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

SPONSORED BY THE

Federal Ministry of Education and Research

Überblick

- Das NGFN-Plus Qualitätskontroll-Experiment
- Fehlerwahrscheinlichkeit und Fehlermodell
- Abschätzung der Fehlerrate für verschiedene Genotypisierungschips
- Analyse von „High-level”-Qualitätsparametern (z. B. *Call Rate, HWE, etc.*) in Hinblick auf die Fehlerrate
- Analyse des Silhouette-Score zur Cluster-Validierung
- Analyse von Intensitätssignalen abgeleiteten Qualitätsparametern („Low-level”-Qualitätsparameter)

Teil 1

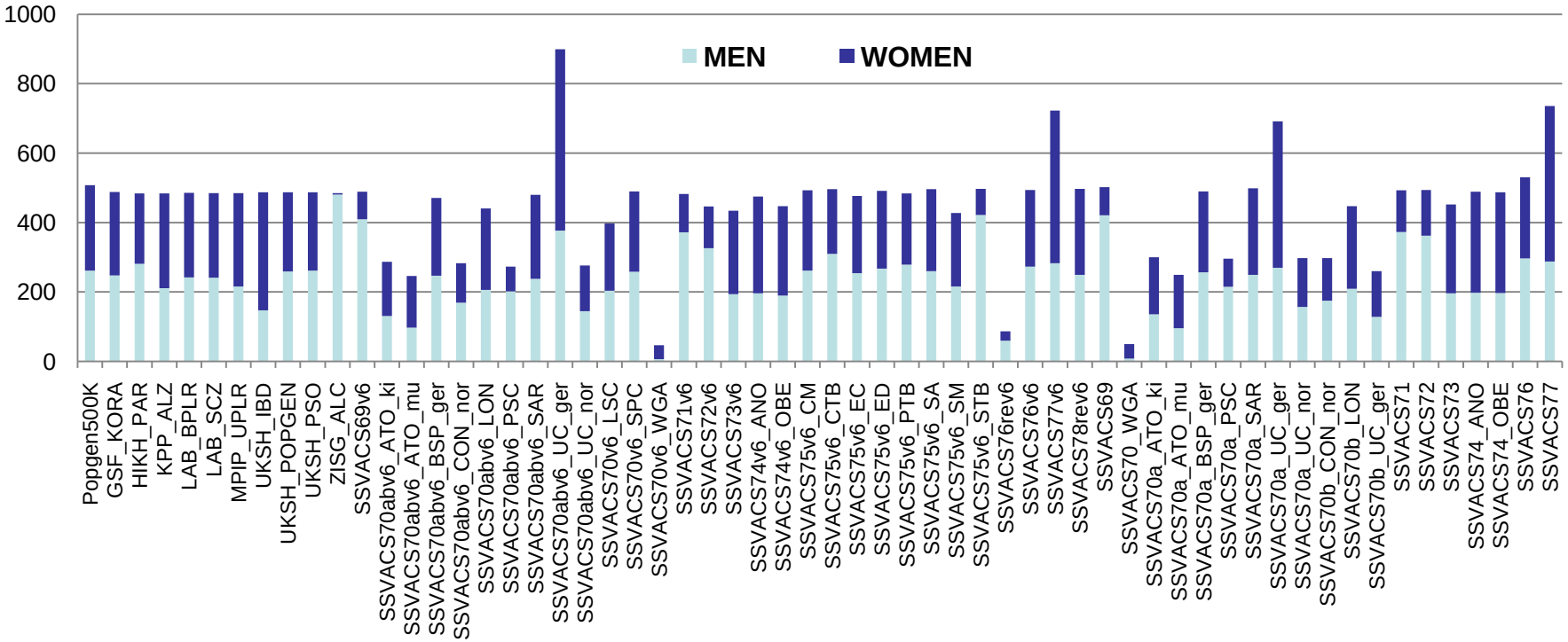
Das NGFN-Plus Qualitätskontroll-Experiment

SPONSORED BY THE
Federal Ministry
of Education
and Research

Das NGFN GWAS Add-on Projekt

- Aufgelegt 2007 am Ende der NGFN II Periode
- Ziel: Identifizierung der genetischen Ursachen von über 40 komplex genetischen Krankheiten mittels Hochdurchsatz-Genotypisierung
- Insgesamt: 69 Einzelprojekte
  - 10 genotypisiert mit dem Projekte Illumina HH550k
  - 40 genotypisiert mit dem AGWH SNP Array 6.0
  - 19 genotypisiert mit dem AGWH SNP Array 5.0
- Anzahl genotypisierter Personen: 26.456
- Anzahl Genotypen: > 20 x 10^9
- Ca. 230 - 450 x 10^6 Genotypen pro Projekt

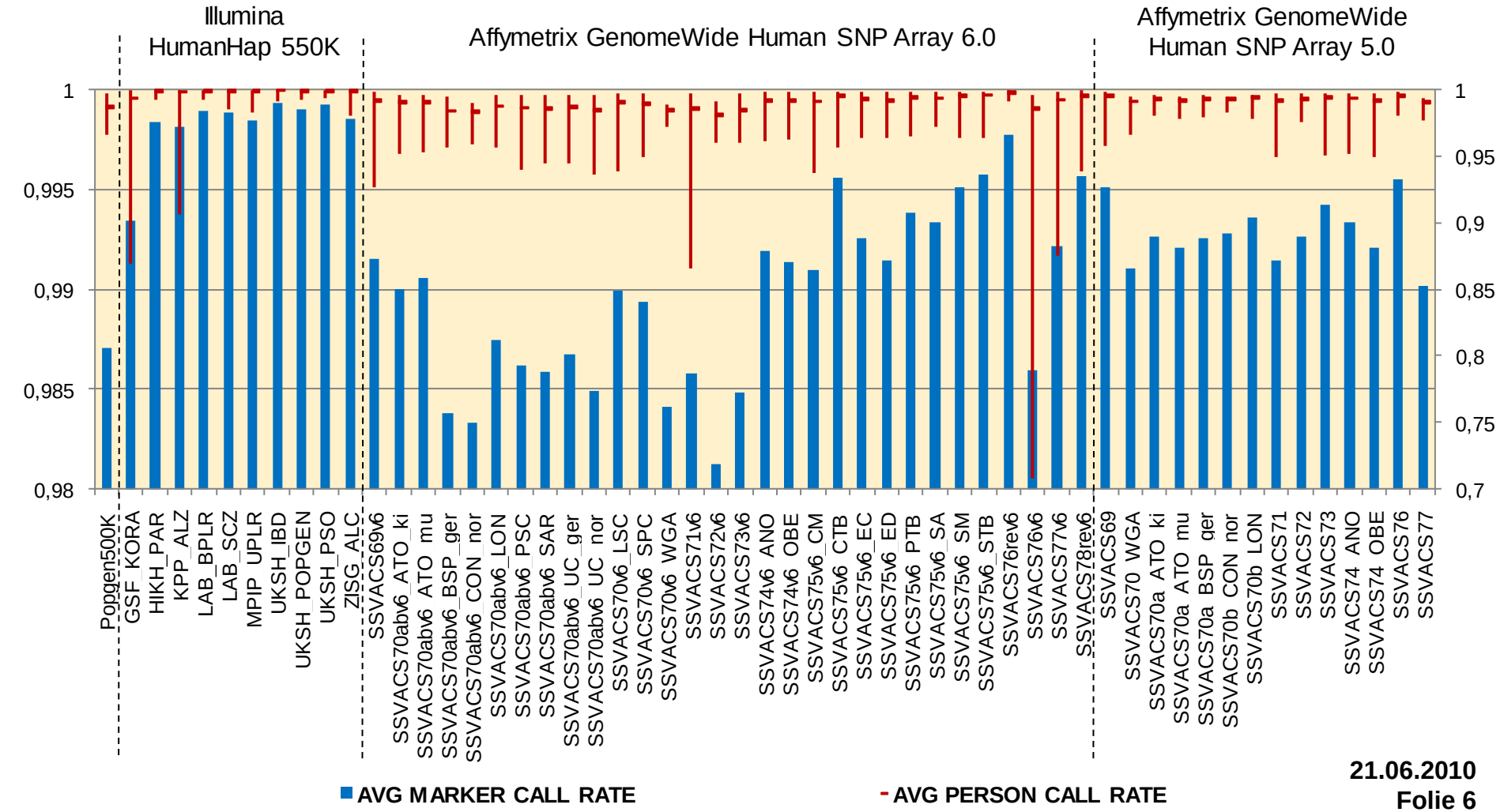
Das NGFN GWAS Add-on Projekt



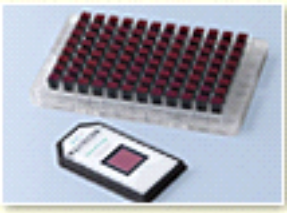
- Durchschnittliche Studiengröße: ~ 500 Personen

Das NGFN-GWAS Add-on Projekt

Marker- und Personen Call-Raten



Das Qualitätskontroll-Experiment



Studiendesign für den AGWH SNP Array

- 9 Qualitätskontroll-Samples (Blutspender, IHT Bonn) mit 5 DNA Aliquots pro Sample
- 2 verschiedene Qualitätskontroll-Samples für jedes Teilprojekt
- 1 DNA-Sample je Mikrotiterplatte (96 wells)
- Im Mittel 6 DNA-Samples (je 3 von der-selben Qualitätskontrolle) je Teilprojekt
- → 12-fache Replikation pro Sample

	A	B	C	D	E	F	G	H	I	J	Sum:
A:		X	X	X	X	X					5
B:							X	X	X	X	5
C:				X	X	[X]	X				5
D:								X	X	X	5
E:						X	X	X			5
F:									X	X	5
G:								X	X		5
H:										X	5
I:										X	5
J:											5
Sum:											50

A: IHT-06001	25SLHENGFN2453527	25SLHENGFN2453177	25SLHENGFN2453856	25SLHENGFN2453001	25SLHENGFN2454117
B: IHT-06005	25SLHENGFN2453643	25SLHENGFN2454515	25SLHENGFN2454499	25SLHENGFN2454473	25SLHENGFN2454481
…	…	…	…	…	…

Das Qualitätskontroll-Experiment

Studiendesign für den Illumina HH550k Chip

- 4 Qualitätskontroll-Samples (Blutspender, IHT Bonn) mit 4 DNA Aliquots pro Sample
- → 12-fache Replikation je Samples

	A	B	C	E	F	H	Sum:
A:		XX	X	X			4
B:			X	X			4
C:				X	X		3
E:					X		3
F:							
Sum:							14

A: IHT 06001	**FEMALE**	25SLHENGFN2453711	25SLHENGFN2453713	25SLHENGFN2453756	25SLHENGFN2453710
B: IHT 06005	**MALE**	25SLHENGFN2453705	25SLHENGFN2453703	25SLHENGFN2453702	25SLHENGFN2453704
…	…	…	…	…	…

Studienzweig Genotypisierungschip	Anzahl der in das Qualitätsexperiment eingegangenen Genotypen
AGWH SNP Array 5.0	39,338,000
AGWH SNP Array 6.0	141,090,606
Illumina HumanHap 550k	27,265,204

Das Qualitätskontroll-Experiment

Überblick über die Qualitätskontrollen (DNA-Samples)

