Informationsveranstaltung „Qualitätsmanagement für Hochdurchsatz-Genotypisierung“

Fehlererkennung und Fehlerkorrektur von Hochdurchsatz-Genotypisierungsdaten

21. Juni 2010, Berlin

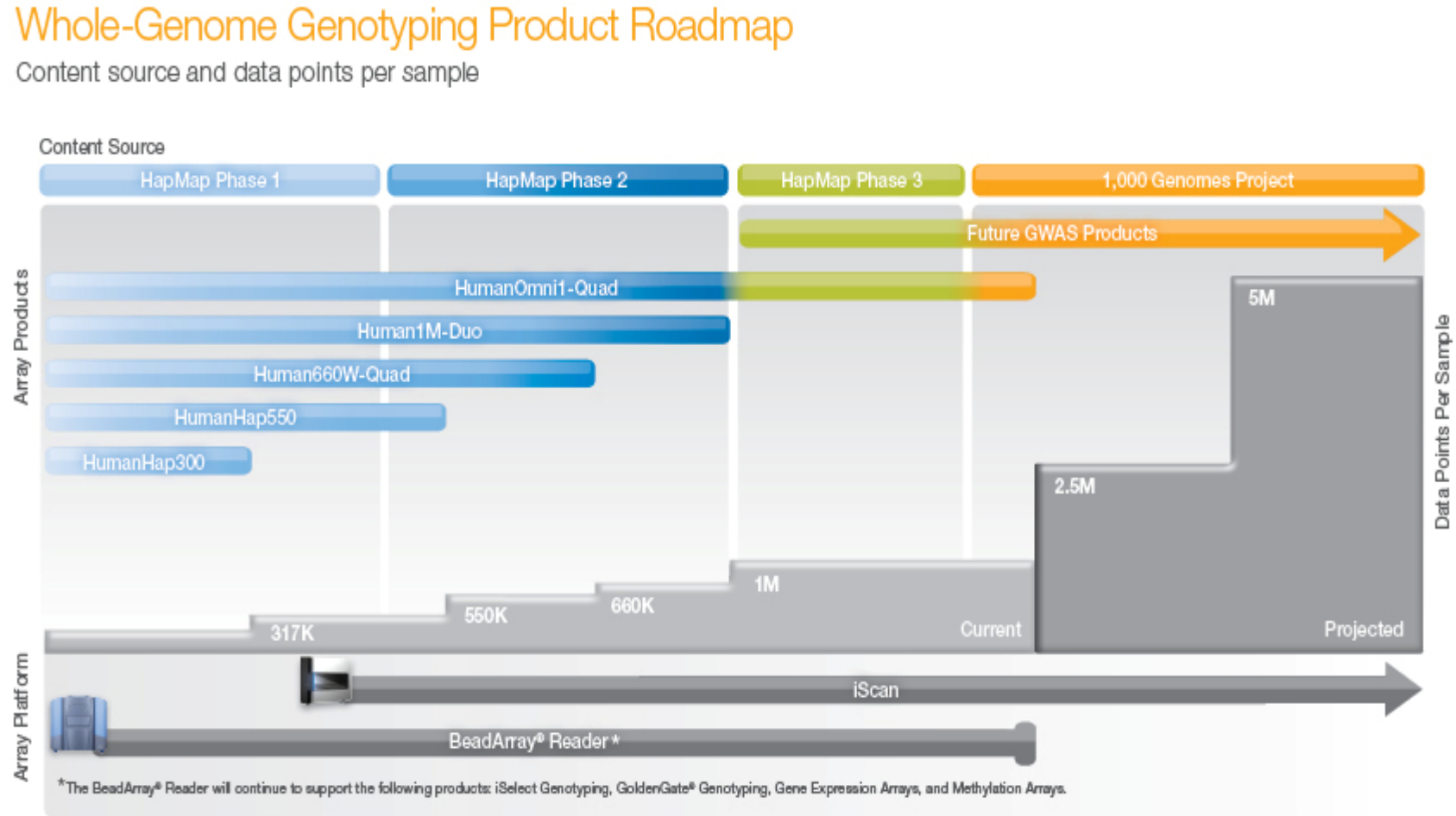
Michael Steffens & Thomas F. Wienker

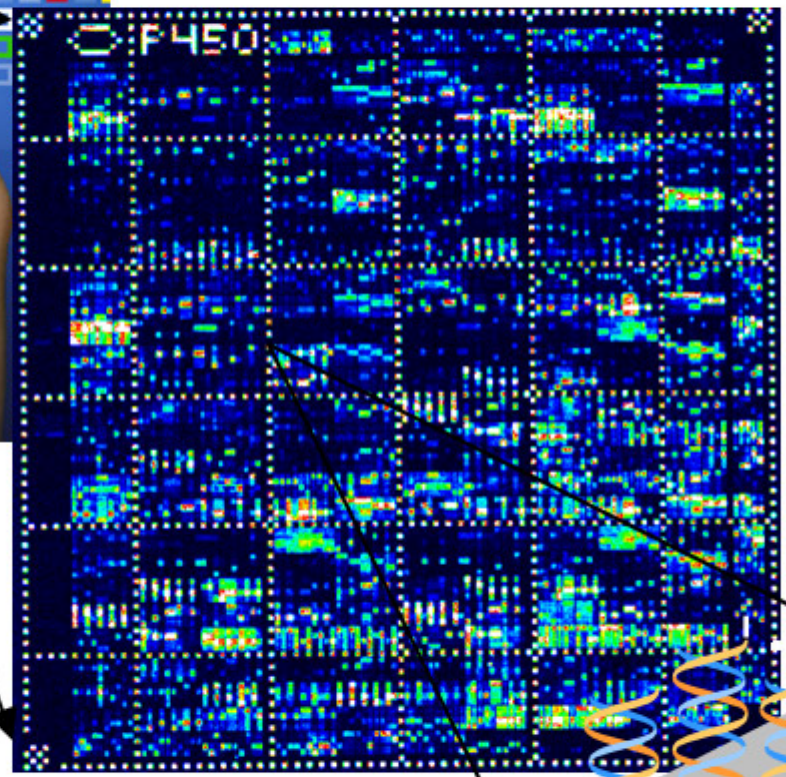
Institut für Medizinische Biometrie,

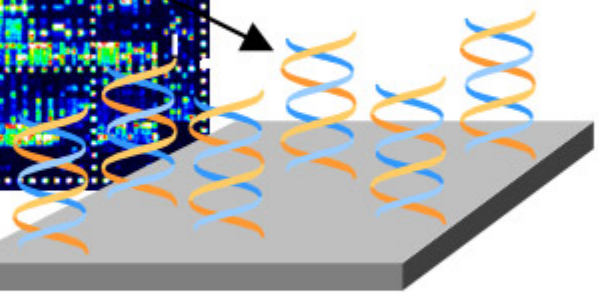
Informatik und Epidemiologie (IMBIE)

Universität Bonn

Illumina "Omni" Roadmap







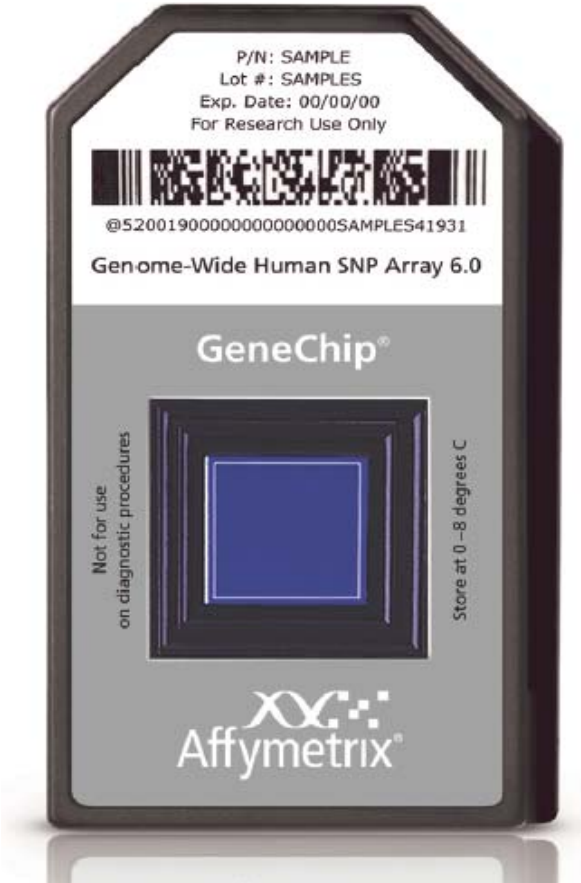
Powered by Affymetrix

The Way Ahead™

Each 20 μm^2 cell on the array can contain 10^7 DNA fragments, or "probes"

