Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing

Hansoo Park, Jong-Il Kim, Young Seok Ju, Omer Gokcumen, Ryan E. Mills, Sheehyun Kim, Seungbok Lee, Dongwhan Suh ,Dongwan Hong, Hyunseok Peter Kang, Yun Joo Yoo, Jong-Yeon Shin, Hyun-Jin Kim, Maryam Yavartanoo, Young Wha Chang, Jung-Sook Ha, Wilson Chong, Ga-Ram Hwang, Katayoon Darvishi, HyeRan Kim, Song Ju Yang, Kap-Seok Yang, Hyungtae Kim, Matthew E. Hurles, Stephen W. Scherer, Nigel P. Carter, Chris Tyler-Smith, Charles Lee & Jeong-Sun Seo

Nature Genetics: doi:10.1038/ng.555

Supplementary Materials

Nature Genetics: doi:10.1038/ng.555

Nature Genetics: doi:10.1038/ng.555

A. Suppl. Table1. Filtering set training for aCGH with AK1 genome sequence data.

Filter ID	Filtering Condition												Optimization Score		
	CNV < 5000bp						CNV >= 5000bp								
	(1) minimum log2 ratio for low CNV	(2) minimum log2 ratio for middle CNV	(3) minimum log2 ratio for high CNV	threshold p-value (1)	threshold p-value (2)	threshold p-value (3)	(4) minimum log2 ratio for low CNV	(5) minimum log2 ratio for middle CNV	(6) minimum log2 ratio for high CNV	threshold p-value (4)	threshold p-value (5)	threshold p-value (6)	relative sensitivity	PPV[a]	sum of the two
final optimized filter	0.35	0.50	0.70	1.00E-21	1.00E-14	1.00E-08	0.3	0.5	0.7	1.00E-21	1.00E-08	1.00E-08	0.845	0.840	1.685
filter1	0.2	0.50	0.70	1.00E-21	1.00E-14	1.00E-08	0.3	0.5	0.7	1.00E-21	1.00E-08	1.00E-08	0.860	0.793	1.653
filter2	0.25	0.50	0.70	1.00E-21	1.00E-14	1.00E-08	0.3	0.5	0.7	1.00E-21	1.00E-08	1.00E-08	0.857	0.805	1.662
filter3	0.3	0.50	0.70	1.00E-21	1.00E-14	1.00E-08	0.3	0.5	0.7	1.00E-21	1.00E-08	1.00E-08	0.852	0.824	1.676
filter4	0.4	0.50	0.70	1.00E-21	1.00E-14	1.00E-08	0.3	0.5	0.7	1.00E-21	1.00E-08	1.00E-08	0.832	0.850	1.682
filter5	0.45	0.50	0.70	1.00E-21	1.00E-14	1.00E-08	0.3	0.5	0.7	1.00E-21	1.00E-08	1.00E-08	0.825	0.855	1.680
filter6	0.5	0.50	0.70	1.00E-21	1.00E-14	1.00E-08	0.3	0.5	0.7	1.00E-21	1.00E-08	1.00E-08	0.820	0.860	1.680
filter7	0.35	0.50	0.70	1.00E-19	1.00E-14	1.00E-08	0.3	0.5	0.7	1.00E-21	1.00E-08	1.00E-08	0.845	0.833	1.678
filter8	0.35	0.50	0.70	1.00E-17	1.00E-14	1.00E-08	0.3	0.5	0.7	1.00E-21	1.00E-08	1.00E-08	0.849	0.814	1.663
filter9	0.35	0.50	0.70	1.00E-15	1.00E-14	1.00E-08	0.3	0.5	0.7	1.00E-21	1.00E-08	1.00E-08	0.849	0.814	1.663
filter10	0.35	0.50	0.70	1.00E-23	1.00E-14	1.00E-08	0.3	0.5	0.7	1.00E-21	1.00E-08	1.00E-08	0.838	0.847	1.685
filter11	0.35	0.50	0.70	1.00E-25	1.00E-14	1.00E-08	0.3	0.5	0.7	1.00E-21	1.00E-08	1.00E-08	0.835	0.854	1.689
filter12	0.35	0.50	0.70	1.00E-27	1.00E-14	1.00E-08	0.3	0.5	0.7	1.00E-21	1.00E-08	1.00E-08	0.829	0.856	1.685
filter13	0.35	0.50	0.70	1.00E-21	1.00E-12	1.00E-08	0.3	0.5	0.7	1.00E-21	1.00E-08	1.00E-08	0.854	0.797	1.651
filter14	0.35	0.50	0.70	1.00E-21	1.00E-10	1.00E-08	0.3	0.5	0.7	1.00E-21	1.00E-08	1.00E-08	0.872	0.721	1.594
filter15	0.35	0.50	0.70	1.00E-21	1.00E-08	1.00E-08	0.3	0.5	0.7	1.00E-21	1.00E-08	1.00E-08	0.920	0.578	1.497
filter16	0.35	0.50	0.70	1.00E-21	1.00E-16	1.00E-08	0.3	0.5	0.7	1.00E-21	1.00E-08	1.00E-08	0.820	0.860	1.680
filter17	0.35	0.50	0.70	1.00E-21	1.00E-18	1.00E-08	0.3	0.5	0.7	1.00E-21	1.00E-08	1.00E-08	0.811	0.872	1.683
filter18	0.35	0.50	0.70	1.00E-21	1.00E-20	1.00E-08	0.3	0.5	0.7	1.00E-21	1.00E-08	1.00E-08	0.804	0.876	1.681
filter19	0.35	0.50	0.70	1.00E-21	1.00E-14	1.00E-07	0.3	0.5	0.7	1.00E-21	1.00E-08	1.00E-08	0.852	0.834	1.686

PPV[a] ,Positive predictive value

Nature Genetics: doi:10.1038/ng.555

A. Suppl. Table1. Filtering set training for aCGH with AK1 genome sequence data *(continued).*

Filter ID	Filtering Condition												Optimization Score		
	CNV < 5000bp						CNV >= 5000bp								
	(1) minimum log2 ratio for low CNV	(2) minimum log2 ratio for middle CNV	(3) minimum log2 ratio for high CNV	threshold p-value (1)	threshold p-value (2)	threshold p-value (3)	(4) minimum log2 ratio for low CNV	(5) minimum log2 ratio for middle CNV	(6) minimum log2 ratio for high CNV	threshold p-value (4)	threshold p-value (5)	threshold p-value (6)	relative sensitivity	PPV	sum of the two
final optimized filter	0.35	0.50	0.70	1.00E-21	1.00E-14	1.00E-08	0.3	0.5	0.7	1.00E-21	1.00E-08	1.00E-08	0.845	0.840	1.685
filter20	0.35	0.50	0.70	1.00E-21	1.00E-14	1.00E-10	0.3	0.5	0.7	1.00E-21	1.00E-08	1.00E-08	0.840	0.848	1.688
filter21	0.35	0.50	0.70	1.00E-21	1.00E-14	1.00E-12	0.3	0.5	0.7	1.00E-21	1.00E-08	1.00E-08	0.836	0.859	1.695
filter22	0.35	0.50	0.70	1.00E-21	1.00E-14	1.00E-14	0.3	0.5	0.7	1.00E-21	1.00E-08	1.00E-08	0.829	0.863	1.692
filter23	0.35	0.50	0.70	1.00E-21	1.00E-14	1.00E-08	0.2	0.5	0.7	1.00E-21	1.00E-08	1.00E-08	0.857	0.790	1.647
filter24	0.35	0.50	0.70	1.00E-21	1.00E-14	1.00E-08	0.25	0.5	0.7	1.00E-21	1.00E-08	1.00E-08	0.852	0.829	1.680
filter25	0.35	0.50	0.70	1.00E-21	1.00E-14	1.00E-08	0.35	0.5	0.7	1.00E-21	1.00E-08	1.00E-08	0.840	0.843	1.683
filter26	0.35	0.50	0.70	1.00E-21	1.00E-14	1.00E-08	0.4	0.5	0.7	1.00E-21	1.00E-08	1.00E-08	0.831	0.841	1.672
filter27	0.35	0.50	0.70	1.00E-21	1.00E-14	1.00E-08	0.45	0.5	0.7	1.00E-21	1.00E-08	1.00E-08	0.820	0.841	1.660
filter28	0.35	0.50	0.70	1.00E-21	1.00E-14	1.00E-08	0.5	0.5	0.7	1.00E-21	1.00E-08	1.00E-08	0.802	0.838	1.639
filter29	0.35	0.50	0.70	1.00E-21	1.00E-14	1.00E-08	0.3	0.5	0.7	1.00E-19	1.00E-08	1.00E-08	0.847	0.837	1.684
filter30	0.35	0.50	0.70	1.00E-21	1.00E-14	1.00E-08	0.3	0.5	0.7	1.00E-17	1.00E-08	1.00E-08	0.850	0.833	1.683
filter31	0.35	0.50	0.70	1.00E-21	1.00E-14	1.00E-08	0.3	0.5	0.7	1.00E-15	1.00E-08	1.00E-08	0.850	0.833	1.683
filter32	0.35	0.50	0.70	1.00E-21	1.00E-14	1.00E-08	0.3	0.5	0.7	1.00E-23	1.00E-08	1.00E-08	0.840	0.840	1.681
filter33	0.35	0.50	0.70	1.00E-21	1.00E-14	1.00E-08	0.3	0.5	0.7	1.00E-25	1.00E-08	1.00E-08	0.839	0.840	1.679
filter34	0.35	0.50	0.70	1.00E-21	1.00E-14	1.00E-08	0.3	0.5	0.7	1.00E-27	1.00E-08	1.00E-08	0.838	0.840	1.678
filter35	0.35	0.50	0.70	1.00E-21	1.00E-14	1.00E-08	0.3	0.5	0.7	1.00E-21	1.00E-07	1.00E-08	0.846	0.833	1.679
filter36	0.35	0.50	0.70	1.00E-21	1.00E-14	1.00E-08	0.3	0.5	0.7	1.00E-21	1.00E-10	1.00E-08	0.843	0.844	1.688
filter37	0.35	0.50	0.70	1.00E-21	1.00E-14	1.00E-08	0.3	0.5	0.7	1.00E-21	1.00E-14	1.00E-08	0.842	0.849	1.691
filter38	0.35	0.50	0.70	1.00E-21	1.00E-14	1.00E-08	0.3	0.5	0.7	1.00E-21	1.00E-08	1.00E-07	0.846	0.840	1.686

Nature Genetics: doi:10.1038/ng.555

B. Suppl. Table 2. Final filter conditions for CNV calling.

.

a. Optimized filter conditions of CNV calls for AK1

Criteria	Subset	p-value	TP[a]	FP[b]	PPV	relative sensitivity	overall PPV	overall relative sensitivity
Length <5000bp, minimum llog2 ratiol ≥0.35	0.5>|log2 ratio| ≥0.35	<=E-21	18	20	0.474	0.462	0.840	0.845
	0.7>|log2 ratio| ≥0.5	<=E-14	83	42	0.664	0.546		
	|log2 ratio| ≥0.7	<=E-8	322	35	0.902	0.985		
Length ≥ 5000bp, minimum llog2 ratiol ≥0.30	0.5>|log2 ratio| ≥0.3	<=E-21	31	4	0.886	0.674		
	0.7>|log2 ratio| ≥0.5	<=E-8	61	13	0.824	0.984		
	|log2 ratio| ≥0.7	<=E-8	94	2	0.979	0.989		

TP[a], True Positive; FP[b]; False Positive

b. Validation of filter conditions of CNV calls for AK2

Criteria	Subset	p-value	TP	FP	PPV	overall PPV
Length <5000bp, minimum llog2 ratiol ≥0.35	0.5>|log2 ratio| ≥0.35	<=E-21	13	4	0.765	0.855
	0.7>|log2 ratio| ≥0.5	<=E-14	54	19	0.740	
	|log2 ratio| ≥0.7	<=E-8	326	56	0.853	
Length ≥5000bp, minimum llog2 ratiol ≥0.30	0.5>|log2 ratio| ≥0.3	<=E-21	23	7	0.767	
	0.7>|log2 ratio| ≥0.5	<=E-8	42	7	0.857	
	|log2 ratio| ≥0.7	<=E-8	109	3	0.973	

Nature Genetics: doi:10.1038/ng.555

C. Suppl. Table 3. Absolute CNVs of 30 Asians

See SuppTable3_Absolute_CNVS_20099.*xls*

Depicted below is a preview for the part of this file.

index	sample	chr	length	start	stop	absolute log2ratio	gene_annotation
1	NA18592	1	42575	13065132	13107706	0.388	CDS:LOC440563:promoter:LOC440563:utr:LOC440563
2	NA18592	1	1756	14309636	14311391	-0.69	intergenic
3	NA18592	1	65031	17080164	17145194	0.35	CDS:CROCC:promoter:CROCC:utr:CROCC:intron:CROCC
4	NA18592	1	763	17549141	17549903	-2.25	intron:PADI4
5	NA18592	1	977	23609362	23610338	-1.16	intron:TCEA3
6	NA18592	1	2594	24393400	24395993	-0.81	intergenic
7	NA18592	1	1776	31492567	31494342	-0.87	intergenic
8	NA18592	1	1949	54864767	54866715	-0.82	intron:ACOT11
9	NA18592	1	3646	54864917	54868562	-2.335	intron:ACOT11
10	NA18592	1	901	58516499	58517399	-2.953	intergenic
11	NA18592	1	1223	58927955	58929177	0.92	CDS:MYSM1:intron:MYSM1
12	NA18592	1	910	59878829	59879738	0.413	CDS:FGGY:intron:FGGY
13	NA18592	1	1048	61855369	61856416	-0.404	intergenic
14	NA18592	1	45875	72538815	72584689	-1.141	intergenic
15	NA18592	1	10329	75246063	75256391	-0.94	intergenic
16	NA18592	1	3648	86173442	86177089	-0.97	intron:COL24A1
17	NA18592	1	2841	94060915	94063755	-1.987	intergenic
18	NA18592	1	2887	95295609	95298495	-0.92	intron:ALG14
19	NA18592	1	496	104244674	104245169	-0.375	intergenic
20	NA18592	1	1072	105056655	105057726	-4.04	intergenic
21	NA18592	1	6574	105817666	105824239	-0.94	intergenic

Nature Genetics: doi:10.1038/ng.555

Supplementary Table 4. Summary of statistics for absolute CNVs in the 30 Asians studied

