ORIGINAL ARTICLE

Regions of homozygosity in three Southeast Asian populations

Shu-Mei Teo^{1,2,3,4,5}, Chee-Seng Ku^{1,2,5}, Agus Salim², Nasheen Naidoo¹, Kee-Seng Chia^{1,2,4} and Yudi Pawitan⁴

The genomes of outbred populations were first shown in 2006 to contain regions of homozygosity (ROHs) of several megabases. Further studies have also investigated the characteristics of ROHs in healthy individuals in various populations but there are no studies on Singapore populations to date. This study aims to identify and investigate the characteristics of ROHs in three Singapore populations. A total of 268 samples (96 Chinese, 89 Malays and 83 Indians) are genotyped on Illumina Human 1 M Beadchip and Affymetrix Genome-Wide Human SNP Array 6.0. We use the PennCNV algorithm to detect ROHs. We report an abundance of ROHs (≥ 500 kb), with an average of more than one hundred regions per individual. On average, the Indian population has the lowest number of ROHs and smallest total length of ROHs per individual compared with the Chinese and Malay populations. We further investigate the relationship between the occurrence of ROHs and haplotype frequency, regional linkage disequilibrium (LD) and positive selection. Based on the results of this data set, we find that the frequency of occurrence of ROHs is positively associated with haplotype frequency and regional LD. The majority of regions detected for recent positive selection and regions with differential LD between populations overlap with the ROH loci. When we consider both the location of the ROHs and the allelic form of the ROHs, we are able to separate the populations by principal component analysis, demonstrating that ROHs contain information on population structure and the demographic history of a population.

Journal of Human Genetics (2012) 57, 101–108; doi:10.1038/jhg.2011.132; published online 1 December 2011

Keywords: PennCNV; regions of homozygosity; Singapore; Southeast Asian populations

INTRODUCTION

A region of homozygosity (ROH) is defined as a continuous stretch of DNA sequence without heterozygosity in the diploid state. All genetic variations such as single-nucleotide polymorphisms (SNPs) or microsatellites within the homologous DNA segments have two identical alleles that create homozygosity.¹ Currently, there is no consensus or standardized criteria to define an ROH. Previous studies focused on ROHs larger than 1 Mb which could have led to an underestimation of the true extent of homozygosity in the human genome,^{2,3} whereas more recent studies define an ROH at a minimum length of 500 kb,⁴ with the intention of avoiding this underestimation. This is of relevance as shorter ROHs are now also thought to be associated with complex phenotypes.⁴

The genomes of outbred populations were first shown in 2006 to contain ROHs of several megabases.^{2,3,5} Their location is markedly nonrandom, where different individuals share similar region boundaries. Some loci are caused by a single common haplotype, whereas others are a consequence of several common haplotypes that could be

markedly disparate.⁶ Several mechanisms for the occurrence of ROHs have been suggested, including uniparental isodisomy (a chromosomal abnormality where a child inherits two identical copies of a chromosome from one parent and none from the other) and autozygosity (where a child inherits the same common ancestral haplotype chromosomal segment from both parents). Studies have found no significant violation of Mendelian transmission in these areas and concluded autozygosity as the most likely cause for the majority of ROHs observed.^{7,8}

Previous studies have also investigated the population characteristics of ROHs in healthy individuals^{9–11} and performed association analyses to identify ROHs that are associated with complex diseases and traits using a case-control study design.^{4,12,13} However, the majority of these studies are conducted on European populations, and only a few on Asian populations. This study aims to identify and characterize ROHs in three Singapore populations, and to investigate their relationship to linkage disequilibrium (LD), haplotype frequency and positive selection.

¹Centre for Molecular Epidemiology, National University of Singapore, Singapore; ²Department 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 Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

⁵These authors contributed equally to this work.

Correspondence: S-M Teo, NUS Graduate School for Integrative Sciences and Engineering, National University of Singapore, Singapore.

E-mail: g0801862@nus.edu.sg

or Dr Y Pawitan, Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, PO Box 281, 17177 Stockholm, Sweden.