Data Sheet

Genome-Wide Human SNP Array 6.0

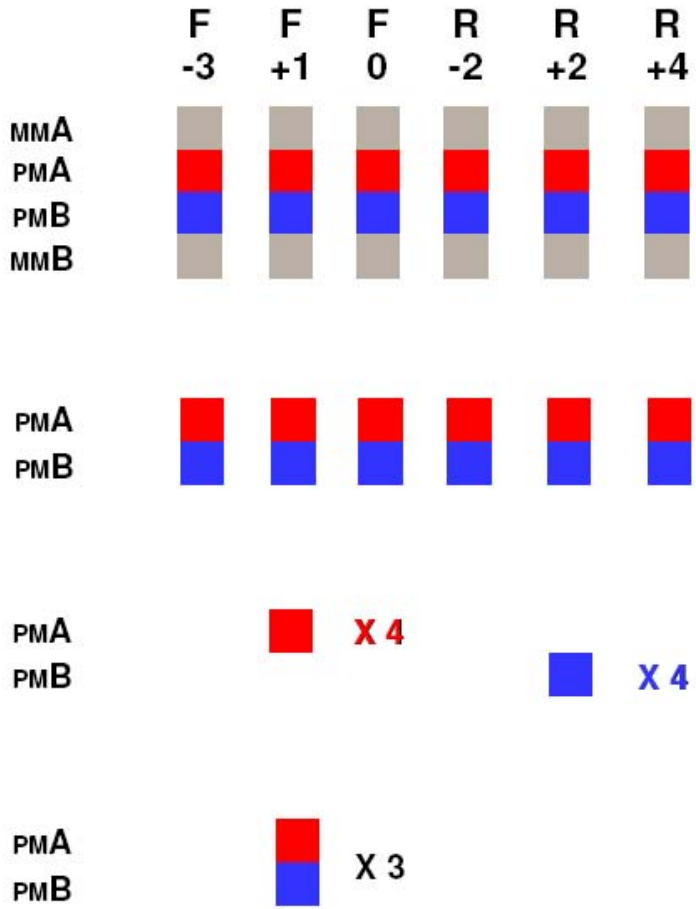


Product information	
Number of SNPs on the array	906,600
Number of non-polymorphic probes for copy number detection	946,000
Number of arrays	1
DNA required	500 ng
Average Birdseed call rate (0.1)	>99 percent*
Average minor allele frequency (MAF)	19.6% in HapMap Caucasians 18.2% in HapMap Asians 20.6% in HapMap Africans
Average heterozygosity	26.7% in HapMap Caucasians 24.6% in HapMap Asians 28.5% in HapMap Africans
PCR primers	1 per sample
Instrumentation	GeneChip® Scanner 3000 7G with AutoLoader
Throughput	>40 million genotypes per day, per scanner with three GeneChip® Fluidics Station 450s

* Average scan from four independent studies

AFFYMETRIX
The Way Ahead.™

Evolution in SNP Probes



# probes	Array
24	500K
12	Python "A"
8	5.0
6	6.0

courtesy Affymetrix, Inc. (U.S.), and Peter Nuernberg, CCG Cologne

Affymetrix Genome–Wide Human SNP Array 5.0

using cdf-file GenomeWideSNP_5.r2.cdf

grid: X [0..2165] x Y [0..2165] = 2166 x 2166 = 4.691.556 dots

probeset-id	probeset-type	features	loci	redundancy
SNP_A-<nnn nnn>	genotyping	3.526.352	440.794	x8
AFFX-SNP_<nnn nnn>	genotyping	81.504 59.064 22.440	3.022 2.461 561	 x24 x40
AFFX-5Q-<nnn>	expression	240	4	x60
AFFX-barcode<A...T>	expression	320	20	x16
CN_<nnn nnn>	unknown	417.269	417.269	x1
SUM		4.025.685	861.109	

feature recruitment [features/dots] = 4.025.685 / 4.61.556 = 85.8 %

Affymetrix Genome–Wide Human SNP Array 5.0

using 4 different cdf-files: GenomeWideSNP_5.xxx.rr.cdf

grid: X [0..2165] x Y [0..2165] = 2166 x 2166 = 4.691.556 dots

probeset-id	probeset-type	GW_SNP_5.cdf loci / features	GW_SNP_5.Full.cdf loci / features	GW_SNP_5.r2.cdf loci / features	GW_SNP_5.Full.r2.cdf loci / features
SNP_A-<nnn nnn>	genotyping	440.794 / 3.526.352	500.568 / 4.004.544	440.794 / 3.526.352	500.568 / 4.004.544
AFFX-SNP_<nnn nnn>	genotyping	3.022 / 81.504	3.022 / 81.504	3.022 / 81.504	3.022 / 81.504
AFFX-5Q-<nnn>	expression	4 / 240	4 / 240	4 / 240	4 / 240
AFFX-barcode<A..T>	expression	20 / 320	20 / 320	20 / 320	20 / 320
AFFX-RandomGC<nn>	expression		23 / 8.463		23 / 8.463
AFFX-AdditionalGC<nn>	expression		22 / 1.721		22 / 1.721
CN_<nnn nnn>	unknown		340.742 / 419.288	417.269 / 417.269	417.269 / 417.269
SUM		443.840 / 3.608.416	844.401 / 4.516.080	861.109 / 4.025.685	920.928 / 4.514.061
feature recruitment		76.9 %	96.3 %	85.8 %	96.2 %

Affymetrix Genome–Wide Human SNP Array 6.0

using cdf-file GenomeWideSNP_6.cdf

grid: X [0..2679] x Y [0..2571] = 2680 x 2572 = 6.892.960 dots

probeset-id	probeset-type	features	loci	redundancy
SNP_A–<nnnn>	GenoType	5.660.710 4.776.270 884.440	906.600 796.045 110.555	 x6 x8
AFFX–SNP_<nnnn>	GenoType	81.504 59.064 22.440	3.022 2.461 561	 x24 x40
AFFX–5Q–<nnn>	Expression	240	4	x60
AFR_<nnn>[_NP;_SB]	Expression	9.752	198 x 3 = 594	x30, x15; x2,x1
RandomGC<nn>	Expression	8.463	23	x9…x489
CN_<nnn nnn>	Copynumber	945.826	945.826	x1
SUM		6.706.495	1.856.069	

feature recruitment [features/dots] = 97.3 %

Affymetrix Genome–Wide Human SNP Array 6.0

intensities extracted from CEL-file using apt-cel-extract.exe (with cdf-file)

	probe_id	x	y	probe_type	probeset_id	probeset_type	block	a52053~1.CEL
0000001:	probe_id	x	y	probe_type	probeset_id	probeset_type	block	a52053~1.CEL
0000002:	3466655	1414	1293	pm	AFFX-5Q-123	Expression	0	892
0000003:	3466656	1415	1293	mm	AFFX-5Q-123	Expression	0	211
0000004:	3463975	1414	1292	pm	AFFX-5Q-123	Expression	0	640
0000005:	3463976	1415	1292	mm	AFFX-5Q-123	Expression	0	172
0000006:	3461295	1414	1291	pm	AFFX-5Q-123	Expression	0	718
0000007:	3461296	1415	1291	mm	AFFX-5Q-123	Expression	0	195
0000008:	3458615	1414	1290	pm	AFFX-5Q-123	Expression	0	965
0000009:	3458616	1415	1290	mm	AFFX-5Q-123	Expression	0	149
0000010:	3455935	1414	1289	pm	AFFX-5Q-123	Expression	0	898
...								
6706487:	3247473	1992	1211	pm	CN_943505	Copynumber	0	460
6706488:	4316216	1415	1610	pm	CN_943502	Copynumber	0	1143
6706489:	3348335	1014	1249	pm	CN_943503	Copynumber	0	829
6706490:	5423012	1371	2023	pm	CN_943504	Copynumber	0	1482
6706491:	3522442	921	1314	pm	CN_943511	Copynumber	0	614
6706492:	6245764	1363	2330	pm	CN_943506	Copynumber	0	301
6706493:	656599	2678	244	pm	CN_943507	Copynumber	0	868
6706494:	2308895	1414	861	pm	CN_943508	Copynumber	0	778
6706495:	3208252	291	1197	pm	CN_943512	Copynumber	0	745
6706496:	1818309	1268	678	pm	CN_954736	Copynumber	0	3434
EOF								

```
00000001:  #num_probe_sets  1856069
00000002:
00000003:  #probe_set_name  AFFX-5Q-123
00000004:  #probe_set_type  expression
00000005:  ...

13385617:  ...
13385618:  #probe_set_name  SNP_A-8475161
13385619:  #probe_set_type  genotyping
13385620:  #num_groups      2
13385621:
13385622:  #group_index     0
13385623:  #group_name      A
13385624:  #direction       Sense
13385625:  #num_cells       3
13385626:  #cell_header     x     y
13385627:                   1017  2037
13385628:                   2025  1920
13385629:                   801    691
13385630:
13385631:  #group_index     1
13385632:  #group_name      C
13385633:  #direction       Sense
13385634:  #num_cells       3
13385635:  #cell_header     x     y
13385636:                   1016  2037
13385637:                   2024  1920
13385638:                    800   691
13385639:
13385640:  #probe_set_name  SNP_A-847516
13385641:  #probe_set_type  genotyping
13385642:  ...

30757884:  ...
30757885:  #probe_set_name  CN_954736
30757886:  #probe_set_type  unknown
30757887:  #num_groups      1
30757888:
30757889:  #group_index     0
30757890:  #group_name      CN_954736
30757891:  #direction       Antisense
30757892:  #num_cells       1
30757893:  #cell_header     x     y     expos  pbase  tbase  atom
30757894:                   1268   678  599    t      a      1
30757895:
EOF
```

```
>>      apt-cdf-export  -c GenomeWideSNP_6.cdf
```

```
>>      apt-cel-extract.exe  -o intensities.txt  *.CEL
```

```
00000001:  probe_id  x     y     probe_type  probeset_id     probeset_type  block  a52053~1.CEL
00000002:  3466655   1414  1293  pm          AFFX-5Q-123     Expression     0          892
00000003:  ...

03761102:  ...
03761103:  5460178   1017  2037  pm          SNP_A-8475161   GenoType       0          466
03761104:  5147626   2025  1920  pm          SNP_A-8475161   GenoType       0          529
03761105:  1852682    801   691  pm          SNP_A-8475161   GenoType       0          413

03761106:  5460177   1016  2037  pm          SNP_A-8475161   GenoType       1         1183
03761107:  5147625   2024  1920  pm          SNP_A-8475161   GenoType       1         1244
03761108:  1852681    800   691  pm          SNP_A-8475161   GenoType       1         1409
03761108:  ...

06706495:  ...
06706496:  1818309   1268   678  pm          CN_954736       Copynumber     0         3434
EOF
```