Sample Id	Origin	CN Gain				CN Loss			
		# of Segments	Length (Mb)	# of genes	Examples of clinically important genes	# of Segments	Length (Mb)	# of genes	Examples of clinically important genes
NA18592	CHB	154	4.86	101	CLPS,DMBT1,IRF4,LPA,NBAS	482	5.04	227	ADAMTS14,DAZL,GSTT2,MCTP2,PGA3,4,5
NA18547	CHB	181	5.15	134	IRF4,LPA,PIK3CA,PRKRA,TPPP	475	4.85	203	ADAMTS14,DAZL,MCTP2,PGA3,4,5
NA18526	CHB	205	4.41	181	IRS2,LPA,PIK3CA	472	6.03	216	ADAMTS14,DMBT1,LY9,MCTP2,PGA3,4,5,RHD
NA18570	CHB	172	3.69	118	CEL,CLPS,IRF4,LPA,PRKRA	457	4.94	214	ADAMTS14,GHR,GSTT2,LY9,PGA3,4,5
NA18566	CHB	124	3.64	80	CES1,LPA,MUC20,PIK3CA	481	5.50	221	ADAMTS14,CFH,DAZL,DMBT1,GHR,MGAM,PGA3,4,5,PRSS2
NA18542	CHB	260	8.05	291	ADAMTS14,CLPS,EBF3,FOXC1,HYLS1,IRS2,PRKRA,TPPP	545	5.55	211	MGAM,PGA3,4,5
NA18537	CHB	117	3.27	84	CES1,CLPS,LPA,PIK3CA,PRKRA	480	4.03	204	ADAMTS14,CFH,DAZL,LY9,MCTP2,PGA3,4,5
NA18564	CHB	135	4.62	124	CES1,NBAS,PRKRA,SKI	477	4.10	193	ADAMTS14,DAZL,DMBT1,GSTT2,LY9,MCTP2,MGAM,NAIP,PGA3,4,5
NA18552	CHB	155	4.36	101	CEL,CLPS,DMBT1,IRS2,LPA,NAIP,PIK3CA,PRKRA	455	6.12	210	ADAMTS14,CFH,DAZL,MGAM,PGA3,4,5,RHD
NA18582	CHB	170	5.75	142	LPA,NBEA,PRKRA,TPPP	467	4.87	220	ADAMTS14,CFH,DAZL,GHR,MCTP2,MGAM,PGA3,4,5
NA18947	JPT	165	4.25	113	CEL,LPA,MGAM,PIK3CA,TPPP	488	5.58	240	ADAMTS14,CNR2,DAZL,MCTP2,MGAM,NBEA,PGA3,4,5
NA18972	JPT	277	5.12	272	IRX1,LPA,MUC20,MUC4,NAIP,PRKRA,TPPP	473	5.69	223	ADAMTS14,CNR2,DAZL,MCTP2,PGA3,4,5
NA18942	JPT	114	3.99	83	CES1,IRF4,LPA,NBEA,PIK3CA,PRKRA	598	10.36	424	ADAMTS14,CFH,DAZL,DMBT1,EBF3,IRS2,MCTP2,MUC20,MUC4,PGA3,4,5,TPPP
NA18949	JPT	176	5.16	145	CES1,EBF3,FOXC1,LPA,PRKRA,SKI	464	5.68	218	ADAMTS14,DAZL,MGAM,PGA3,4,5
NA18951	JPT	139	3.65	83	IRF4,KRT34,NBAS,PRKRA,SKI	450	4.09	170	ADAMTS14,CNR2,DAZL,DMBT1,LY9,MGAM,PGA3,4,5
NA18973	JPT	278	14.60	387	ADAMTS14,CEL,CES1,EBF3,HYLS1,IRS2,LPA,PITX1,TPPP	504	6.93	200	DAZL,MGAM,PGA3,4,5
NA18969	JPT	345	13.70	466	ADAMTS14,EBF3,HYLS1,NBAS,SKI	553	9.54	226	ADAMTS14,DAZL,LY9,MCTP2,MGAM,PGA3,4,5
NA18968	JPT	216	12.80	343	CES1,CLPS,DMBT1,EBF3,HYLS1,IRF4,IRS2,IRX1,NBEA,PIK3CA,PITX1	454	4.08	168	DAZL,NAIP,PGA3,4,5
NA18997	JPT	201	5.10	197	CCL4,MCTP2,MUC20,MUC4,PRKRA	542	4.37	223	ADAMTS14,CFH,DAZL,MGAM
NA18999	JPT	170	4.25	119	CCL4,CES1,FOXC1,LPA,PRKRA	504	4.47	207	ADAMTS14,DAZL,MCTP2,MGAM,PGA3,4,5,PIK3CA
AK2	KRS	134	3.66	94	DMBT1,LPA,NAIP	480	6.16	230	ADAMTS14,CNR2,DAZL,MCTP2,PGA3,4,5
AK4	KRS	252	4.28	120	IRF4,KRT34,LPA,MGAM,PRKRA	460	5.64	225	ADAMTS14,CEL,DAZL,MCTP2,PGA3,4,5,PRKRA
AK6	KRS	245	6.83	273	CEL,CLPS,EBF3,FOXC1,IRS2,IRX1,LPA,NBAS,PIK3CA,PITX1,PRKRA,SKI,TPPP	469	5.00	200	ADAMTS14,CNR2,DAZL,DMBT1,PGA3,4,5
AK8	KRS	154	4.55	98	CCL4,IRF4,LPA,NBAS,PRKRA,TPPP	469	5.63	212	ADAMTS14,DAZL,DMBT1,PGA3,4,5,PRSS2
AK10	KRS	194	6.22	204	DMBT1,FOXC1,IRF4,IRS2,IRX1,LPA,PITX1	455	5.39	199	ADAMTS14,DAZL,MCTP2,MUC20,PGA3,4,5
AK12	KRS	139	4.22	98	CEL,CLPS,EBF3,FOXC1,IRS2,IRX1,LPA,SKI,TPPP	457	5.99	206	ADAMTS14,DAZL,MCTP2,MUC20,PGA3,4,5
AK14	KRS	176	4.58	151	CCL4,CES1,CLPS,DMBT1,LPA,MUC20,MUC4,PRKRA,RHD,TPPP	454	5.08	204	ADAMTS14,DAZL,GHR,GSTT2,MCTP2,MGAM,PGA3,4,5
AK16	KRS	129	3.04	92	IRF4,KRT34,LPA,MUC20,MUC4,PITX1,PRKRA,TPPP	461	6.07	219	ADAMTS14,DAZL,RHD
AK18	KRS	161	5.30	108	CES1,CLPS,EBF3,LPA,PRKRA,RHD	467	5.82	217	ADAMTS14,DAZL,DMBT1,GHR,MCTP2,PGA3,4,5,PIK3CA
AK20	KRS	164	3.39	97	IRF4,LPA,PIK3CA,PRKRA	604	10.22	433	ADAMTS14,DAZL,DMBT1,PGA3,4,5,PRSS2

ADAMTS14, ADAM metallopeptidase with thrombospondin type 1 motif, 1; **CCL4**, chemokine (C-C motif) ligand 3; **CEL**, carboxyl ester lipase (bile salt-stimulated lipase); **CES1**, carboxylesterase 1 (monocyte/macrophage serine esterase 1); **CFH** , complement factor H; **CLPS**, colipase, pancreatic; **CNR2**, cannabinoid receptor 2 (macrophage) =CB2; **DAZL**, deleted in azospermia-like; **DMBT1**, deleted in malignant brain tumors 1; **EBF3**, early B-cell factor 3; **FOXC1**, forkhead box C1; GMDS GDP-mannose 4,6-dehydratase; **GHR**, growth hormone receptor; **GSTT2**, DDT D-dopachrome tautomerase; glutathione S-transferase theta 2; **HYLS1**, hydrolethalus syndrome 1; **IRF4**, interferon regulatory factor 4; **IRS2**, insulin receptor substrate 2; **IRX1**, iroquois homeobox 1; **KRT34**, keratin 34; **LPA**, lipoprotein, Lp(a); **LY9**, lymphocyte antigen 9; **MCTP2**, multiple C2 domains, transmembrane 2; **MGAM**, maltase-glucoamylase (alpha-glucosidase); **MUC4**, mucin 4; **MUC20**, mucin 20, cell surface associated; **NAIP**, NLR family, apoptosis inhibitory protein; **NBAS**, neuroblastoma amplified sequence; **NBEA**, neurobeachin; **PGA3,PGA4,PGA5**, pepsinogen 3,4,5; **PIK3CA**, PIK3CA phosphoinositide-3-kinase, catalytic, alpha polypeptide; **PITX1**, paired-like homeodomain 1; **PRKRA**, protein kinase, interferon-inducible double stranded RNA dependent activator; **PRSS2**, protease, serine, 2 (trypsin 2); **RHD**, RhD; **SKI**, v-ski sarcoma viral oncogene homolog (avian); **TPPP**, tubulin polymerization promoting protein

Nature Genetics: doi:10.1038/ng.555

E. Suppl. Table 5. List of primers for qPCR and breakpoint sequencing experiments

(a) Quantitative PCR

See SuppTable5_qPCR_primers_revision.*xls*

Depicted below is a preview for the part of this file.

Index	Forward Primer	Reverse Primer	Length PCR_product(bp)	Chr	Product_Start	Product_Stop
1	AGAGGCAAAGGCTAGGTTCTTATT	AACATGCTTCATCATCAGAGTGAG	78	1	62,428,839	62,428,916
2	TGACTCCTAAGACAAGGCTGTATG	CTCCTTGGCTGTAGTAAAATCTCC	78	1	108,536,366	108,536,443
3	CTCCCTATTTACTTGACTGCCTGT	TGTGGTGGAAGGTAGAGTTCAATA	89	1	150,842,083	150,842,171
4	CTGAGTAGTTCCTCCTTTGGTGTT	CTGACACTGATTTCTTCATTCCAG	71	2	33,080,098	33,080,168
5	GAGAATGTGTTTCTACTGGGGACT	CTATCGGCTCCTGAGGAATATTTA	92	2	89,028,624	89,028,715
6	TAGTAAGATTCAGAGCCTGACCTG	AAGCTGTGACGATATTTGTAGCTG	71	2	89,288,976	89,289,046
7	TGTTTTTCCCTAGCAGACCTTATC	TGTCCTATGTGTGGAGTGTTCTTT	98	2	89,644,666	89,644,763
8	CCTGTCACCTCTTCTGATTACCTT	CCCTTTGGTATACCTGTTTATTGC	100	2	99,470,685	99,470,784
9	TCCTACACACTGTTTTCATCACCT	GTAATGGAGGGCTTTGAAAGTCTA	103	2	176,978,395	176,978,497
10	GAAGTTCAGAGTTCAGCTTCTTGG	GGCATCAACACTCTTTGATATGTG	75	3	32,079,870	32,079,944
11	AAGACCAAGCAAAGTAAGAAGTGG	TCGTTGTATCTAAGAGGTGGGATT	75	3	196,946,815	196,946,889
12	GATGTGCTACTTAGACTCCACGAA	CTGGTAGCATCCTGAGAATAATGA	144	4	6,734,301	6,734,444
13	CCTATAATGGCATGTGACAAAGAG	TTACCAGAAATGGTGCTACATGAC	125	4	70,237,125	70,237,249
14	CAAGTGAGAATTTCTGGCAGTG	GTTGATCTAACTGGCAACCACA	102	4	165,424,175	165,424,276
15	GGGAGGAAGAATAGAGAGAGGAAC	ATAAGTCTAGACACAGGGGTGAGG	129	5	60,038,322	60,038,450
16	GAAGTGAAACAACAGTACCCTGTG	ATATTAAGAGTCCCGAGACAGTGG	136	5	170,063,220	170,063,355
17	GGAAAACAGAGTAGCACTCTAGGC	TTGACACGACTAGTAACCAAGGAA	78	6	281,744	281,821
18	GGGGTAACATAGGGTTATAAAGCA	CAGTAATTCAACTGTCTTCCCAGA	112	8	39,499,757	39,499,868
19	AGGTGAGACCAGTTCCTTGTAATC	GTCTCTTGGCTTTAGACCAGTTGT	116	9	112,068,988	112,069,103
20	CAGAGTAGGGAGTCGGTTGTCTAT	GAGGTTCAGCTTCATCACAGTAGA	133	10	124,338,181	124,338,313
21	AAATAGATAAGCCCACTCCTCCAC	CCCCAATATCTCACAACACAGTAG	127	11	11,780,381	11,780,507

Nature Genetics: doi:10.1038/ng.555

E. Suppl. Table 5. List of primers for qPCR and breakpoint sequencing experiments

(b) Breakpoint PCR and sequencing

Index	Forward primer	Reverse primer
1	TCACCAGCTCCTAAAATCCAAT	CTTTTCACACAGTTGCTTGGAG
2	TCCTTTCAATCACTTTGAGCTG	TCTCTTGATCCTCTTGCTCCTC
3	CTCCTCTCCTAACCCTGGAAGT	TGGAATAAGGTCCCAATAGGAG
4	GAGACAGCACAAAACAACAAGC	AGCTTGCTGCCTTTAGTCAAAC
5	CTGGAAGCAATTAAGCCACTCT	TGCCTCTATAAGTTTGTGTGACG
6	AAAGAGTGGTTTTAGCCTTTGC	TCCTTTTTAAGCGCTAGGTCAG
7	AACCTTTTGGTGGCTATTGAGA	TAGCAAGGATTCAAGACCCTGT
8	ACATGCCTTCCAGGCTATAGTG	ACCAATGTTGAAATGTCACAGG
9	TTTACCTTGAGGCCACTGAAAT	TTCTGACTCAGCATTTCTGCAT
10	GCTGATGACTGTCCCTTTATCC	CAGTTTCACCATTTCTTACAGCAG
11	GTCAGCACCAAATCTTCTTAGAAAC	GAATGCCAATGTAACAGAATGG
12	CAGTCACCAACCAGATGAAAGA	TCAGAGAAAGCATGACTCAGGA
13	GTTGACTTGAGACCATTGTGGA	AACAGTGTCCAGTGACATGTCTTA
14	TAGTGTTTGCATGGGAGGAAG	GTCCAGCAGATTCACATAATGG
15	CCTGCTAGTGCTTCTCTTCTCC	CATCTTCCTTCCTCCTCCTTTT
16	GTTGGACAAGGCTACACACAAA	TCACTCTCACTCTCCCAGATCA
17	TAGTGGAATTTGGTCCCTGACT	AAAAGAAGGTTGTATGGCAGGA
18	ACAGGCTATTTGGAATTCAAGC	GGGTCATAGTAGGCAGCTCAGT
19	GAATTCATCCTCCATGTTCCAT	ATCCTGTTGGCATATTTTGCTC
20	CGTGTGAATGACATCAGCCTAT	ATGCTGGACTGCAGAGTAAACA
21	TGAGCAGCAGTGATTGCTTAAT	TCAGGGAGTTGTAATGCAAAGA
22	GTCTCCTGACAGTGCCATACAA	AGAAGCAAACGTTGAAAAGAGG
23	AAACCCACTCCTCCTCTTTCTC	ACTCAGGGTCAAGCAATTAGGA
24	TTATATCCCCAGAGAGCTTTGC	GATGTGGCTTTTCCTGAGTAGG
25	GACCCCTGTAATTTTGGAGAGA	CTGAGCTCTGCCTCAATCAGTA
26	GCATGGTAGGATTTGGACTCTC	ATGGAACTCATTTCCTTGTGCT
27	GCTATGAACCCGTACCTTTTTG	GGGAAATATACAAGGCAAAGGA
28	AGACAAAAAGAAGGTGCCAAAG	AACTTGCGAAGTTACCAAAGGA
29	AGCCACCATCTCATAATTCACA	CCTAAACCTCTCATCCATCAGG
30	AAATTTCAGAGGTCACCCCTTT	GGAGCTTGGTGTCCTATCTCAC
31	CCTCATCTCTCTGGTCTGAAGG	ACCCTCAGCATTTTTATCCTCA
32	GTTTGGCAGCTTCAGAAAAACT	CTGGGCCTAGTTAAAAAGTAAAGG
33	AAATTAGATCAATGCCCTGCAC	GTGGTCAAATCTTCCTGGACTC
34	TTGGTACAACGTGAGGTGAGAC	TGATTGTCTGGCTGAAAACAAG
35	TGCCTCTTTCAAACCAGAGATT	TTGAAAGAATATGTCCCTGGTC
36	ACTGTGAGGAAGCTCACAATCC	TTGGCCACTATTCCCTTTCTTA
37	TTCATCACTCCCTCTAACAGCA	ATCTGGGCCATCGTATAAGAGA
38	ACCTCAGACTTGGGTGTTCAGT	GGTGATTCCCTGCTCAAATACA
39	GGACAAAAAGGAACAGGTTCTG	CCAACCTTCTTTCCTTCATCAC
40	CAGGATCTGGACCTGTCCTTAC	TCCATTCCAGTACAAGAAGCAC
41	AGAGGTACTTGATTGCCTCTGG	GGACTTCTGAGGCTTGAAGAAA
42	CAAGCATGACTGGTAAAATTGG	AAAAGCCACATAGTGCTACCAAG

Nature Genetics: doi:10.1038/ng.555

F. Suppl. Table 6. Summary of qPCR and breakpoint sequencing validation studies

(a) Quantitative PCR

		Agilent 24M aCGH			
		CN gain	CN normal	CN loss	Overall
qPCR	CN gain	594	61	27	682
	CN normal	19	593	17	629
	CN loss	15	25	530	570
	Overall	628	679	574	1881