E-mail: Yudi.Pawitan@ki.se

Received 21 April 2011; revised 30 September 2011; accepted 24 October 2011; published online 1 December 2011

MATERIALS AND METHODS

Data

We use data from the Singapore Genome Variation Project (SGVP),¹⁴ where a total of 292 DNA samples (consisting of 99 Chinese, 98 Malays and 95 Indians) are genotyped using the Illumina Human 1 M Beadchip and the Affymetrix Genome-Wide Human SNP Array 6.0. The characteristics of copy number variations of these populations have been investigated and reported.¹⁵ The Chinese, Indians and Malays in Singapore descended from immigrants from neighboring countries such as China (mainly from southern provinces such as Fujian and Guangdong), India (majority from south-eastern India), Indonesia and Malaysia. The detailed information on the sources of DNA samples, demographic data of the samples, sample selection, and the origin and migration history of the three Singapore populations have been described in previous publications.^{14,15} A total of 268 samples (consisting of 96 Chinese, 89 Malays and 83 Indians) are used in the subsequent analysis after removing samples on the basis of high rates of SNP missingness (greater than 2%), excessive heterozygosity or cryptic relatedness by excessive identity-by-states. Population membership is ascertained on the basis that all four grandparents belong to the same population, and samples that display either evidence of admixture or clear evidence of discordance between self-reported and genetically inferred population membership are excluded.

SNP genotypes are obtained from the SGVP website (<http://www.nus-cme.org.sg/sgvp/>). These SNPs have undergone a series of quality control measures,¹⁴ including removing SNPs with SNP missingness $\geq 5\%$ and P -value < 0.001 for a test of departure of Hardy–Weinberg Equilibrium (HWE), resulting in ~ 1.58 million SNPs per population remaining. Quality control measures were conducted separately for each of the populations.

Identification of individual-specific ROHs

Individual-specific ROHs are identified using the PennCNV algorithm¹⁶ for the Illumina and Affymetrix arrays based on the log R ratio and B allele frequency for each sample. The ROHs identified by PennCNV are copy neutral events, meaning that one-copy deletions are excluded. We exclude regions < 500 kb. To further filter regions that may be called erroneously by PennCNV, we check the SNPs genotypes for the number of heterozygous genotypes within the region. Ideally, we would expect no heterozygous genotypes in the region, but we allow for some heterozygosity that may be due to genotyping errors or other causes.

We investigate the effect of allowing some heterozygosity on the relationship between ROH and LD. From a simulation (see Supplementary Methods, ‘Simulation’ section), we observe that ROH detection is very sensitive to heterozygosity present either due to mutation or genotyping errors, whereas the LD in the region is largely preserved despite the mutations introduced. By not allowing any heterozygosity, we miss detecting older ROHs in many individuals and this affects the formation of the common regions. So, to capture the LD/haplotype structure using ROHs, it is important to allow a small percentage of heterozygosity.

We use a binomial probability upper bound to calculate a confidence score for each region (see Supplementary Methods, ‘Confidence scores calculation’ section). The confidence score takes into account the amount of heterozygosity, as well as the SNP density, and is an indication of how confident we are that the ROH is true. In general, the confidence scores for regions detected by the Affymetrix platform are lower than that detected by the Illumina platform (see Supplementary Methods Figure S1). We decide to use the Illumina platform with more than 1 million SNPs for ROHs detection but still use the combined genotypes from 1.58 million SNPs from both platforms in the calculation of confidence scores. Several summary statistics are computed to describe and compare the characteristics of ROHs in the three Singapore populations.

Identification of common ROHs

We identify common ROH loci using a previously published method.¹⁷ We define common loci as regions with consecutive probes where at least 5% of the subjects (that is, 13 subjects) have individual regions that overlap with the probes. Occasionally, individual regions within a common locus can show considerable variations in their boundaries, resulting in a heterogeneous region. To refine the identified common loci, we form clusters of regions by

requiring all individual regions within a cluster to overlap by at least 80%. For each common locus, individual regions are said to be concordant if it overlaps with at least 80% of the length of the locus. Common loci with < 2 concordant individuals or < 500 kb or having a SNP density < 0.2 (SNP per kb) are discarded. The common loci are further refined as the intersection of the concordant regions. We perform population comparisons and test of departure of HWE for each locus. For each set of tests, we account for multiple comparisons using the false discovery rate,¹⁸ with results or discoveries considered interesting at false discovery rate of < 0.01 .