#	project	chip	MTP	IHT-06001	IHT-06005	IHT-06006	IHT-06008	IHT-06009	IHT-06011	IHT-06012	IHT-060013	IHT-06018	IHT-06020	total
1.	SSVACS69	AWGH SNP Array 5.0	6	0	0	0	3+	0	0	0	3+	0	0	6
2.	SSVACS69v6	AWGH SNP Array 6.0	6	0	0	0	3+	0	0	0	3+	0	0	6
3.	SSVACS70a_ATO_ki	AWGH SNP Array 5.0	4	0	0	2+	2+	0	0	0	0	0	0	4
4.	SSVACS70a_ATO_mu	AWGH SNP Array 5.0	3	0	0	0	2+	0	0	0	0	1+	0	3
5.	SSVACS70a_BSP_ger	AWGH SNP Array 5.0	6	0	0	0	0	3+	3+	0	0	0	0	6
6.	SSVACS70a_PSC	AWGH SNP Array 5.0	3+1T	1+	2+	0	0	0	0	0	0	0	0	3
7.	SSVACS70a_SAR	AWGH SNP Array 5.0	6	0	3+	0	0	0	0	0	0	3+	0	6
8.	SSVACS70a_UC_ger	AWGH SNP Array 5.0	8	0	1+	0	0	4+	0	3+	0	0	0	8
9.	SSVACS70a_UC_nor	AWGH SNP Array 5.0	3+1T	0	3+	0	0	0	0	0	0	0	0	3
10.	SSVACS70abv6_ATO_ki	AWGH SNP Array 6.0	4	0	0	2+	2+	0	0	0	0	0	0	4
11.	SSVACS70abv6_ATO_mu	AWGH SNP Array 6.0	3	0	0	0	2+	0	0	0	0	1+	0	3
12.	SSVACS70abv6_BSP_ger	AWGH SNP Array 6.0	4+2T	0	0	0	0	3+	2+	0	0	0	0	5
13.	SSVACS70abv6_CON_nor	AWGH SNP Array 6.0	4	0	0	4+	0	0	0	0	0	0	0	4
14.	SSVACS70abv6_LON	AWGH SNP Array 6.0	6	0	3+	0	0	0	0	3+	0	0	0	6
15.	SSVACS70abv6_PSC	AWGH SNP Array 6.0	3+1T	1+	1+	0	0	0	0	0	0	0	0	2
16.	SSVACS70abv6_SAR	AWGH SNP Array 6.0	6	0	3+	0	0	0	0	0	0	2+	0	6
17.	SSVACS70abv6_UC_ger	AWGH SNP Array 6.0	9+2T	0	1+	0	0	4+	0	5+	0	0	0	10
18.	SSVACS70abv6_UC_nor	AWGH SNP Array 6.0	2+2T	0	3+	0	0	0	0	0	0	0	0	3
19.	SSVACS70b_CON_nor	AWGH SNP Array 5.0	4	0	0	4+	0	0	0	0	0	0	0	4
20.	SSVACS70b_LON	AWGH SNP Array 5.0	6	0	3+	0	0	0	0	3+	0	0	0	6
21.	SSVACS70b_UC_ger	AWGH SNP Array 5.0	3	0	0	0	0	1+	0	2+	0	0	0	3
22.	SSVACS70v6_LSC	AWGH SNP Array 6.0	6	0	0	3+	0	0	3+	0	0	0	0	6
23.	SSVACS70v6_SPC	AWGH SNP Array 6.0	6	0	0	0	0	3+	3+	0	0	0	0	6
24.	SSVACS71	AWGH SNP Array 5.0	6	0	3+	0	0	0	0	0	3+	0	0	6
25.	SSVACS71v6	AWGH SNP Array 6.0	6	0	3+	0	0	0	0	0	3+	0	0	6
26.	SSVACS72	AWGH SNP Array 5.0	6	0	0	3+	0	3+	0	0	0	0	0	6
27.	SSVACS72v6	AWGH SNP Array 6.0	6	0	0	2+	0	3+	0	0	0	0	0	6
28.	SSVACS73	AWGH SNP Array 5.0	6	0	0	3+	0	0	0	3+	0	0	0	6
29.	SSVACS73v6	AWGH SNP Array 6.0	6	0	0	3+	0	0	0	3+	0	0	0	6
30.	SSVACS74_ANO	AWGH SNP Array 5.0	5+1T	2+	0	0	3+	0	0	0	0	0	0	5
31.	SSVACS74_OBE	AWGH SNP Array 5.0	6	0	0	0	3+	0	0	0	0	0	3+	6
32.	SSVACS74v6_ANO	AWGH SNP Array 6.0	5+1T	2+	0	0	3+	0	0	0	0	0	0	5
33.	SSVACS74v6_OBE	AWGH SNP Array 6.0	6	0	0	0	3+	0	0	0	0	0	3+	6
34.	SSVACS75v6_CM	AWGH SNP Array 6.0	5+2T	0	0	0	0	0	4+	0	0	0	2+	6
35.	SSVACS75v6_CTB	AWGH SNP Array 6.0	5+2T	3+	0	0	0	0	3+	0	0	0	0	6
36.	SSVACS75v6_EC	AWGH SNP Array 6.0	4+3T	0	0	0	0	0	0	0	2+	0	4+	6
37.	SSVACS75v6_ED	AWGH SNP Array 6.0	6	0	0	0	0	0	0	0	0	3+	3+	6
38.	SSVACS75v6_PTB	AWGH SNP Array 6.0	6	0	0	0	0	0	0	3+	3+	0	0	6
39.	SSVACS75v6_SA	AWGH SNP Array 6.0	5+2T	0	0	0	0	0	2+	0	0	3+	0	5
40.	SSVACS75v6_SM	AWGH SNP Array 6.0	5	0	0	0	0	3+	0	0	2+	0	0	5
41.	SSVACS75v6_STB	AWGH SNP Array 6.0	5+2T	0	0	0	0	0	0	3+	0	3+	0	6
42.	SSVACS76	AWGH SNP Array 5.0	5+11T	3+	0	0	0	3+	0	0	0	0	0	5
43.	SSVACS76v6	AWGH SNP Array 6.0	5+1T	2+	0	0	0	3+	0	0	0	0	0	5
44.	SSVACS77	AWGH SNP Array 5.0	8+1T	3+	0	3+	0	0	0	0	0	0	0	6
45.	SSVACS77v6	AWGH SNP Array 6.0	8+1T	3+	0	3+	0	0	0	0	0	0	0	6
46.	SSVACS78rev6	AWGH SNP Array 6.0	6	0	6+	0	0	0	0	0	0	0	6+	12
47.	HIKH_PAR	Illumina HumanHap550K	6	3-	0	0	0	3-	0	0	0	0	0	6
48.	KPP_ALZ	Illumina HumanHap550K	5+1T	0	2-	3-	0	0	0	0	0	0	0	5
49.	LAB_BPLR	Illumina HumanHap550K	6	3-	3-	0	0	0	0	0	0	0	0	6
50.	LAB_SCZ	Illumina HumanHap550K	6	3-	3-	0	0	0	0	0	0	0	0	6
51.	MPIP_UPLR	Illumina HumanHap550K	6	3-	0	3-	0	0	0	0	0	0	0	6
52.	UKSH_IBD	Illumina HumanHap550K	5+1T	0	0	2-	0	0	3-	0	0	0	0	5
53.	UKSH_POPGEN	Illumina HumanHap550K	5+1T	0	0	0	0	3-	2-	0	0	0	0	5
54.	UKSH_PSO	Illumina HumanHap550K	5+1T	0	0	3-	0	3-	0	0	0	0	0	5
55.	ZISG_ALC	Illumina HumanHap550K	6	0	3-	0	0	3-	0	0	0	0	0	6

Berechnung von Fehlerraten

Fehlermodelle

		wahrer Genotyp		
		0	1	2
beobachteter Genotyp	0	$1 - \varepsilon$	$\varepsilon / 2$	$\varepsilon / 2$
	1	$\varepsilon / 2$	$1 - \varepsilon$	$\varepsilon / 2$
	2	$\varepsilon / 2$	$\varepsilon / 2$	$1 - \varepsilon$

Einfaches Fehlermodell mit einem Parameter (einer Fehlerrate) je Marker.
Die Wahrscheinlichkeit hängt nicht vom zugrundeliegenden wahren Genotyp ab.

		wahrer Genotyp		
		0	1	2
beobachteter Genotyp	0	$(1 - \varepsilon)^2$	$\varepsilon (1 - \varepsilon)$	ε^2
	1	$2 \varepsilon (1 - \varepsilon)$	$\varepsilon^2 + (1 - \varepsilon)^2$	$2 \varepsilon (1 - \varepsilon)$
	2	ε^2	$\varepsilon (1 - \varepsilon)$	$(1 - \varepsilon)^2$

Biologisch begründetes Fehlermodell mit einem Parameter. Die Fehlerrate ist proportional zur Transitionsrate der Allele.

		wahrer Genotyp		
		0	1	2
beobachteter Genotyp	0	$1 - \varepsilon_1 - \varepsilon_2$	ε_3	ε_5
	1	ε_1	$1 - \varepsilon_3 - \varepsilon_4$	ε_6
	2	ε_2	ε_4	$1 - \varepsilon_5 - \varepsilon_6$

Vollständiges Fehlermodell mit einem Parameter für jeden möglichen Fehlertyp (6 Freiheitsgrade). *(Verwendetes Fehlermodell zur Bestimmung der Fehlerraten für die verschiedenen Variationstypen)*.

Berechnung von Fehlerraten

Fehlermodelle

		wahrer Genotyp		
		0	1	2
beobachteter Genotyp	0	$1 - \varepsilon_0$	ε_{10}	0
	1	ε_0	$1 - (\varepsilon_{10} + \varepsilon_{12})$	ε_2
	2	0	ε_{12}	$1 - \varepsilon_2$

Referenz-Fehlermodell von *Scheet und Stephens* mit 4 Parametern (Fehlerraten) pro Marker.

		wahrer Genotyp		
		0	1	2
beobachteter Genotyp	0	$1 - \varepsilon_0$	ε_1	0
	1	ε^0	$1 - 2\,\varepsilon_1$	ε_0
	2	0	ε_1	$1 - \varepsilon_0$

2 - Parametermodell mit unterschiedlichen Fehlerraten für homozygote und heterozygote Genotypen.

		wahrer Genotyp		
		0	1	2
beobachteter Genotyp	0	$1 - \varepsilon_0$	$\varepsilon_1 / 2$	$\varepsilon_0 / 2$
	1	$\varepsilon_0 / 2$	$1 - \varepsilon_1$	$\varepsilon_0 / 2$
	2	$\varepsilon_0 / 2$	$\varepsilon_1 / 2$	$1 - \varepsilon_0$

Modifiziertes 2 - Parametermodell. Im Gegensatz zum 2-Parametermodell können auch homozygote Genotypen beobachtet werden, wenn der wahre Genotyp dem homozygoten Genotyp des anderen Allels entspricht.