Correct call	1717
Incorrect call	164
Correct call rate	91.28%

Nature Genetics: doi:10.1038/ng.555

F. Suppl. Table 6. Summary of qPCR and breakpoint sequencing validation studies (b) Breakpoints sequencing

Index	Sample	Chr	True Start	True End	Size (bp)	Start Difference (bp)	End Difference (bp)
1	AK6	1	84,484,593	84,488,463	3,871	126	15
2	AK14	1	105,056,556	105,057,713	1,158	99	13
3	NA18564	2	51,827,327	51,827,745	419	-38	46
4	AK10	2	108,221,850	108,222,714	865	-108	-72
5	AK8	3	26,425,973	26,427,303	1,331	-30	-64
6	NA18968	3	78,862,108	78,862,409	302	-112	74
7	AK18	3	133,190,943	133,196,075	5,133	119	-108
8	NA18542	3	153,247,620	153,248,061	442	-170	-86
9	NA18968	3	167,470,179	167,470,516	338	-71	154
10	AK12	3	191,220,038	191,223,219	3,182	116	11
11	AK18	4	30,623,047	30,624,073	1,027	-29	68
12	NA18949	4	43,446,564	43,446,887	324	-133	-18
13	NA18942	4	165,422,493	165,425,670	3,178	-24	6
14	NA18997	5	97,427,318	97,428,518	1,201	-22	-80
15	AK6	5	127,363,899	127,364,827	929	-33	-33
16	AK6	5	162,794,351	162,795,870	1,520	9	27
17	NA18542	5	170,062,496	170,063,968	1,473	-52	-33
18	AK6	6	22,158,817	22,162,220	3,404	135	112
19	NA18526	7	131,923,553	131,924,090	538	-23	-65
20	NA18552	8	62,197,914	62,198,447	534	-24	20
21	AK14	10	20,036,712	20,038,183	1,472	53	-13
22	AK6	10	66,976,938	66,985,301	8,364	-57	-73
23	AK10	10	107,940,672	107,941,586	915	289	-80
24	NA18537	10	130,726,861	130,727,265	405	-76	-31
25	AK4	12	49,259,982	49,261,778	1,797	271	-11
26	AK18	13	38,832,183	38,833,482	1,300	206	51
27	NA18942	13	108,159,746	108,160,439	694	78	-152
28	AK10	14	21,951,506	21,952,100	595	99	48
29	NA18542	14	38,074,269	38,074,779	511	-11	-13
30	AK18	14	81,568,863	81,573,084	4,222	-25	136
31	NA18973	14	84,366,861	84,371,909	5,049	-51	-3
32	NA18592	15	37,531,682	37,532,152	471	10	10
33	AK10	15	44,647,999	44,648,461	463	-178	-64
34	AK10	15	99,159,012	99,159,896	885	154	-33
35	NA18582	17	27,130,737	27,131,657	921	193	-237
36	AK10	18	33,560,058	33,560,631	574	-19	-31
37	AK4	18	45,948,975	45,952,385	3,411	0	130
38	NA18564	18	48,716,563	48,717,029	467	-34	19
39	NA18564	18	53,097,735	53,099,716	1,982	380	53
40	NA18552	18	72,476,184	72,476,990	807	-73	-375
41	NA18999	19	59,548,033	59,548,601	569	-170	-158
42	AK8	21	28,634,908	28,635,998	1,091	-143	-26

Nature Genetics: doi:10.1038/ng.555

G. Suppl. Table 7. List of 5,177 CNVE identified in 30 Asians

See *SuppTable7_5177CNVE_EthnicComparison.xls*

Depicted below is a preview for the part of this file.

Index	CNVE	Individual	chr	length	start	stop	log2ratio	Total	CHB	JPT	KOR	gene_annotation	Is_validated?	is_potentially Asian_specific?	GSV_CNVR
1	Asians30_CNVR_1.1	NA18947,NA18972,NA18564,	1	103530	736271	839800	0.43,0.33,0.51,0.48	4	1	3	0	utr:NCRNA00115,LOC64	YES	NO	CNVR5.1
2	Asians30_CNVR_1.2	NA18942,NA18542	1	4097	934543	938639	-0.81,0.78	2	1	1	0	promoter:ISG15	YES	YES	-
3	Asians30_CNVR_1.3	NA18968,NA18969	1	548883	802861	1351743	0.4,0.36	2	0	2	0	CDS:SAMD11,NOC2L,KI	YES	YES	-
4	Asians30_CNVR_1.4	NA18542	1	2403	923968	926370	0.82	1	1	0	0	CDS:HES4:promoter:HES	NO	YES	-
5	Asians30_CNVR_1.5	NA18942	1	2436	981285	983720	-0.67	1	0	1	0	promoter:AGRN:utr:AGR	NO	YES	-
6	Asians30_CNVR_1.6	NA18542	1	943	1102010	1102952	0.73	1	1	0	0	intron:TTLL10	NO	YES	-
7	Asians30_CNVR_1.7	NA18526	1	5421	1154213	1159633	0.55	1	1	0	0	CDS:B3GALT6:promoter	NO	YES	-
8	Asians30_CNVR_2.1	AK10	1	3552	1433831	1437382	0.97	1	0	0	1	promoter:ATAD3A	NO	YES	-
9	Asians30_CNVR_2.2	NA18969	1	19859	1433831	1453689	0.45	1	0	1	0	CDS:ATAD3A:promoter:	NO	YES	-
10	Asians30_CNVR_3.1	NA18942	1	705	1542495	1543199	-0.71	1	0	1	0	intron:MIB2	NO	YES	-
11	Asians30_CNVR_4.1	NA18566,NA18542,NA18582	1	81781	1575237	1657017	0.33,0.76,0.59	3	3	0	0	CDS:CDC2L1,LOC7286	YES	NO	CNVR17.1
12	Asians30_CNVR_4.2	NA18537,NA18552	1	45586	1624860	1670445	0.81,-0.37	2	2	0	0	CDS:CDC2L1,CDC2L2,S	YES	NO	CNVR17.1
13	Asians30_CNVR_4.3	NA18973	1	15613	1625061	1640673	0.53	1	0	1	0	CDS:CDC2L1,CDC2L2:u	YES	NO	CNVR17.1
14	Asians30_CNVR_4.4	AK18	1	19289	1637729	1657017	-0.42	1	0	0	1	CDS:CDC2L1,CDC2L2,S	YES	NO	CNVR17.1
15	Asians30_CNVR_5.1	NA18542	1	2290	1978127	1980416	0.78	1	1	0	0	intron:PRKCZ	NO	YES	-
16	Asians30_CNVR_6.1	AK12,NA18951,NA18969,NA1	1	4453	2227388	2231840	0.43,0.43,0.56,0.4,0.4	5	1	3	1	CDS:SKI:promoter:SKI:ut	YES	YES	-
17	Asians30_CNVR_6.2	AK6	1	32093	2224073	2256165	0.44	1	0	0	1	CDS:SKI,MORN1:promo	NO	YES	-
18	Asians30_CNVR_7.1	AK6	1	16227	2310992	2327218	0.4	1	0	0	1	CDS:MORN1,RER1,PEX1	NO	YES	-
19	Asians30_CNVR_7.2	NA18542	1	1255	2324440	2325694	0.55	1	1	0	0	promoter:PEX10:utr:RER	NO	YES	-
20	Asians30_CNVR_8.1	NA18969	1	70098	2403604	2473701	0.39	1	0	1	0	CDS:PLCH2,PANK4,HES	NO	YES	-
21	Asians30_CNVR_8.2	AK6	1	30859	2442099	2472957	0.39	1	0	0	1	CDS:PANK4,HES5:prom	NO	YES	-
22	Asians30_CNVR_9.1	AK20,NA18972	1	1650	2476434	2478083	-0.73,-0.66	2	0	1	1	intergenic	YES	YES	-
23	Asians30_CNVR_10.1	NA18942	1	1103	2480304	2481406	-0.93	1	0	1	0	CDS:TNFRSF14:intron:TI	NO	YES	-
24	Asians30_CNVR_10.2	NA18972	1	444	2480963	2481406	-0.7	1	0	1	0	CDS:TNFRSF14:intron:TI	NO	YES	-
25	Asians30_CNVR_11.1	NA18973	1	2974	2570328	2573301	0.559	1	0	1	0	intergenic	NO	YES	-
26	Asians30_CNVR_12.1	AK6	1	18856	2965717	2984572	0.4	1	0	0	1	CDS:PRDM16:promoter:	NO	YES	-

Nature Genetics: doi:10.1038/ng.555

H. Suppl. Table 8. List of OMIM genes in identified CNVs

See SuppTable8_1843_OMIMgene.xls

Depicted below is a preview for the part of this file.

Gene	Chr	GeneStatus* (See Doc.)	context	Gene	OMIM_Num	Method* (See Doc.)	Comment	Disease
40057	16	P	CDS	Septin 1	609062	REc	-	-
40061	22	C	CDS	Septin 5	609062	REn, Ch	just 5' of GP1BB	-
40066	2	P	CDS	Septin 10	609062	REc	pseudogene on 8q22.1-q12	-
A2BP1	16	C	intron	Ataxin 2-binding protein 1	609062	Ch, A, REc	-	-
AATK	17	C	CDS	Apoptosis-associated tyrosine kinase	609062	REa, A	-	-
ABCA1	9	C	intron	ATP-binding cassette 1	600046	A, REc	-	HDL deficiency, type 2, 604091 (3)
ABCA3	16	P	UTR	ATP-binding cassette-3	601615	REc	-	Surfactant metabolism dysfunction, pulmonary, 3, 610921 (3)
ABCA7	19	P	intron	ATP-binding cassette, subfamily A, member 7	609062	REc	-	-
ABCB10	1	P	CDS	ATP-binding cassette, subfamily B, member 10	609062	REc	pseudogene on 15q13-q14	-
ABCC11	16	C	intron	ATP-binding cassette, subfamily C, member 11	607040	R, REc, Fd	-	[Earwax, wet/dry], 117800 (3)
ABCC2	10	P	CDS	ATP-binding cassette, subfamily C, member 2	601107	A	-	Dubin-Johnson syndrome, 237500 (3)
ABCG8	2	P	intron	ATP-binding cassette, subfamily G, member 8	605460	REc	-	Gallbladder disease 4, 611465 (3)
ABR	17	C	CDS	Active BCR-related gene	609062	A	-	-
ACACA	17	C	intron	Acetyl-Coenzyme A carboxylase, alpha	200350	A	proximal to q21.33; others pu	Acetyl-CoA carboxylase deficiency (1)
ACBD3	1	P	intron	Acyl-Coenzyme A binding domain containing 3	609062	R, REc	-	-
ACCN1	17	P	intron	Amiloride-sensitive cation channel 1, neuronal (d	609062	A	-	-
ACOT11	1	C	CDS	Acyl-CoA thioesterase 11	609062	REa, R, H	-	-
ACOT2	14	P	CDS	Acyl-CoA thioesterase 2	609062	REc, R	-	-
ACOT7	1	P	UTR	Acyl-CoA thioesterase 7	609062	REc	-	-
ACR	22	C	CDS	Acrosin	102480	REa, Ch	-	Male infertility due to acrosin deficiency (2) (?)
ACSL1	4	C	UTR	Acyl-CoA synthetase long-chain family member	609062	REb, A	-	-
ACTA2	10	C	intron	Actin, alpha-2, smooth muscle, aorta	102620	REa, A	-	Aortic aneurysm, familial thoracic 6, 611788 (3)
ACTG1	17	C	CDS	Actin, gamma-1	102560	REa, A, Fd	-	Deafness, autosomal dominant 20/26, 604717 (3)
ADAM28	8	C	promoter	A disintegrin and metalloproteinase domain 28	609062	REc	-	-
ADAMTS13	9	P	promoter	A disintegrin-like and metalloprotease with thron	604134	Fd, REc	-	Thrombotic thrombocytopenic purpura, familial, 274150 (3)
ADAMTS9	3	P	intron	A disintegrin-like and metalloproteinase with thr	609062	A, Psh	-	-
ADAMTSL1	9	P	intron	ADAMTS-like protein 1	609062	REc, H	-	-
ADAMTSL3	15	P	intron	ADAMTS-like protein 3	609062	REc, H	-	-
ADARB2	10	P	CDS	Adenosine deaminase, RNA-specific, B2 (homolo	609062	REa	-	-
ADCY5	3	C	intron	Adenylate cyclase-5	609062	REa, A	-	-

Nature Genetics: doi:10.1038/ng.555

I. Suppl. Table 9. List of microRNAs overlapping the personal CNVs identified in the study

See SuppTable9_miRNA.xls

Depicted below is a preview for the part of this file.

Sample	Chr	CNV Start	CNV Stop	log2ratio	Accession_#	miRNA_ID	miRNA Start	miRNA stop
NA18592	2	132,719,667	132,766,153	3.391	MI0003566	hsa-mir-560	132,731,971	132,732,065
NA18592	22	21,484,527	21,577,178	-1.242	MI0003665	hsa-mir-650	21,495,270	21,495,365
NA18547	2	132,719,667	132,766,153	3.264	MI0003566	hsa-mir-560	132,731,971	132,732,065
NA18547	8	145,062,114	145,100,595	0.37	MI0003669	hsa-mir-661	145,091,347	145,091,435
NA18547	17	76,247,436	77,313,922	-0.529	MI0003681	hsa-mir-657	76,713,671	76,713,768
NA18547	17	76,247,436	77,313,922	-0.529	MI0000814	hsa-mir-338	76,714,278	76,714,344
NA18547	22	21,484,621	21,576,911	-1.875	MI0003665	hsa-mir-650	21,495,270	21,495,365
NA18947	2	132,719,667	132,766,153	3.385	MI0003566	hsa-mir-560	132,731,971	132,732,065
NA18947	17	76,247,436	77,313,922	-0.632	MI0003681	hsa-mir-657	76,713,671	76,713,768
NA18947	17	76,247,436	77,313,922	-0.632	MI0000814	hsa-mir-338	76,714,278	76,714,344
NA18947	22	20,999,466	21,577,178	0.348	MI0003665	hsa-mir-650	21,495,270	21,495,365
NA18972	2	132,726,662	132,755,838	3.884	MI0003566	hsa-mir-560	132,731,971	132,732,065
NA18972	3	196,822,900	196,973,214	0.355	MI0003577	hsa-mir-570	196,911,452	196,911,548
NA18972	17	76,247,436	77,313,922	-0.559	MI0003681	hsa-mir-657	76,713,671	76,713,768
NA18972	17	76,247,436	77,313,922	-0.559	MI0000814	hsa-mir-338	76,714,278	76,714,344
NA18972	22	21,494,330	21,495,768	-4.601	MI0003665	hsa-mir-650	21,495,270	21,495,365
NA18526	17	76,247,436	77,313,922	-0.406	MI0003681	hsa-mir-657	76,713,671	76,713,768
NA18526	17	76,247,436	77,313,922	-0.406	MI0000814	hsa-mir-338	76,714,278	76,714,344
NA18526	22	21,484,791	21,570,549	-3.037	MI0003665	hsa-mir-650	21,495,270	21,495,365
NA18570	2	132,719,667	132,766,153	3.295	MI0003566	hsa-mir-560	132,731,971	132,732,065
NA18570	17	76,247,436	77,313,922	-0.434	MI0003681	hsa-mir-657	76,713,671	76,713,768

Nature Genetics: doi:10.1038/ng.555

J. Suppl. Table 10. List of fusion gene overlapping the personal CNVs identified in this study

See SuppTable10_fusion_gene_list.xls

Depicted below is a preview for the part of this file.

