

ORIGINAL ARTICLE

A population-based study of copy number variants and regions of homozygosity in healthy Swedish individuals

Shu-Mei Teo^{1,2,3,5}, Chee-Seng Ku^{2,5}, Nasheen Naidoo², Per Hall¹, Kee-Seng Chia^{1,2,4}, Agus Salim⁴ and Yudi Pawitan¹

The abundance of copy number variants (CNVs) and regions of homozygosity (ROHs) have been well documented in previous studies. In addition, their roles in complex diseases and traits have since been increasingly appreciated. However, only a limited amount of CNV and ROH data is currently available for the Swedish population. We conducted a population-based study to detect and characterize CNVs and ROHs in 87 randomly selected healthy Swedish individuals using the Affymetrix SNP Array 6.0. More than 600 CNV loci were detected in the population using two different CNV-detection algorithms (PennCNV and Birdsuite). A total of 196 loci were consistently identified by both algorithms, suggesting their reliability. Numerous disease-associated and pharmacogenetics-related genes were found to be overlapping with common CNV loci such as CFHR1/R3, LCE3B/3C, UGT2B17 and GSTT1. Correlation analysis between copy number polymorphisms (CNPs) and genome-wide association studies-identified single-nucleotide polymorphisms also indicates the potential roles of several CNPs as causal variants for diseases and traits such as body mass index, Crohn's disease and multiple sclerosis. In addition, we also identified a total of 14 815 ROHs ≥ 500 kb or 2814 ROHs ≥ 1 Mb in the Swedish individuals with an average of 170 and 32 regions detected per individual respectively. Approximately 141 Mb or 4.92% of the genome is homozygous in each individual of the Swedish population. This is the first population-based study to investigate the population characteristics of CNVs and ROHs in the Swedish population. This study found many CNV loci that warrant further investigation, and also highlighted the abundance and importance of investigating ROHs for their associations with complex diseases and traits.

Journal of Human Genetics (2011) 56, 524–533; doi:10.1038/jhg.2011.52; published online 2 June 2011

Keywords: Affymetrix SNP Array 6.0; Birdsuite; copy number variants; PennCNV; regions of homozygosity; Swedish population

INTRODUCTION

There is a growing body of copy number variant (CNV) maps covering different world populations.^{1–5} Most of these newer studies used high-resolution methods for detecting CNVs, such as the Affymetrix SNP Array 6.0, which has a higher density of single-nucleotide polymorphism (SNP) and copy number probes than previous microarray-based methods. This has led to an improved performance of microarray-based methods to detect smaller CNVs (<50 kb).^{1,6} In contrast, previous studies have used much lower resolution arrays, such as the bacterial artificial chromosome (BAC) clone or oligonucleotide comparative genomic hybridization arrays and SNP genotyping arrays.^{7–10} Currently, there is only one CNV-detection study in a Swedish population,¹⁰ but this was performed in a small sample size of 33 individuals and used a low-resolution 32-K bacterial artificial chromosome clone microarray. This has hampered the study from detecting less common and smaller CNVs and from estimating the population frequency of CNVs. The ability to

detect smaller CNVs is critical as they are more numerous than the larger CNVs.¹¹

In addition, the study by Díaz de Ståhl *et al.*¹⁰ was unable to detect regions of homozygosity (ROHs) as the bacterial artificial chromosome clone microarray was unable to generate allelic intensity data. Research on ROHs has started to gain impetus, as evidenced by the increasing number of publications after the first study by Gibson *et al.*¹² reported the abundance of ROHs in the human genome of outbred populations. Further studies have investigated the population characteristics of ROHs in healthy individuals,^{13–15} and also performed association analyses to identify ROHs that are associated with complex diseases and traits in a case–control study design.^{16–18}

To circumvent the limitations of the previous study by Díaz de Ståhl *et al.*,¹⁰ we conducted a study in a Swedish population by genotyping 100 individuals using the Affymetrix SNP Array 6.0 (Affymetrix, Santa Clara, CA, USA). The main aim of this study was to perform a more comprehensive detection of CNVs and ROHs in the Swedish

¹Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden; ²Centre for Molecular Epidemiology, Department of Epidemiology and Public Health, Yong Loo Lin School of Medicine, National University of Singapore, Singapore; ³NUS Graduate School for Integrative Sciences and Engineering, National University of Singapore, Singapore and ⁴Department of Epidemiology and Public Health, Yong Loo Lin School of Medicine, National University of Singapore, Singapore

⁵Joint first author.

Correspondence: C-S Ku, Center for Molecular Epidemiology, National University of Singapore, Singapore 117597, Singapore.

E-mail: csikcs@nus.edu.sg or Professor Y Pawitan, Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, PO Box 281, 17177 Stockholm, Sweden.

E-mail: Yudi.Pawitan@ki.se

Received 18 January 2011; revised 12 April 2011; accepted 25 April 2011; published online 2 June 2011

population and to describe their population characteristics. Although several studies have been performed to detect and characterize CNVs and ROHs in multiple European populations, these studies have also documented the genetic differences among these populations.^{14,15,19} The extension of the International HapMap Project to include an additional seven populations in Phase III further suggests that multiple populations from diverse ancestries or different geographical locations are needed to study their population genetics.²⁰ These previous studies have justified the need for a population-based study to characterize CNVs and ROHs in healthy Swedish individuals. We also compared the Swedish population with the HapMap phase III populations using principal component analysis.

MATERIALS AND METHODS

Samples and genotyping platform

A total of 100 randomly selected healthy Swedish individuals volunteering as controls in case-control studies were studied. Peripheral blood samples of the participants for genomic DNA extraction were drawn and stored at the Karolinska Biobank. Identities of the participants were kept anonymous and no personal identifiers were used. All 100 samples were genotyped using the Affymetrix Genome-Wide Human SNP Array 6.0 as per the manufacturer's protocol. Two samples were removed from further analysis because their genotype call rates were below 98% and the remaining 98 samples were used for CNV detection.

CNV-detection algorithms and analyses

CNV calling using PennCNV. We used two CNV-detection algorithms, namely PennCNV²¹ and Birdsuite,²² for both comparison and validation. This study focused only on the CNVs in the 22 autosomes because of the inaccuracy of Birdsuite to detect CNVs in sex chromosomes. Log *R* ratio and B allele frequency were calculated according to the PennCNV algorithm (http://www.openbioinformatics.org/penncnv/penncnv_tutorial_affygw6.html). Smaller CNVs (<1 kb) were also included in our analysis, as PennCNV by default does not limit its detection to CNVs >1 kb in size. We applied a set of filtering criteria as recommended by the algorithm, namely Log *R* ratio-s.d >0.35, B allele frequency-median >0.55, B allele frequency-median <0.45 and B allele frequency-drift >0.006 to exclude samples with poor quality of signal intensity data (<http://www.openbioinformatics.org/penncnv/>). This resulted in a further exclusion of 11 samples, with the final set for analysis consisting of 87 samples. For each sample, PennCNV generated a list of CNVs with their confidence scores. The confidence score is a log Bayes factor that measures the likelihood that the locus harbors an abnormal copy number. A confidence score of ≥ 10 has been recommended as the threshold to classify reliable CNVs. Therefore, we retained all CNVs called with confidence scores ≥ 10 for subsequent analyses. Although the confidence score is only a statistical measure of a true positive, our previous study⁵ found that CNVs with a higher confidence score are more likely to be detected consistently across two genotyping platforms. Therefore, this justifies our decision to retain only reliable CNVs called with a sufficient degree of confidence.

Construction of CNV loci using PennCNV output. The CNVs called by PennCNV were shown to overlap across samples. Thus, we merged or grouped these individual CNV calls into discrete, non-overlapping loci, with the boundaries of each locus determined by the union of all CNVs that belonged to that particular locus. This construction of CNV loci was needed to estimate the population frequencies and these steps were performed using the methods that we have developed previously.^{5,23} We classified the status of these CNV loci into three categories, 'del' (loci containing deletions), 'dup' (loci containing duplications) and 'del/dup' (loci containing both deletions and duplications).