Quantification of regional LD

The two most widely used measures to quantify the amount of LD between two markers are the D' and r^2 statistics.¹⁹ Here, instead of LD between two markers, we are interested in the amount of LD in a region. For all SNPs in a region, we calculate the pairwise D' (and r^2). We perform eigenvalue decomposition on the D' (r^2) matrix and calculate the percentage explained by the first eigenvalue (y). This percentage will take values between $100/n$ and 100, where n is the number of (polymorphic) SNPs in the region. To make the percentages comparable across regions with different number of SNPs, we scale it such that the value varies between 0 and 1. So, $y^* = (y - 100/n) / (100 - 100/n)$. The higher the value of y^* , the stronger the LD in the region.

Haplotypes in ROH loci

For each common locus, we use phased genotypes (using the program fastPHASE version 1.3, see Supplementary Methods in Teo *et al.*¹⁴ for details on the choice of parameters for phasing) to determine the different haplotypes present in the three populations. To reduce the dimension of the data, we consider only the top three most frequent haplotypes and combine the others as ‘other haplotypes’, that is, we categorize each region into four alleles (top three most common haplotypes and ‘other’ haplotypes). Each individual has two alleles for each region. For convenience, we will refer to the alleles as A, B, C and D.

Identification of regions with differential LD between populations

We use a previously published method, varLD,^{20,21} to identify regions with differential LD between populations. Briefly, the method tests for equality between two LD matrices for a user-defined window size, shifting each window one SNP at a time. We calculate the varLD score for a window size of 50 SNPs for the signed r^2 matrices.²¹ For each pair of populations, a region is considered to have differential LD if consecutive positions are above the 95th percentile of the genome-wide varLD score. We restrict to regions > 500 kb for comparison with ROHs. We exclude the region if it overlaps by $> 50\%$ with copy number variations previously reported for the same set of individuals,¹⁴ as LD measures for regions that encapsulate copy number variations may not be reliable.²¹

RESULTS

Summary statistics of individual ROHs

We discard regions whose confidence scores are below the 25th percentile of the confidence scores. Table 1 summarizes the characteristics of ROHs. On average, the Indian population has lower number of ROHs compared with the Chinese and Malay populations. There is wide inter-individual difference in the number of ROHs, which ranges from 98 (sample 334_01 and 461_01) to 241 (sample 81_01). More than 80% of the ROHs are < 1 Mb in length. The largest ROH spans a length of ~ 68.5 Mb, and is detected in one Indian individual (sample 408_01) in Chromosome 3. A total of 32 ROHs larger than 10 Mb are detected (Table 2). Interestingly, three Indian samples (397_01, 290_01 and 408_01) have five or more of these ‘extremely long’ ROHs. Figure 1 plots the number of ROHs versus the total length of ROHs in each individual. We see clusters of the three populations, indicating that number and length of ROHs differ among populations. This result was also observed by Kirin *et al.*²²

Table 1 Characteristics of ROHs in three Singapore populations (using 1 029 591 SNPs from the Illumina 1 M platform)

Characteristics	Chinese (n=96)	Malay (n=89)	Indian (n=83)
<i>Number of ROHs per individual</i>			
Mean	207	179	126
Median	206	178	127
Minimum	157	123	98
Maximum	241	228	173
<i>Length of ROHs (in kb)</i>			
Mean	800.9	806.1	879.6
Median	670.5	672.8	666.3
Maximum	23 230	21 850	68 500
<i>Total length of ROHs per individual (in Mb)</i>			
Mean	166.1	144.6	111.2
Median	165.7	143.4	100.5
Minimum	115.4	90.49	73.91
Maximum	195.4	191.9	315.6
<i>Size distribution of ROHs (proportion, %)</i>			
500kb–1 Mb	83.0	82.5	83.5
≥ 1 Mb	17.0	17.5	16.5

Abbreviations: ROHs, regions of homozygosity; SNPs, single-nucleotide polymorphisms.