Index	Fusion Gene	Sample	Chr	Length	StartPos	StopPos	Log2Ratio	Left Gene	Strand	StartPos	StopPos	RightGene	Strand	StartPos	StopPos
1	PCDHA8-PCDHA9	NA18592	5	15,795	140,203,406	140,219,200	-0.499	PCDHA8	+	140,201,091	140,203,535	PCDHA9	+	140,207,541	140,372,113
2	PCDHA8-PCDHA10	NA18592	5	15,795	140,203,406	140,219,200	-0.499	PCDHA8	+	140,201,091	140,203,535	PCDHA10	+	140,215,818	140,372,113
3	HLA-DRB5-HLA-DRB6	NA18592	6	23,305	32,605,201	32,628,505	-1.561	HLA-DRB5	-	32,593,132	32,605,984	HLA-DRB6	-	32,628,468	32,635,757
4	OR51A4-OR51A2	NA18592	11	8,669	4,924,706	4,933,374	-0.85	OR51A4	-	4,923,965	4,924,906	OR51A2	-	4,932,578	4,933,519
5	PCDHA8-PCDHA9	NA18547	5	15,788	140,203,413	140,219,200	-0.473	PCDHA8	+	140,201,091	140,203,535	PCDHA9	+	140,207,541	140,372,113
6	PCDHA8-PCDHA10	NA18547	5	15,788	140,203,413	140,219,200	-0.473	PCDHA8	+	140,201,091	140,203,535	PCDHA10	+	140,215,818	140,372,113
7	PCDHA8-PCDHA9	NA18947	5	15,788	140,203,413	140,219,200	-0.496	PCDHA8	+	140,201,091	140,203,535	PCDHA9	+	140,207,541	140,372,113
8	PCDHA8-PCDHA10	NA18947	5	15,788	140,203,413	140,219,200	-0.496	PCDHA8	+	140,201,091	140,203,535	PCDHA10	+	140,215,818	140,372,113
9	GSTTP1-GSTTP2	NA18947	22	49,240	22,676,441	22,725,680	-0.675	GSTTP1	-	22,670,595	22,677,258	GSTTP2	-	22,715,938	22,731,899
10	APOBEC3A-APOBEC3B	NA18947	22	30,133	37,686,393	37,716,525	-2.03	APOBEC3A	+	37,683,473	37,689,134	APOBEC3B	+	37,708,351	37,718,729
11	PCDHA8-PCDHA9	NA18972	5	15,788	140,203,413	140,219,200	-0.442	PCDHA8	+	140,201,091	140,203,535	PCDHA9	+	140,207,541	140,372,113
12	PCDHA8-PCDHA10	NA18972	5	15,788	140,203,413	140,219,200	-0.442	PCDHA8	+	140,201,091	140,203,535	PCDHA10	+	140,215,818	140,372,113
13	PCDHA8-PCDHA9	NA18526	5	15,788	140,203,413	140,219,200	-0.301	PCDHA8	+	140,201,091	140,203,535	PCDHA9	+	140,207,541	140,372,113
14	PCDHA8-PCDHA10	NA18526	5	15,788	140,203,413	140,219,200	-0.301	PCDHA8	+	140,201,091	140,203,535	PCDHA10	+	140,215,818	140,372,113
15	LOC646227-BTNL3	NA18526	5	21,388	180,341,884	180,363,271	-2.526	LOC646227	+	180,341,824	180,345,858	BTNL3	+	180,348,507	180,366,333
16	C4B-C4A	NA18526	6	32,950	32,071,612	32,104,561	-0.52	C4B	+	32,057,813	32,078,436	C4A	+	32,090,550	32,111,173
17	OR51A4-OR51A2	NA18526	11	8,669	4,924,706	4,933,374	-0.93	OR51A4	-	4,923,965	4,924,906	OR51A2	-	4,932,578	4,933,519
18	APOBEC3A-APOBEC3B	NA18526	22	30,242	37,686,284	37,716,525	-0.7	APOBEC3A	+	37,683,473	37,689,134	APOBEC3B	+	37,708,351	37,718,729
19	PCDHA8-PCDHA9	NA18570	5	15,795	140,203,406	140,219,200	-0.485	PCDHA8	+	140,201,091	140,203,535	PCDHA9	+	140,207,541	140,372,113
20	PCDHA8-PCDHA10	NA18570	5	15,795	140,203,406	140,219,200	-0.485	PCDHA8	+	140,201,091	140,203,535	PCDHA10	+	140,215,818	140,372,113
21	PRB1-PRB2	NA18570	12	37,652	11,398,210	11,435,861	-0.486	PRB1	-	11,396,024	11,399,791	PRB2	-	11,435,743	11,439,765
22	PIGZ-MFI2	NA18942	3	77,622	198,161,095	198,238,716	-0.32	PIGZ	-	198,157,611	198,180,101	MFI2	-	198,214,553	198,241,083
23	PIGZ-MFI2	NA18942	3	77,622	198,161,095	198,238,716	-0.32	PIGZ	-	198,157,611	198,180,101	MFI2	-	198,230,221	198,241,083
24	WHSC2-POLN	NA18942	4	89,201	1,962,765	2,051,965	-0.34	WHSC2	-	1,954,241	1,980,757	POLN	-	2,043,443	2,200,756

Nature Genetics: doi:10.1038/ng.555

K. Suppl. Table 11. Modified PANTHER ontology analysis

See SuppTable11_GeneOntology.xls

Depicted below is a preview for the part of this file.

Copy Number Gain		
Modified_Panther_BiologicalProcess	count	percent
Cell adhesion	4	1.3%
Develop. Proc.	39	13.1%
Immunity and defense	30	10.1%
Miscellaneous	93	31.3%
Nucleic Acid Metabolism	57	19.2%
Protein Processing	21	7.1%
Sensory perception	19	6.4%
Signal Transduction	34	11.4%
SUM	**297**	

Copy Number Loss		
Modified_Panther_BiologicalProcess	count	percent
Cell adhesion	14	8.2%
Develop. Proc.	7	4.1%
Immunity and defense	37	21.6%
Miscellaneous	51	29.8%
Nucleic Acid Metabolism	4	2.3%
Protein Processing	15	8.8%
Sensory Perception	17	9.9%
Signal transduction	26	15.2%
SUM	**171**	

Nature Genetics: doi:10.1038/ng.555

L. Suppl. Table 12. Examples of Genes showing different copy number status between this study and Conrad *et al*.[1]

Gene Name	Absolute CN status in 30 Asians (this study)	Absolute CN status in 90 Asians (Conrad *et al*.)
RHD	3/30 CN Loss 2/30 CN Gain	88/88 CN Gain
SIRPB1	29/30 CN Loss	38/90 CN Gain
CR1	29/30 CN Loss	88/90 CN Gain
PGA3, PGA4, PGA5	28/30 CN Loss	0/90 CN Loss
NOTCH2	30/30 CN Gain	0/90 CN Gain
OR4S1	27/30 CN Gain	0/90 CN Gain
FAM21A, FAM21B	30/30 CN Loss	0/89 CN Loss
MUC6	30/30 CN Gain	1/87 CN Gain
UPK3B	30/30 CN Loss	31/77 CN Loss 15/77 CN Gain

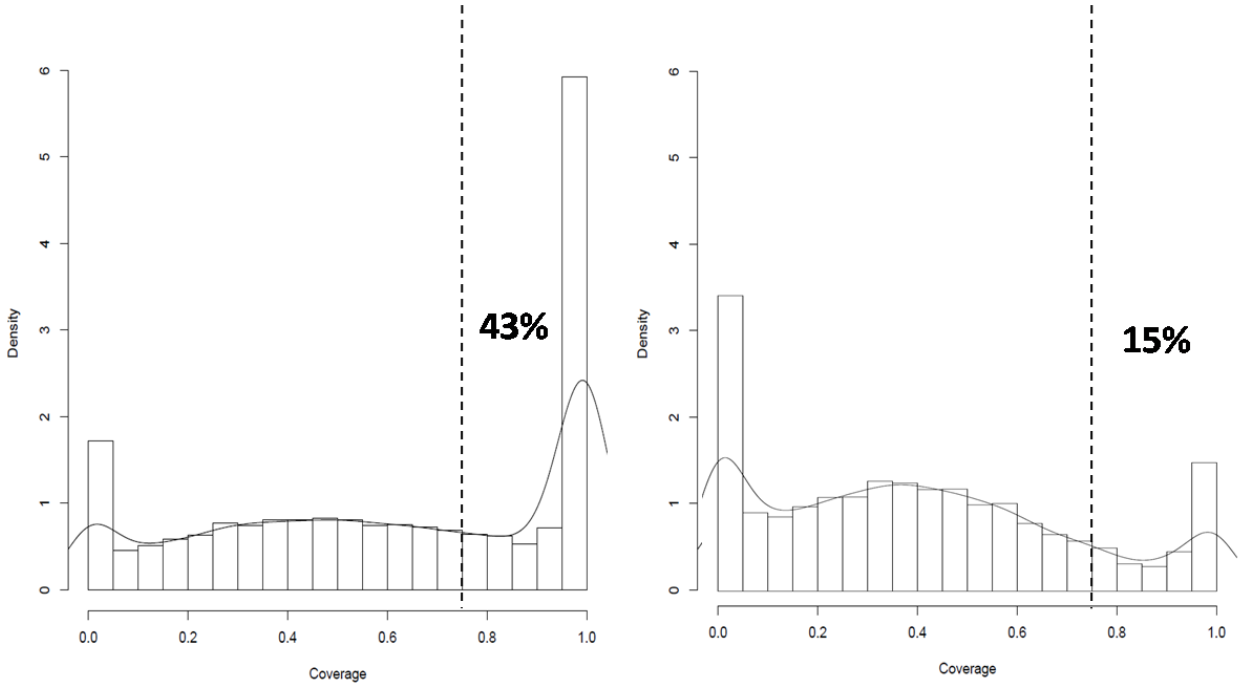
Nature Genetics: doi:10.1038/ng.555

M. Suppl. Figure 1. Content of repetitive sequence for the Agilent 24M array set and the NimbleGen 42M array set.

For the CNVs that are called by a single platform, we have used the respective probe distributions to filter out any CNVs that “cannot” be called by either of the platform due to lack of enough probes.

X-axis is the density (frequency/total number of CNVEs). Y-axis is the coverage of the CNVEs with repeatmasker + Segmental duplications. 0.8 means that 80% of those CNVEs covered with CNVs. The dashed line represents 0.75 and the percentage close to that line is the percentage of CNVEs, 75% or more of which is covered with repeats or segmental duplications.

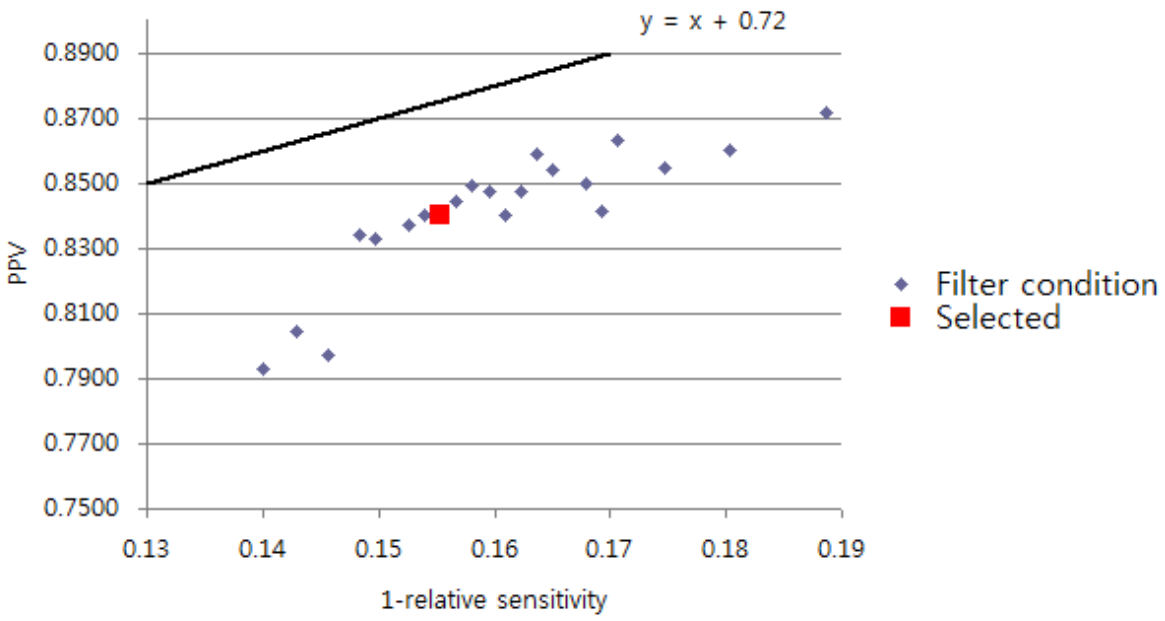
[NimbleGen 42M] [Agilent 24M]



Nature Genetics: doi:10.1038/ng.555

N.Suppl. Figure 2. Modified Receiver Operator curve (ROC) analysis for filter training using data for AK1.

We used "modified ROC curve", using PPV for the Y-axis and (1-relative sensitivity) in X-axis because this is more advantageous for training our filter conditions. The predominance of non-CNV areas in the genome results in artificially high values for specificity with a limited distribution. PPV was used for the Y-axis to better discriminate between the performances of different parameters.



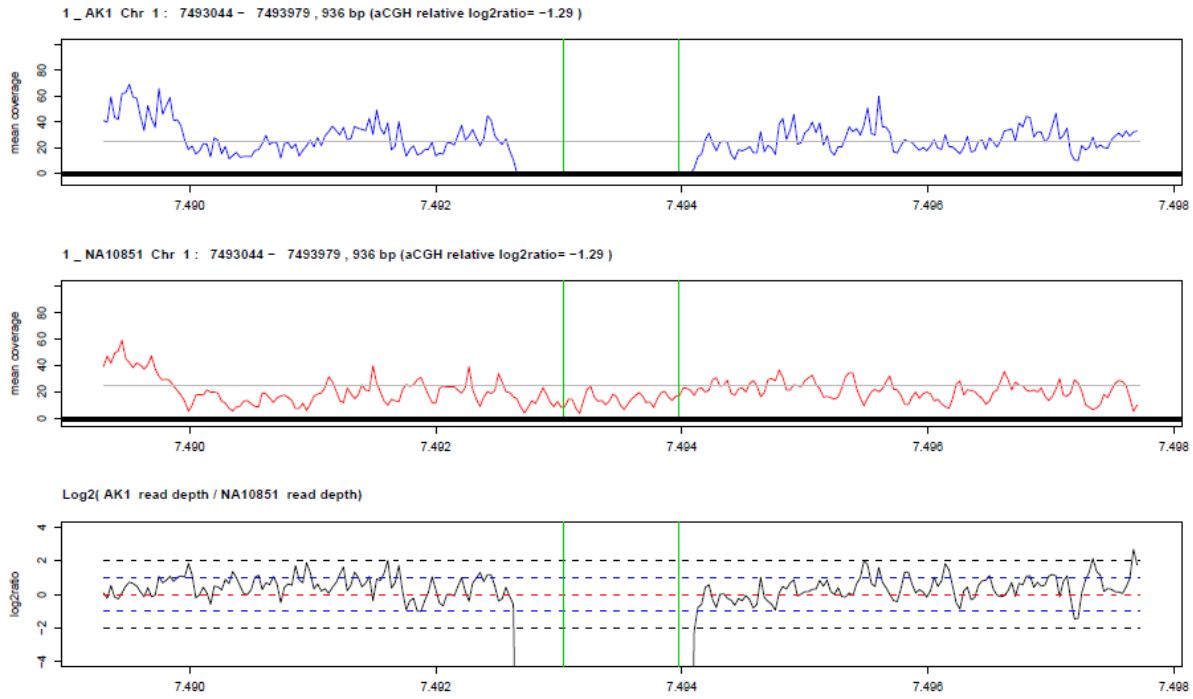
Nature Genetics: doi:10.1038/ng.555

O.Suppl. Figure 3. Read-depth information for 721 validated CNVs in AK1 using data for AK1 and NA10851

Top panel, Read depth of AK1; Middle panel, Read depth of NA10851; Bottom panel, logarithm of read-depth ratio (AK1 read-depth/NA10851 read-depth)

See *SuppFig3_ReadDepth.pdf*

Depicted below is a preview for the part of this file



Nature Genetics: doi:10.1038/ng.555

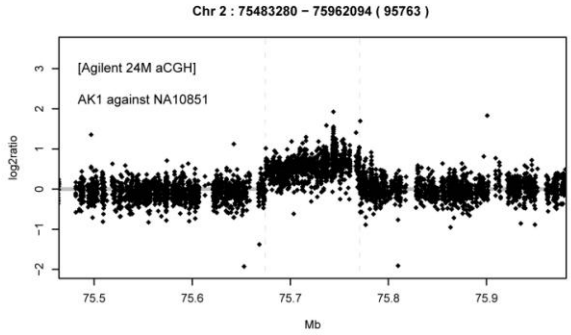
P.Suppl. Figure 4. Application of the absolute CNV calling algorithm and confirmation of results using read-depth sequence information

In regions where the reference sample possesses two copies, absolute log2 ratio of the segment is identical to its relative log2 ratio.