Berechnung von Fehlerraten

Fehlerraten für die Affymetrix GenomeWide Human SNP Arrays

AGWH SNP Array 5.0

n = 3.533.469

		wahrer Genotyp AA	AC	CC
beobachteter Genotyp	AA	0,999257026	0,000616686	0,000033168
	AC	0,000670169	0,998762917	0,001198470
	CC	0,000072805	0,000620396	0,998768362

n =13.843.194

		wahrer Genotyp AA	AG	GG
beobachteter Genotyp	AA	0,999316478	0,000497671	0,000021901
	AG	0,000640071	0,998927437	0,001205702
	GG	0,000043451	0,000574892	0,998772397

n = 2.554.366

		wahrer Genotyp AA	AT	TT
beobachteter Genotyp	AA	0,999212000	0,000639988	0,000045442
	AT	0,000767998	0,998795095	0,001083278
	TT	0,000020002	0,000564918	0,998871280

AGWH SNP Array 6.0

n = 12.532.072

		wahrer Genotyp AA	AC	CC
beobachteter Genotyp	AA	0,991680565	0,006997736	0,001745060
	AC	0,006467633	0,985502933	0,006536337
	CC	0,001851800	0,007499329	0,991718602

n = 47.721.838

		wahrer Genotyp AA	AG	GG
beobachteter Genotyp	AA	0,991756958	0,006846679	0,001694787
	AG	0,006359841	0,985692457	0,006622274
	GG	0,001883200	0,007460863	0,991682937

n = 8.938.344

		wahrer Genotyp AA	AT	TT
beobachteter Genotyp	AA	0,991881655	0,007006647	0,001897422
	AT	0,006344276	0,986170047	0,006273185
	TT	0,001774067	0,006823305	0,991829391

Berechnung von Fehlerraten

Fehlerraten für die Affymetrix GenomeWide Human SNP Arrays

AGWH SNP Array 5.0

n = 3.876.710

		wahrer Genotyp		
		CC	CG	GG
beobachteter Genotyp	CC	0,999022283	0,000505138	0,000069224
	CG	0,000940740	0,998913927	0,001087718
	GG	0,000036977	0,000580936	0,998843058

n = 13.016.401

		wahrer Genotyp		
		CC	CT	TT
beobachteter Genotyp	CC	0,999156284	0,000598561	0,000053234
	CT	0,000822346	0,998840781	0,000752955
	TT	0,000021370	0,000560659	0,999193811

n = 2.513.860

		wahrer Genotyp		
		GG	GT	TT
beobachteter Genotyp	GG	0,999088985	0,000557975	0,000048494
	GT	0,000894830	0,998795154	0,000797791
	TT	0,000016185	0,000646871	0,999153714

AGWH SNP Array 6.0

n = 12.844.811

		wahrer Genotyp		
		CC	CG	GG
beobachteter Genotyp	CC	0,991338215	0,006971320	0,001742112
	CG	0,006749566	0,985710434	0,006614098
	GG	0,001912217	0,007318244	0,991643788

n = 49.441.210

		wahrer Genotyp		
		CC	CT	TT
beobachteter Genotyp	CC	0,991707697	0,007555624	0,001960166
	CT	0,006606332	0,985743137	0,006366356
	TT	0,001685970	0,006701238	0,991673476

n = 9.612.331

		wahrer Genotyp		
		GG	GT	TT
beobachteter Genotyp	GG	0,992099633	0,007490192	0,001984291
	GT	0,006287102	0,985674343	0,006486723
	TT	0,001613264	0,006835464	0,991596521

Berechnung von Fehlerraten

Fehlerraten für den Illumina HumanHap 550k

n = 2.505.304

beobachteter Genotyp	wahrer Genotyp: AA	AC	CC
AA	0,99991464	0,000001250	0
AC	0,00008536	0,99987495	0,000051944
CC	0	0,000011254	0,999948056

n = 2.585.002

beobachteter Genotyp	wahrer Genotyp: GG	GT	TT
GG	0,999935565	0,000010953	0
GT	0,000064435	0,999986427	0,000011720
TT	0	0,000002620	0,999988280

n = 10.715.910

beobachteter Genotyp	wahrer Genotyp: AA	AG	GG
AA	0,999996805	0,000004430	0
AG	0,000003195	0,999983317	0,000055588
GG	0	0,000012253	0,999944412

n = 11.458.988

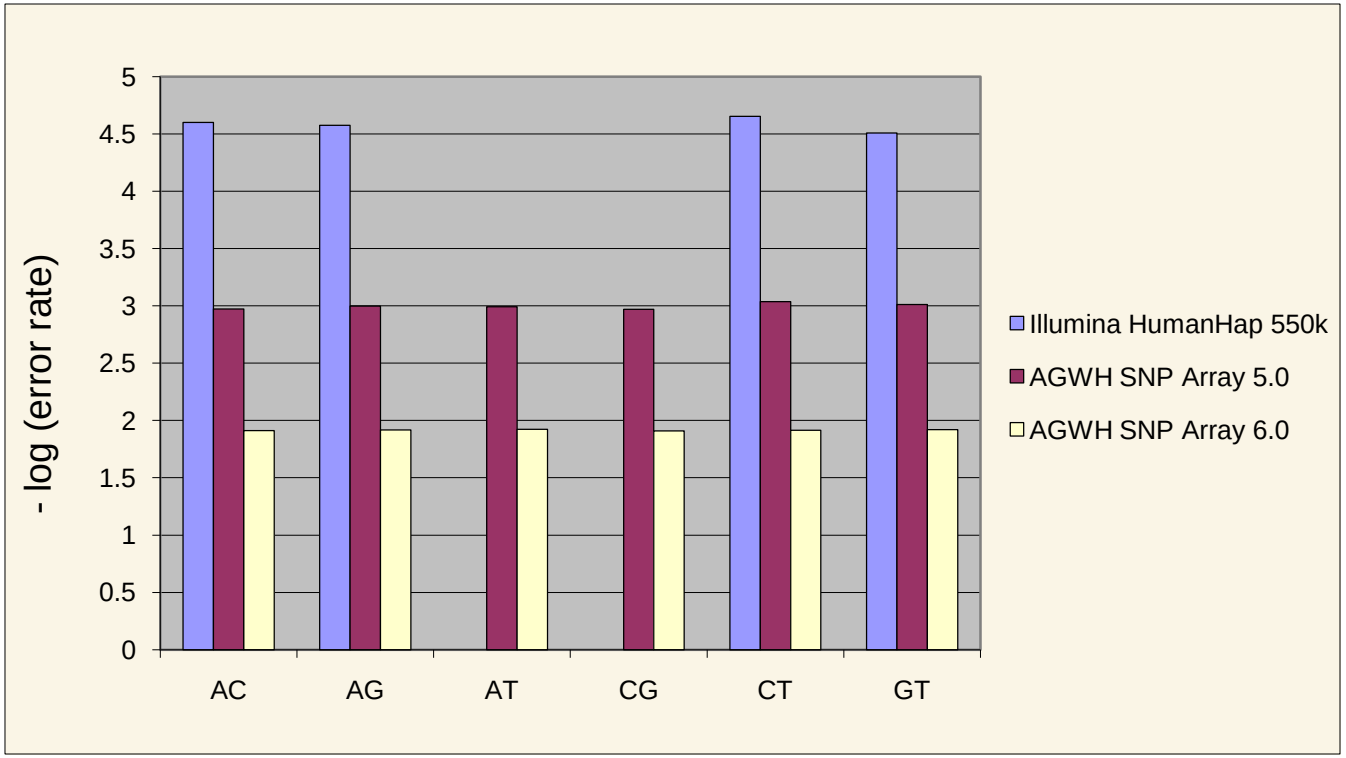
beobachteter Genotyp	wahrer Genotyp: CC	CT	TT
CC	0,999950746	0,000006546	0,000000272
CT	0,000049011	0,999991795	0,000005425
TT	0,000000243	0,000001659	0,999994303

Der Illumina HumanHap 550k enthält keine Marker vom AT- oder CG-Variationstyp.

Vergleich von Fehlerraten

Fehlerraten getrennt nach Variationstypen

„Einfaches Fehlermodell“



Vergleich von Fehlerraten

Fehlerrate pro Genotypisierungschip

GENOTYPISIERUNGSCHIP	Anzahl beobachteter Genotypen | Anzahl falscher Genotypen	FEHLERRATE *bestimmt mittels Majoritätsentscheidung*
AGWH SNP Array 5.0	39.338.000 | 38.800	0,0009863240 = 0,0986 % (99,9014 %)
AGWH SNP Array 6.0	141.090.606 | 1.389.732	0,00984993 = 0,9849 % (99,0151%)
Illumina HumanHap 550k	27.265.204 | 682	0,0000250136 = 0,0025 % (99,9975 %)

UNTERE GRENZE

GENOTYPISIERUNSCHIP	FEHLERRATE *1–Parametermodell (Scheet et al.)* (Fehlerrate pro Genotyp)	FEHLERRATE *Biologisch begründetes Fehlermodell* (Fehlerrate pro Allel)
AGWH SNP Array 5.0	0,0009975	0,0005119
AGWH SNP Array 6.0	0,0098500	0,0055991
Illumina HumanHap 550k	0,0000250	0,0000125

Berechnung von Fehlerraten

Schlussfolgerung

- Der AGWH SNP Array 6.0 besitzt die höchste Fehlerrate (0,98%) von den drei analysierten Genotypisierungschips. Der Illumina HumanHap 550k weist die niedrigste Fehlerrate auf (0,0025%).
- Die Fehlerrate für homozygote / gegen-homozygote Transitionen ist wesentlich niedriger als für homozygote / heterozygote oder heterozygote / homozygote Transitionen. (*Die Wahrscheinlichkeit, dass ein Allel fehltypisiert ist, ist wesentlich höher, als dass beide Allele fehltbestimmt sind.*)
- Die Fehlerrate scheint nahezu unabhängig vom Variationstyp zu sein.
- Die Auswahl des Fehlermodells spielt keine entscheidende Rolle bei der Berechnung der Fehlerrate für die einzelnen Genotypisierungschips.