Data cleaning at different levels

Low level data:

- feature data from hybridization probes
- fluorescence signal intensities ($\mathbb{R}^k$, k = 6 to 8)

Intermediate level data:

- normalized signal intensities = cluster coordinates ($\mathbb{R}^2$)
- cluster assignments = genotype calls

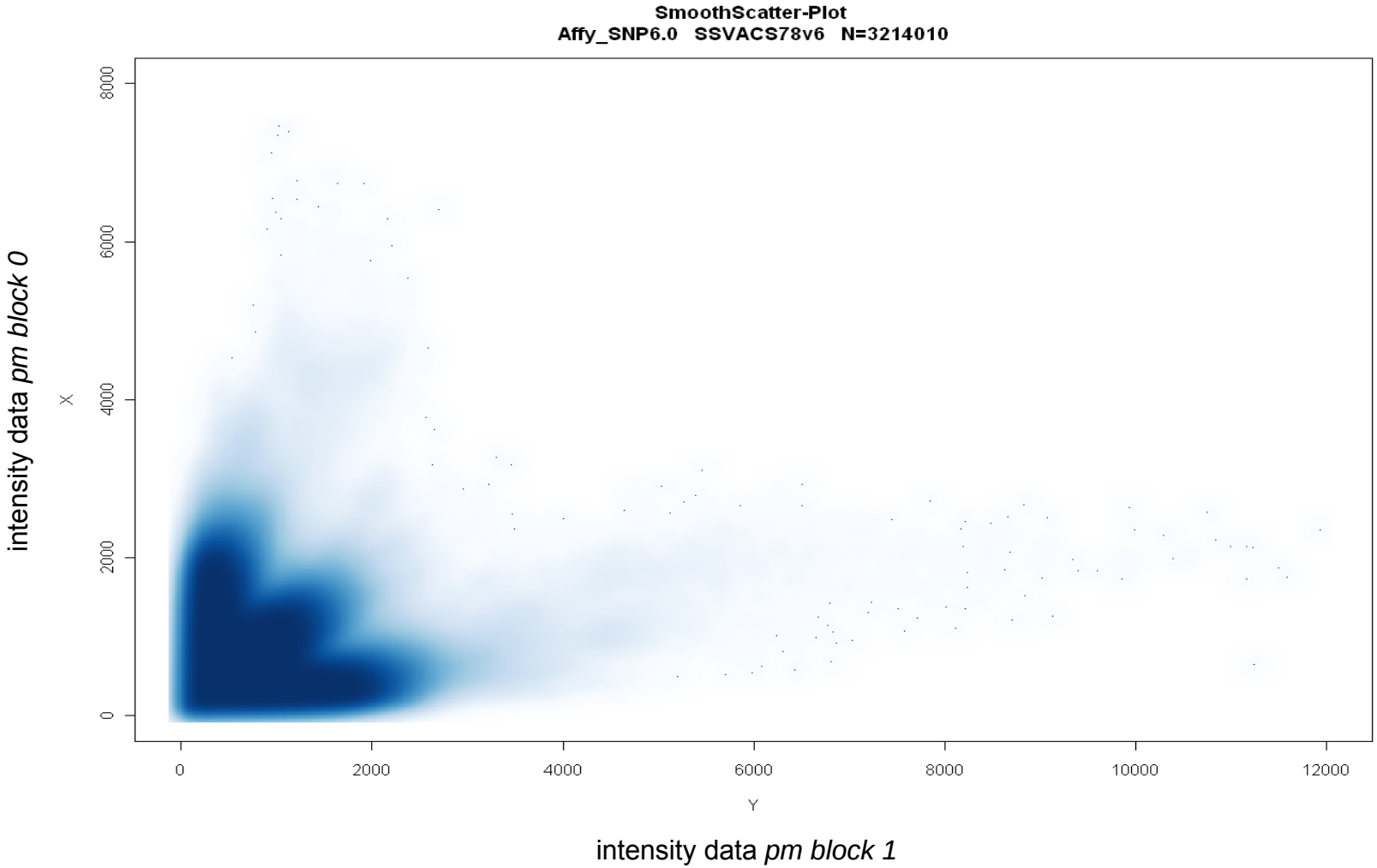
High level data:

- single SNP stats (HWE, HET-rate, IBS, etc.)
- multilocus stats (LD-based, improbable haplotypes, etc.)

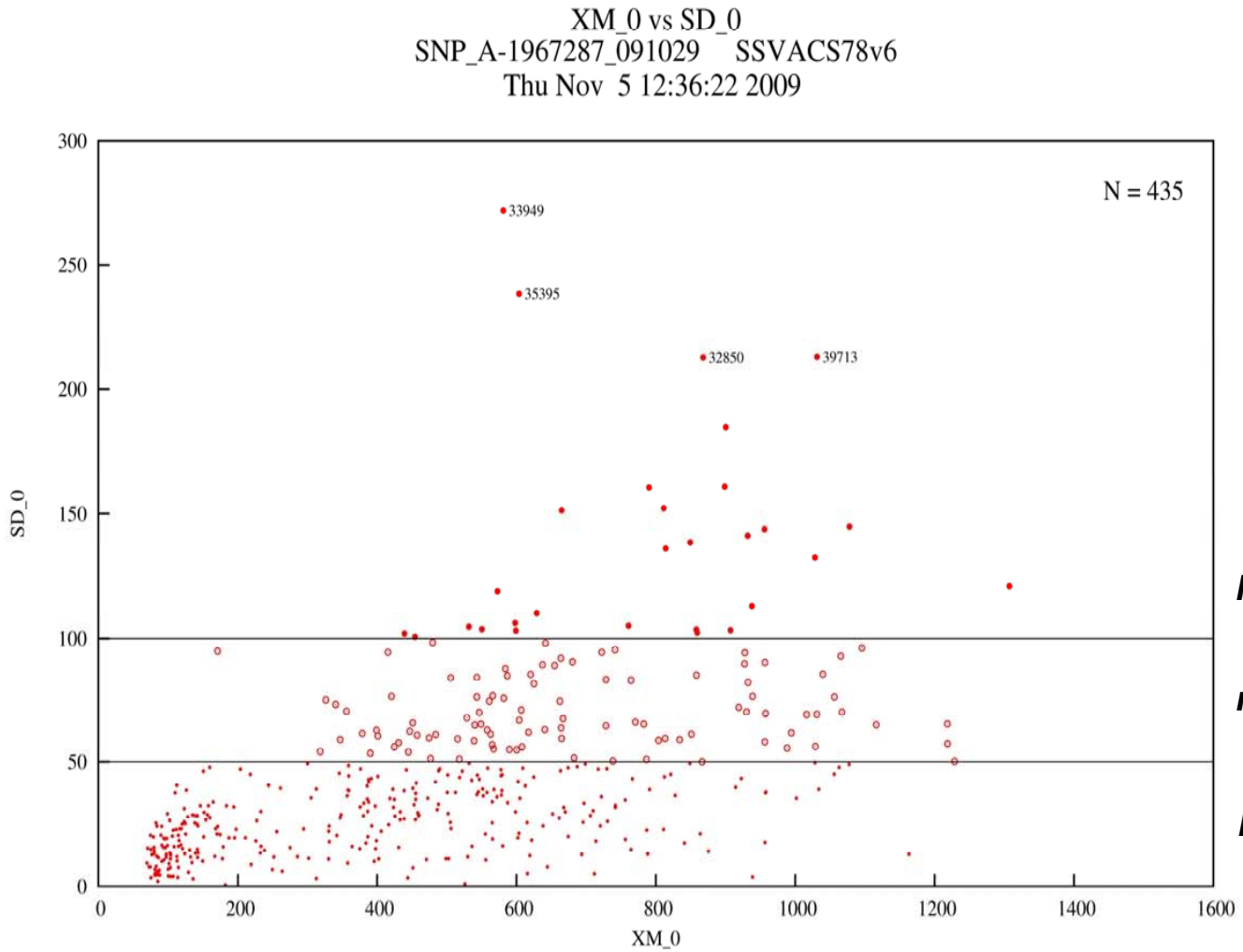
Question: Is there a correspondency between lower level and higher level QC-parameters?