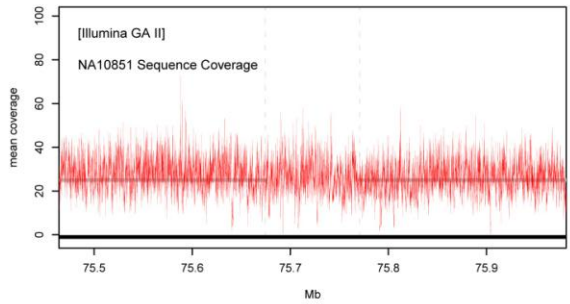
For complete loss regions of NA10851, test sample log2 ratios are generally very high and unstable. In other cases where the reference sample has a copy number loss (of 1 copy) or gain, array CGH will yield a positive call if the test sample has two copies. If the copy number of a given genomic segment in the test sample is identical to the corresponding genomic segment in NA10851, this CNV will not be detected by aCGH.

Nature Genetics: doi:10.1038/ng.555

(a) An example of an overt CN gain

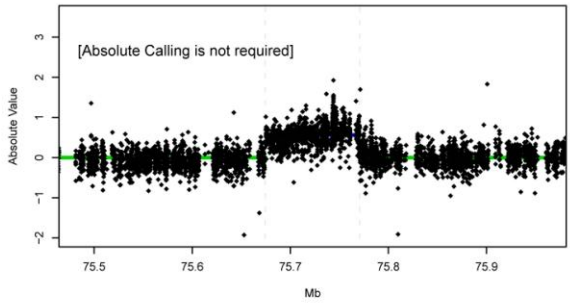


CN gain is identified in aCGH.

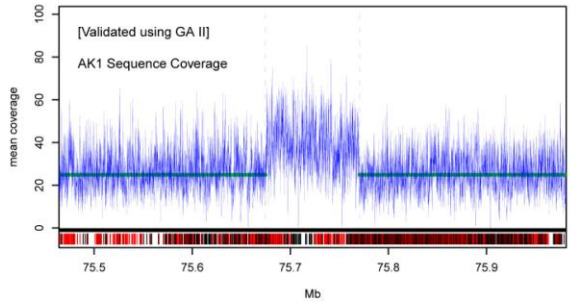


NA10851 has no CNV on the region.
(by read-depth and aCGH information)

Overt call.



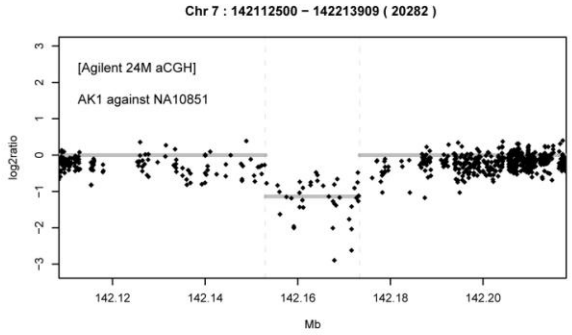
Absolute calling is not required.



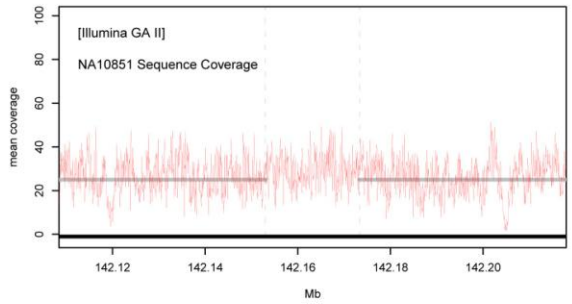
Absolute CN gain for the test sample is confirmed by read-depth of AK1 sequencing.

Nature Genetics: doi:10.1038/ng.555

(b) An example of an overt CN loss

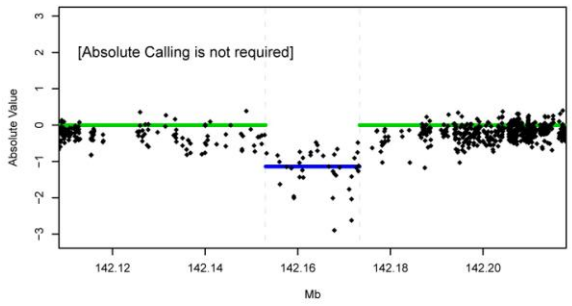


CN loss is identified in aCGH.

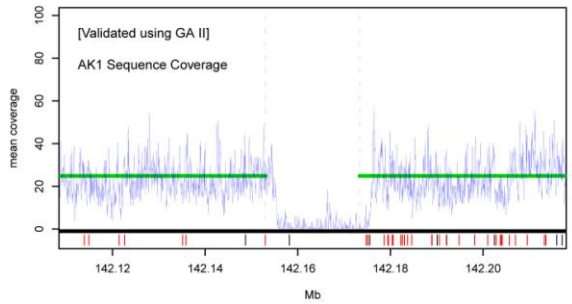


NA10851 has no CNV on the region.
(by read-depth and aCGH information)

Overt call.



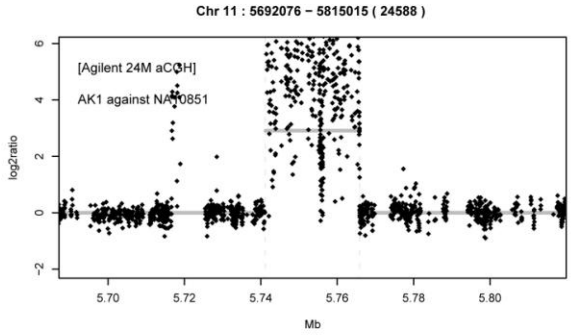
Absolute calling is not required.



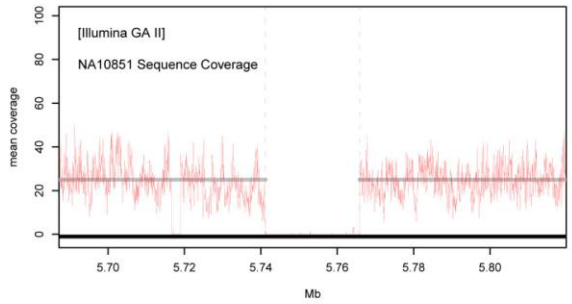
Absolute CN loss for the test sample is confirmed
by read-depth of AK1 sequencing

Nature Genetics: doi:10.1038/ng.555

(c) An example of an obscure CN gain (removed by absolute calling algorithm)

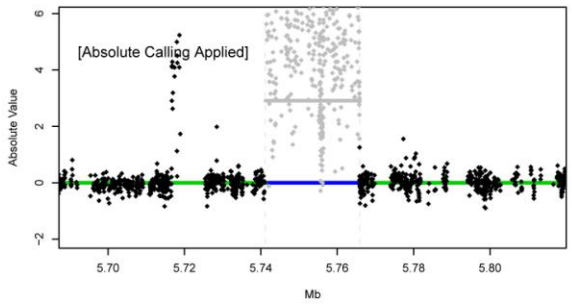


High CN gain is identified in aCGH



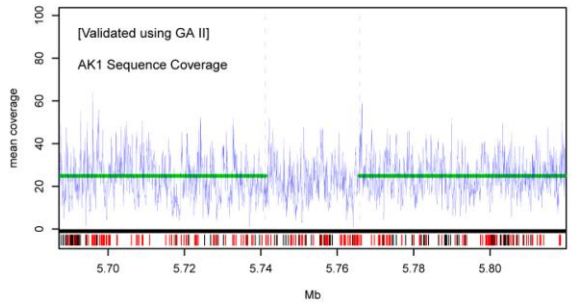
NA10851 has complete CN loss on the region.

Obscure call.



Absolute calling is necessary and applied.

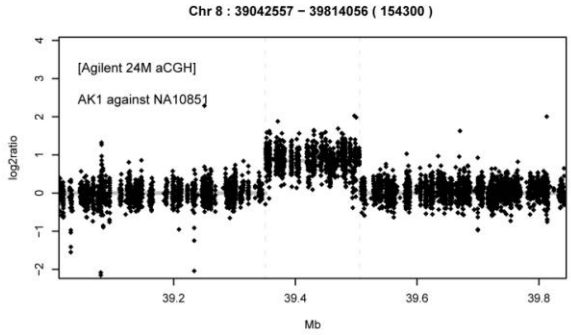
Corrected absolute log2ratio shows no evidence of CN gain for the test sample.



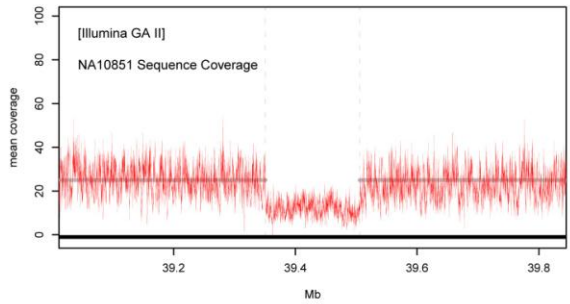
CN normal is confirmed by read-depth of AK1 sequencing.

Nature Genetics: doi:10.1038/ng.555

(d) An example of an obscure CN gain (removed by absolute calling algorithm)

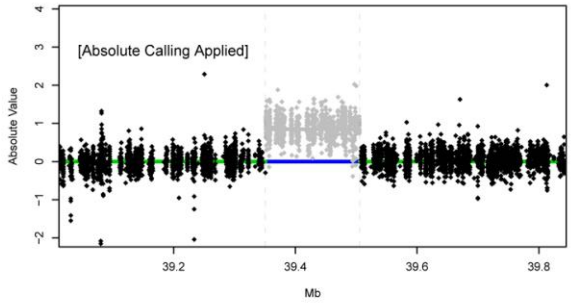


CN gain is identified in aCGH.



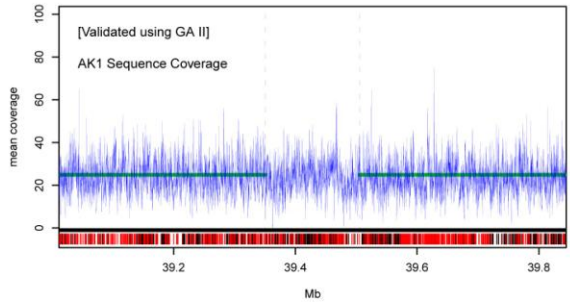
NA10851 has heterozygous CN loss on the region.

Obscure call.



Absolute calling is necessary and applied.

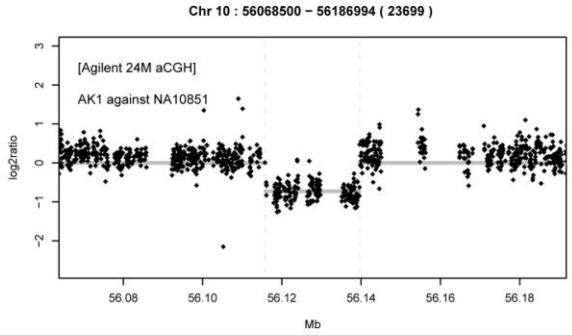
Corrected absolute log2ratio shows no evidence of CN gain for the test sample.



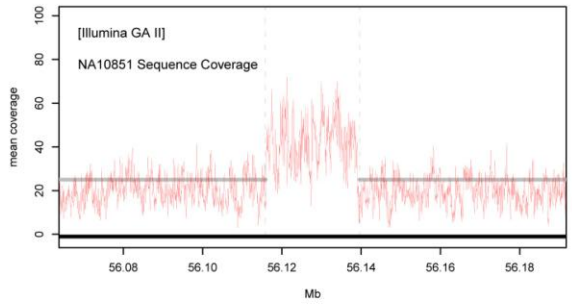
CN normal is confirmed by read-depth of AK1 sequencing.

Nature Genetics: doi:10.1038/ng.555

(e) An example of an obscure CN loss (removed by absolute calling algorithm)

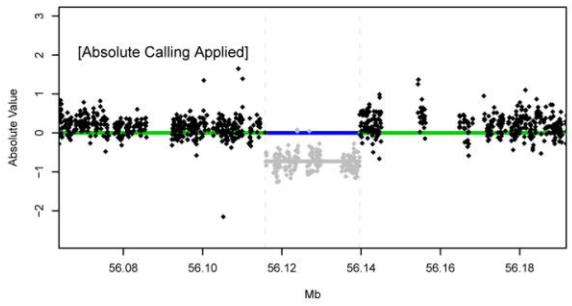


CN loss is identified in aCGH.



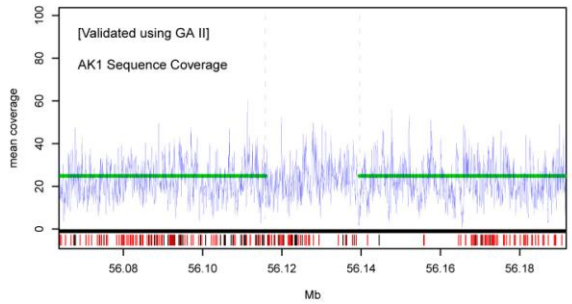
NA10851 has CN gain on the region.

Obscure call.



Absolute calling is necessary and applied.

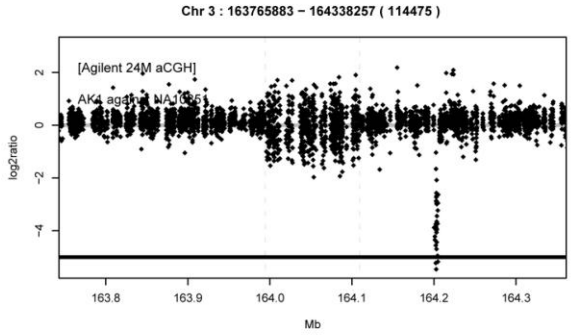
Corrected absolute log2ratio shows no evidence of CN loss for the test sample.



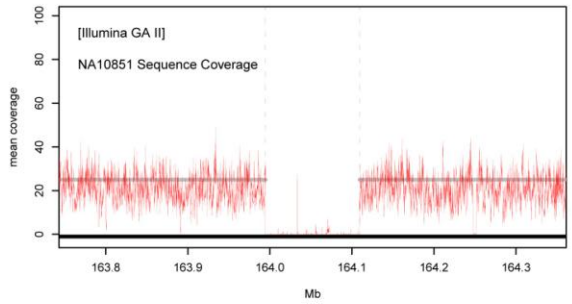
CN normal is confirmed by read-depth of AK1 sequencing.

Nature Genetics: doi:10.1038/ng.555

(f) An example of a covert CN loss

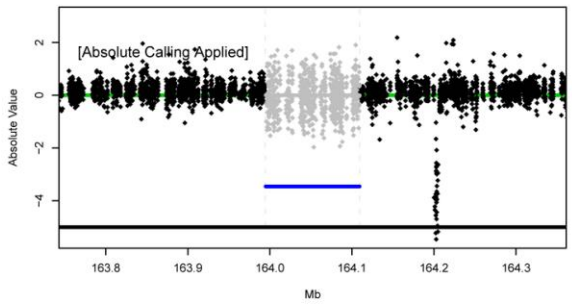


aCGH reported no CNV in this region.



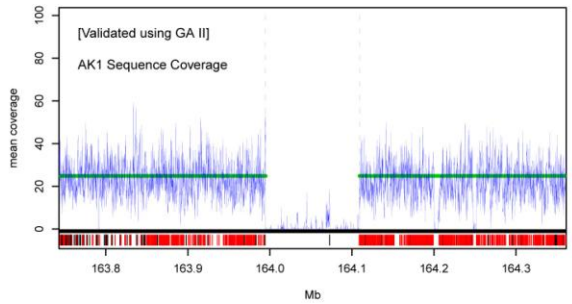
NA10851 has complete CN loss on the region.