Teil 2

Analyse von „High-level”-Qualitätsparametern

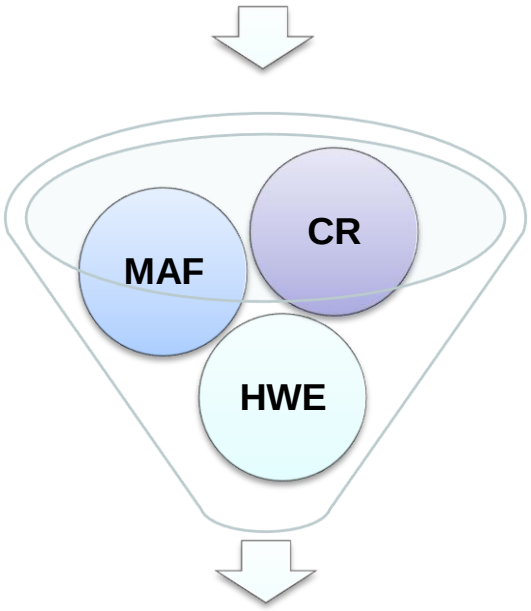
für Hochdurchsatz-Genotypisierungsdaten

SPONSORED BY THE

Federal Ministry
of Education
and Research

Systematische Analyse von Qualitätsparametern

Genotypdaten / Assoziationsergebnisse



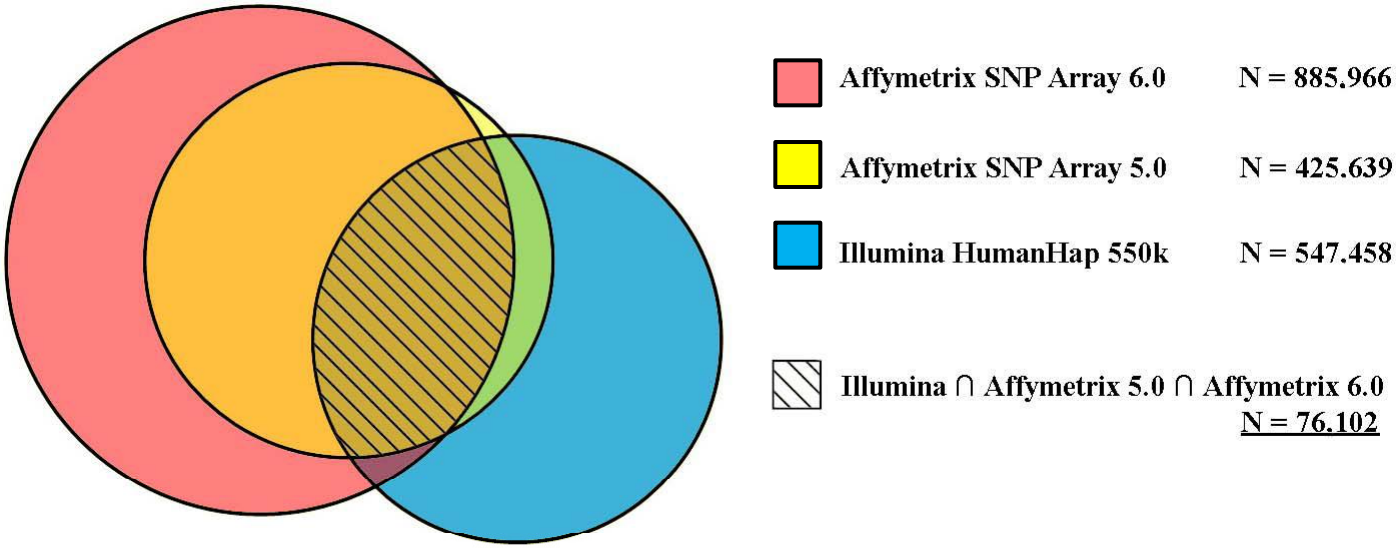
Reduktion falsch positiver und negativer Assoziationsbefunde

Szenarios

- ***Qualitätsparameter***
  - *Callrate:* 0,90, 0,95, 0,98, 0,99, 1,00
  - *MAF:* 0,0001, 0,001, 0,01, 0,05, 0,1
  - *HWE p-Wert:* 0,00001, 0,0001, 0,001, 0,01, 0,05
- ***Fehlermodelle***
  - Einfaches Fehlermodell
  - Biologisch begründetes Fehlermodell
- ***Markersätze***
  - Kompletter Markersatz
  - Schnittmenge der autosomalen Markersätze der drei getesteten Genotypisierungschips

SNP-Content der DNA-Mikroarrays

Schnittmenge der autosmomalen Marker



Systematische Analyse von Qualitätsparametern

Kompletter Markersatz, Einfaches & Biologisch begründetes Fehlermodell

		Call Rate >= 0		Call Rate >= 0,90		all Rate >= 0,95		Call Rate >= 0,98		Call Rate >= 0,99		Call Rate = 1	
		error rate	count	error rate	count	error rate	count	error rate	count	error rate	count	error rate	count
Simple Error Model	Affymetrix v5	0,0009975	39.338.000	0,0005006	38.601.049	0,0004145	38.267.237	0,0002956	37.124.213	0,0002217	35.498.817	0,0001131	27.056.355
	Affymetrix v6	0,0098500	141.090.606	0,0090626	138.876.733	0,0086127	135.354.700	0,0081332	126.211.876	0,0077855	115.405.053	0,0076808	69.758.170
	Illumina	0,0000250	27.265.204	0,0000250	27.246.731	0,0000224	27.209.593	0,0000147	27.005.541	0,0000094	26.584.941	0,0000035	23.530.999
Biological Justified Error Model	Affymetrix v5	0,0005119	39.338.000	0,0002518	38.601.049	0,0002084	38.267.237	0,0001487	37.124.213	0,0001115	35.498.817	0,0000570	27.056.355
	Affymetrix v6	0,0055991	141.085.452	0,0051693	138.876.733	0,0049317	135.354.700	0,0046828	126.211.876	0,0044964	115.405.053	0,0044433	69.758.170
	Illumina	0,0000125	27.265.204	0,0000125	27.246.731	0,0000112	27.209.593	0,0000074	27.005.541	0,0000048	26.584.941	0,0000018	23.530.999

		MAF >=0		MAF >=0,0001		MAF >=0,001		MAF >= 0,01		MAF >= 0,05		MAF >= 0,10	
		error rate	count	error rate	count	error rate	count	error rate	count	error rate	count	error rate	count
Simple Error Model	Affymetrix v5	0,0009975	39.338.000	0,0009808	36.565.020	0,0009787	36.281.694	0,0009589	33.821.023	0,0008546	30.335.839	0,0007421	26.240.821
	Affymetrix v6	0,0098500	141.090.606	0,0104092	132.575.155	0,0104437	132.102.542	0,0109865	122.192.303	0,0115516	109.168.043	0,0122035	93.302.213
	Illumina	0,0000250	27.265.204	0,0000255	26.634.563	0,0000255	26.634.563	0,0000256	26.118.901	0,0000259	24.706.789	0,0000264	21.481.684
Biological Justified Error Model	Affymetrix v5	0,0005119	39.338.000	0,0004971	36.565.020	0,0004956	36.281.694	0,0004831	33.821.023	0,0004303	30.335.839	0,0003737	26.240.821
	Affymetrix v6	0,0055991	141.090.606	0,0059138	132.575.155	0,0059334	132.102.542	0,0062424	122.192.303	0,0065940	109.168.043	0,0070195	93.302.213
	Illumina	0,0000125	27.265.204	0,0000128	26.634.563	0,0000128	26.634.563	0,0000129	26.118.901	0,0000130	24.706.789	0,0000133	21.481.684

		HWE p >= 0		HWE p >= 0,00001		HWE p >= 0,0001		HWE p >= 0,001		HWE p >= 0,01		HWE p >= 0,05	
		error rate	count	error rate	count	error rate	count	error rate	count	error rate	count	error rate	count
Simple Error Model	Affymetrix v5	0,0009975	39.338.000	0,0007401	37.757.730	0,0007203	37.698.572	0,0006956	37.565.874	0,0006599	37.079.886	0,0006188	35.635.382
	Affymetrix v6	0,0098500	141.090.606	0,0089097	133.872.996	0,0088494	133.520.084	0,0087791	132.863.433	0,0086927	130.892.246	0,0085715	125.522.638
	Illumina	0,0000250	27.265.204	0,0000245	27.251.521	0,0000244	27.240.325	0,0000242	27.196.735	0,0000236	26.937.533	0,0000230	25.922.570
Biological Justified Error Model	Affymetrix v5	0,0005119	39.338.000	0,0003776	37.757.730	0,0003678	37.698.572	0,0003554	37.565.874	0,0003375	25.922.570	0,0003173	35.635.382
	Affymetrix v6	0,0055991	141.090.606	0,0050058	133.872.996	0,0049752	133.520.084	0,0049395	132.863.433	0,0048956	130.892.246	0,0048316	125.522.638
	Illumina	0,0000125	27.265.204	0,0000123	27.251.521	0,0000123	27.240.325	0,0000121	27.196.735	0,0000118	26.937.533	0,0000115	25.922.570

Systematische Analyse von Qualitätsparametern

Schnittmenge der autosomalen Marker, Einfaches & Biologisch begründetes Fehlermodell

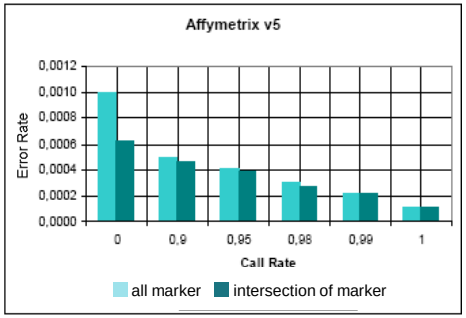
		Call Rate >= 0		Call Rate >= 0,90		all Rate >= 0,95		Call Rate >= 0,98		Call Rate >= 0,99		Call Rate = 1	
		error rate	count	error rate	count	error rate	count	error rate	count	error rate	count	error rate	count
Simple Error Model	Affymetrix v5	0,0006270	6.977.749	0,0004653	6.715.914	0,0003896	6.661.652	0,0002774	6.464.646	0,0002129	6.179.917	0,0001060	4.648.756
	Affymetrix v6	0,0106273	11.801.676	0,0101713	11.690.110	0,0099062	11.506.932	0,0096606	10.993.617	0,0094791	10.334.522	0,0097535	6.783.409
	Illumina	0,0000163	3.796.222	0,0000163	3.795.619	0,0000150	3.792.853	0,0000103	3.776.078	0,0000072	3.737.587	0,0000043	3.383.635
Biological Justified Error Model	Affymetrix v5	0,0003149	6.977.749	0,0002332	6.715.914	0,0001950	6.661.652	0,0001388	6.464.646	0,0001065	6.179.917	0,0000530	4.648.756
	Affymetrix v6	0,0059541	11.801.676	0,0057215	11.690.110	0,0055897	11.506.932	0,0054733	10.993.617	0,0053821	10.334.522	0,0055506	6.783.409
	Illumina	0,0000083	3.796.222	0,0000083	3.795.619	0,0000076	3.792.853	0,0000053	3.776.078	0,0000037	3.737.587	0,0000024	3.383.635

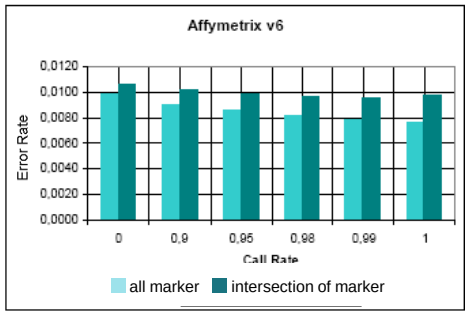
		MAF >=0		MAF >=0,0001		MAF >=0,001		MAF >= 0,01		MAF >= 0,05		MAF >= 0,1	
		error rate	count	error rate	count	error rate	count	error rate	count	error rate	count	error rate	count
Simple Error Model	Affymetrix v5	0,0006270	6.977.749	0,0006163	6.651.927	0,0006163	6.651.927	0,0006138	6.492.807	0,0005824	6.151.920	0,0005393	5.372.987
	Affymetrix v6	0,0106273	11.801.676	0,0107713	11.609.877	0,0107713	11.609.877	0,0109711	11.251.943	0,0113167	10.535.163	0,0118948	9.212.553
	Illumina	0,0000163	3.796.222	0,0000167	3.718.418	0,0000167	3.718.418	0,0000170	3.649.114	0,0000176	3.460.740	0,0000182	3.024.531
Biological Justified Error Model	Affymetrix v5	0,0003149	6.977.749	0,0003091	6.651.927	0,0003091	6.651.927	0,0003074	6.492.807	0,0002917	6.151.920	0,0002700	5.372.987
	Affymetrix v6	0,0059541	11.801.676	0,0060364	11.609.877	0,0060364	11.609.877	0,0061551	11.251.943	0,0063695	10.535.163	0,0067398	9.212.553
	Illumina	0,0000083	3.796.222	0,0000085	3.718.418	0,0000085	3.718.418	0,0000086	3.649.114	0,0000090	3.460.740	0,0000093	3.024.531