Copy number polymorphism (CNP) calling using Canary (Birdsuite). Birdsuite software²² was also used to analyze the Affymetrix SNP Array 6.0 data. There are two components in the software for detecting copy number changes, namely Canary and Birdseye. Canary was used to determine the integer copy number at each of the predefined 1316 CNPs. The term 'CNPs' used by

McCarroll *et al.*¹ is to describe common CNV loci. These 1316 CNPs were found in more than one HapMap II individual and their sizes were also accurately determined. Therefore, we used the Canary component in Birdsuite to determine the integer copy number of the 1316 CNPs in the 87 Swedish samples. These 1316 CNPs are distributed in all the autosomes and sex (X and Y) chromosomes. However, 25 CNPs located in the sex chromosomes were removed because the CNP calling in these chromosomes was less accurate. Thus, the results reported in this study comprised only 1291 CNPs in the 22 autosomes. Confidence statistics generated for the CNPs were also used to identify poor-quality calls, and only integer copy numbers detected with high confidence as recommended by the software (confidence score >0.1) were used for subsequent analyses.

Correlation analysis of CNPs. We performed a correlation analysis of CNPs and the nearby SNPs. Because the sizes of the CNPs were previously accurately determined by McCarroll *et al.*,¹ we restricted the analysis to only the CNPs detected by Canary. For each of the 1291 CNPs, SNPs within a 200-kb window from the start and end positions of the CNP were considered. We used the squared Pearson's correlation (r^2) for correlation analysis. The genotype calling of the Affymetrix SNP Array 6.0 was carried out using Birdsuite. In addition, to investigate the potential associations of CNPs with human diseases and traits, the same methods of r^2 calculations for the 1291 autosomal CNPs and the SNPs that were identified by genome-wide association studies (GWAS) were adopted. The list of GWAS-SNPs was downloaded from the National Human Genome Research Institute website (<http://www.genome.gov/gwastudies/>) on 26 October 2010.

CNV calling using Birdseye (Birdsuite). In addition to PennCNV, we also used another algorithm, Birdseye, to analyze the same set of data as different algorithms tend to have different sensitivities and specificities for detection of CNVs in different regions throughout the genome. As such, CNV loci detected by PennCNV and Birdseye can be cross-validated. Therefore, we used the Birdseye component in Birdsuite to detect additional CNVs throughout the genome, which was not restricted to the 1316 predefined CNPs. Similarly, only CNVs in autosomal chromosomes were used because of the inaccuracy of Birdseye in the sex chromosomes. CNVs with low confidence, as recommended by the software (confidence score ≤ 5), were removed from subsequent analysis.

Construction of CNV loci using Birdseye output. We also constructed CNV loci based on the Birdseye output using methods similar to those applied to the PennCNV output. The cutoff for the confidence score used by PennCNV (≥ 10) and Birdseye (≥ 5) was recommended by both algorithms. This allowed for greater comparability between the CNV loci detected by these two algorithms.

Comparison of CNV loci detected by PennCNV and Birdsuite. The CNV loci identified by PennCNV and Birdseye were compared as a 'validation' step. We used a 'reciprocal 50% overlapping' method to compare the CNV loci detected by these two algorithms and considered a CNV locus 'found' by both algorithms when this locus was detected in both PennCNV and Birdseye with an overlap of $\geq 50\%$ of their lengths.

Novel CNV loci. To identify novel CNV loci, we compared the CNV loci detected by PennCNV and Birdseye with the data from the Database of Genomic Variants (DGV).²⁴ We used the latest data from the DGV (variation.hg18.v8.aug.2009.txt and indel.hg18.v8.aug.2009.txt) downloaded from the DGV Website (<http://projects.tcag.ca/variation/>). A CNV locus identified by PennCNV and Birdseye was considered novel if it did not share at least 50% of its length with any CNV loci cataloged in the DGV. All the downstream analyses after PennCNV and Birdsuite were performed using the statistical software package R (<http://www.r-project.org/>).

Comparison with HapMap phase III populations

The CEL files of the Affymetrix SNP Array 6.0 for the seven populations in HapMap phase III project were downloaded from the ftp site (ftp://ftp.ncbi.nlm.nih.gov/hapmap/raw_data/hapmap3_affy6.0/). The HapMap phase III populations studied are people of African ancestry in the southwestern USA (ASW), the Chinese community in Metropolitan Denver, Colorado, USA (CHD), Gujarati Indians in Houston, Texas, USA (GIH), the Luhya in Webuye,

Kenya (LWK), people of Mexican ancestry in Los Angeles, California, USA (MEX), the Maasai in Kinyawa, Kenya (MKK) and the Tuscans in Italy (TSI). All the samples were analyzed using Canary similarly to the analysis of the Swedish population. Only unrelated samples were included in our study, that is, family-related samples were removed using the 'relationships' file provided by the International HapMap Project. After the sample exclusion step, a total of 594 unrelated samples from the seven HapMap III populations were analyzed: ASW ($n=52$), CHD ($n=89$), GIH ($n=89$), LWK ($n=90$), MEX ($n=53$), MKK ($n=132$) and TSI ($n=89$). We performed principal component analysis to compare the Swedish population with the HapMap phase III populations using the CNP output generated by Canary.

ROH-detection algorithms and analyses

In addition to CNVs, we also detected ROHs using PennCNV in the 22 autosomes of the 87 Swedish individuals. However, we only focused on ROHs ≥ 500 kb, as this cutoff was adopted in a previous study.¹⁸ For each of these we confirmed that they are ROHs by determining the genotypes of the SNPs that fall within each region. We then calculated the percentage of heterozygosity (number of heterozygotes/total number of heterozygotes and homozygotes). We also calculated the percentage of missingness (number of missingness/total number of SNPs in each ROH). First, we used an arbitrary cutoff of the median of the percentage of heterozygosity (2.5%) to allow for some heterozygote calls resulting from calling or genotyping errors. As a result, we removed half of the ROHs with a percentage $> 2.5\%$. Second, we removed ROHs with $> 1\%$ for the missingness, to remove regions where genotype calling was problematic. Finally, for the remaining ROHs, we also ensured a density of one SNP per 10 kb to exclude those ROHs that could be spuriously detected by a sparse number of SNPs. As such, for a 500-kb ROH, a minimum of 50 SNPs is required. These three criteria were used as the filters to exclude less reliable ROHs. Several summary statistics were then computed to describe the characteristics of ROHs in the Swedish population.

RESULTS

Characteristics of CNVs identified by PennCNV

After filtering unreliable CNV calls, an average of approximately 36 CNVs per individual with a ratio of deletions to duplications of approximately 2.6:1 was discovered (Supplementary Table 1). The number of CNVs per individual ranged from 22 to 65. The median size of a CNV was 28.6 kb and approximately 66% of the CNVs were < 50 kb and 26% were < 10 kb (Supplementary Figure 1). The median size of deletions was approximately fourfold smaller than the median size of duplications.

Characteristics of CNV loci identified by PennCNV

We merged overlapping CNVs to construct CNV loci and identified 623 loci, of which 476 loci contained deletions ('del-loci'), 102 loci contained duplications ('dup-loci') and 45 loci contained both deletions and duplications ('del/dup-loci'; Table 1). These 623 loci covered approximately 61.52 Mb of the nucleotide sequence and the sum of the lengths for del-loci (19.83 Mb) was smaller than that for dup-loci (25.80 Mb). Similarly for the individual CNVs (Supplementary Table 1), the average size of del-loci (41.66 kb) was much smaller than that of dup-loci (252.93 kb; Table 1). More than 77% of the del-loci were < 50 kb, and in comparison only 22.55% of dup-loci were within this size range. The majority (62.75%) of dup-loci ranged from 50 to 500 kb. In summary, there were far more del-loci, but their sizes tended to be smaller than those of dup-loci. A list of the 623 loci is shown in Supplementary Table 2.