Affymetrix Genome–Wide Human SNP Array 6.0

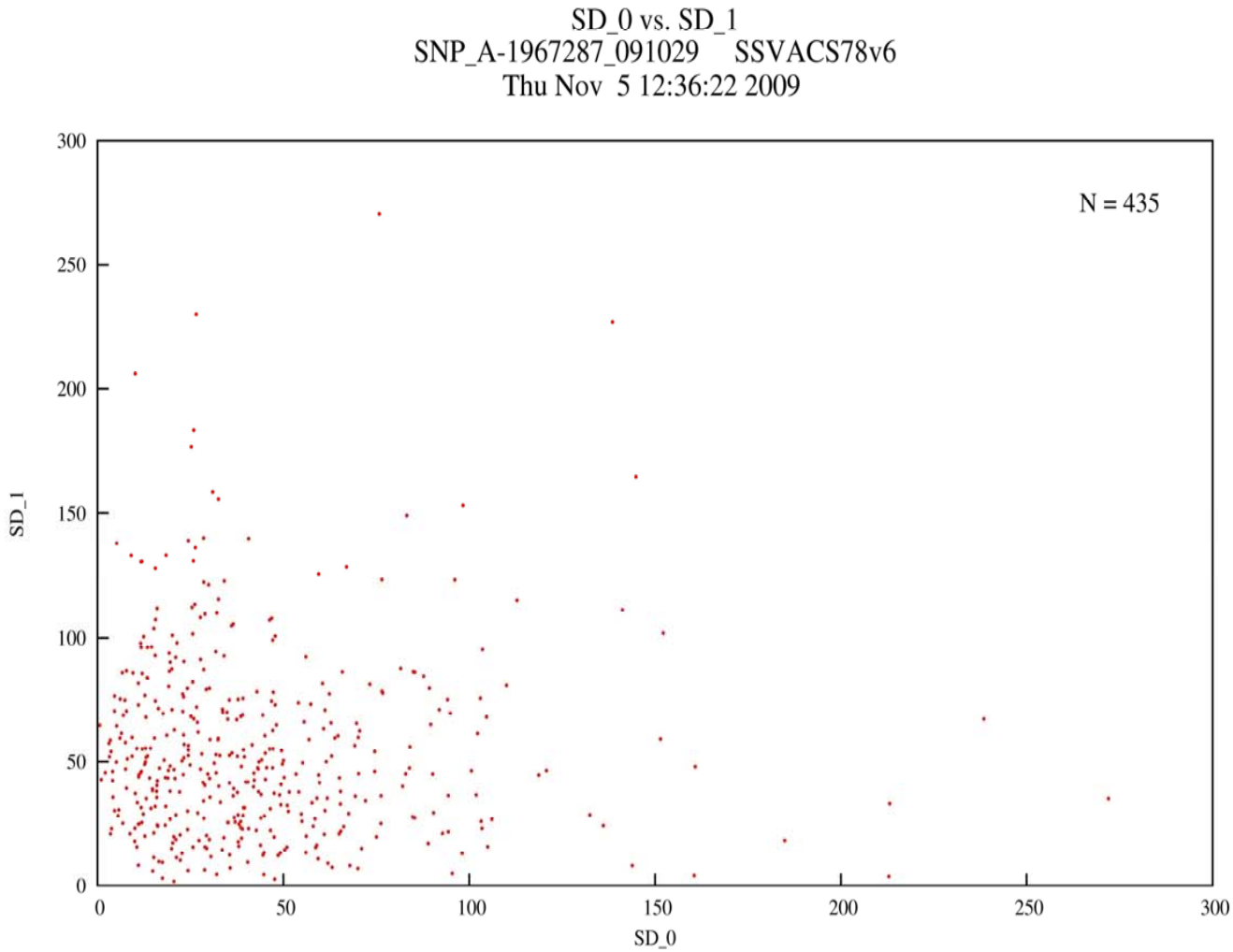
raw fluorescence intensities of all SNP features



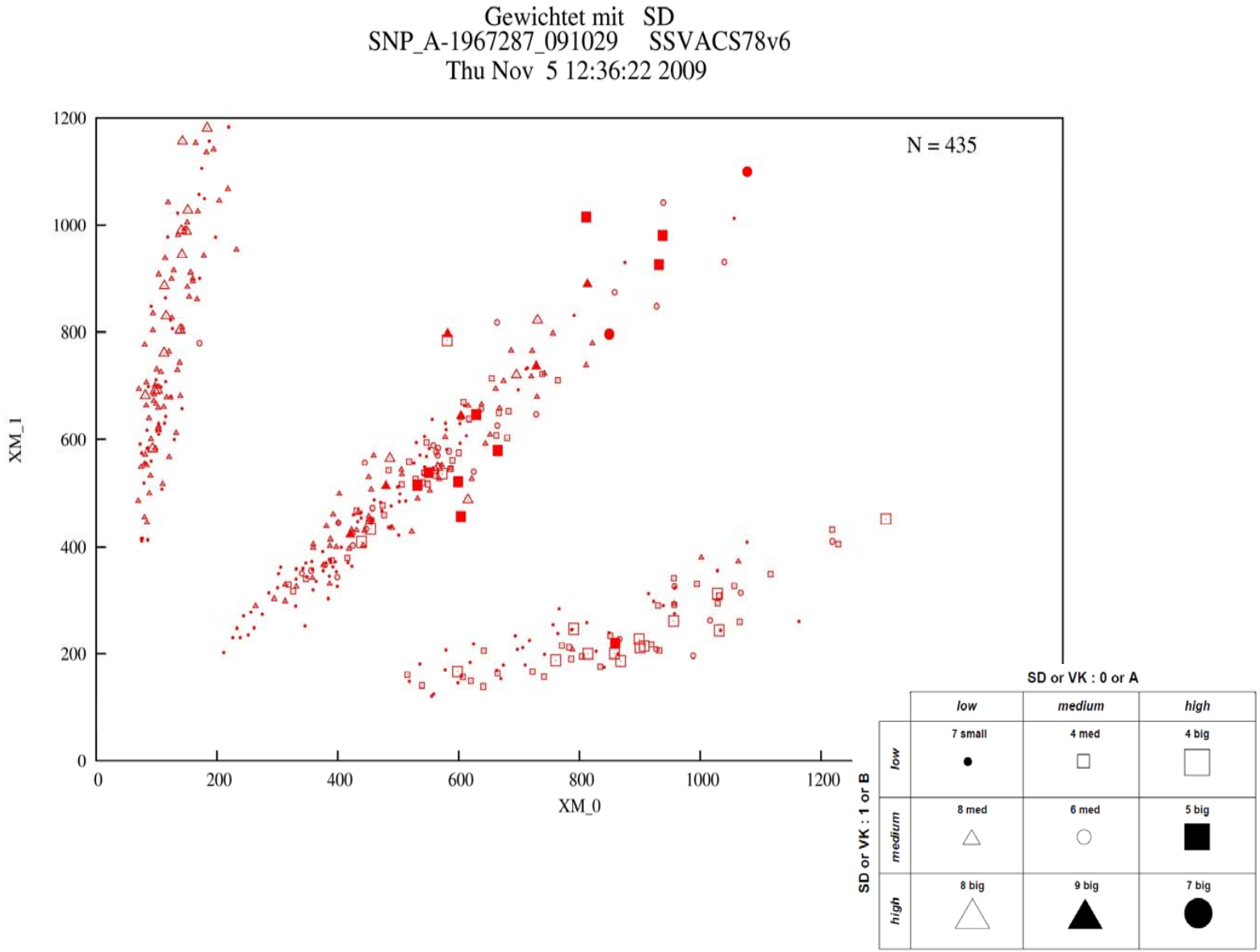
Affymetrix Genome-Wide Human SNP Array 6.0 : low level data (1)



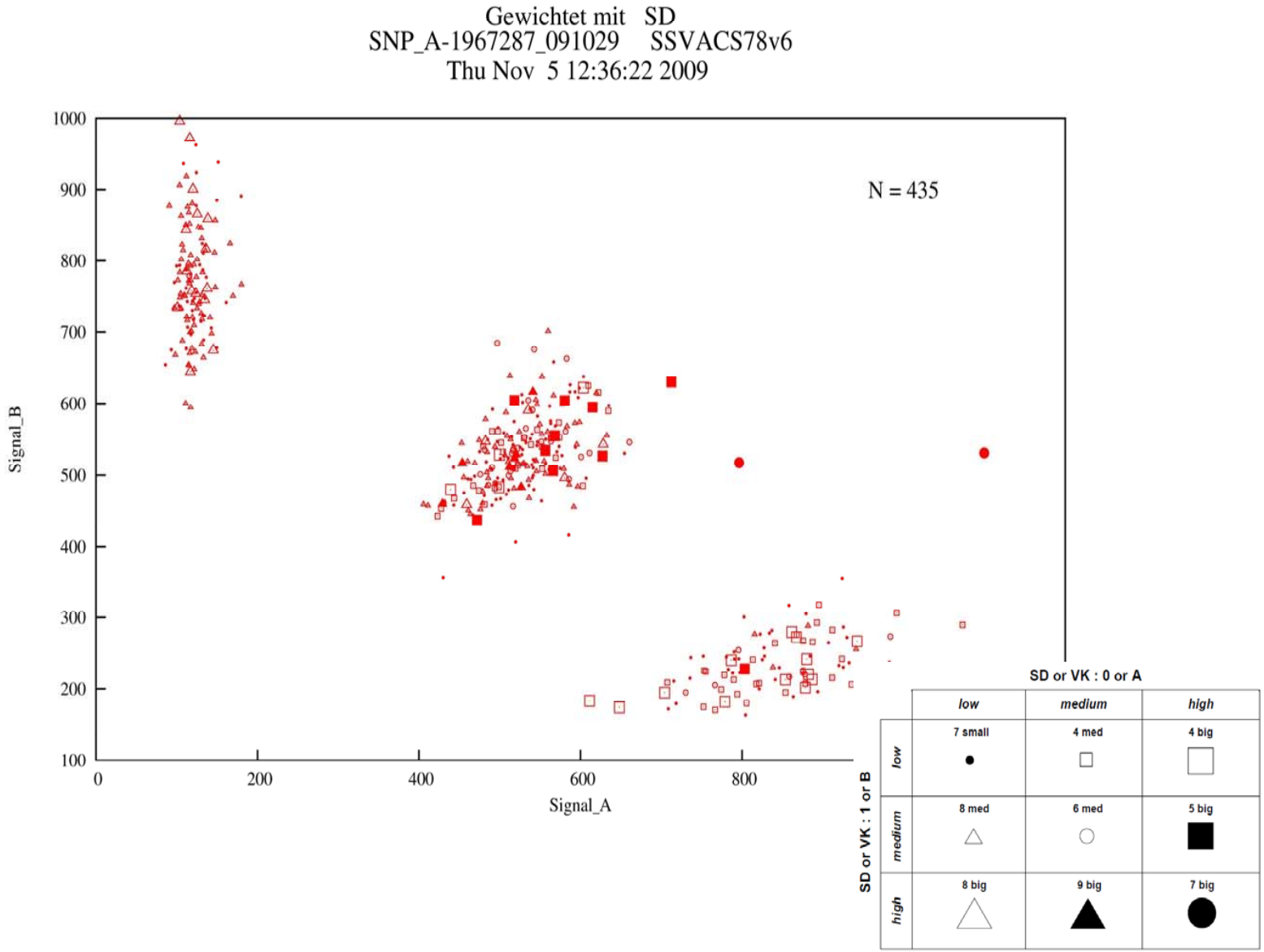
Affymetrix Genome-Wide Human SNP Array 6.0 : low level data (2)