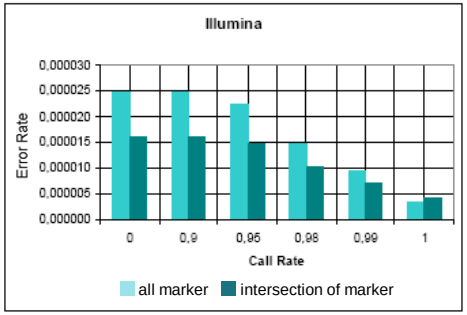
		HWE p >= 0		HWE p >= 0,00001		HWE p >= 0,0001		HWE p >= 0,001		HWE p >= 0,01		HWE p >= 0,05	
		error rate	count	error rate	count	error rate	count	error rate	count	error rate	count	error rate	count
Simple Error Model	Affymetrix v5	0,0006270	6.977.749	0,0004938	6.701.419	0,0004882	6.692.943	0,0004768	6.672.387	0,0004588	6.582.614	0,0004427	6.299.388
	Affymetrix v6	0,0106273	11.801.676	0,0099541	11.688.174	0,0099189	11.665.963	0,0098739	11.618.782	0,0098238	11.450.938	0,0097724	10.947.471
	Illumina	0,0000163	3.796.222	0,0000150	3.794.822	0,0000145	3.793.468	0,0000145	3.788.089	0,0000144	3.751.985	0,0000141	3.610.696
Biological Justified Error Model	Affymetrix v5	0,0003149	6.977.749	0,0002478	6.701.419	0,0002449	6.692.943	0,0002393	6.672.387	0,0002302	6.582.614	0,0002223	6.299.388
	Affymetrix v6	0,0059541	11.801.676	0,0056111	11.688.174	0,0055941	11.665.963	0,0055722	11.618.782	0,0055492	11.450.938	0,0055244	10.947.471
	Illumina	0,0000083	3.796.222	0,0000076	3.794.822	0,0000074	3.793.468	0,0000074	3.788.089	0,0000073	3.751.985	0,0000072	3.610.696

Systematische Analyse von Qualitätsparametern

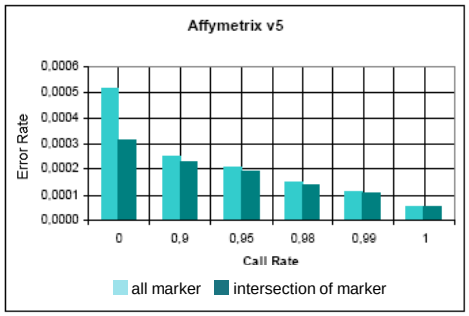
Callrate, Einfaches Fehlermodell

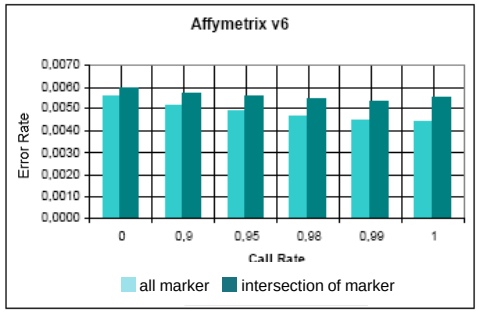


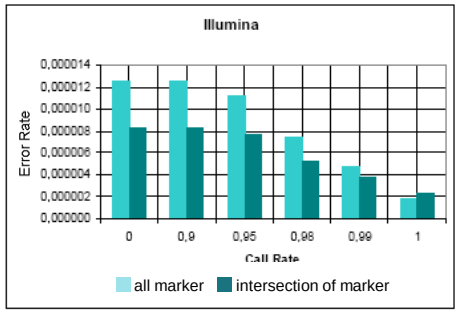




Callrate, Biologisch begründetes Fehlermodell

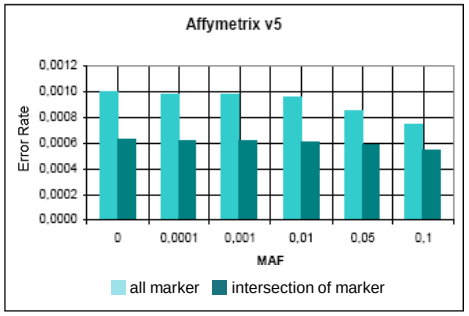


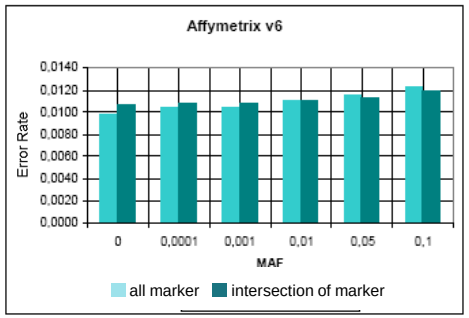


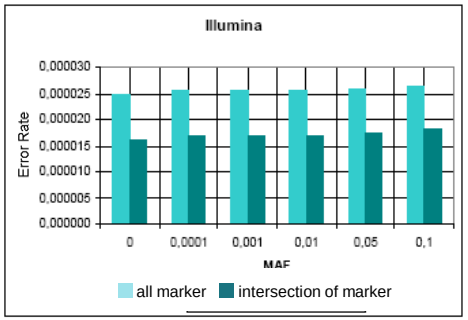


Systematische Analyse von Qualitätsparametern

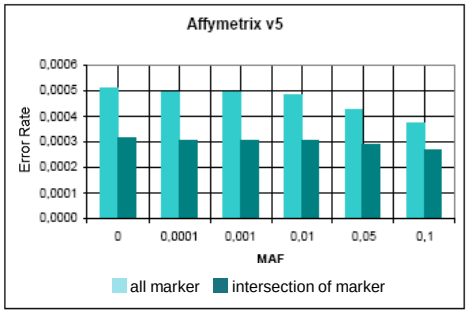
MAF, Einfaches Fehlermodell

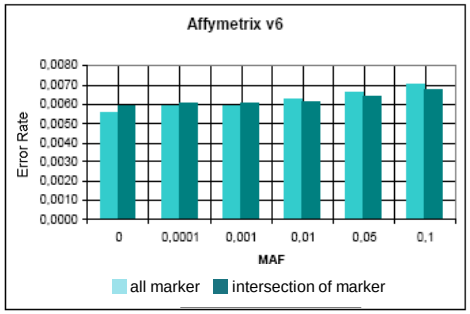


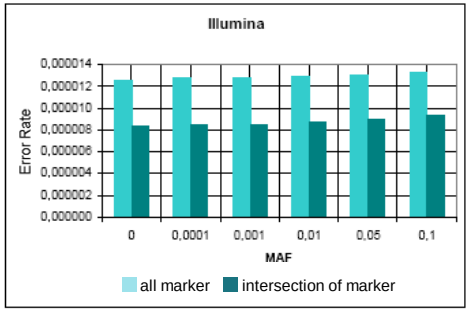




MAF, Biologisch begründetes Fehlermodell

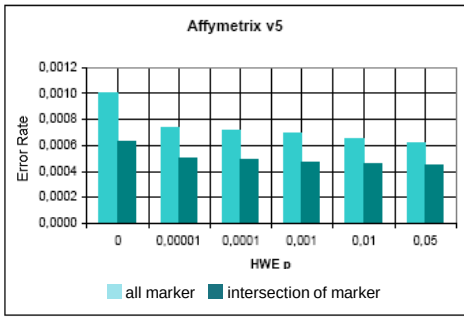


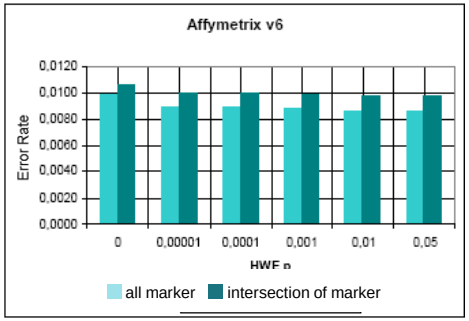


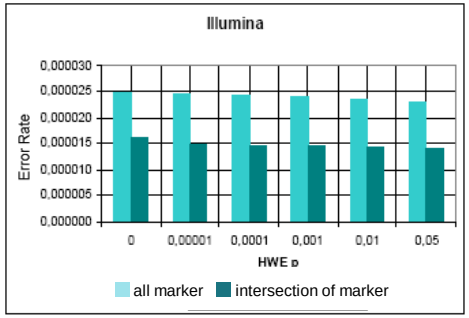


Systematische Analyse von Qualitätsparametern

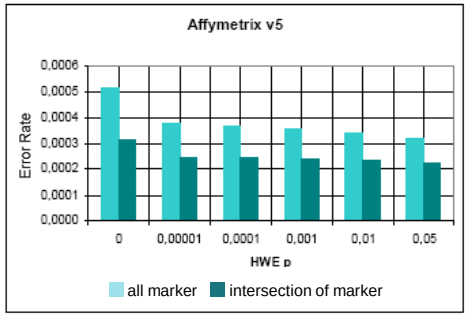
HWE (p-Wert), Einfaches Fehlermodell

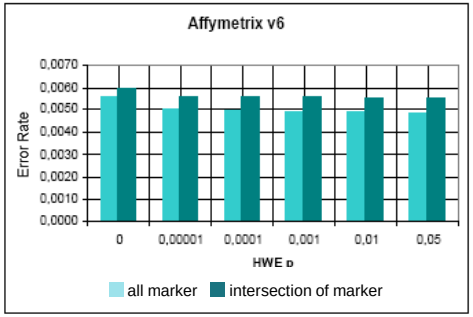


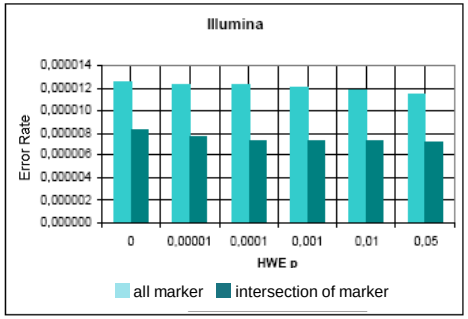




HWE (p-Wert), Biologisch begründetes Fehlermodell







Systematische Analyse von Qualitätsparametern

Kompletter Markersatz, Einfaches Fehlermodell

		Call Rate	1.00		1.00		1.00	
		HWE p	----		0.05		0.05	
		MAF	----		----		0.1	
			error rate	count	error rate	count	error rate	count
Simple Error Model	Affymetrix v5		0.0001131	27,056,355	0.0001107	25,125,055	0.0001195	15,887,356
	Affymetrix v6		0.0076808	69,758,170	0.0076417	63,895,223	0.0106205	40,943,733
	Illumina		0.0000035	23,530,999	0.0000034	22,402,295	0.0000031	17,336,614

Schlussfolgerung

- Effekt der “High-Level”-Qualitätsparameter auf die Fehlerrate: CR >> HWE, MAF
- Im Allgemeinen führt ein restriktiverer Schwellenwert eines Qualitätskriteriums zu einer niedrigeren Fehlerrate (Ausnahme: MAF, Affymetrix SNP 6.0).
- Die Verwendung mehrerer Qualitätskriterien macht Sinn, insbesondere für den HH550k Chip.
- „High Level”-Qualitätsparameter sind nahezu ineffizient, die Fehlerrate des AGWH SNP 6.0 zu senken.