Of the 623 CNV loci, 268 loci were detected in ≥ 2 individuals (Table 1). The remaining loci were detected in only one individual; these loci were not necessarily 'singleton loci' as we only studied

Table 1 Summary statistics of CNV loci constructed from PennCNV output

Summary statistics of CNV loci (PennCNV output)	Total	Del	Dup
Number of CNV loci	623	476 (76.40%) ^a	102 (16.37%) ^a
Number of CNV loci detected in ≥ 2 individuals	268 (43.02%) ^b	194 (40.76%) ^b	29 (28.43%) ^b
Sum of the length of loci (Mb)	61.52	19.83	25.80
Average length per locus (kb)	98.75	41.66	252.93
Average number of markers per locus	58	34	141
<i>Size distribution</i>			
< 10 kb	141 (22.63%)	132 (27.73%)	6 (5.88%)
≥ 10 – < 50 kb	265 (42.54%)	236 (49.58%)	17 (16.67%)
≥ 50 – < 100 kb	79 (12.68%)	54 (11.34%)	21 (20.59%)
≥ 100 – < 500 kb	110 (17.66%)	52 (10.92%)	43 (42.16%)
≥ 500 kb	28 (4.49%)	2 (0.42%)	15 (14.71%)
<i>Overlapping with DGV</i>			
CNV loci that overlap	388 (62.28%)	298 (62.61%)	54 (52.94%)
CNV loci that did not overlap	235 (37.72%)	178 (37.39%)	48 (47.06%)
<i>Overlapping with UCSC genes</i>			
CNV loci that overlap	202 (32.42%)	135 (28.36%)	51 (50.00%)
CNV loci that did not overlap	421 (67.58%)	341 (71.64%)	51 (50.00%)
<i>Overlapping with CNV loci from Birdseye data and consistent in CNV status that is, del/dup/del+dup</i>			
CNV loci that overlap	196 (31.46%)	160 (33.61%)	30 (29.41%)
CNV loci that did not overlap	427 (68.54%)	316 (66.39%)	72 (70.59%)

Abbreviations: CNV, copy number variant; DGV, database of genomic variants; UCSC, University of California Santa Cruz genes.

^aThe percentage was calculated by dividing 623 loci.

^bThe percentage was calculated by dividing 623, 476 and 102 loci, respectively.

Note: As there are only 45 CNV loci (7.22%) with status del+dup, the summary statistics of these loci are not shown in the table. A full colour version of this Table is available at the Journal of Human Genetics Journal online.

87 individuals. The proportion of del-loci detected in ≥ 2 individuals (40.76%) was much higher than the proportion for dup-loci (28.43%). Among the high-frequency CNV loci (loci that were detected in multiple individuals), several overlapped with disease-related genes such as *WWOX* and *ERBB4* (gastric and pancreatic cancers and melanoma)^{25–27} and *CACNA1C* (bipolar disorder)²⁸ or drug-metabolizing genes such as *GSTT1*²⁹ (Supplementary Table 2). For example, a deletion locus overlapping with *WWOX* (a tumor suppressor gene) was detected in 24 of the 87 individuals (27.6%), and a deletion locus encompassing *GSTT1* was deleted at a population frequency of 13.8%. In addition, the proportion of del-loci encompassing the UCSC genes (28.36%) was much lower than dup-loci (50.00%) overall.

Detection of CNVs using microarrays is usually plagued with poor specificity or a high false-positive rate. In an effort to validate the 623 CNV loci constructed from the PennCNV output, we compared them with the CNV loci detected by Birdseye. We found 196 loci (31.46%) with $\geq 50\%$ reciprocal overlap with the Birdseye data and the status of ‘del’, ‘dup’ and ‘del/dup’ of the 196 loci were consistent with the Birdseye data. For the remaining 427 CNV loci that were not confirmed by Birdseye data, we found that 247 loci had been cataloged in the DGV (please see Materials and methods). Therefore, by applying two different ways of validation, 443 (71.1%) of the 623 CNV loci detected by PennCNV were considered reliable in this study (Table 1).

Characteristics of CNPs identified by Canary (Birdsuite)

Approximately 49.81% of the 1291 autosomal CNPs were non-polymorphic in the Swedish population (Supplementary Table 3). The population frequency distribution pattern of the 1291 CNPs is shown in Supplementary Figure 2. Among the polymorphic loci (648 CNPs) and non-polymorphic CNPs (643 loci) in the Swedish population, 289 loci (44.60%) and 255 loci (39.66%) overlapped with genes or entries from the UCSC annotation of the human genome, respectively. No substantial difference was observed between the polymorphic and non-polymorphic loci.

The majority of the 648 polymorphic CNPs were biallelic (545 CNPs or 84.1%), of which the integer copy numbers were either exclusively deletions, that is, copy number of 0 or 1 (387 CNPs or 59.7%), or exclusively duplications, that is, copy number of 3 or 4 (158 CNPs or 24.4%). Among the biallelic 545 CNPs, only one showed significant deviation from HWE at an FDR < 0.01 .

Numerous CNPs were found to overlap with important known disease- or pharmacogenetics-related genes (Table 2). The frequencies of these CNPs ranged from relatively uncommon (2.78% for CNP118) to completely polymorphic (100% for CNP88). For example, CNP88 overlapped with *GSTM1* and *GSTM2* was found to be completely deleted in the Swedish population, where all except one carried two-copy deletions. However, it is noteworthy that in approximately half of the sample (47 individuals), the integer copy numbers were successfully determined with high confidence scores. In addition, high deletion frequencies were also found for CNPs overlapping with other GST enzymes such as *GSTT1* (60.00%), *GSTT2*, *GSTT2B* and *GSTTP1* (98.65%). Two-copy deletion was common for these enzymes—17.6% of the individuals for *GSTT1* (CNP2560) and 43.2% for the other GST enzymes (CNP2559).

Besides these phase II metabolizing enzymes, several disease-associated genes were also found to overlap with these CNPs, such as the FCG receptor genes (autoimmune or inflammatory diseases),³⁰ *TP63*³¹ and *WWOX*²⁶ (lung adenocarcinoma, gastric, pancreatic and other cancers), *CFHR3* and *CFHR1* (age-related macular degeneration),³² *UGT2B17* (prostate cancer and graft-versus-host disease),^{33,34}

Table 2 CNPs that overlap with important and known disease- and pharmacogenetics-related genes

CNP ID	CN=0	CN=1	CN=2	CN=3	CN=4	Frequency	Chromosome	Start	End	Length	UCSC gene (disease/trait)
118	0	1	70	0	1	2.78	1	159 778 034	159 906 183	128 149	FCGR3A, FCGR2B, FCGR2C, FCGR3B (autoimmune or inflammatory diseases)
11164	0	1	83	2	0	3.49	6	162 658 558	162 660 430	1872	PARK2, parkin (Parkinson's disease)
530	1	10	71	0	0	13.41	3	190 846 372	190 847 332	960	TP63 (cancers)
147	3	31	53	0	0	39.08	1	194 997 658	195 068 695	71 037	CFHR3, CFHR1 (age-related macular degeneration)
603	8	33	46	0	0	47.13	4	69 043 083	69 168 574	125 491	UGT2B17 (prostate cancer, graft-versus-host disease)
2560	15	36	34	0	0	60.00	22	22 680 529	22 726 814	46 285	GSTT1 (phase II metabolizing enzyme)
2203	20	46	17	1	0	79.76	16	76 929 941	76 942 266	12 325	WWOX (cancers)
109	33	39	15	0	0	82.76	1	150 822 330	150 853 218	30 888	LCE3C, LCE3B (psoriasis)
2559	32	41	1	0	0	98.65	22	22 613 016	22 670 785	57 769	GSTT2, GSTT2B, GSTTP1 (phase II metabolizing enzyme)
88	46	1	0	0	0	100.00	1	110 025 907	110 044 476	18 569	GSTM2, GSTM1 (phase II metabolizing enzyme)

Abbreviations: CNPs, copy number polymorphisms; UCSC, University of California Santa Cruz genes. A full colour version of this Table is available at the Journal of Human Genetics Journal online.

