products to old (or new) Illumina chips

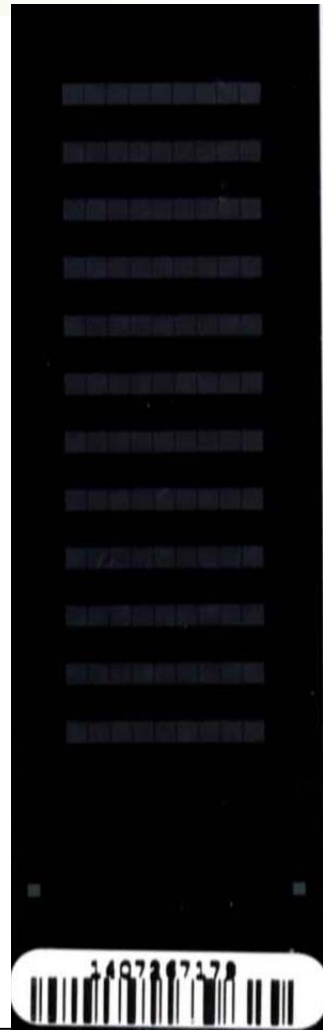
Wash

Collect hybridized DNA by denaturation

PCR Amplify

Submit to 2nd Gen Sequencing

Sequence analysis



Approach

1. Classification
2. Identification of non canonical clusters
   - 2.1 Manually
   - 2.2 By software
3. Interpretation
   - 3.1 Sequencing of selected samples
     - 3.1.1 Classical sequencing
     - 3.1.2 2nd Gen Sequencing around the SNPs loci
   - 3.2 In silico analyses
     - 3.2.1 Searching for other SNP variants in direct neighbourhood of SNPs with non canonical clusters
     - 3.2.2 Searching for known CNVs and SV variants in direct neighbourhood

Dokumentation von Hochdurchsatz-Genotypisierungsdaten

Material-, Daten- und Informationsfluss
(vereinfacht und exemplarisch)