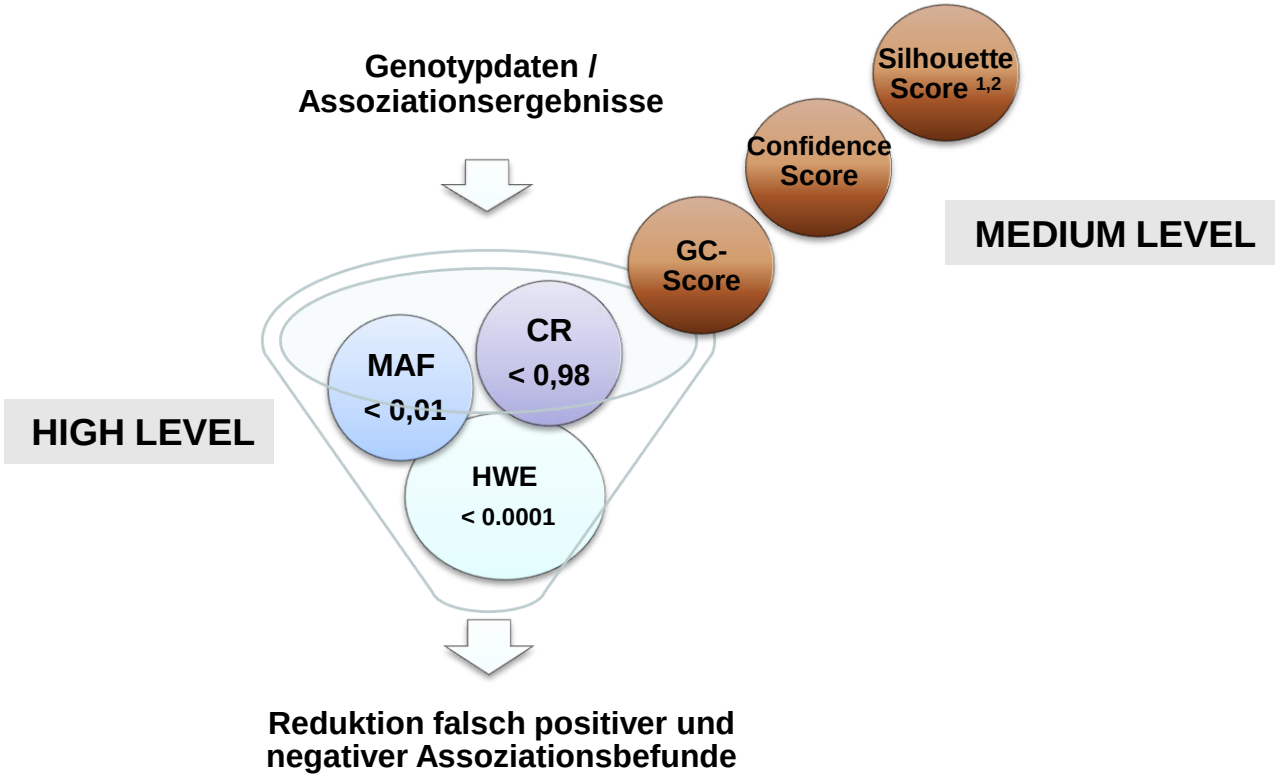
Teil 3

Analyse des Silhouette-Score zur Cluster-Validierung

SPONSORED BY THE

Federal Ministry
of Education
and Research

High, Medium und Low Level-Qualitätskriterien



1) *Lovmar L, Ahlford A, Jonsson M, Syvänen AC. (2005) „Silhouette scores for assessment of SNP genotype clusters". BMC Genomics. Mar 10;6(1):35.*

2) *Rousseeuw, P. (1987). "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis". J Comput Appl Math 20, 1, 53-65*

3) *Attia J et al. (2010) „Detecting genotyping error using measures of degree of Hardy-Weinberg disequilibrium". Stat Appl Genet Mol Biol. 5;9(1).*

Medium (Second) Level-Qualitätskriterien

Getestetes Programm: **„ClusterA”** (*Lovmar et al*. (*2005*))

- Verbesserung der Qualitätskontrolle von Genotypisierungsdaten unabhängig von bereits bestehenden Qualitätsparametern
- generiert für jeden Marker einen Silhouette-Score zwischen -1 und 1
- alle Genotypen eines Markers mit einem Score > 0,65 gelten als valide
- SNP Assays mit einem Score < 0,25 werden als falsch betrachtet

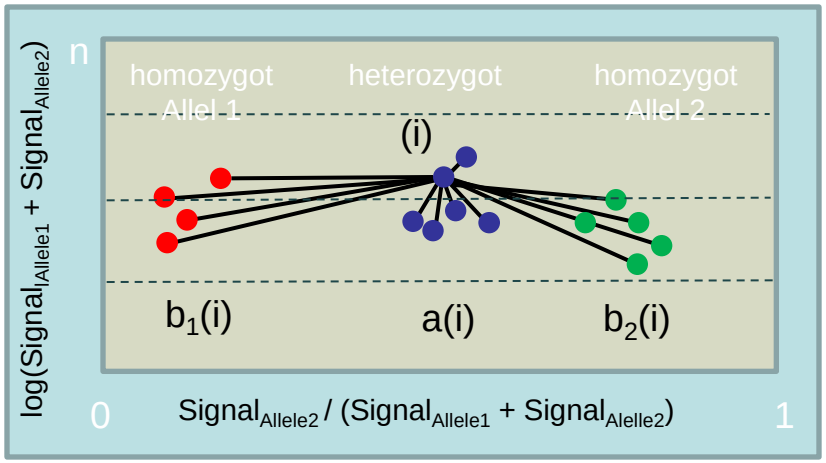
Technische Einzelheiten:

- „ClusterA” gibt einen Silhouette-Score nur in Verbindung mit einer gleichzeitigen Neuberechnung eines Clusterplots aus. Es existiert keine Option, den Silhouette-Score für eine bereits existierende Klassifizierung zu bestimmen.
- läuft nur unter Windows Betriebssystemen, da in Visual Basic geschrieben
- kein Aufruf von der Kommandozeile aus möglich; keine parallele Ausführung
- lange Laufzeit aufgrund des impliziten Clustering

→ Implementierung eines eigenen Programms (“**SILHOUETTE**”) zur effizienten Berechnung des Silhouette-Scores aus den Positionskoordinaten eines Clusterplots.

Berechnung des Silhouette-Score

1. Für jeden Datenpunkt i des Clusterplots wird dessen ***Silhouette*** s(i) wie folgt bestimmt:



$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

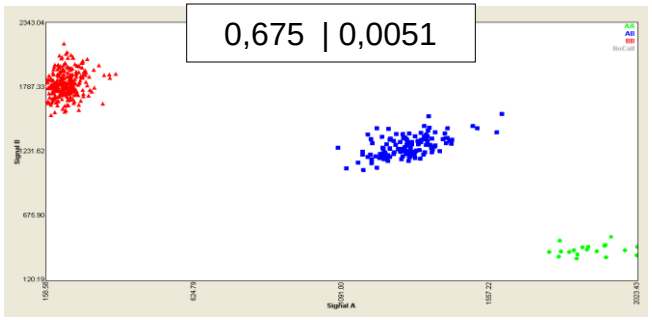
$$b(i) = \min\{b_1(i), b_2(i)\}$$

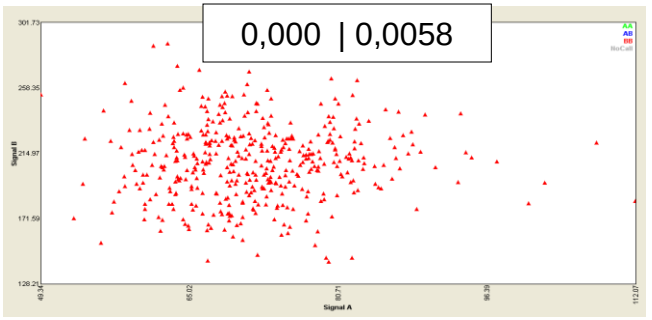
a(i) := mittlere Distanz von i zu allen anderen Punkten desselben Clusters
b(i) := mittlere Distanz von i zu allen Punkten des nächstgelegenen Clusters, entweder $b_1(i)$ oder $b_2(i)$

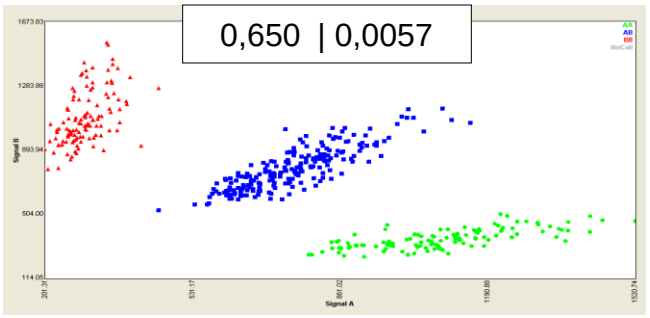
2. Die ***mittlere Silhouettenweite*** entspricht dem Mittelwert aller s(i) eines Genotyp-Clusters.
3. Der ***Silhouette-Score*** (*average overall silhouette width*) eines Clusterplots entspricht dem gewichteten Mittel der *mittleren Silhouettenweiten* über alle Genotyp-Cluster.