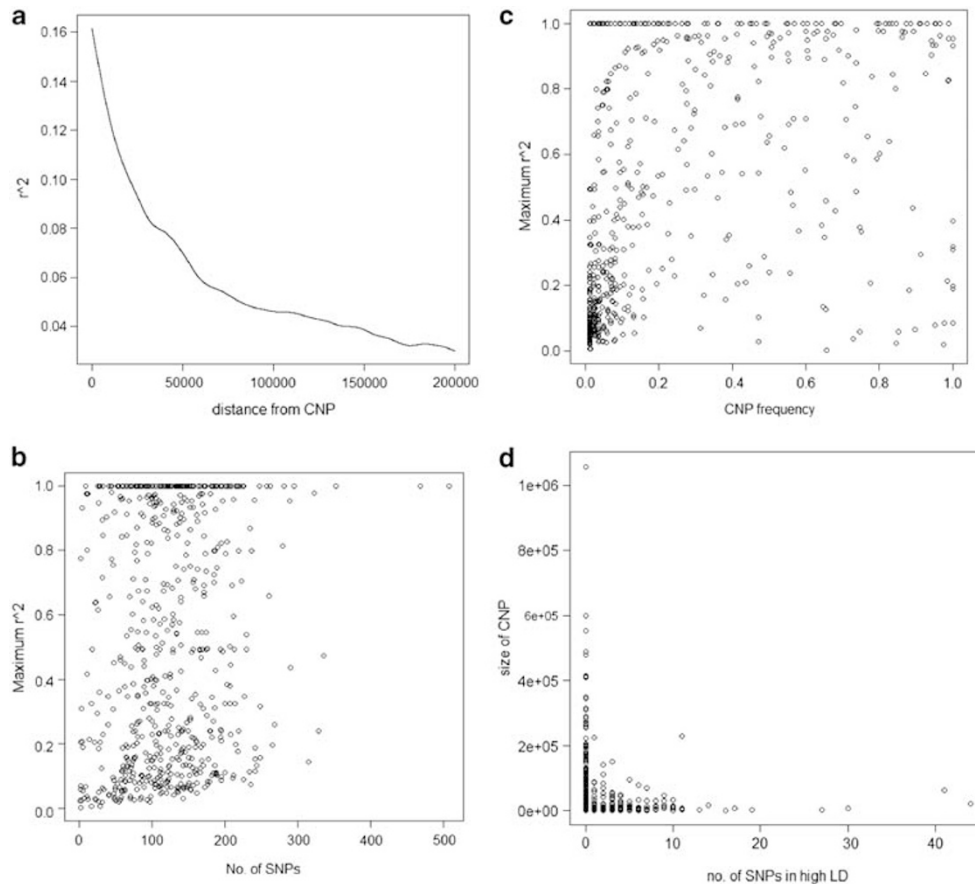


Figure 1 (a) The correlation between the r^2 and the distance between copy number polymorphism (CNP) and single-nucleotide polymorphism (SNP). (b) Maximum r^2 of CNP versus number of nearby SNPs in 200-kb windows. (c) Maximum r^2 of CNP versus CNP frequency. (d) Number of SNPs in strong correlation with the size of CNPs.

and *LCE3C* and *LCE3B* (psoriasis and rheumatoid arthritis) among others.^{35,36} The high deletion frequency of loci overlapping with *LCE3C* and *LCE3B* (82.76%), *UGT2B17* (47.13%) and *WWOX* (79.76%) requires further studies to investigate their associations with complex diseases such as psoriasis, rheumatoid arthritis and graft-versus-host disease for hematopoietic stem cell transplantation patients. For example, the mismatch of the copy numbers of *UGT2B17* was found to be associated with graft-versus-host disease in patients with hematopoietic stem cell transplantation.³⁴ Deletion of *UGT2B17* was also associated with an increased risk for prostate cancer.³³

Correlation analyses between CNPs and nearby SNPs

To study the correlation patterns with SNPs, we calculated the r^2 between the 648 polymorphic CNPs and nearby SNPs within a 200-kb window from the start and end positions of the CNP. The proportion of the CNPs with at least one SNP in strong correlation ($r^2 > 0.8$) was 31.9%, that is, 207 CNPs were found to be in strong correlation with at least one SNP. The median and maximum numbers of SNPs that were in strong correlation with the 207 CNPs were 3 and 44, respectively. This suggests that half of the 207 CNPs can be tagged by more than three SNPs and some of the CNPs were tagged by tens of SNPs. These results suggest that the majority of CNPs were not being well tagged by the nearby SNPs in the Affymetrix SNP Array 6.0. The strength of the r^2 value decreases with distance between the CNP and SNP (Figure 1a). We further investigated whether CNPs that were not well tagged tend to be located in the genomic regions where

SNP markers are sparse. The correlation patterns do not appear to be affected by the number of nearby SNPs and the frequencies of CNPs (Figures 1b and c). In other words, there was no apparent difference in the number of nearby SNPs and the frequencies of CNPs between (a) the CNPs that were in strong correlation ($r^2 > 0.8$) and (b) CNPs that were not in strong correlation with SNPs (Figures 1b and c). However, smaller-sized CNPs were generally in strong correlation with more SNPs than the larger CNPs (Figure 1d).

Correlation analyses between CNPs and GWAS-SNPs

To investigate the potential role of CNPs in the etiology of complex diseases or traits, we computed the r^2 between CNPs and the SNPs on the NHGRI GWAS Catalog (<http://www.genome.gov/gwastudies/>). Of the > 3000 GWAS-SNPs that have been found to be associated with various complex diseases and traits, only eight GWAS-SNPs were found to be in strong correlation with six CNPs (Table 3). Following the methods of Conrad *et al.*,² we define in our analysis a strong correlation as $r^2 > 0.5$. These eight SNPs were reported to be associated with five diseases or traits, namely body mass index, childhood acute lymphoblastic leukemia, early-onset myocardial infarction, Crohn's disease and multiple sclerosis. Several SNPs were in strong correlation with a single CNP, for example, three SNPs (rs13361189, rs1000113 and rs11747270) were found to be in strong correlation with CNP874.

The most notable SNP was rs2815752 near the *NEGR1* gene (associated with body mass index), which was in perfect correlation ($r^2 = 1$) with CNP60. This locus is a 42-kb deletion located in

Table 3 Correlation between CNPs and GWAS-SNPs at $r^2 > 0.5$

CNP ID	Chromosome	Start position	End position	Length	GWAS-SNP	r^2 value	Gene	Complex disease/trait
60	1	72 541 504	72 583 736	42 232	rs2815752	1	NEGR1	BMI
147	1	194 997 658	195 068 695	71 037	rs6428370	0.647399825	Intergenic	Acute lymphoblastic leukemia (childhood)
333	2	203 608 045	203 610 291	2246	rs6725887	0.84632626	WDR12	Myocardial infarction (early onset)
874	5	150 185 693	150 198 797	13 104	rs13361189	0.927251567	IRGM	Crohn's disease
874	5	150 185 693	150 198 797	13 104	rs1000113	0.927251567	IRGM	Crohn's disease
874	5	150 185 693	150 198 797	13 104	rs11747270	0.927251567	IRGM	Crohn's disease
877	5	155 409 350	155 415 307	5957	rs4704970	1	SGCD	Multiple sclerosis
933	6	32 539 530	326 81 749	142 219	rs3129934	0.664781909	HLA-DRB1	Multiple sclerosis