Affymetrix Genome-Wide Human SNP Array 6.0 : low level data (3)



Affymetrix Genome-Wide Human SNP Array 6.0 : low level data (4)

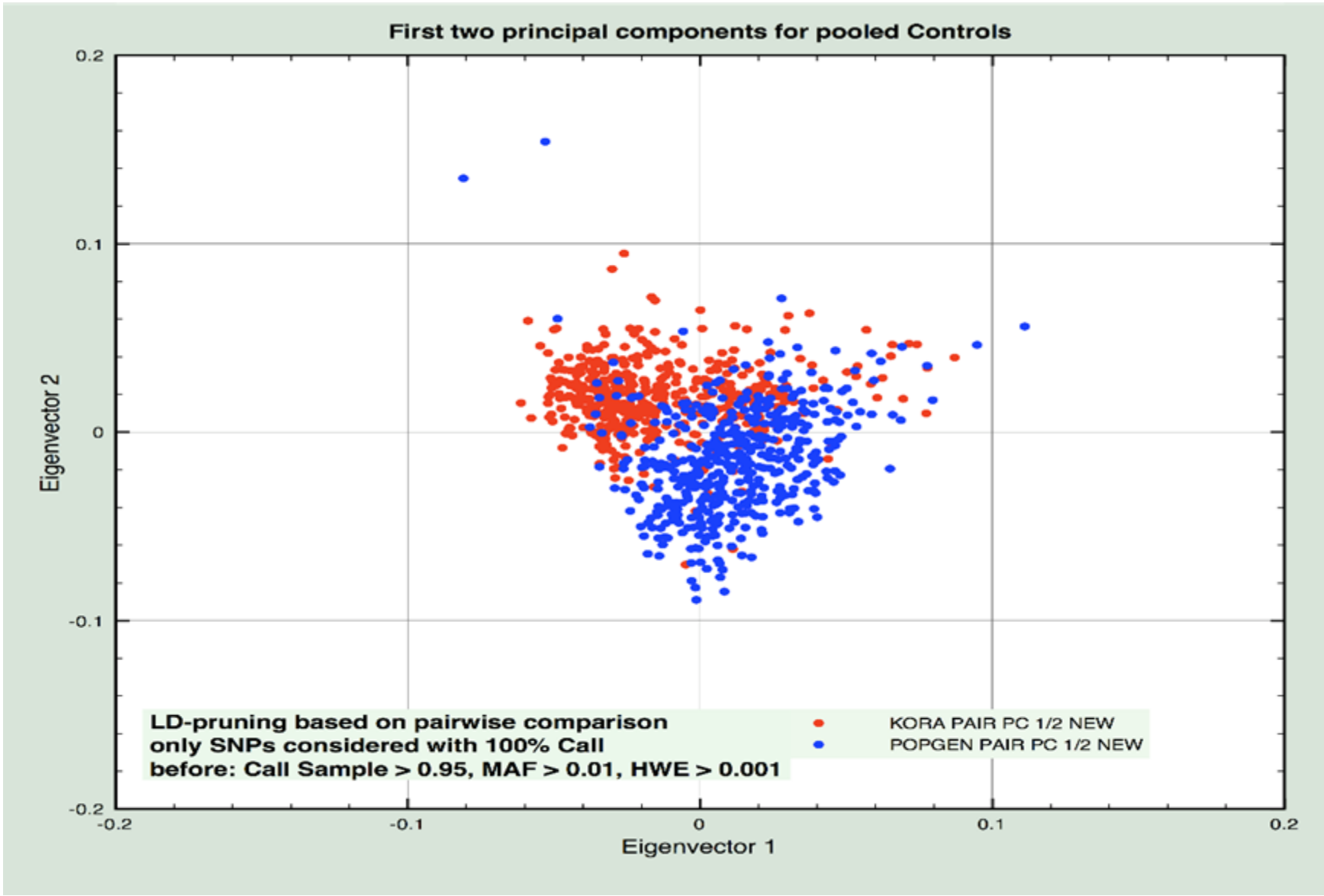


High Level Data

Aggregated Data

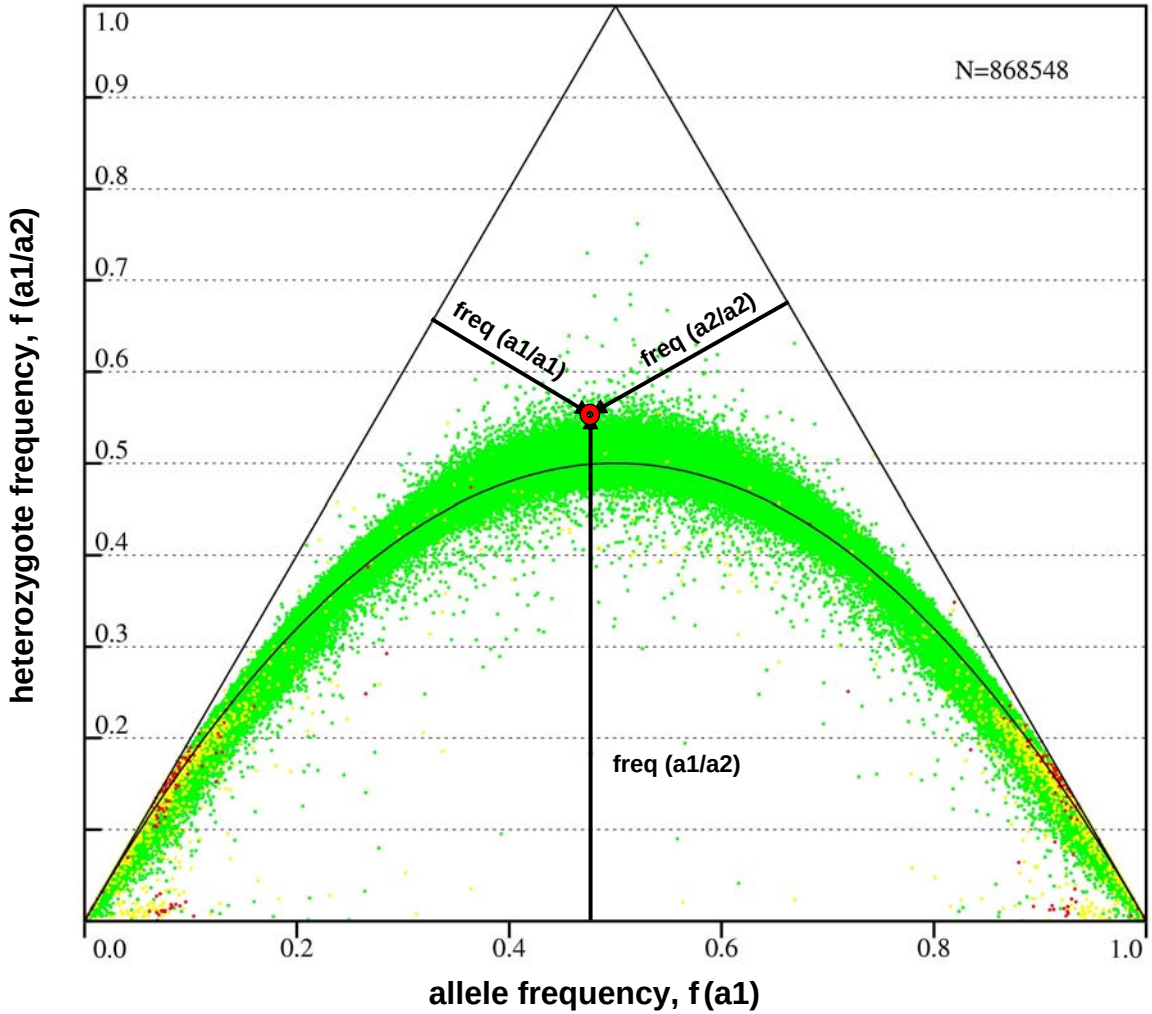
Genotype Data organized in Subsamples (Cases and Controls)