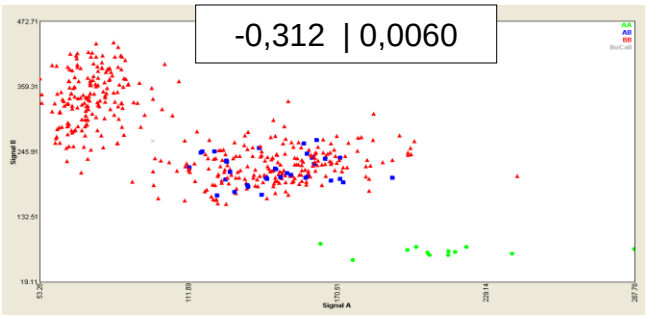
Beispiele von Cluster-Plots

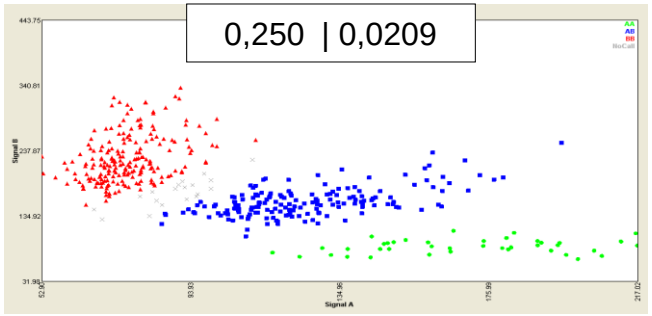
Silhouette- & Confidence-Score

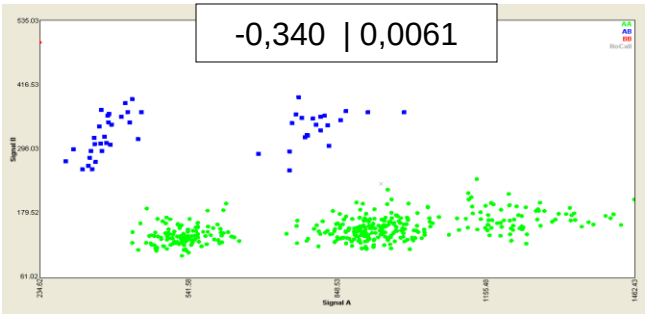




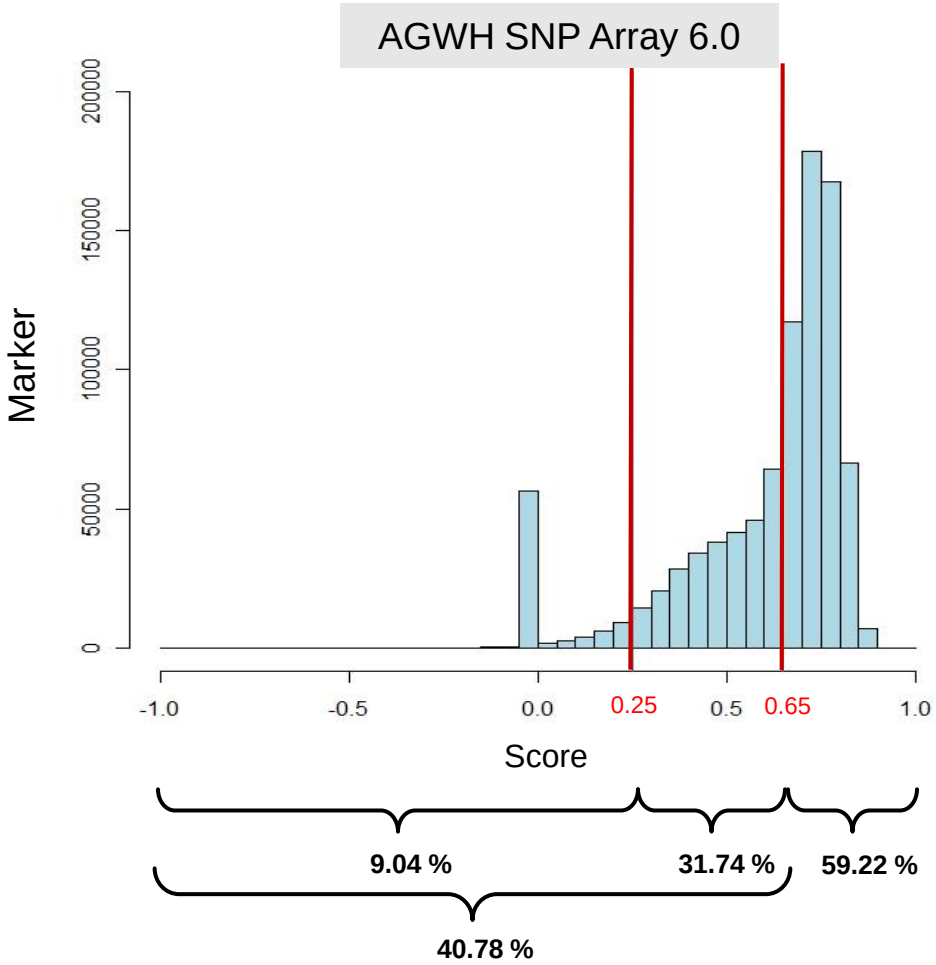








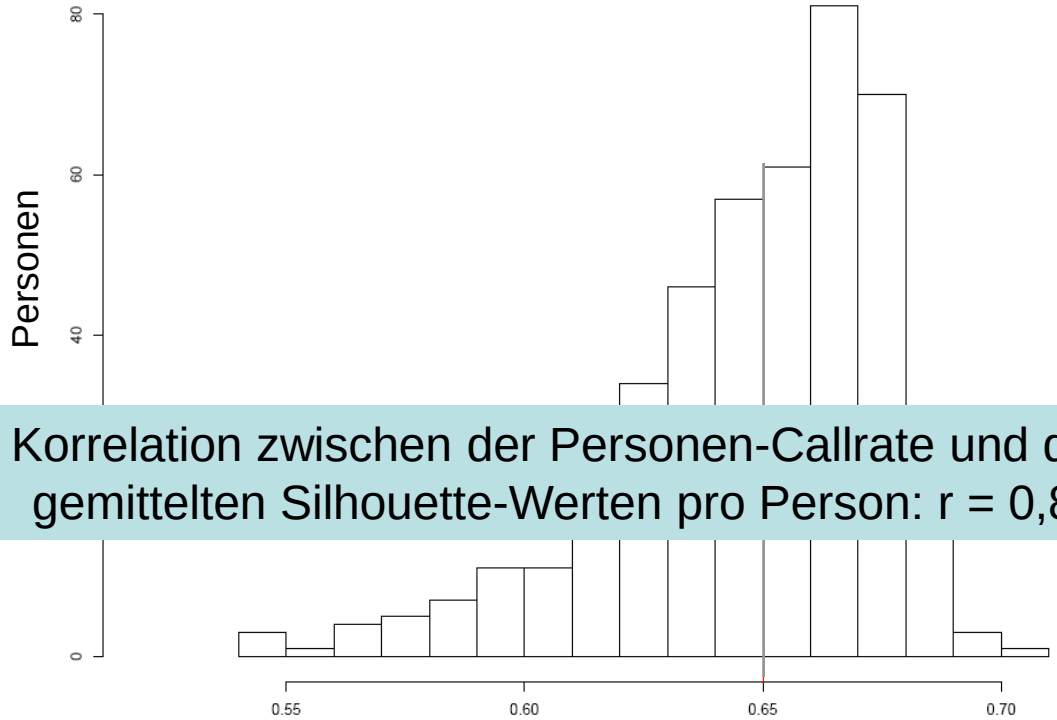
Histogramm des Silhouette-Score



n = 906600

Histogramm der gemittelten Silhouette-Werte pro Person

AGWH SNP Array 6.0



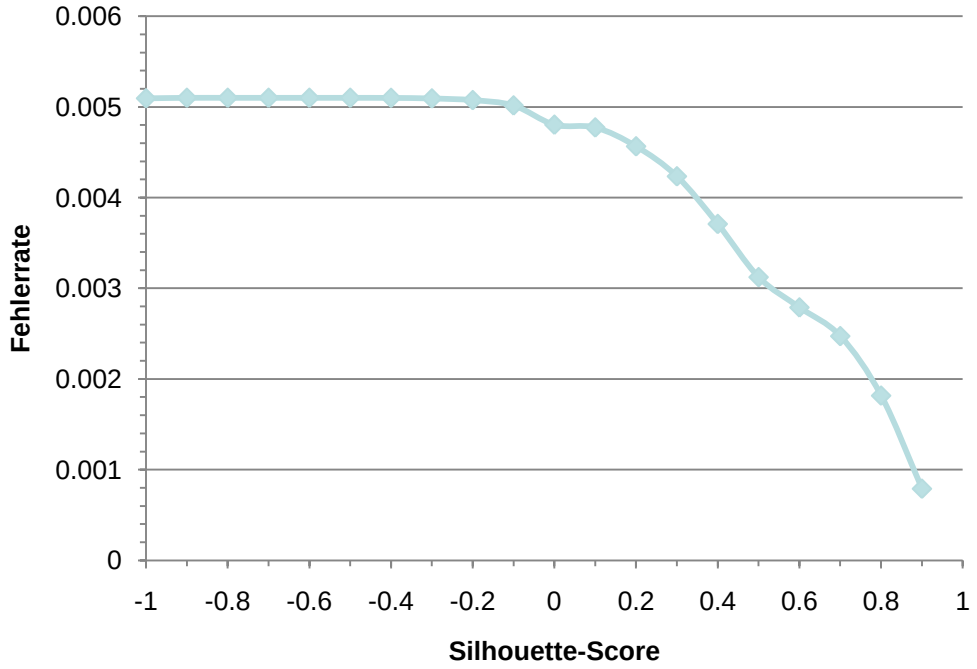
n = 435

Korrelation zwischen der Personen-Callrate und den gemittelten Silhouette-Werten pro Person: r = 0,81

gemittelte Silhouette-Werte über alle Marker pro Person

Analyse des Silhouette-Score

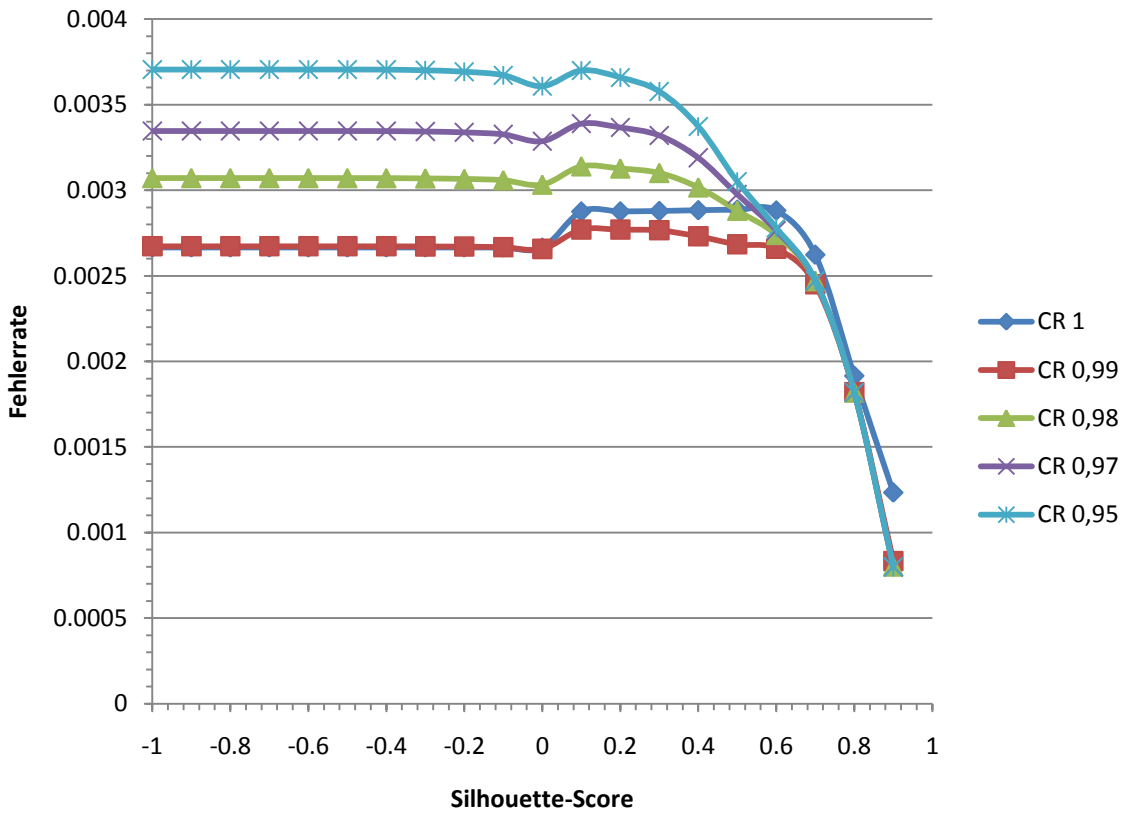
AGWH SNP Array 6.0



Silhouette-Score	Anzahl der in die Fehlerratenberechnung eingegangenen Genotypen
-1	141.986.389
-0,9	141.986.389
-0,8	141.986.389
-0,7	141.986.389
-0,6	141.986.343
-0,5	141.985.958
-0,4	141.983.046
-0,3	141.965.489
-0,2	141.901.535
-0,1	141.689.672
0	141.129.613
0,1	134.612.937
0,2	132.118.584
0,3	126.771.135
0,4	115.910.898
0,5	99.188.626
0,6	78.965.479
0,7	49.164.353
0,8	6.195.148
0,9	20.513