Abbreviations: BMI, body mass index; CNPs, copy number polymorphisms; GWAS, genome-wide association studies; SNP, single-nucleotide polymorphism. A full colour version of this Table is available at the *Journal of Human Genetics* Journal online.

chromosome 1 that did not overlap with any of the UCSC genes and is located only 1.3 kb away from the SNP. The total deletion frequency in the Swedish population was high (Table 3 and Supplementary Table 4), of which 51.72% were one-copy deletions and 29.89% were two-copy deletions. CNP874 was found to be in nearly perfect correlation ($r^2=0.93$) with three GWAS-SNPs located near the *IRGM* gene, which is associated with Crohn's disease. However, in comparison with CNP60, the total deletion frequency for CNP874 was much lower, with only 11.90% one-copy deletions and 1.19% two-copy deletions. This locus spans 13 kb in chromosome 5 and does not overlap with any of the UCSC genes. The three GWAS-SNPs were located 4.8 kb (rs13361189), 21.4 kb (rs1000113) and 40.2 kb (rs11747270) away from the deletion. The CNP877 locus is implicated in multiple sclerosis, where it is in perfect correlation with the GWAS-SNP (rs4704970). None of the individuals were deleted in both copies, and 32.56% were one-copy deletions. The other CNPs were implicated in childhood acute lymphoblastic leukemia (CNP147) and early-onset myocardial infarction (CNP333). Interestingly, all the CNPs found to be in strong correlation with GWAS-SNPs had only deletions in the loci.

Characteristics of CNV loci identified by Birdseye (Birdsuite)

Similar to the PennCNV output analysis, we also merged overlapping CNVs to construct CNV loci for the Birdseye data and identified 641 loci, of which 451 were del-loci, 102 were dup-loci and the remaining 31 were del/dup-loci (Table 4). The proportion of del-loci (76.40%) identified by PennCNV data was higher than that for the Birdseye data (70.36%). In comparison, the Birdseye data identified a higher proportion of dup-loci (24.80%) than the PennCNV data (16.37%). However, these differences are not substantial.

The 641 loci identified by the Birdseye data cover approximately 35.23 Mb of the nucleotide sequence, and the sum of the length for del-loci (13.10 Mb) is smaller than that for dup-loci (15.06 Mb). Similar to PennCNV data, the average size of del-loci (29.04 kb) is much smaller than that of the dup-loci (94.70 kb). However, substantial differences were observed for these parameters between the PennCNV and Birdseye data (Tables 1 and 4). For example, the sum of lengths covering CNV loci detected by the PennCNV data (61.52 Mb) was approximately twice that for the Birdseye data (35.23 Mb), while they have an almost similar number of CNV loci.

More than 60% of del-loci were <10 kb, and in comparison, only 18.24% of dup-loci fall within this size range. The majority (52.20%) of dup-loci ranged from 10 to 100 kb. In summary, there were more del-loci, but their sizes tended to be smaller than those of the dup-loci. This is in agreement with the PennCNV data. However, the size distribution pattern of the CNV loci for the Birdseye data is skewed towards the 'smaller' end compared with the PennCNV data. This is apparent when comparing the proportions in the first two strata:

(a) <10 kb and (b) ≥ 10 –<50 kb between the two sets of data (Tables 1 and 4). The list of the 641 loci is shown in Supplementary Table 2.

Of the 641 CNV loci, 280 loci were detected in ≥ 2 individuals (Table 4), and the remaining loci in only one individual. The proportion of del-loci detected in ≥ 2 individuals (43.90%) was much higher than the proportion for dup-loci (32.08%). Among the high-frequency CNV loci (loci detected in multiple individuals), several overlapped with disease-associated or pharmacogenetics-related genes such as *WVOX* and *GSTT1*, which have also been observed in the PennCNV data (Supplementary Table 2). Furthermore, the deletion frequencies were comparable between the Birdseye and PennCNV data. For example, a deletion locus overlapped with *WVOX* was also found in the Birdseye data. It was detected in 29 of the 87 individuals (33.33%), and a deletion locus encompassing *GSTT1* was deleted at a population frequency of 11.49%. Among the 196 CNV loci (160 del-loci, 30 dup-loci and 6 del/dup-loci) that were detected by both the Birdseye and PennCNV data and consistent in their CNV status, only 21 loci differed significantly (FDR <0.01) in their frequencies estimated by both sets of data. In addition, the proportion of del-loci encompassing UCSC genes (24.83%) was much lower than dup-loci (45.28%); this finding is again consistent with the PennCNV data.

For the CNV loci detected with the Birdseye data, we also performed the 'validation' steps for overlap with the PennCNV data and the DGV. As mentioned earlier, we found 196 loci with $\geq 50\%$ reciprocal overlap between the Birdseye and PennCNV data. For the remaining 445 CNV loci that were not confirmed by PennCNV data, we found that 322 loci have been cataloged in the DGV (please see Materials and methods). Therefore, by applying two different ways of validation, 518 (80.81%) of the 641 CNV loci detected by Birdseye were considered reliable in this study (Table 4).

Comparison with HapMap phase III populations

The principal component analysis showed distinct clusters for populations with different ancestries. The first two principal components (PC1 and PC2) separated the African (ASW, MKK and LWK) and non-African (CHD, GIH, MEX, SWED and TSI) populations (Figure 2a). This suggests that the CNP profiles of the African populations were substantially different from those of the non-African populations. From the second and fourth principal components (PC2 and PC4), three distinct clusters were observed (Figure 2b). The three African populations remained as a distinct cluster; however, CHD was separated from the European populations (MEX, SWED and TSI) and the Gujarati Indians (GIH). This indicates that the CNP profile of Gujarati Indians in Houston (Texas, USA) resembles that of the European populations. Principal component analysis was also performed by restricting only the 'European cluster' populations

Table 4 Summary statistics of CNV loci constructed from Birdseye (Birdsuite) output

Summary statistics of CNV loci (Birdseye output)	Total	Del	Dup
Number of CNV loci	641	451 (70.36%) ^a	159 (24.80%) ^a
Number of CNV loci detected in ≥ 2 individuals	280 (43.68%) ^b	198 (43.90%) ^b	51 (32.08%) ^b
Sum of the length of loci	35.23 Mb	13.10 Mb	15.06 Mb
Average length per locus	54.96 kb	29.04 kb	94.70 kb
Average number of markers per locus	30	22	42
<i>Size distribution</i>			
< 10 kb	303 (47.27%)	272 (60.31%)	29 (18.24%)
≥ 10 –< 50 kb	193 (30.11%)	119 (26.39%)	63 (39.62%)
≥ 50 –< 100 kb	52 (8.11%)	27 (5.99%)	20 (12.58%)
≥ 100 –< 500 kb	79 (12.32%)	31 (6.87%)	40 (25.16%)
≥ 500 kb	14 (2.18%)	2 (0.44%)	7 (4.40%)
<i>Overlapping with DGV</i>			
CNV loci that overlap	465 (72.54%)	335 (74.28%)	106 (66.67%)
CNV loci that did not overlap	176 (27.46%)	116 (25.72%)	53 (33.33%)
<i>Overlapping with UCSC genes</i>			
CNV loci that overlap	202 (31.51%)	112 (24.83%)	72 (45.28%)
CNV loci that did not overlap	439 (68.49%)	339 (75.17%)	87 (54.72%)
<i>Overlapping with CNV loci constructed from Birdseye and consistent in CNV status, that is, del/dup/del+dup</i>			
CNV loci that overlap	196 (30.58%)	160 (35.48%)	30 (18.87%)
CNV loci that did not overlap	445 (69.42%)	291 (64.52%)	129 (81.13%)

Abbreviations: CNV, copy number variant; DGV, database of genomic variants; UCSC, University of California Santa Cruz genes.

^aThe percentage was calculated by dividing 641 loci.

^bThe percentage was calculated by dividing 641, 451 and 159 loci, respectively.