A candidate for covert (cryptic) call.



Absolute calling is necessary and applied.

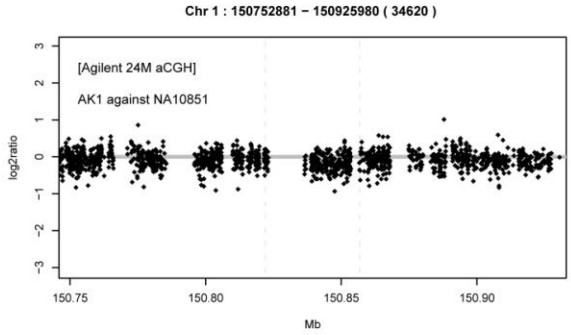
Corrected absolute log2ratio shows complete CN loss for the test sample.



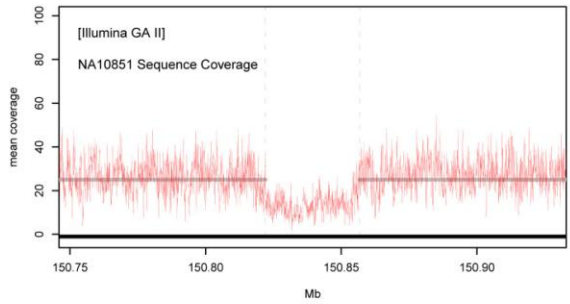
The complete CN loss of the test sample is confirmed by read-depth of AK1 sequencing.

Nature Genetics: doi:10.1038/ng.555

(g) An example of a covert CN loss

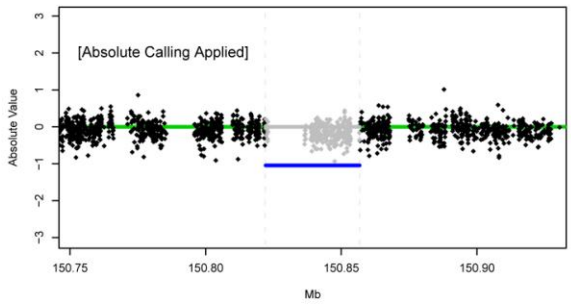


aCGH reported no CNV in this region.



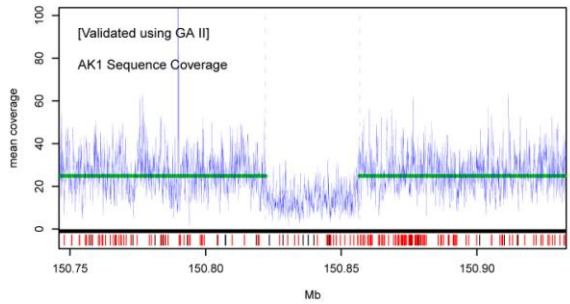
NA10851 has heterozygous CN loss on the region.

A candidate for covert (cryptic) call.



Absolute calling is necessary and applied.

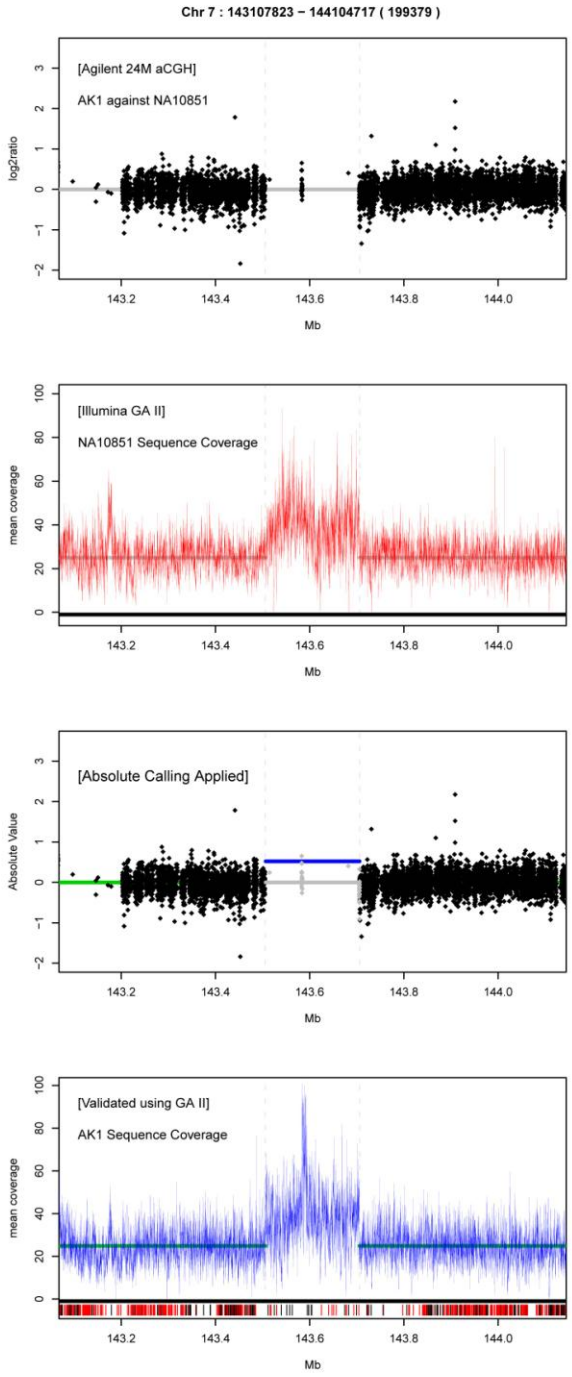
Corrected absolute log2ratio shows heterozygous CN loss for the test sample.



The heterozygous CN loss of the test sample is confirmed by read-depth of sequencing.

Nature Genetics: doi:10.1038/ng.555

(h) An example of a covert CN gain



aCGH reported no CNV in this region.

NA10851 has CN gain on the region.

A candidate for covert (cryptic) call.

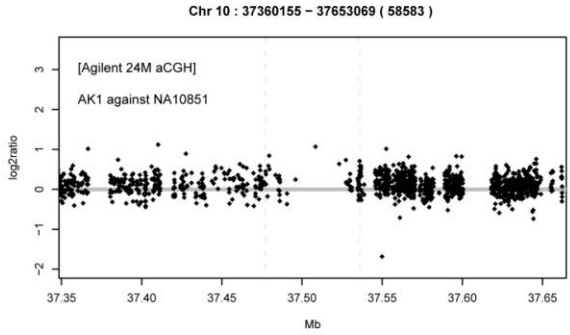
Absolute calling is necessary and applied.

Corrected absolute log2ratio shows CN gain for the test sample.

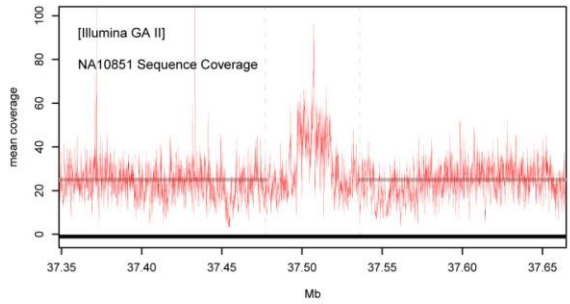
The CN gain of the test sample is confirmed by read-depth of sequencing.

Nature Genetics: doi:10.1038/ng.555

(i) An example of a covert CN gain

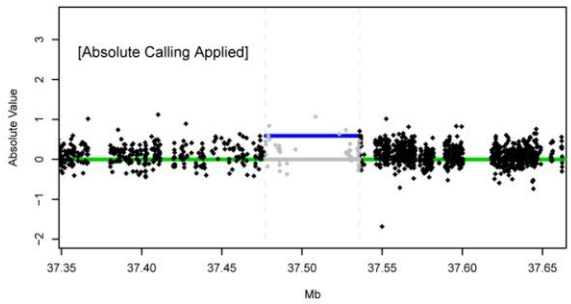


aCGH reported no CNV in this region.



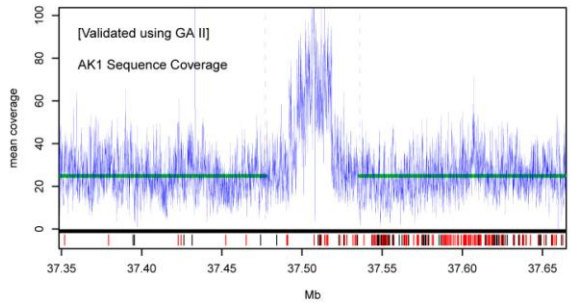
NA10851 has CN gain on the region.

A candidate for covert (cryptic) call.



Absolute calling is necessary and applied.

Corrected absolute log2ratio shows CN gain for the test sample.

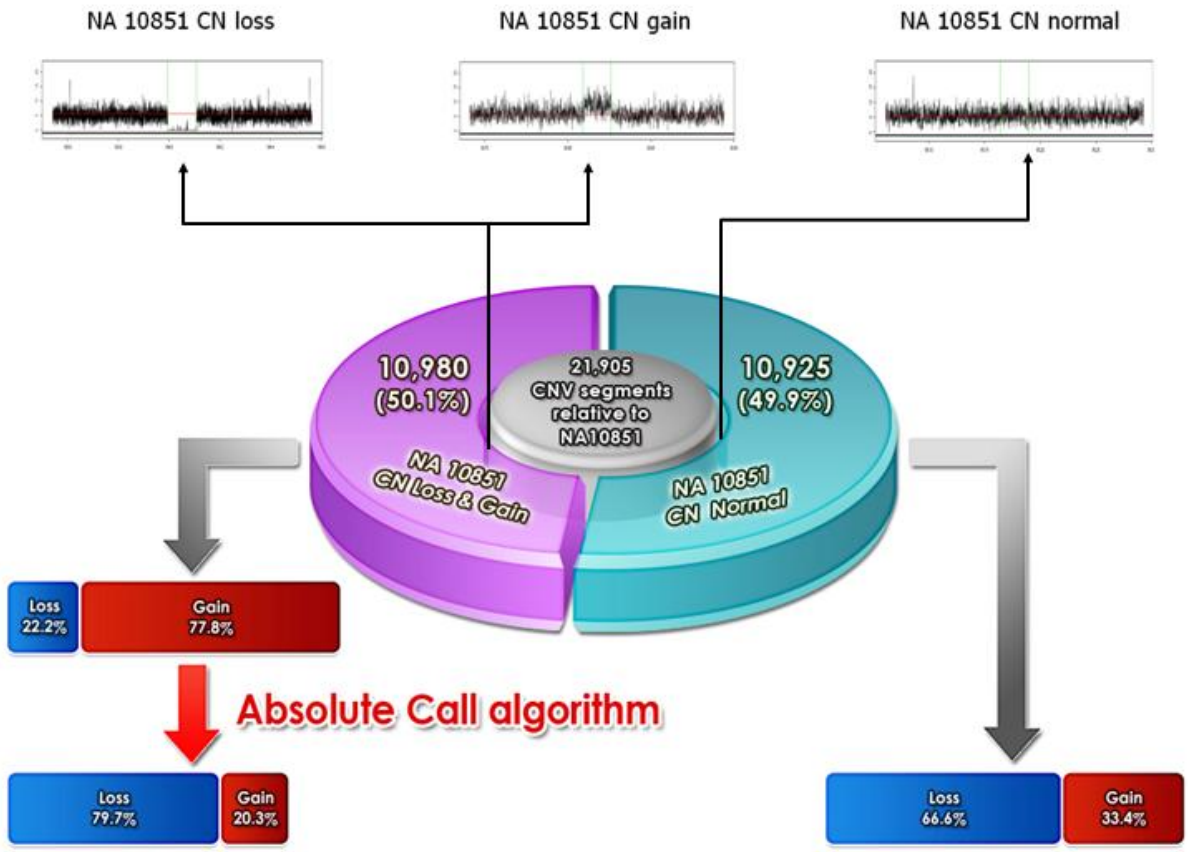


The CN gain of the test sample is confirmed by read-depth of sequencing.

Nature Genetics: doi:10.1038/ng.555

Q. Suppl. Figure 5. Copy number loss is the predominant type of human copy number variation.

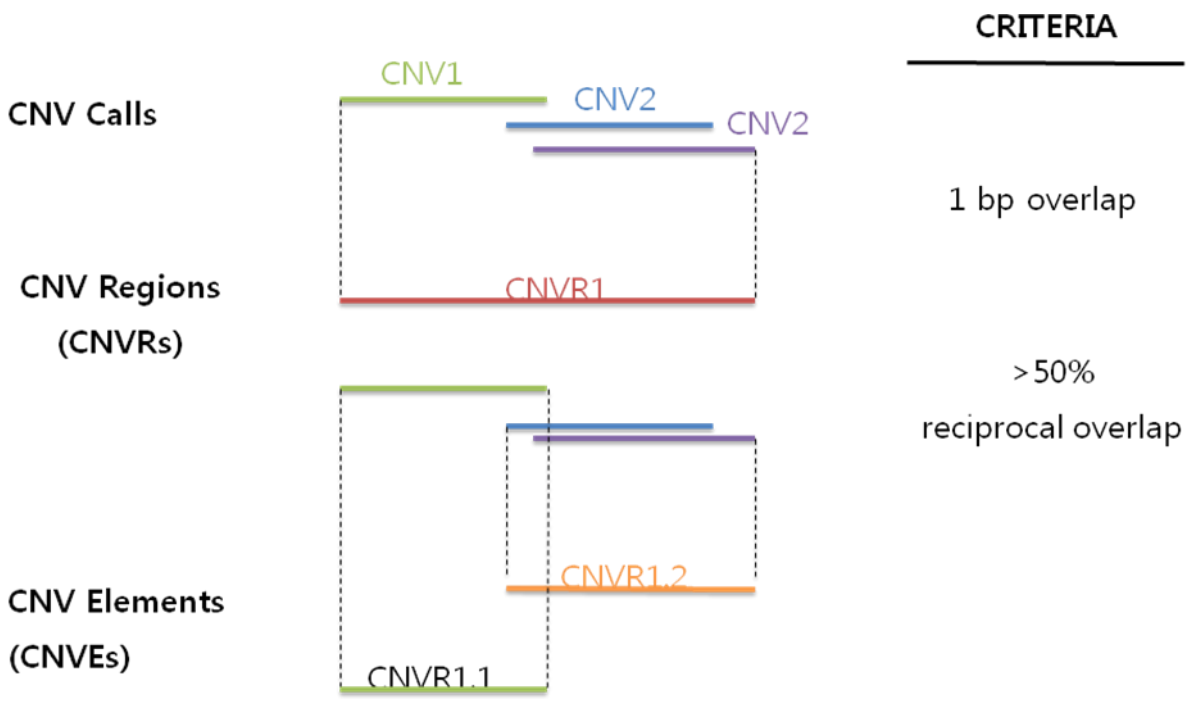
The proportion of CN loss and gain is disparate in overt calls (CNV segments in NA10851 CN normal regions) vs. obscure calls (CNV segments in NA10851 CNV regions). Applying the absolute calling algorithm to the obscure calls increases corrects the total ratio of CN loss significantly, which is more consistent with previous studies.