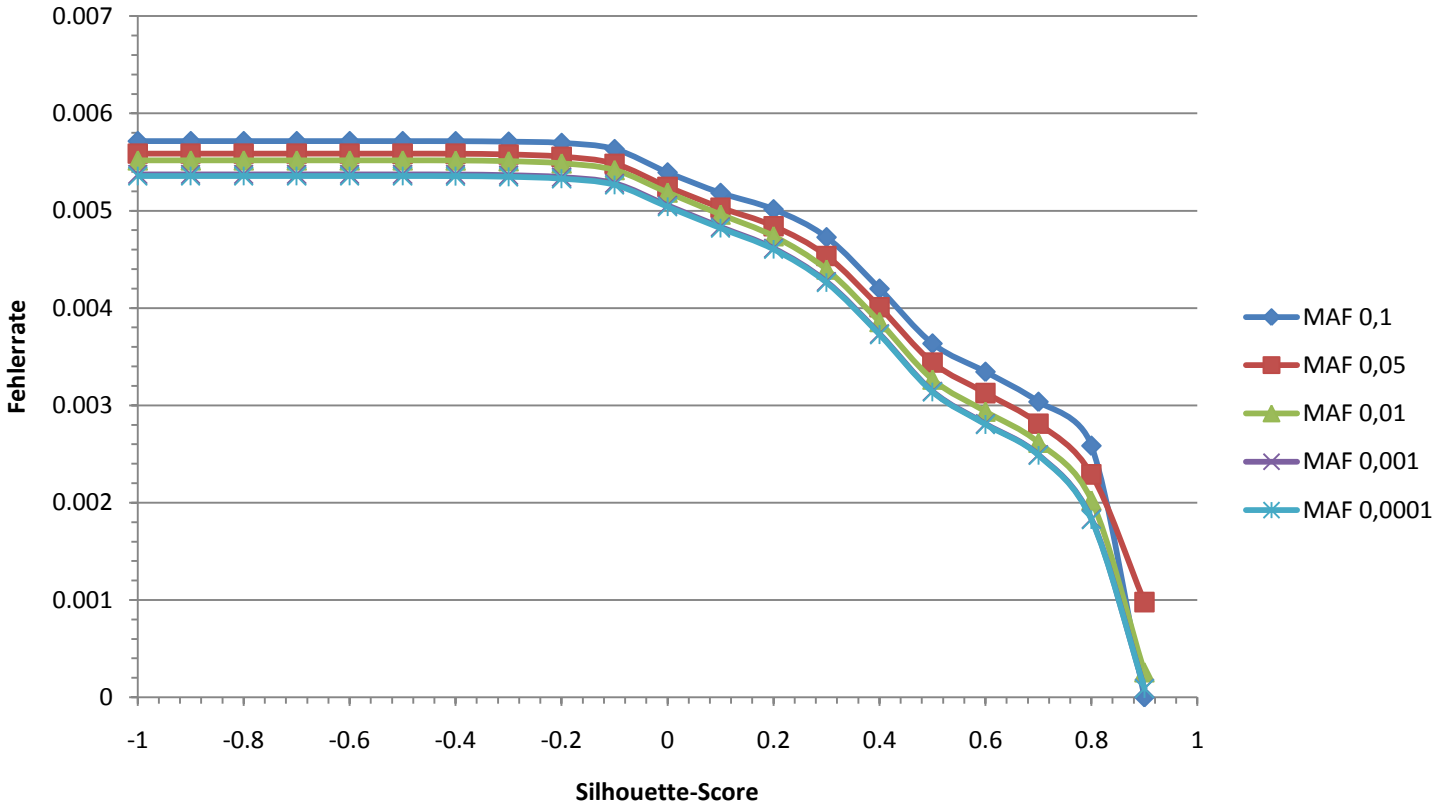
Analyse des Silhouette-Score
in Kombination mit der Callrate als QC-Parameter

AGWH SNP Array 6.0

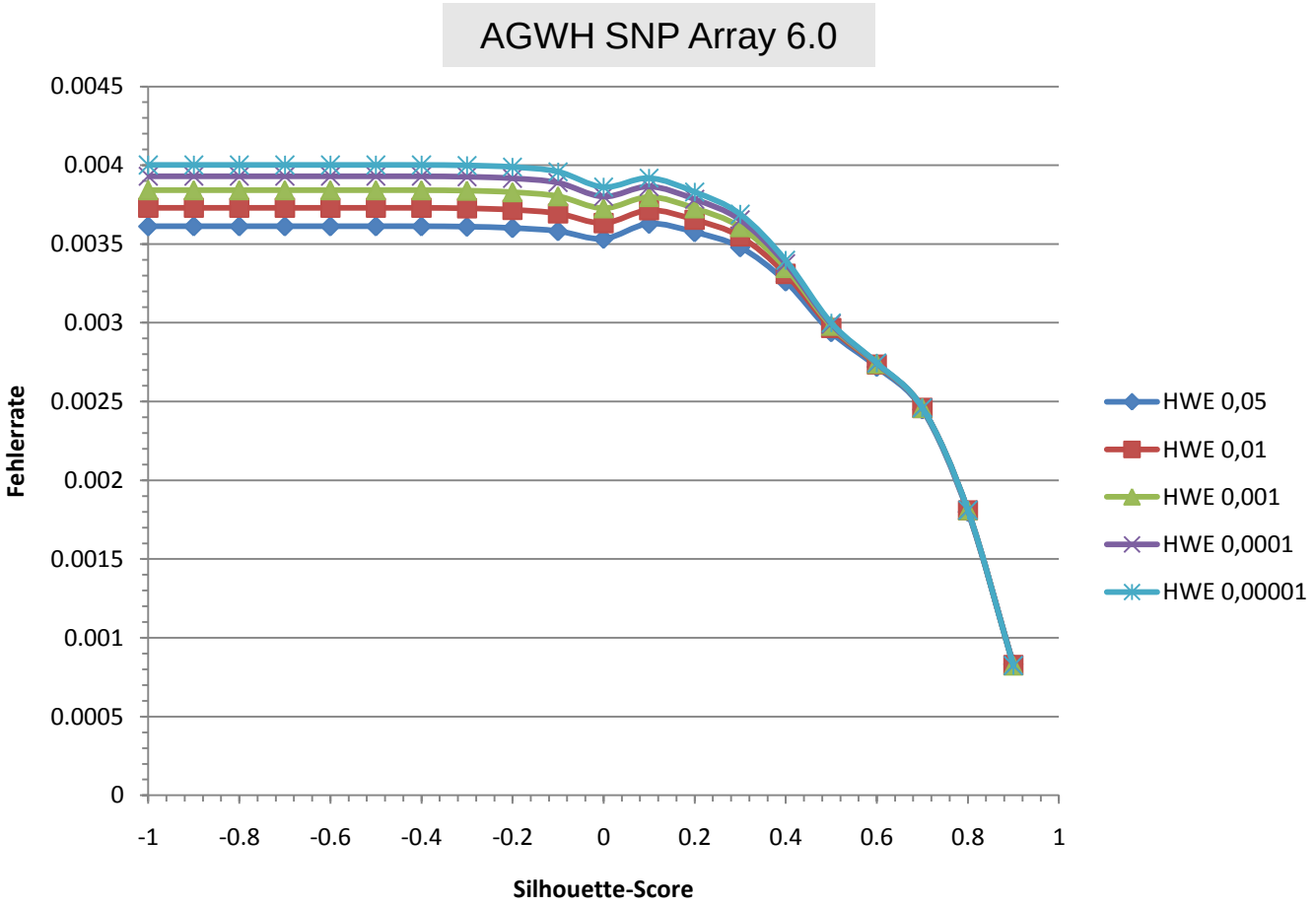


Analyse des Silhouette-Score
in Kombination mit der MAF als QC-Parameter

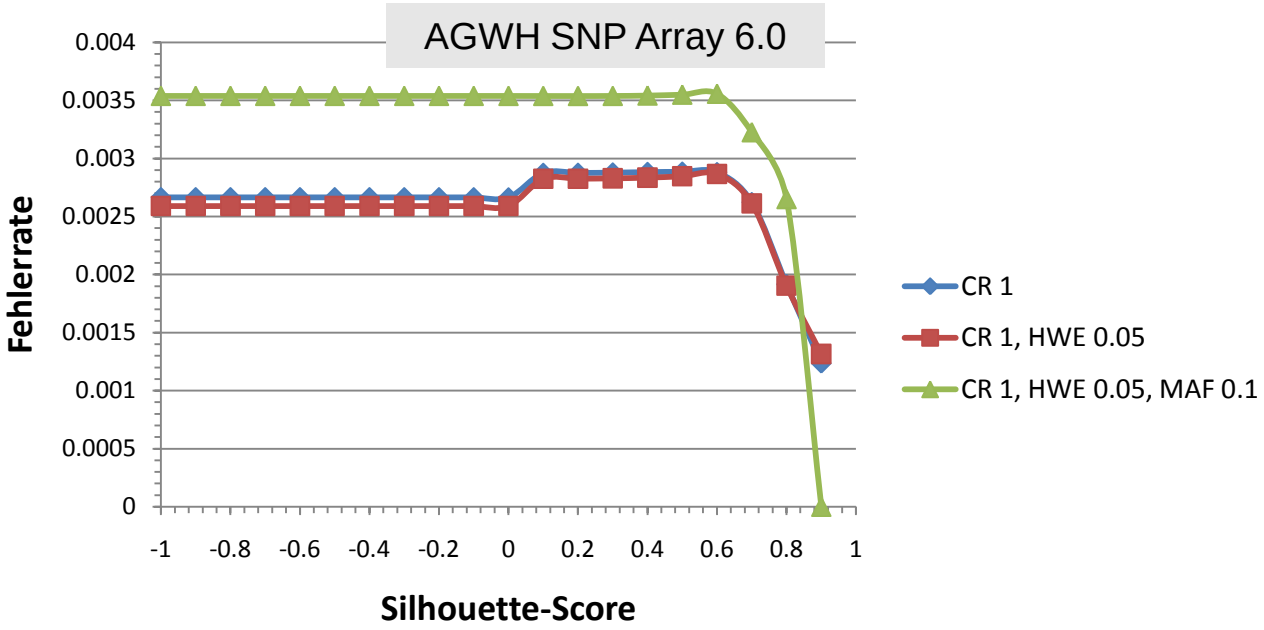
AGWH SNP Array 6.0



Analyse des Silhouette-Score
in Kombination mit HWE QC-Kriterium



Analyse des Silhouette-Score



Schlussfolgerung

- Der Silhouette-Score kann fehltypisierte Genotypen identifizieren, die nicht mit den klassischen „High-Level"-Qualitätskriterien erkannt werden.

Danksagung

Daniela Holler

Marina Angisch

Sudesha Johann

Caroline Pawlak

SPONSORED BY THE
Federal Ministry
of Education
and Research