Note: as there are only 31 CNV loci (4.84%) with status del+dup, the summary statistics of these loci were not shown in the table.

A full colour version of this Table is available at the Journal of Human Genetics Journal online.

(GIH, MEX, SWED and TSI) in PC2 versus PC4 (Figure 2b). More interestingly, we also found that the CNP profile of the Swedish population was substantially different from that of the other populations such as GIH and MEX, but it was also appreciably different from that of TSI (Figure 2c). These differences further justify the need to detect and characterize the CNV/CNP profile of the Swedish population.

Characteristics of ROHs

By restricting ROHs to ≥ 500 kb, a total of 14 815 regions were found in the 87 Swedish individuals with an average of 170 ROHs (Supplementary Table 5). The number of ROHs ranged from 105 to 220. The majority of these ROHs were < 1 Mb in length (Supplementary Figure 3). However, by restricting ROHs to ≥ 1 Mb, 2814 ROHs with an average of 32 ROHs per individual were found. The median size of the ROHs was approximately 686 kb, with the largest ROH spanning a length of approximately 25 Mb in chromosome 11. This ROH contained 9034 homozygotes, 29 heterozygotes and 2 missing genotypes, and had a density of 3.6 SNPs per 10 kb. The second largest ROH was 12 Mb in length and was detected in a different individual. This ROH contained 1571 homozygotes and 19 heterozygotes and had a density of 1.3 SNPs per 10 kb. The sum of the length of ROHs in each individual (that is, the total length of all the ROHs in one individual) was then computed. It ranged from approximately 87 to 179 Mb with a median and mean of approximately 141 Mb, respectively. This finding suggests that, on average, 141 Mb or 4.92% of the human genome (2867 Mb) was homozygous in these Swedish individuals (Table 5).

The distribution pattern of these ROHs in the 22 autosomes was also studied. The larger chromosomes (chromosomes 1–8) tended to

have a higher average number of ROHs per individual (Table 5). For example, these chromosomes had an average number of > 9 ROHs per individual, and in contrast, an average number of < 5 ROHs per individual was detected in chromosomes 16–22. As a result, chromosomes 1–8 also had a higher average sum of length of ROHs per individual (> 7 Mb) than the smaller chromosomes, that is, < 4 Mb for chromosomes 16–22. However, this pattern was less obvious when the parameters were adjusted for the sizes of the chromosomes. For example, the proportion of the chromosome encompassed by ROHs for the largest chromosome 1 (4.78%) was smaller than that for the other chromosomes such as chromosome 17 (5.14%). An apparent trend is not observed for the proportion of the chromosome encompassed by ROHs across the 22 autosomes. However, chromosomes 3, 4, 8 and 12 tended to have the highest proportions (5.90–6.16%), and, in contrast, chromosomes 16, 19, 21 and 22 had the lowest proportions (1.76–2.59%). These results were not due to differences in the density of SNPs across the 22 autosomes, as we found no substantial differences in the density of SNPs across the chromosomes (except for chromosome 19, which had a density of < 2 SNPs per 10 kb when compared with the other chromosomes). Although chromosomes 3 and 4 had $> 6\%$ of the proportion of the chromosome encompassed by ROHs, the density of SNPs of these chromosomes was similar to that of chromosome 16, where only approximately 2% of this chromosome was covered by ROHs (Table 5).

DISCUSSION

In this study, > 600 CNV loci were detected in the Swedish population using two different CNV-detection algorithms, that is, PennCNV (623 loci) and Birdsuite (641 loci). From these, 196 loci were consistently identified by both algorithms, suggesting their reliability. In addition,

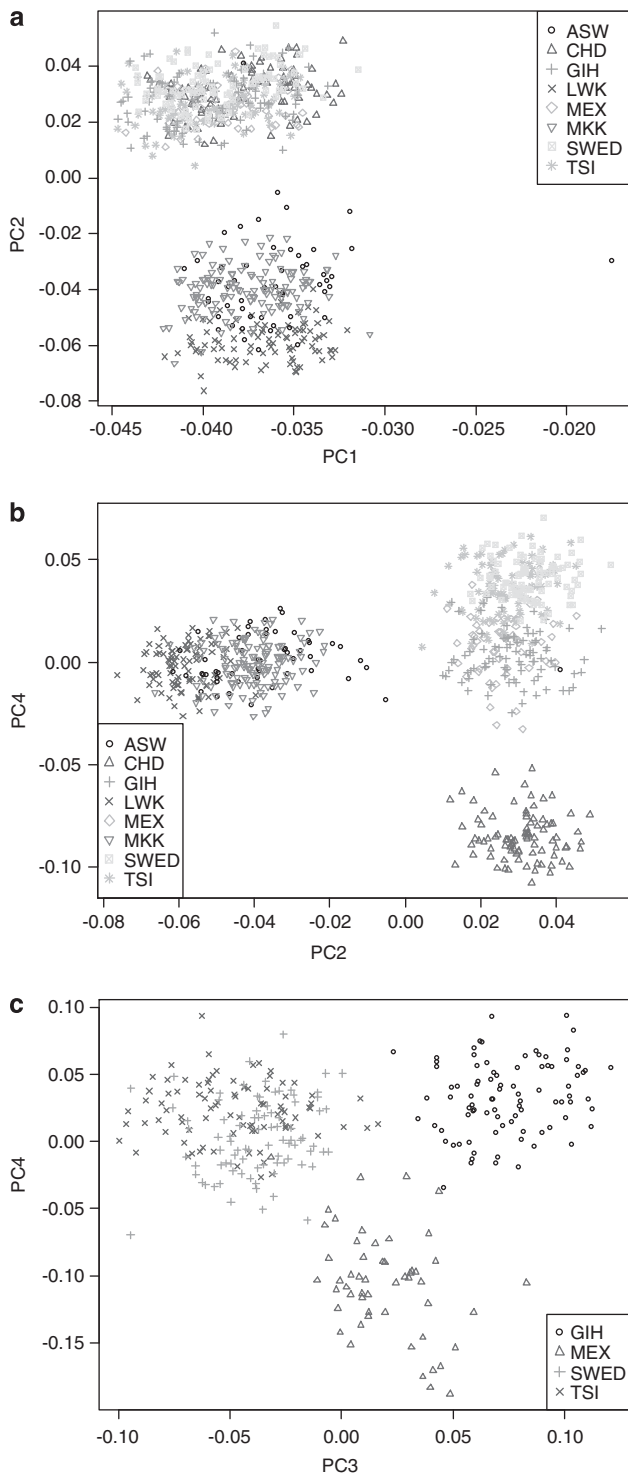


Figure 2 Principal component analysis comparing the populations. (a) Swedish and HapMap phase III populations—PC 1 versus PC 2. (b) Swedish and HapMap phase III populations—PC 2 versus PC 4. (c) Swedish and three HapMap phase III populations (GIH, MEX and TSI)—PC 3 versus PC 4.

we also identified a total of 14 815 ROHs ≥ 500 kb or 2814 ROHs ≥ 1 Mb in the Swedish individuals with an average of 170 and 32 regions detected per individual, respectively.

CNVs have been increasingly recognized as a significant source of genetic variation or diversity in human populations. Detection of

CNVs using SNP genotyping arrays is more cost-effective and affordable for population-based studies as compared with sequencing-based methods, which are limited to only a few individuals.^{37–39} This has enabled our study to investigate the population characteristics of CNVs. Although > 600 CNV loci were identified, only 268 were detected in at least two individuals by PennCNV. Similarly, Birdseye also found 280 common CNV loci in the 87 Swedish individuals. More importantly, these common CNV loci were found to encompass several disease-related and important drug-metabolizing genes, suggesting that these loci warrant further characterization and study for their associations with the relevant diseases or traits.