EIGENSTRAT / EIGENSOFT Analysis

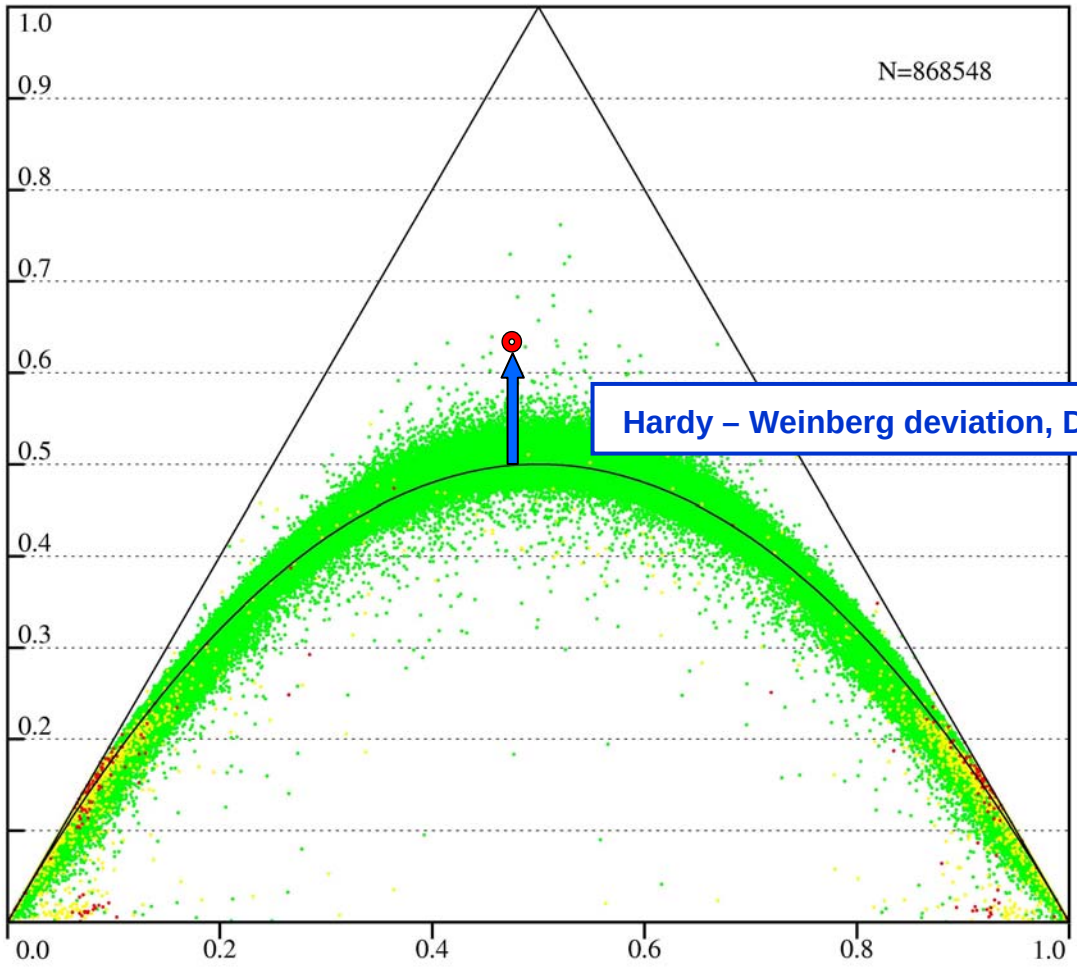


Price et al. (2006) Nat Genet 38(8):904-909; Patterson et al. (2006) PLoS Genetics 2(12): e190

GWAS: de Finetti Plot

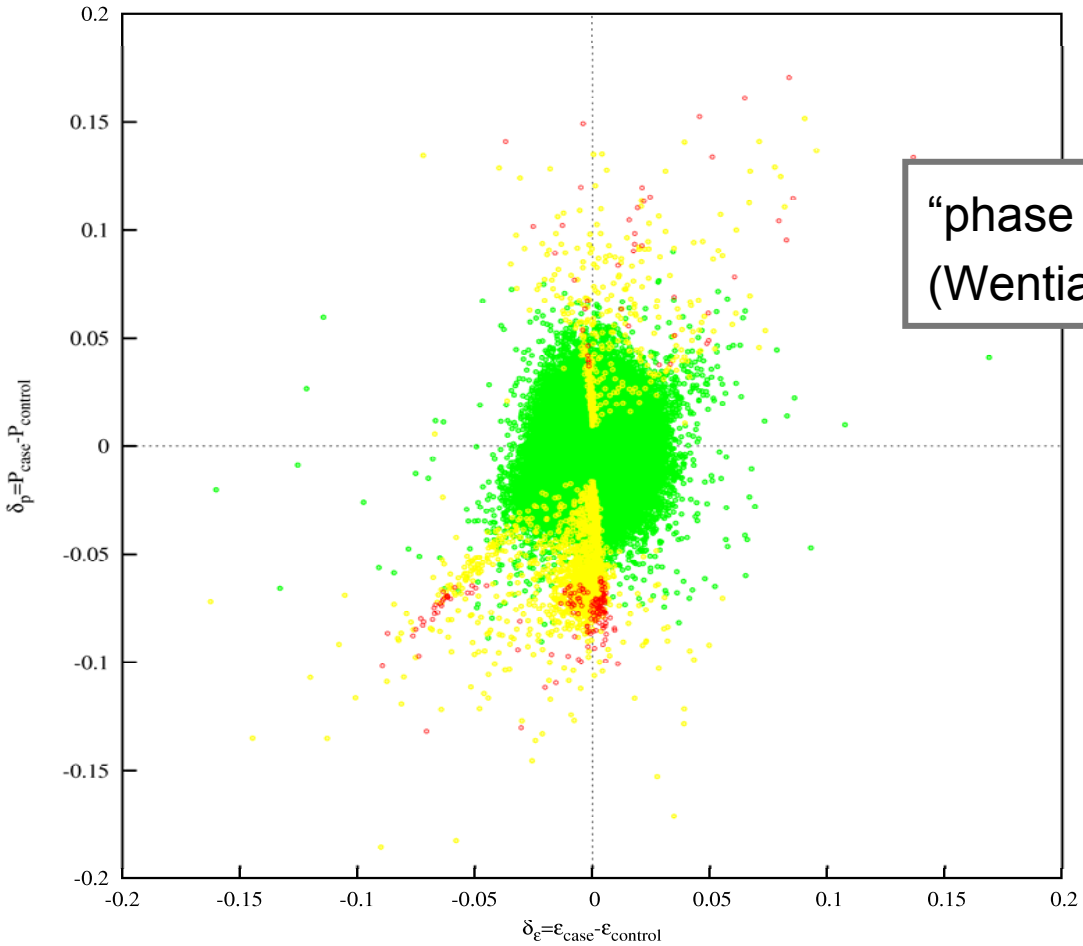


GWAS: de Finetti Plot



Hardy – Weinberg deviation, D_{HWE} or ε

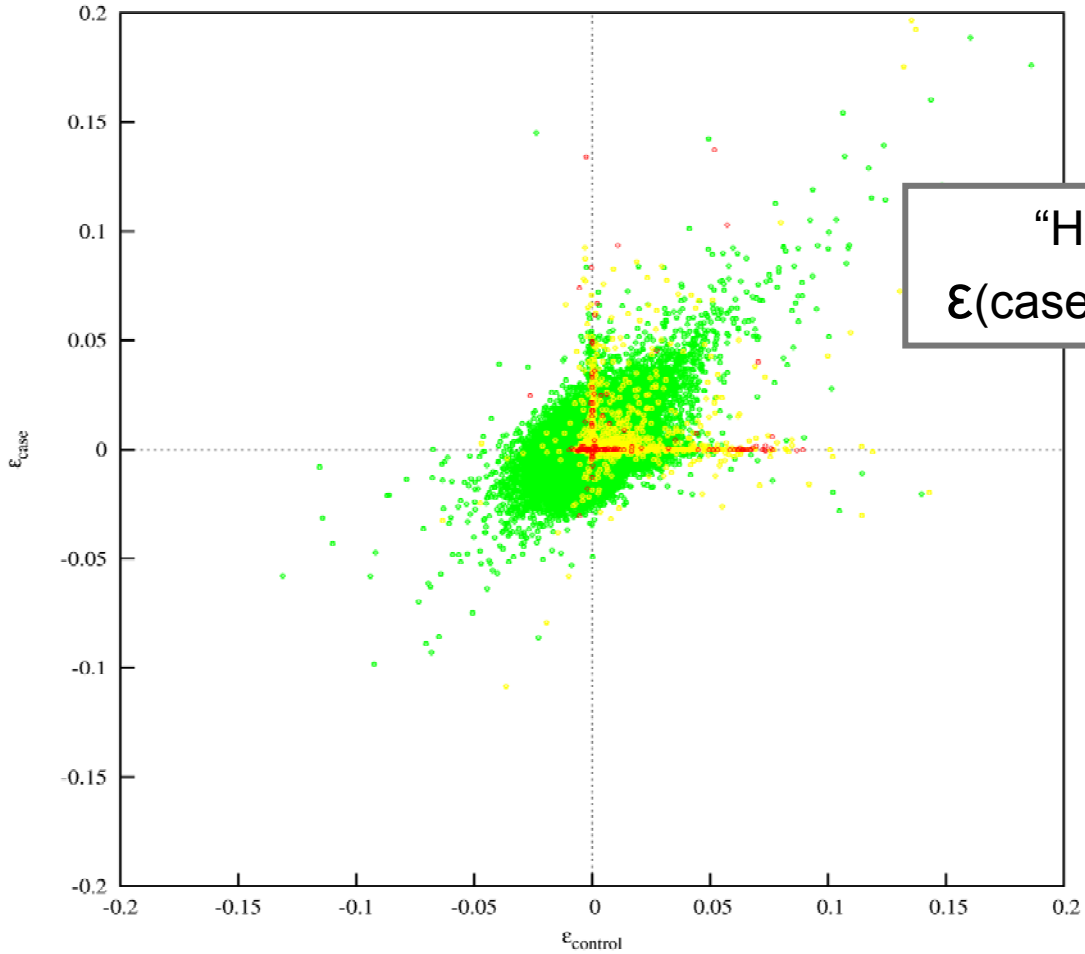
Case – Control Quality Analysis



"phase diagram"
(Wentian Li et al., 2009)

p-value <10e-40	**N = 195**	**• red**
10e-40 ... 10e-12	**N = 2,458**	**• yellow**
>10e-12	**N = 865,895**	**• green**