Nature Genetics: doi:10.1038/ng.555

R.Suppl. Figure 6. Definition of CNVs, CNVR, and CNVE

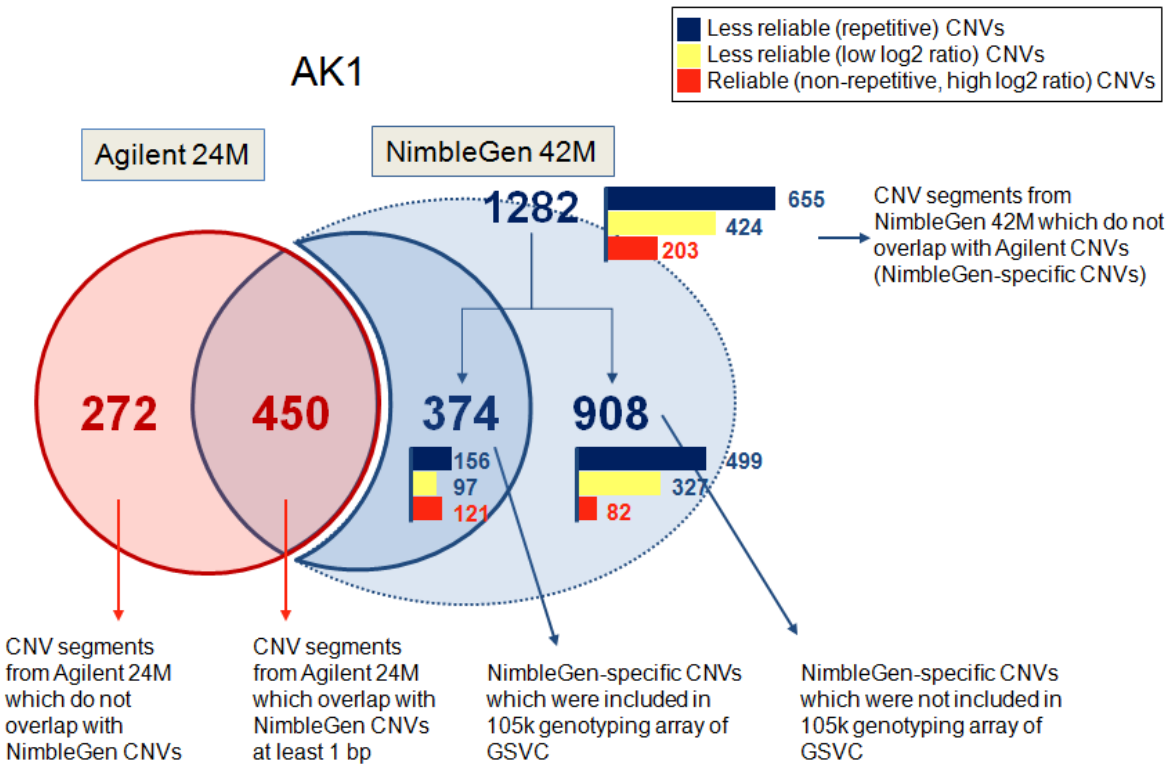
CNV calls with any overlap are combined into CNV regions, while CNV elements are composed of CNV calls that have more than 50% of their sequence in common.



Nature Genetics: doi:10.1038/ng.555

S. Suppl. Figure 7. A Comparison between the Agilent 24M array platform and the NimbleGen 42M array platform using genomic DNA from AK1

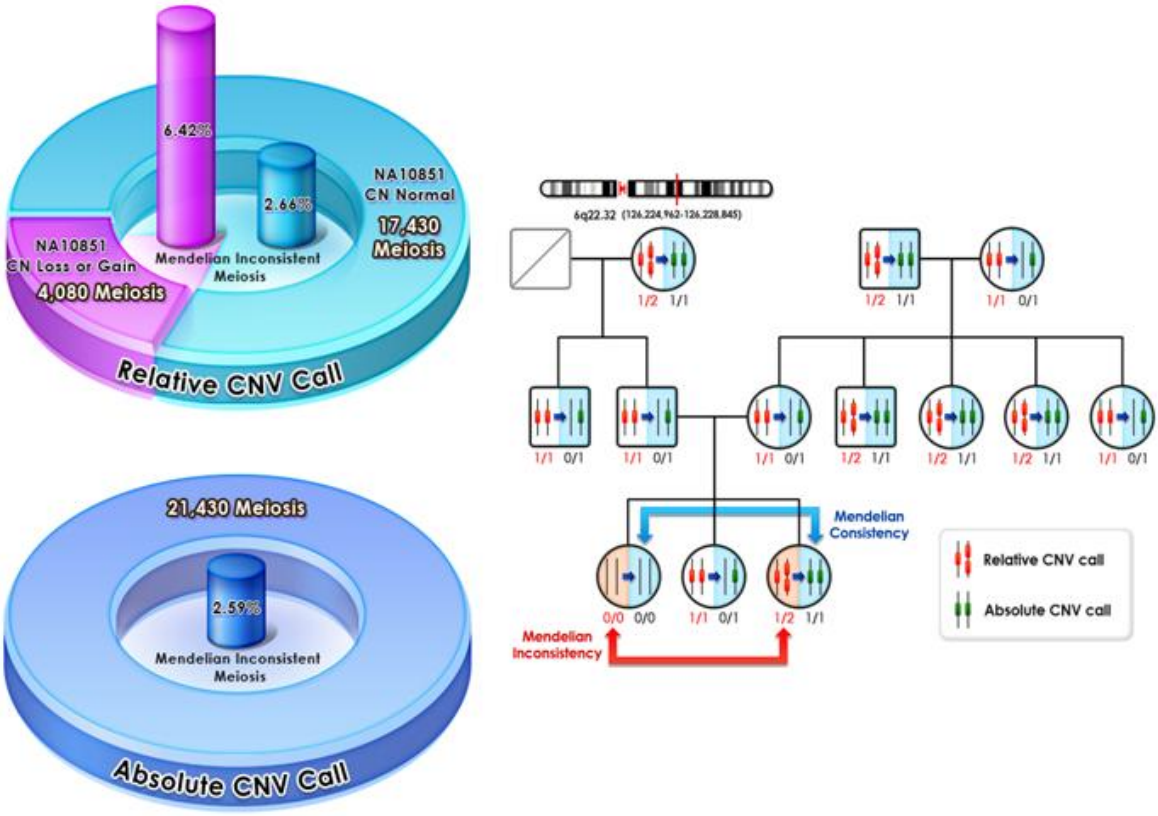
Left circle of the Venn diagram represents 722 CNV segments obtained from Agilent 24M aCGH (Agilent CNVs) and right circle represents CNV segments found by the NimbleGen 42M platform (NimbleGen CNVs). Intersection of the two circles represents 450 Agilent CNVs which have at least one bp overlap with the NimbleGen CNVS. Outer circle on the right with blue color denotes 1,282 NimbleGen CNVs which do not overlap with Agilent CNVs at all (NimbleGen-specific CNVs). Dotted right circle represents NimbleGen-specific CNVs which are included in 105k genotyping array set. Right panel indicates that 1,282 NimbleGen-specific CNVS can be divided into three classes as described in the figure (See **Supplementary Note** for detailed explanation).



Nature Genetics: doi:10.1038/ng.555

T. Suppl. Figure 8. Mendelian inconsistency of CNVs in a large Mongolian family using 180k probe aCGH array

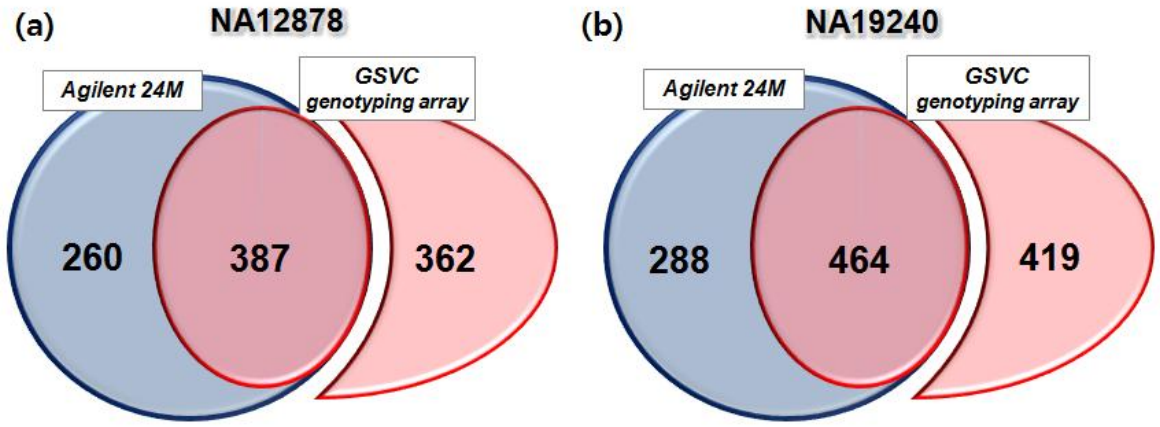
Investigation of Mendelian inconsistency in a Mongolian family, pre and post corrections for NA10851 copy number status. 6.42% of the meioses were Mendelian inconsistent using relative copy number data. 2.59% of the meioses were Mendelian inconsistent using absolute copy number data. An example of relative and absolute genotype calls for a CNV on chromosome 6 in this Mongolian family is shown on the right.



Nature Genetics: doi:10.1038/ng.555

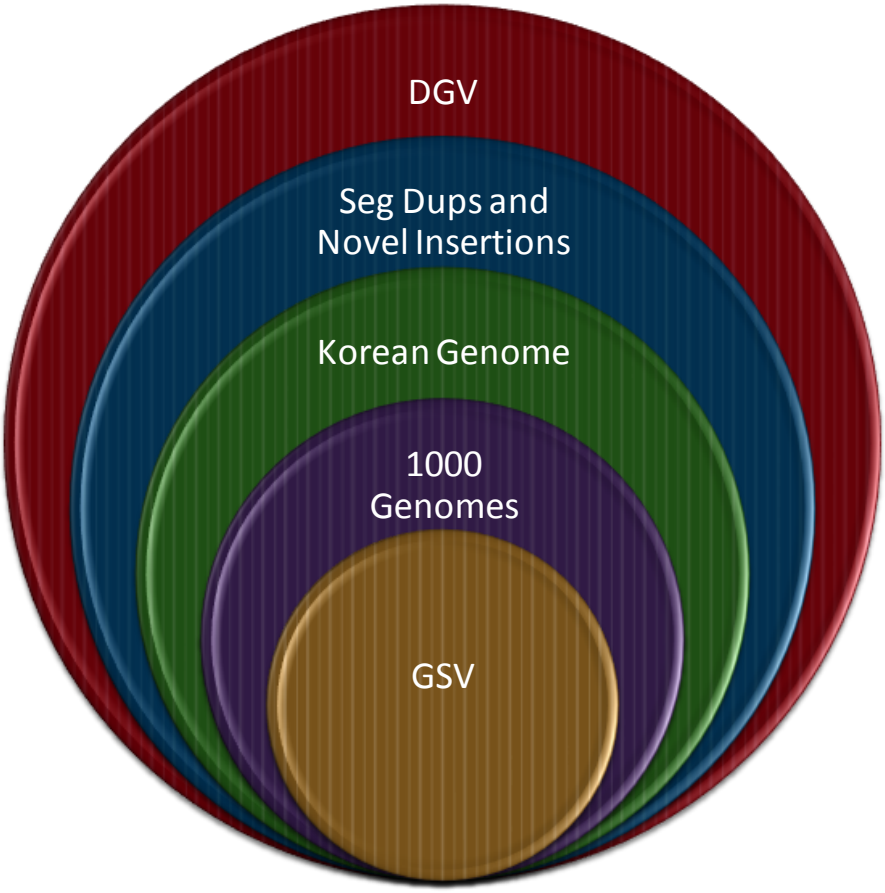
U. Suppl. Figure 9. Comparison of CNV calls made with the Agilent 24M aCGH platform and data from a 105k CNV genotyping platform by Genome Structural Variation Consortium (GSVC)

CNV segments of NA12878 (a, HapMap sample with European origin) and NA19240 (b, HapMap sample with African origin) obtained from Agilent 24M (in this study) and from 105k CNV genotyping array (GSVC[1]) were compared by one base pair overlap. Left circle of the Venn diagram represents CNV segments obtained from Agilent 24M aCGH in this study (Gemomic Medicine Institute(GMI) CNVs) and right circle represents CNV segments found by 105K platform by GSVC[1] (GSVC CNVs). Intersection of two circles represents GMI CNVs which have at least one bp overlap with the GSVC CNVs. Right outer circle with red color denotes GSVC CNVs which do not overlap with GMI CNVs at all (See Supplementary Note for detailed explanation).



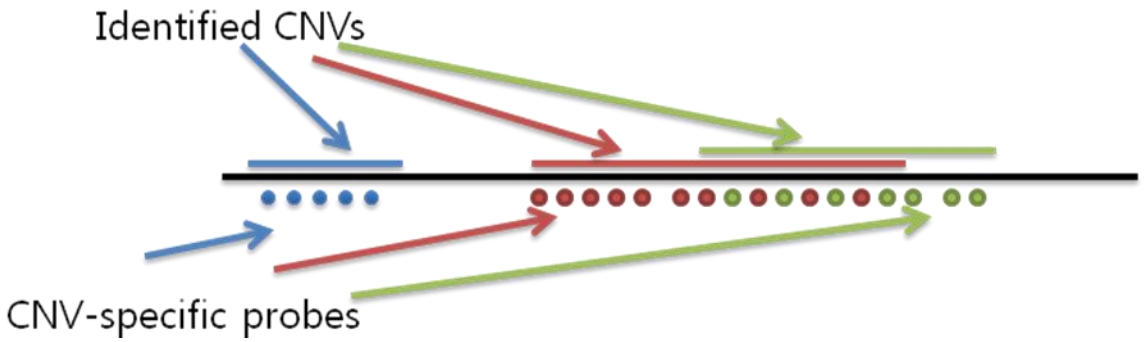
Nature Genetics: doi:10.1038/ng.555

V. Suppl. Figure 10. The hierarchical selection of CNVRs which were included on the 180k probe aCGH array. CNV regions in higher tier (inner) data sets were given deferential preference with regards to size and breakpoints when overlapped with lower tier databases.



Nature Genetics: doi:10.1038/ng.555

W. Suppl. Figure 11. The distribution of probes within the targeted CNVRs in 180k probe aCGH array. For each CNVR, there are between 6 and 9 distinct probes. Where possible, each probe corresponds to a single CNVR to ensure a clear-cut analysis of overlapping CNVRs.



Nature Genetics: doi:10.1038/ng.555

X. Suppl. Note

Platform difference

The platform and design used to build the 24Million CNV identification array set in this study are comparable to, but different from, the 42Million NimbleGen array set used in a previous study by Conrad *et al*[1]. A large portion of the human genome consists of moderately and highly repetitive sequences, where identifying CNVs using hybridization methods is less feasible due to the lack of high quality oligonucleotide probes. To design a high performance CNV identification array set, we excluded low quality probes using a homology filter. As a consequence, most of the moderately and highly repetitive DNA sequences and some segmental duplications were not included on the 24Million Agilent CNV identification array set. This resulted in an effectively smaller portion of the genome being assayed for CNVs with a lower false positive rate. The median inter-probe distance (in the interrogated areas) for the 24Million Agilent array set was 40 bp, which is significantly smaller than that of the 42Million NimbleGen array platform at 50 bp. In contrast, the NimbleGen 42M array set includes a large number of probes in moderately – highly repetitive genomic regions (e.g., 43% of the probes are in highly repetitive regions), since their probes were designed to be evenly distributed throughout the entire human genome (Supplementary Figure 1). Hence, the Agilent 24Million array set platform interrogates a smaller portion of the genome at higher resolution, whereas the NimbleGen 42Million array set platform interrogates a larger section of the genome that includes a majority of repeats and segmental duplications, at a slightly lower resolution but with more uniformly distributed probes.

To compare these two array sets, the genomic DNA from AK1 were analyzed on both platforms (Supplementary Figure 7). The Agilent 24M platform revealed 722 CNVs. For the NimbleGen 42M platform, the filter conditions used in Conrad *et al.* were applied (i.e., ≥10 consecutive probes with an average log2 ratio ≥ 0.1, and ≤ -0.25 for CN gains and losses, respectively)[1], resulting in 1,829 CNVs. 62.3% (n=450) of the CNVs identified in AK1 by the Agilent 24M platform overlapped ≥ 1 bp with

Nature Genetics: doi:10.1038/ng.555

CNVs identified by the NimbleGen 42M platform. A substantial number of CNVs (n=1,282) were specific to the NimbleGen platform. Further analysis revealed that 655 (51.1%) of these CNVs were in moderately – highly repetitive regions (hence no probes on the Agilent 24M array platform were available to interrogate these regions). Moreover, 424 (33.1%) of the CNVs showed had log2 ratios which did not meet the more stringent filter criteria established for the Agilent 24M platform (Supplementary Table 2). Consequently, only 203 NimbleGen-specific CNVs were relevant to this comparison. The genotyping array of Conrad *et al.*[1] seems to have culled the CNV regions identified from their discovery studies, thereby targeting a selected subset and including fewer repetitive regions. Hence, only 30% (374/1282) of the NimbleGen 42M specific CNVs were included in the genotyping arrays, perhaps reflecting decreased confidence in the remainder of these putative CNV loci.