We applied two different algorithms to detect CNV loci as a validation step; 196 loci were found by both the algorithms and these loci were also consistent in their CNV status ('del', 'dup' or 'del+dup'). In the majority of the 196 loci, the population frequencies were also in good agreement between PennCNV and Birdseye data, indicating that these CNV loci are highly reliable. In addition, most of the CNV loci detected by PennCNV ($> 70\%$) and Birdseye ($> 80\%$) can be 'validated' by comparing them with each other and with the DGV. The proportion of CNV loci overlapping with the DGV was approximately 62% and 72% for PennCNV and Birdseye, respectively. These percentages could be overestimated because of the false-positive entries in the DGV. Of the 196 CNV loci that were identified by both algorithms, 53 loci had not been previously cataloged in the DGV, which represents a subset of reliable novel CNV loci identified in our study. The list of CNV loci in the DGV is not as yet complete as results from only 42 published studies were documented as of November 2010 (<http://projects.tcag.ca/variation/>).

On performing the correlation analysis between CNPs and GWAS-SNPs, our results also indicated that several CNPs could be potential causal variants because of their strong correlation with the GWAS-SNPs. Notably, the strong correlation between the CNPs and the GWAS-SNPs near NEGR1 and IRGM for body mass index and Crohn's disease, respectively, are consistent with previous studies.^{40,41}

Our study has a higher sensitivity than the study by Díaz de Ståhl *et al.*,¹⁰ which only detected an average of 15 CNVs per individual compared with our study, which detected an average of 36 CNVs per individual. An average of 4 clones per CNV was detected in the Díaz de Ståhl *et al.* study, whereas in our study, each CNV was detected by an average of 51 markers (Supplementary Table 1). The ability to detect smaller CNVs was also demonstrated in our study, because the average size of CNVs detected by Díaz de Ståhl *et al.* was approximately 3.5-fold (358 kb) larger than that in our study. Although Díaz de Ståhl *et al.* also clustered individual overlapping CNVs into loci, their analysis was performed using data from different ancestries (33 Europeans, 24 Africans and 14 Asians), whereas the CNV loci constructed in our study were based entirely on the data from 87 Swedish individuals. Therefore, our list of CNV loci and their frequencies was more representative of the Swedish population.

We did not compare our results with existing data from published studies because of the methodological issues in CNV and ROH detection in the different studies. As different studies have used different platforms, quality control criteria and methods to construct CNV loci and detect ROHs, comparisons with published studies would not be valid. Therefore, we would need to analyze the data from different populations with same analytical procedure. Furthermore, such a comparison is beyond the scope of the current paper and will be addressed in a future publication. However, to provide some preliminary insight into the population differences, we compared the CNP profiles of the Swedish population with the HapMap phase III populations. This comparison was appropriate as

Table 5 Distribution pattern of ROHs across the 22 autosomes

Chromosome	Total number of ROHs	Average number of ROHs per individual	Sum of length of ROHs (bp)	Average sum of length of ROHs per individual (bp)	Chromosome size (bp) ^a	Proportion (%) of chromosome encompassed by ROHs	Number of SNPs in Affymetrix 6.0	Density of SNPs per 10 kb
1	1243	14.3	1 029 256 231	11 830 531	247 249 719	4.78	73469	3.0
2	1491	17.1	1 223 537 523	14 063 650	242 951 149	5.79	75933	3.1
3	1256	14.4	1 069 972 110	12 298 530	199 501 827	6.16	62316	3.1
4	1246	14.3	1 015 875 656	11 676 732	191 273 063	6.10	57561	3.0
5	1021	11.7	859 950 902	9 884 493	180 857 866	5.47	57967	3.2
6	1008	11.6	834 180 388	9 588 280	170 899 992	5.61	57855	3.4
7	811	9.3	632 768 685	7 273 203	158 821 424	4.58	48419	3.0
8	896	10.3	762 529 281	8 764 704	146 274 826	5.99	50019	3.4
9	566	6.5	439 197 494	5 048 247	140 273 252	3.60	42710	3.0
10	722	8.3	612 229 774	7 037 124	135 374 737	5.20	49608	3.7
11	722	8.3	650 352 277	7 475 314	134 452 384	5.56	45944	3.4
12	725	8.3	679 233 723	7 807 284	132 349 534	5.90	43833	3.3
13	482	5.5	360 268 323	4 141 015	114 142 980	3.63	35158	3.1
14	571	6.6	448 210 796	5 151 848	106 368 585	4.84	28942	2.7
15	438	5.0	371 570 656	4 270 927	100 338 915	4.26	26905	2.7
16	192	2.2	159 973 057	1 838 771	88 827 254	2.07	28658	3.2
17	428	4.9	352 288 646	4 049 295	78 774 742	5.14	21347	2.7
18	330	3.8	234 464 335	2 694 992	76 117 153	3.54	27219	3.6
19	184	2.1	143 788 195	1 652 738	63 811 651	2.59	12419	1.9
20	271	3.1	220 116 198	2 530 071	62 435 964	4.05	23487	3.8
21	100	1.1	71 684 424	823 959	46 944 323	1.76	12948	2.8
22	112	1.3	100 622 242	1 156 577	49 691 432	2.33	12059	2.4

Abbreviations: ROHs, regions of homozygosity; SNPs, single-nucleotide polymorphisms; UCSC, University of California Santa Cruz genes.

^aThe size of chromosome was obtained from UCSC Genome Browser.

A full colour version of this Table is available at the Journal of Human Genetics Journal online.

we analyzed the CNP output for the HapMap III populations generated by Canary similar to the Swedish population output. As expected, the results of our analysis showed that the CNP profile of the Swedish population was substantially different from that of the African populations (ASW, MKK and LWK) and CHD. More interestingly, the CNP profile of the Swedish population was also considerably different from that of other European populations (MEX and TSI) and GIH. This further supports the importance of delineating the population characteristics of CNVs/CNPs in the Swedish population.

There are a number of limitations when using SNP genotyping arrays to detect CNVs and ROHs, and the CNV and ROH list reported in our study is not complete. Future studies will require higher sensitivity methods and larger sample sizes for a more thorough detection of CNVs and ROHs. Nevertheless, this is the first population-based study to investigate the population characteristics of CNVs and ROHs in the Swedish population. This study found many reliable CNV loci and also highlighted numerous loci that warrant further investigation for their medical or pharmacogenetic importance. The abundance of ROHs detected in the human genome also suggests the importance of studying their associations with complex phenotypes.

ACKNOWLEDGEMENTS

The Yong Loo Lin School of Medicine, the Life Science Institute and the Office of Deputy President (Research and Technology), National University of Singapore. We also acknowledge the support of the Genome Institute of Singapore, and Agency for Science, Technology and Research, Singapore.

- McCarroll, S. A., Kuruwilla, F. G., Korn, J. M., Cawley, S., Nemes, J., Wysoker, A. *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat. Genet.* **40**, 1166–1174 (2008).
- Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704–712 (2010).