Summary statistics of common ROHs

We identify 1256 common ROH loci in all three populations (Supplementary Table 1), where 90% of the loci overlap with UCSC genes (<http://genome.ucsc.edu/>), and 292 (23%) overlap with genes listed in the Online Mendelian Inheritance in Man Morbid Map (<ftp://ftp.ncbi.nih.gov/repository/OMIM/ARCHIVE/morbidmap>). For each locus, we test for differences among the three populations in terms of ROH frequencies and haplotype frequencies, and 47 loci (<4%) differ significantly in frequencies while 899 loci (69%) differ significantly in haplotype frequencies among the populations. Approximately 52% of the loci are detected in >5% (more common ROH loci) of individuals (Figure 2). Figure 3 shows the length distribution of the ROH loci; ~78% of the ROH loci are ≤1 Mb, and majority of the long ROH loci (>1Mb) are in the range of 1–2 Mb. The proportion of the genome that is in the different ROH length categories differs among the three populations (Figure 4). The Chinese and Malays have more ROHs of shorter lengths compared with the Indians, while the Indians have more ROHs in the longer length categories (>4 Mb).

We compare the common loci we found to that published in previous studies.^{10,23} Two regions are defined to overlap if the regions have a reciprocal overlap of at least 50%. Nothnagel *et al*'s study¹⁰ surveys ROHs in Europeans; we found that all 10 regions listed as 'ROH islands' (meaning they have a high population frequency) in their study overlap with an ROH loci found in this study, suggesting

Table 2 ROHs larger than 10 Mb

Chromosome	Start	End	Length	Sample	Ethnicity
1	120 837 663	143 420 875	22 583 213	108_01	Chinese
6	3 217 193	26 449 280	23 232 088	17_01	Chinese
16	34 034 376	45 968 704	11 934 329	131_01	Chinese
1	120 837 663	143 420 875	22 583 213	218_01	Chinese
8	41 842 707	52 102 021	1 025 9315	465_01	Malay
1	94 915 135	108 531 282	13 616 148	174_01	Malay
11	19 924 676	41 772 573	21 847 898	174_01	Malay
1	67 073 684	90 862 713	23 789 030	290_01	Indian
3	175 758 479	190 839 635	15 081 157	290_01	Indian
4	41 334 756	55 223 410	13 888 655	290_01	Indian
6	112 768 454	147 227 544	34 459 091	290_01	Indian
11	86 911 515	131 748 067	44 836 553	290_01	Indian
14	71 970 357	88 634 741	16 664 385	290_01	Indian
17	152 362	10 559 477	10 407 116	290_01	Indian
3	102 253 981	170 758 820	68 504 840	408_01	Indian
6	79 661	14 549 208	14 469 548	408_01	Indian
6	121 892 269	132 743 942	10 851 674	408_01	Indian
13	52 267 631	77 893 750	25 626 120	408_01	Indian
13	78 497 109	94 452 168	15 955 060	408_01	Indian
22	37 722 142	49 582 267	11 860 126	408_01	Indian
13	74 778 668	100 318 094	25 539 427	367_01	Indian
3	158 294 635	169 108 914	10 814 280	361_01	Indian
6	90 065 419	106 409 967	16 344 549	397_01	Indian
7	28 668 234	43 132 968	14 464 735	397_01	Indian
7	80 118 053	105 839 742	25 721 690	397_01	Indian
8	590 729	10 908 015	10 317 287	397_01	Indian
9	33 415 385	45 059 163	11 643 779	397_01	Indian
9	66 448 030	100 731 809	34 283 780	397_01	Indian
10	121 636	24 722 946	24 601 311	397_01	Indian
15	63 308 076	73 720 143	10 412 068	397_01	Indian
7	9 983 924	21 830 396	11 846 473	76_01	Indian
13	47 337 000	72 862 520	25 525 521	76_01	Indian

Abbreviations: ROHs, regions of homozygosity.

Figure 1 Number of ROH versus total length of ROHs in each individual. A full color version of this figure is available at the *Journal of Human Genetics* journal online.

Figure 2 Number of ROH loci in the respective population frequency classes.

that these regions are not specific to Europeans (see Supplementary Methods Table S1). The population frequencies of these ROHs in our populations differ from that reported in Nothnagel *et al.*'s study,¹⁰ but formal testing is inappropriate as the methods used to calculate the frequencies are different.