To further determine how these two aCGH platforms compare in the ultimate CNV calls made, we compared the CNV calls made for NA12878 (CEU) and NA19240 (YRI) using the Agilent 24M platform with those identified by Conrad *et al.*[1] using a 1bp overlap criteria (Supplementary Figure 9). Both of the results showed that ~60% (59.8% and 61.7%, respectively, (Figure 5)) of the CNVs identified by our Agilent 24M platform were also identified with the 105k CNV genotyping platform in Conrad *et al.*[1], and ~40% (40.2 % and 38.3%, respectively) of the CNVs discovered by the Agilent 24M platform in NA12878 and NA19240 were not captured by the 105k CNV genotyping platform. Taken together, these comparisons indicate that ~40% of the Agilent 24M CNV calls may be platform dependent and not captured by the NimbleGen 42M array set or 105k CNV genotyping platforms used in GSVC data[1].

Upon analysis of the CNVs for population stratification / differentiation, 1,630 of the 5,177 CNVEs were also discovered in the CEU and YRI populations by Conrad *et al* (Figure 4a and Supplementary Table 7; 5,177-3,547=1,630). If these are considered to represent platform independent CNVs, we would expect approximately 1,100 (1630 X 40%/60%) CNVs calls to be specific to the Agilent platform. However, we identified 3,547 novel CNV regions in our experiments, indicating that ~70% of these may be specific to Asian ethnicities.

Nature Genetics: doi:10.1038/ng.555

Technologies for CNV detection; aCGH and massively parallel sequencing

Recently, many genomic technologies have been used for detecting copy number variation in the human genome[2-4]. Among these, CGH microarrays have been used by the majority of comprehensive studies[1,3,5-7]. The resolution of aCGH platforms has continuously risen to a point where it is now possible to identify CNVs with sizes of only a few hundred bases.

aCGH identifies CNVs in a test sample by comparison to a reference or control sample. Ideally, the control sample is expected to have a normal two copy value for every region across the genome. However, since no known human genomic sample actually has two copies of every segment of the human reference genome (hg18, assembly build 36.3), we have attempted to ascertain the absolute copy number value for each putative CNV region of interest in NA10851, a commonly used reference individual[1,6-7]. This information can then be used to convert the relative copy number information obtained from aCGH experiments to absolute copy numbers.

Research has recently been performed in identifying CNVs through paired-end mapping and/or observing read depth (coverage) changes by massively parallel sequencing technology, but few have validated these findings in depth[4,8-11]. Resequencing methods do not require any reference sample for detecting CNVs. However, paired-end mapping methods have limitations for identifying smaller copy number losses or larger copy number gains, and some regions in the human genome show coverage drops or ‘spikes’ (due to GC ratio or other uncertain reasons) that can result in many false positive and false negative CNV calls[4,12]. In a previous study, we reported a highly-confident set of CNVs in a Korean individual (AK1), using a combination of next generation sequencing and an earlier version of the 24M array platform[12]. However, only 19.1% of the CNVs detected by this high-resolution aCGH platform could be confirmed by read-depth (sequence coverage). Such a low correlation has also been reported by other groups[9,11].

From these previous reports, we realized the importance of having sequence data of

Nature Genetics: doi:10.1038/ng.555

NA10851, the control individual used in the present study as well as other large-scale CNV studies. Since aCGH reports CNVs relative to NA10851, the log2 ratio of CNV calls are comparable with the sequence read-depth (coverage) ratio of a test sample and NA10851 rather than the read-depth change of the test sample alone. In this study, we compared the aCGH results of AK1 with the read-depth ratio of AK1 to NA10851. This resulted in a much higher correlation rate between aCGH and massively parallel sequencing up to ~90% (Details are explained in section below). This higher correlation rate enabled us to validate aCGH results as a whole using genome sequence data. By using this validation data, we adjusted filter conditions to minimize false positive aCGH CNV calls.

Training the filtering criteria for calling CNV with aCGH data

Array CGH reports a differing number of CNV calls under various filter conditions. The ADM-2 algorithm (Agilent Inc. CA) provides an average log2 ratio and corresponding p-value for each CNV segment.

ADM-2 identified 17,890 unfiltered CNV segments in AK1's genome by aCGH. We then attempted to identify (by visual inspection) significant read-depth ratio (AK1 read depth/NA10851 read depth) change for each putative CNV segment comparing these to the read depth ratio for their flanking genomic regions. Short reads were aligned by single base windows using a random alignment method to calculate genome-wide read-depth. A CNV segment identified by aCGH was considered to be confirmed when a significant change of read depth ratio was obtained for the CNV segment, compared to the flanking regions. Validation of a group of 300 randomly selected segments out of the total 17,890 provided initial filter conditions to minimize false positives (i.e., we empirically determined that each CNV segment should be called by ≥5 consecutive probes with a p-value of $< 10^{-7}$ if |log2 ratio| ≥0.5 and a p-value of $<10^{-17}$ if 0.5> |log2 ratio| ≥0.2). Applying these criteria to the entire 17,890 CNV segments identified by aCGH in AK1 resulted in 1,853 primary filtered CNV segments.

Nature Genetics: doi:10.1038/ng.555

Read-depth ratio plots for all1,853 primary-filtered CNV segments were generated for further filter training. 721 (38.9%) CNV segments showed significant (by visual inspection) read-depth ratio changes which correlated with the aCGH log2 ratio and thus were thought to be final true positive CNVs in AK1 (Supplementary Figure 3). We set filter conditions by systematically adjusting the threshold log2 ratios and p-values to minimize false positives while maximizing true positives. Due to the vast predominance of non-CNV areas in the genome, using 'specificity' in the classical sense results in less discrimination between conditions with different performance. We therefore utilized positive predictive value (PPV) to efficiently resolve these differences. Read-depth ratio information for AK1 to NA10851 was used as the gold standard. Modified ROC curves were generated using PPV and relative sensitivity, and final filter conditions were set where both the PPV and relative sensitivity were substantially high (Supplementary Table 1; Supplementary Figure 2). Final filter conditions gave a positive predictive value and relative sensitivity for CNV detection of 0.840 and 0.845, respectively (Supplementary Table 2).

In order to test our established filter conditions, they were applied to 8,241 raw autosomal CNV segments of AK2, identifying 695 filtered segments. The PPV was 0.855, similar to that observed for AK1 (Supplementary Table 2).

Absolute Call analysis method

We are aware of examples where multiple copies of a gene exist in normal individuals. However, for the sake of simplicity, we will use the words 'diploid' and 'two copies' interchangeably when referring to 'normal' copy number segments of a genome.

While adjusting filter conditions using the AK1 and NA10851 sequence, we realized that only approximately half of the filtered AK1 CNV segments were overt CNV calls, or not associated with the CN gain or loss of NA10851 (Supplementary figure 2b). The other half of filtered AK1 CNV segments were 'obscure calls', which were influenced by the copy number state of the corresponding DNA segment in NA10851. In other words, they were explained by CN gain or loss of NA10851 rather than in

Nature Genetics: doi:10.1038/ng.555

AK1. For example, if the test sample has 2 copies (normal) of a given CNV region and NA10851 has 1 copy of the same genomic region, aCGH identifies this as a relative copy number gain in the test sample (obscure call, Supplementary Figure 4d). This is because aCGH compares test and reference samples as described above. In addition, relative log2 ratios in the CNV segment can be under/overestimated and in extreme cases, CN loss of a test sample can be called as CN gain, or *vice versa*, if the reference sample simultaneously has homozygous deletion or higher CN gain, respectively, in the genomic region. (another example of an obscure call; Figure 2a). On the other hand, when the test and the reference sample have identical copy number states for a genomic region (e.g., each individual has a 1 copy of a genomic segment), aCGH fails to identify the region as a CNV and therefore this is referred to as a "covert" CNV (Supplementary Figure 4f-4i). Obscure and covert calls should be modified and reinstated, respectively, to identify the absolute copy number. In other words, we should identify the absolute copy number state for each CNV region in each person being studied, rather than the relative copy number state compared to a reference sample.

By combining aCGH and next generation sequencing data for the reference sample, NA10851, we were able to design new methods to identify the absolute copy number for all regions in each individual tested. The application of this algorithm is not limited to only this study, but can be implemented to any aCGH experiment using NA10851 as a reference. First, we identified CNV regions in NA10851 itself, using high resolution array CGH data for 30 Asian women (present study), together with 19 CEU women, 20 YRI women and 1 polymorphic discovery resource individual[1] as well as whole genome sequence data for NA10851 (present study). Since CNV regions in NA10851 are likely to cause obscure CNV calls in a test sample unless the test sample has identical CN status to NA10851, they are likely to be more frequently identified CNV regions in these studies. Hence, high frequency CNV regions in these studies are good candidates for CNVs in NA10851. CNV loci where $\geq$ 10 out of 70 individuals showed copy number variations were investigated by sequence read-depth data of NA10851, using 30, 50, 100 or 1000 bp windows for CNVs with sizes of <1 kb, >1 kb, >100 kb and >1 Mb, respectively. Genomic regions

Nature Genetics: doi:10.1038/ng.555

with substantially higher or lower coverage than their flanking regions were considered to have copy number gain or loss. Using these methods, we identified ~550 putative CNV regions in NA10851. This combination of high resolution aCGH and massively parallel sequencing effectively identified validated CNVs in NA10851.

Using the NA10851 CNV data, an absolute calling algorithm was developed. CNV regions in NA10851 were categorized into 0-copy regions, 1-copy regions, copy number gain regions and complex regions, where sequencing read-depth was close to 0, significantly lower than flanking regions, significantly higher than flanking regions, and not evenly distributed, respectively. Different categories of NA10851 CNV regions required different strategies to calculate absolute calls.

a. NA10851 2-copy region.

If a CNV segment of the test sample does not overlap with any CNVs in NA10851, it is considered to be located in the 2 copy region of NA10851. Absolute log2 ratio of the segment is identical to its relative log2 ratio (Supplementary Figures 4a-4b)

b. NA10851 0-copy region

If a CNV segment of test sample overlaps with one of the complete loss regions in NA10851, its log2 ratio becomes unstable since the aCGH intensity of NA10851 is close to 0. Generally, regardless of real copy number status of the test sample, the log2 ratio becomes very large in value and very sensitive to the degree of background noise (Supplementary Figures 4c, 4f). Therefore, the alternative log2 ratio was calculated by taking the ratio of region's signal intensity in the test sample over the average signal intensity of test sample, and substituted for relative log2 ratio. If the alternative absolute log2 ratio value met final filter criteria, it was considered as a positive CNV call.

$$\log_2 \text{ratio}_{abs} = \log_2 \left(\frac{\overline{\text{signal intensity}}_{\text{CNVR, sample}}}{\overline{\text{signal intensity}}_{\text{slide, sample}}} \right)$$

Nature Genetics: doi:10.1038/ng.555

c. NA10851 1-copy or CN gain region

In these cases, if the sample copy number is two (or normal), array CGH will yield a positive call (Supplementary Figures 4d-e, *4*g-4i). However, if the sample copy number is identical to that of NA10851, array CGH will miss the region (Supplementary Figures 4g-i).

If a CNV segment of a sample overlaps one of these regions (obscure call), its log2 ratio should be recalculated, since the copy number of the reference sample is not two (normal). Absolute log2 ratio is calculated using the following formula.

$$\log_2 \text{ratio}_{absolute} = \log_2 \text{ratio}_{relative} + \log_2 \left(\frac{\overline{\text{Coverage}}_{\text{CNV, NA10851}}}{\overline{\text{Coverage}}_{\text{whole-genome, NA10851}}} \right)$$

If the absolute log2 ratio value meets the final filtering criteria, it is included as a positive CNV call (with corrected log2 ratio).

If none of the CNV segments of a sample overlaps one of these regions, the relative log2 ratio for the region is first calculated from normalized aCGH data using log2 ratio of all the probes in the corresponding region. Absolute log2 ratio is then determined as below.

$$\log_2 \text{ratio}_{absolute} = \overline{\log_2 \text{ratio}}_{probes\ in\ region} + \log_2 \left(\frac{\overline{\text{Coverage}}_{\text{CNV, NA10851}}}{\overline{\text{Coverage}}_{\text{whole-genome, NA10851}}} \right)$$

If the absolute log2 ratio value meets the final filter criteria, it is included as a positive CNV call.

d. NA10851 complex CNV region

For segments that overlap complex coverage regions, we convert the relative log2 ratio into an absolute log2 ratio by the following equation

Nature Genetics: doi:10.1038/ng.555

$$\log_2 \text{ratio}_{absolute} = \log_2 \text{ratio}_{relative} + \log_2 \left(\frac{\overline{\text{Coverage}}_{\text{CNV, NA10851}}}{\overline{\text{Coverage}}_{\text{whole-genome, NA10851}}} \right)$$

When we applied the absolute call algorithm to CNV segments from 30 Asian women, 48% (10,558) of the 21,905 total relative CNV segments were candidates for modification, since they overlapped NA10851 CNV regions (Figure 2b). Out of 10,558, 6,197 false positives were removed and 4,361 segments with under/overestimated log2 ratio had their log2 ratio corrected. In addition, 4,139 false negatives were identified. The copy number gain to loss ratio of the obscure calls was corrected from a predominance in gains to a predominance in losses using the absolute call algorithm, consistent with the ratios of gains and losses previously observed in other studies (Supplementary Figure 5).

Gene ontology analysis

Among 5,177 CNVEs, 383 and 1,059 were found to have CN gain and CN loss, respectively, in more than or equal to 10% of 30 Asians and were considered as Asian common CNVEs. When we counted genes in which coding sequences (CDS) overlapped with common CNVEs, 229 and 159 genes were found to be located in CN gain and CN loss regions, respectively. When we utilized PANTHER ontology (http://www.pantherdb.org) using “NCBI *H. sapiens*” option for classifying 229 genes with common CN gains, 184 genes were matched with 26 Biological Process terms in Panther database. We reclassified 26 terms into 8 major categories to simply visualize them in Figure 4b. For common CN losses, 132 of 159 genes were matched with 22 Biological Process terms, which were reclassified into 8 major categories in Figure 4b. Details of genes and gene ontology terms were listed in Supplementary Table 12.

Nature Genetics: doi:10.1038/ng.555

Y.References

1. Conrad, D.F. et al. Origins and functional impact of copy number variation in the human genome. *Nature* (2009).
2. Scherer, S.W. et al. Challenges and standards in integrating surveys of structural variation. *Nat Genet* **39**, S7-15 (2007).
3. Hurles, M.E., Dermitzakis, E.T. & Tyler-Smith, C. The functional impact of structural variation in humans. *Trends Genet* **24**, 238-45 (2008).
4. Bentley, D.R. et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53-9 (2008).
5. Carter, N.P. Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat Genet* **39**, S16-21 (2007).
6. Perry, G.H. et al. The fine-scale and complex architecture of human copy-number variation. *Am J Hum Genet* **82**, 685-95 (2008).
7. Redon, R. et al. Global variation in copy number in the human genome. *Nature* **444**, 444-54 (2006).
8. Alkan, C. et al. Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat Genet* **41**, 1061-7 (2009).
9. Korbel, J.O. et al. Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**, 420-6 (2007).
10. Wang, J. et al. The diploid genome sequence of an Asian individual. *Nature* **456**, 60-5 (2008).
11. Yoon, S., Xuan, Z., Makarov, V., Ye, K. & Sebat, J. Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res* **19**, 1586-92 (2009).
12. Kim, J.I. et al. A highly annotated whole-genome sequence of a Korean individual. *Nature* **460**, 1011-5 (2009).

Nature Genetics: doi:10.1038/ng.555