- Park, H., Kim, J. I., Ju, Y. S., Gokcumen, O., Mills, R. E., Kim, S. *et al.* Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. *Nat. Genet.* **42**, 400–405 (2010).
- Yim, S. H., Kim, T. M., Hu, H. J., Kim, J. H., Kim, B. J., Lee, J. Y. *et al.* Copy number variations in East-Asian population and their evolutionary and functional implications. *Hum. Mol. Genet.* **19**, 1001–1008 (2010).
- Ku, C. S., Pawitan, Y., Sim, X., Ong, R. T., Seielstad, M., Lee, E. J. *et al.* Genomic copy number variations in three Southeast Asian populations. *Hum. Mutat.* **31**, 851–857 (2010).
- Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444–454 (2006).
- Pinto, D., Marshall, C., Feuk, L. & Scherer, S. W. Copy-number variation in control population cohorts. *Hum. Mol. Genet.* **16**, R168–R173 (2007).
- Zogopoulos, G., Ha, K. C., Naqib, F., Moore, S., Kim, H., Montpetit, A. *et al.* Germ-line DNA copy number variation frequencies in a large North American population. *Hum. Genet.* **122**, 345–353 (2007).
- de Smith, A. J., Tsalenko, A., Sampas, N., Scheffer, A., Yamada, N. A., Tsang, P. *et al.* Array CGH analysis of copy number variation identifies 1284 new genes variant in healthy white males: implications for association studies of complex diseases. *Hum. Mol. Genet.* **16**, 2783–2794 (2007).
- Diaz de Ståhl, T., Sandgren, J., Piotrowski, A., Nord, H., Andersson, R., Menzel, U. *et al.* Profiling of copy number variations (CNVs) in healthy individuals from three ethnic groups using a human genome 32K BAC-clone-based array. *Hum. Mutat.* **29**, 398–408 (2008).
- Estivill, X. & Armengol, L. Copy number variants and common disorders: filling the gaps and exploring complexity in genome-wide association studies. *PLoS Genet.* **3**, 1787–1799 (2007).
- Gibson, J., Morton, N. E. & Collins, A. Extended tracts of homozygosity in outbred human populations. *Hum. Mol. Genet.* **15**, 789–795 (2006).
- Li, L. H., Ho, S. F., Chen, C. H., Wei, C. Y., Wong, W. C., Li, L. Y. *et al.* Long contiguous stretches of homozygosity in the human genome. *Hum. Mutat.* **27**, 1115–1121 (2006).
- McQuillan, R., Leutenegger, A. L., Abdel-Rahman, R., Abdel-Rahman, R., Franklin, C. S., Pericic, M. *et al.* Runs of homozygosity in European populations. *Am. J. Hum. Genet.* **83**, 359–372 (2008).
- Nothnagel, M., Lu, T. T., Kayser, M. & Krawczak, M. Genomic and geographic distribution of SNP-defined runs of homozygosity in Europeans. *Hum. Mol. Genet.* **19**, 2927–2935 (2010).
- Lencz, T., Lambert, C., DeRosse, P., Burdick, K. E., Morgan, T. V., Kane, J. M. *et al.* Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. *Proc. Natl Acad. Sci. USA* **104**, 19942–19947 (2007).
- Nalls, M. A., Guerreiro, R. J., Simon-Sanchez, J., Bras, J. T., Traynor, B. J., Gibbs, J. R. *et al.* Extended tracts of homozygosity identify novel candidate genes associated with late-onset Alzheimer's disease. *Neurogenetics* **10**, 183–190 (2009).
- Yang, T. L., Guo, Y., Zhang, L. S., Tian, Q., Yan, H., Papasian, C. J. *et al.* Runs of homozygosity identify a recessive locus 12q21.31 for human adult height. *J. Clin. Endocrinol. Metab.* **95**, 3777–3782 (2010).

- 19 O'Dushlaine, C. T., Morris, D., Moskvina, V., Kirov, G., Consortium, I. S., Gill, M. *et al*. Population structure and genome-wide patterns of variation in Ireland and Britain. *Eur. J. Hum. Genet.* **18**, 1248–1254 (2010).
- 20 International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).
- 21 Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S. F. *et al*. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **17**, 1665–1674 (2007).
- 22 Korn, J. M., Kuruvilla, F. G., McCarroll, S. A., Wysoker, A., Nemesh, J., Cawley, S. *et al*. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat. Genet.* **40**, 1253–1260 (2008).
- 23 Mei, T. S., Salim, A., Calza, S., Seng, K. C., Seng, C. K. & Pawitan, Y. Identification of recurrent regions of copy-number variants across multiple individuals. *BMC Bioinformatics* **11**, 147 (2010).
- 24 Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y. *et al*. Detection of large-scale variation in the human genome. *Nat. Genet.* **36**, 949–951 (2004).
- 25 Aqeilan, R. I., Kuroki, T., Pekarsky, Y., Albagha, O., Trapasso, F., Baffa, R. *et al*. Loss of WWOX expression in gastric carcinoma. *Clin. Cancer Res.* **10**, 3053–3058 (2004).
- 26 Kuroki, T., Yendamuri, S., Trapasso, F., Matsuyama, A., Aqeilan, R. I., Alder, H. *et al*. The tumor suppressor gene WWOX at FRA16D is involved in pancreatic carcinogenesis. *Clin. Cancer Res.* **10**, 2459–2465 (2004).
- 27 Prickett, T. D., Agrawal, N. S., Wei, X., Yates, K. E., Lin, J. C., Wunderlich, J. R. *et al*. Analysis of the tyrosine kinome in melanoma reveals recurrent mutations in ERBB4. *Nat. Genet.* **41**, 1127–1132 (2009).
- 28 Ferreira, M. A., O'Donovan, M. C., Meng, Y. A., Jones, I. R., Ruderfer, D. M., Jones, L. *et al*. Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. *Nat. Genet.* **40**, 1056–1058 (2008).
- 29 Ouahchi, K., Lindeman, N. & Lee, C. Copy number variants and pharmacogenomics. *Pharmacogenomics* **7**, 25–29 (2006).
- 30 Fanciulli, M., Norsworthy, P. J., Petretto, E., Dong, R., Harper, L., Kamesh, L. *et al*. FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *Nat. Genet.* **39**, 721–723 (2007).
- 31 Miki, D., Kubo, M., Takahashi, A., Yoon, K. A., Kim, J., Lee, G. K. *et al*. Variation in TP63 is associated with lung adenocarcinoma susceptibility in Japanese and Korean populations. *Nat. Genet.* **42**, 893–896 (2010).
- 32 Spencer, K. L., Hauser, M. A., Olson, L. M., Schmidt, S., Scott, W. K., Gallins, P. *et al*. Deletion of CFHR3 and CFHR1 genes in age-related macular degeneration. *Hum. Mol. Genet.* **17**, 971–977 (2008).
- 33 Karypidis, A. H., Olsson, M., Andersson, S. O., Rane, A. & Ekström, L. Deletion polymorphism of the UGT2B17 gene is associated with increased risk for prostate cancer and correlated to gene expression in the prostate. *Pharmacogenomics J.* **8**, 147–151 (2008).
- 34 McCarroll, S. A., Bradner, J. E., Turpeinen, H., Volin, L., Martin, P. J., Chylewski, S. D. *et al*. Donor-recipient mismatch for common gene deletion polymorphisms in graft-versus-host disease. *Nat. Genet.* **41**, 1341–1344 (2009).
- 35 Docampo, E., Rabionet, R., Riveira-Muñoz, E., Escaramis, G., Julià, A., Marsal, S. *et al*. Deletion of the late cornified envelope genes, LCE3C and LCE3B, is associated with rheumatoid arthritis. *Arthritis Rheum.* **62**, 1246–1251 (2010).
- 36 de Cid, R., Riveira-Munoz, E., Zeeuwen, P. L., Robarge, J., Liao, W., Dannhauser, E. N. *et al*. Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis. *Nat. Genet.* **41**, 211–215 (2009).
- 37 Wang, J., Wang, W., Li, R., Li, Y., Tian, G., Goodman, L. *et al*. The diploid genome sequence of an Asian individual. *Nature* **456**, 60–65 (2008).
- 38 Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A. *et al*. The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872–876 (2008).
- 39 Korbel, J. O., Urban, A. E., Affourtit, J. P., Godwin, B., Grubert, F., Simons, J. F. *et al*. Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**, 420–426 (2007).
- 40 Willer, C. J., Speliotes, E. K., Loos, R. J., Li, S., Lindgren, C. M., Heid, I. M. *et al*. Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat. Genet.* **41**, 25–34 (2009).
- 41 McCarroll, S. A., Huett, A., Kuballa, P., Chylewski, S. D., Landry, A., Goyette, P. *et al*. Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nat. Genet.* **40**, 1107–1112 (2008).

Supplementary Information accompanies the paper on Journal of Human Genetics website (<http://www.nature.com/jhg>)