Case – Control Quality Analysis

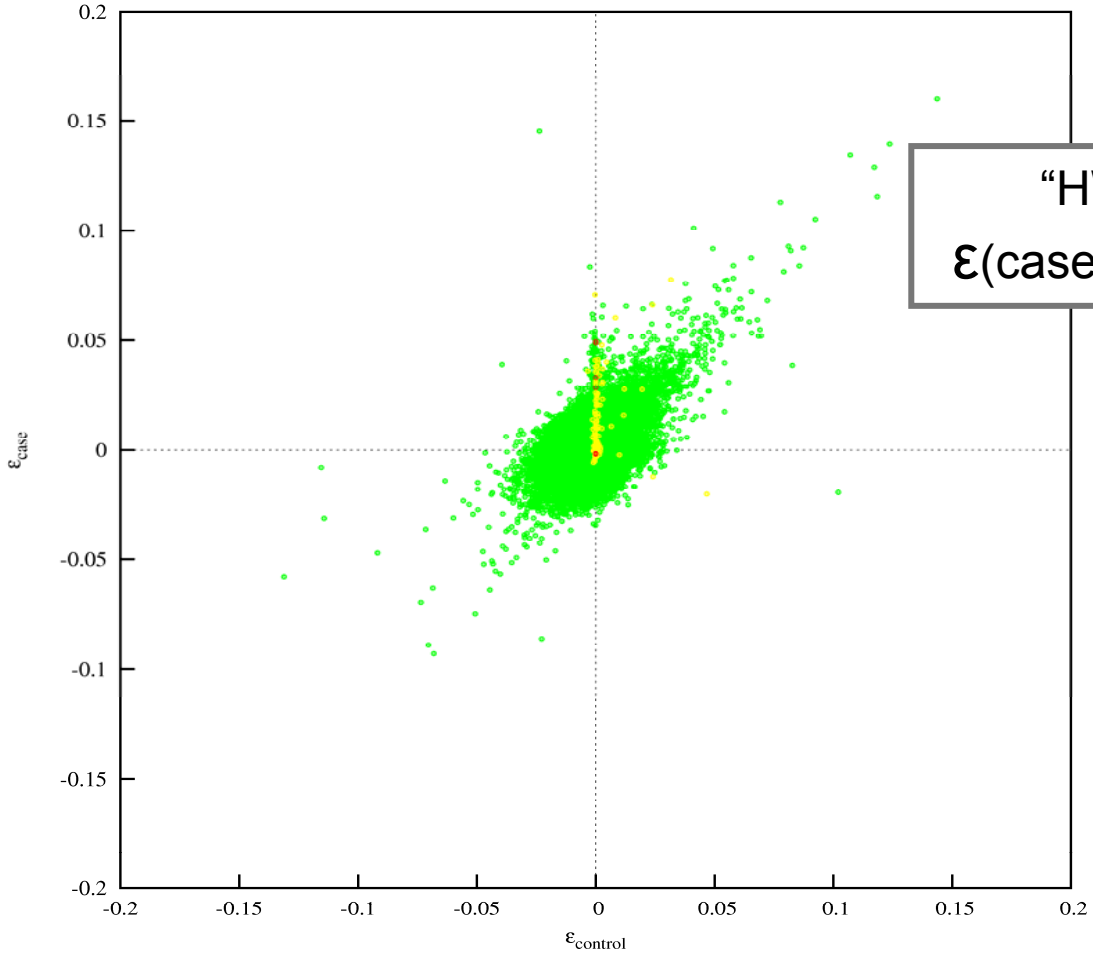


"HWE diagram"
ε(cases) vs. ε(controls)

p-value <10e-40	N = 195	• red
10e-40 ... 10e-12	N = 2,458	• yellow
>10e-12	N = 865,895	• green

Case – Control Quality Analysis

$$0.00 \leq \delta(p_{cases} - p_{controls}) < 0.05$$

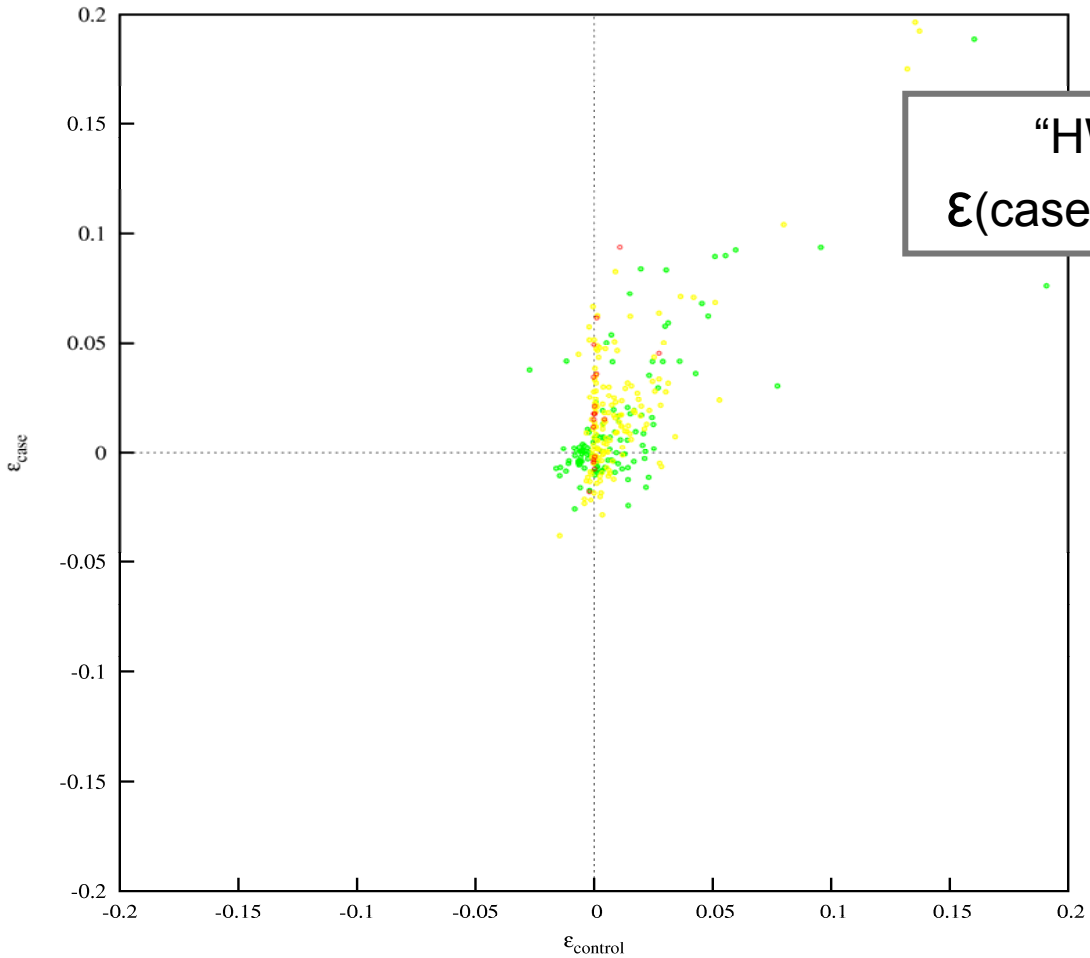


"HWE diagram"
ε(cases) vs. ε(controls)

p-value <10e-40	N = 9	• red
10e-40 ... 10e-12	N = 394	• yellow
>10e-12	N = 422,352	• green

Case – Control Quality Analysis

$$0.05 \leq \delta(p_{cases} - p_{controls}) < 0.10$$

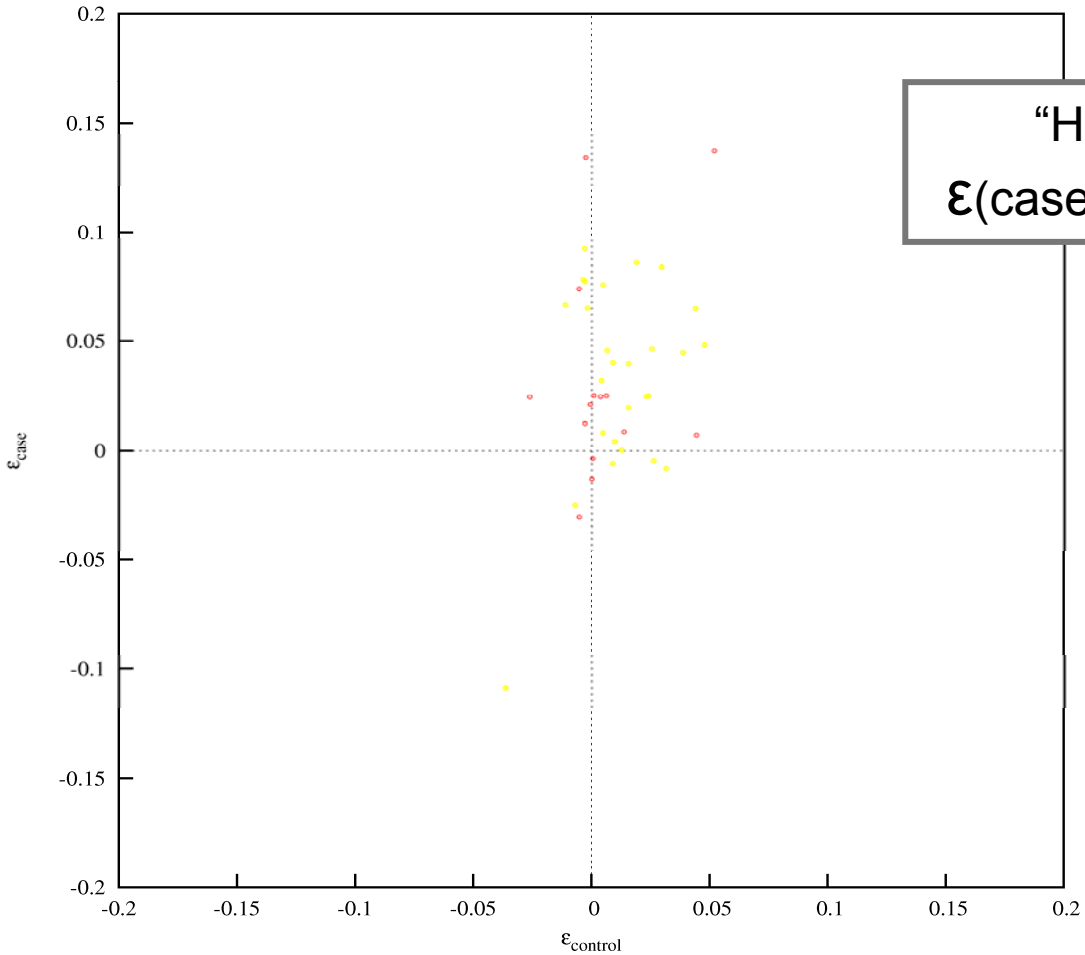


"HWE diagram"
ε(cases) vs. ε(controls)

p-value <10e-40	N = 17	• red
10e-40 ... 10e-12	N = 148	• yellow
>10e-12	N = 126	• green

Case – Control Quality Analysis

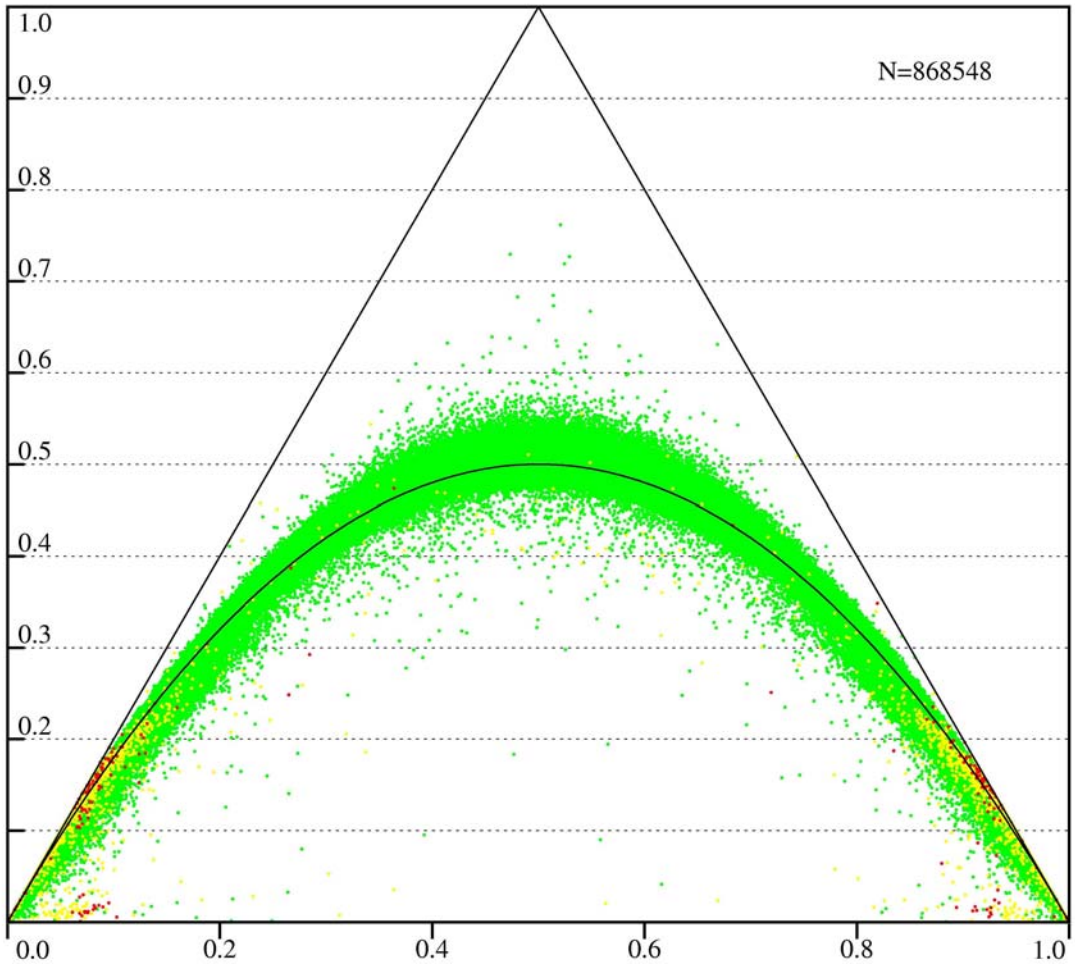
$$0.10 \leq \delta(p_{cases} - p_{controls}) < 0.15$$



"HWE diagram"
ε(cases) vs. ε(controls)

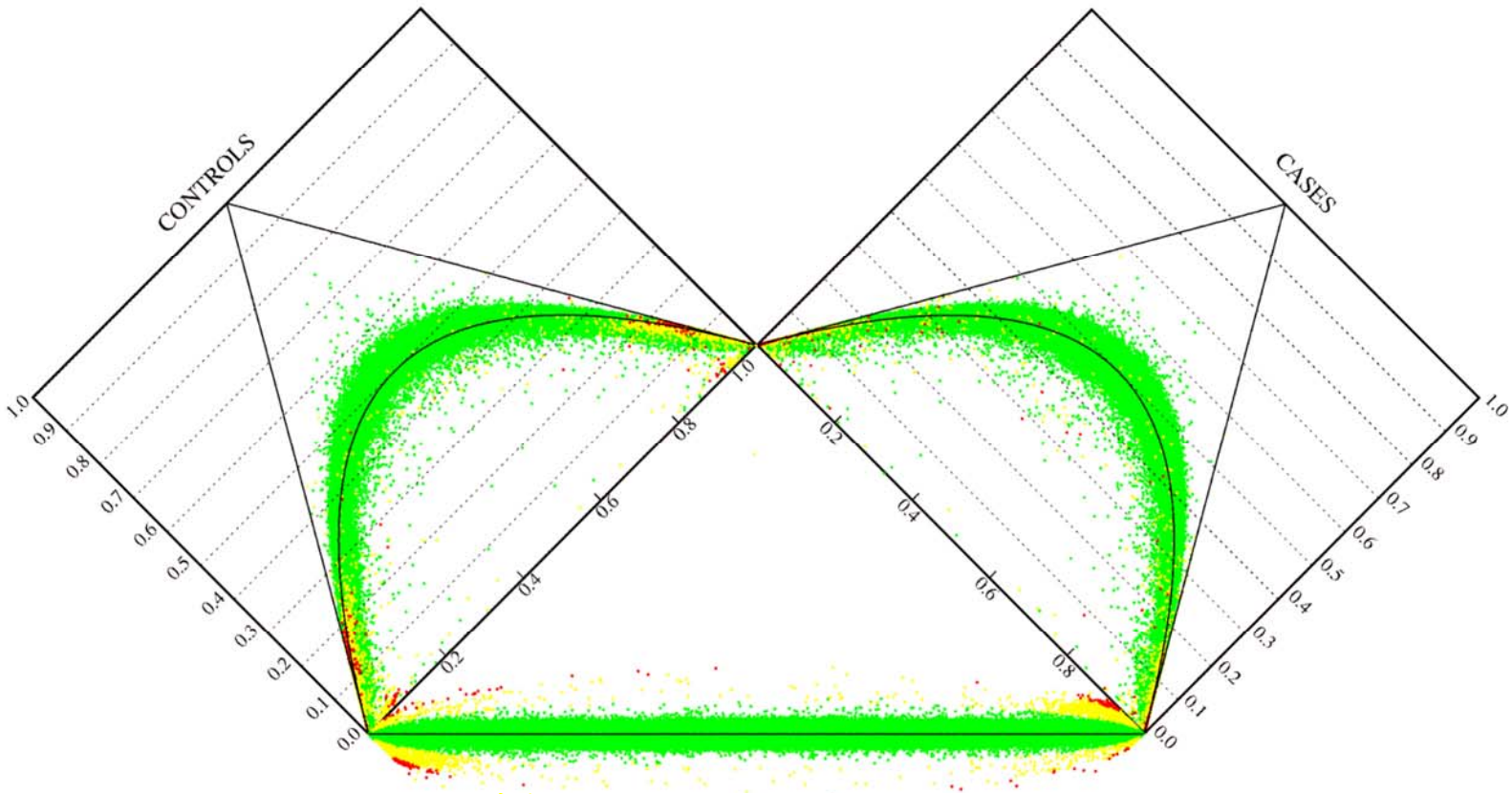
p-value <10e-40	N = 14	• red
10e-40 ... 10e-12	N = 27	• yellow
>10e-12	N = 0	• green

GWAS: de Finetti Plot



P-values (AFD-Test)
$P < 10^{-40}$
$10^{-40} \leq P < 10^{-12}$
$10^{-12} \leq P$

GWAS: de Finetti – Durov Plot



Odds Ratio, AFD_p:	<10e-40	N=195	rot
	10e-40 - 10-e12	N=2458	gelb
	>10e-12	N=865895	gruen
	total	N=868548	

Summary

- High level quality criteria (*call rate, MAF, HWE*) are effective and indispensable:
  - High level quality criteria have demonstrated their usefulness in reducing the error rate.
  - High level quality criteria can not be replaced by medium and low level quality criteria.
- Intermedium level quality criteria (*cluster validation scores*) allow for identification of erroneous / dubious genotypes, which passed the default quality filters implemented by the genotyping manufacturers.
- Further analysis of parametric and non-parametric methods for cluster analysis and cluster validation seems to be necessary and attractive.
- Low level quality criteria (*intensity features / signals*) have not resulted in useful procedures so far.

Acknowledgements

Michael Steffens [1]

Rolf Fimmers [1]

Marina Angisch [1]

Daniela Holler [1]

Sudesha Johann [2]

Caroline Pawlak [2]

(1) Institute for Medical Biometry, Informatics and Epidemiology (IMBIE), University of Bonn

(2) University of Applied Sciences (Fachhochschule) Koblenz, Campus Remagen

Errors happen,
and if not properly dealt with,
this could be the
consequences, but shouldn‘t.

Thank you
for your attention!