Auton *et al.*'s study²³ surveys ROHs in Mexicans, Europeans, East Asians and South Asians; we found that out of 34 high-frequency ROHs (defined as being found in at least 10% of individuals within a population) 11 overlap with an ROH locus found in our study (see Supplementary Methods Table S2). All the regions that overlap are found in the East Asian population, except for one region in Chromosome 4, which is present in all populations. The frequencies of these ROHs are, however, quite low in our population (1–4%).

Association with haplotype frequency and regional LD

Figure 5 shows that the frequency of an ROH is positively associated with the total frequency of the top three haplotypes (correlation of 0.69), and Figure 6 shows that as the frequency of an ROH increases,

Figure 3 Percentage of ROH loci in the respective length classes.

Figure 4 Percentage of ROH loci in the respective length classes. A full color version of this figure is available at the *Journal of Human Genetics* journal online.

so does $\gamma^*_{D'}$ and $\gamma^*_{r^2}$ (figure is shown for the Malay population, similar figures for the Chinese and Indians are shown in Supplementary Methods Figures S2 and S3). If we assume random mating, the homozygosity of any region will be high when there are few haplotypes present at high frequency, thus it reinforces autozygosity as the mechanism for the occurrence of an ROH. These empirical results suggest that there is positive correlation between the frequency of an ROH and the frequency of the common haplotypes, and also between the frequency of an ROH and LD in the region.

Frequency of ROHs and frequency of haplotypes within ROHs

To assess if there is a difference in the location and frequency of ROHs among the populations, we perform principal component analysis

(PCA) using absence/presence of the common ROH loci. For each individual, we check if that individual has an ROH that is concordant with the common ROH. We can view the matrix input for the PCA analysis as a matrix of 1's and 0's where each row corresponds to an individual and each column corresponds to a common loci, so that the (i, j) entry indicates whether individual i has a concordant ROH at locus j . From Figure 7, we see that the Indians are quite well separated from the Chinese and Malays, and that there is some separation between the Chinese and Malays. This implies that the location and frequency of occurrence of ROHs differ among populations.

However, interestingly, populations can share the same (or similar) ROH location, but the common haplotypes driving the ROH can be markedly disparate. One example is a 700-kb ROH in Chromosome 16 (location 30,438,046–31,137,964) that overlaps with the Vitamin K epoxide reductase complex subunit 1 (*VKORC1*) gene (location 31,009,956–31,013,551). Genetic polymorphisms within the gene have been found to correlate with differences in warfarin dosage and response in many studies.^{24–26} In the Singapore populations, the Indians were observed to display warfarin resistance, thus requiring a higher dose as compared with the Chinese and Malays.^{26–29} There is no significant difference in ROH frequencies among the populations (ROH frequencies of 21, 13 and 20% for the Chinese, Malays and Indians, respectively). However, if we examine the

haplotypes in this region, there is significant difference. Fisher's exact test performed on the frequencies of the top three most frequent haplotypes results in a P -value $< 10^{-6}$. In particular, the difference in haplotype frequencies of the Indians differs markedly from the Chinese and Malays. This is highlighted in Table 3, where haplotype A dominates in the Chinese and Malays but is almost absent in the Indians, while haplotype B dominates in the Indians but is almost absent in the Chinese and Malays. Haplotypes A and B differ at 104 locations out of the 158 SNPs in this region.

We also perform PCA on the allele counts of the haplotypes as described in the section 'Haplotypes in ROH loci'. The first two components separates the Indians from the Chinese and Malays while the third component further separates the Chinese from the Malays (see Figure 8). This suggests that ROH loci contain much genetic ancestral haplotype information of a population.

Testing departure from HWE

Using the estimated frequencies of the top three haplotypes, we are able to calculate the expected frequencies of the corresponding genotypes. For the observed frequencies, we use the unphased genotypes. For each individual, we can identify the haplotypes without phase information when all the SNPs in the region are homozygous (removing SNPs where we had allowed heterozygosity in the detection

Figure 5 Frequency of ROH loci versus total frequency of top three haplotypes.

Figure 7 Principal component 2 versus principal component 1 using absence/presence of 1256 common ROHs. A full color version of this figure is available at the *Journal of Human Genetics* journal online.

Figure 6 Regional LD versus frequency of ROH based on (a) D' matrix and (b) r^2 matrix. These results are based on the Malay population.

Table 3 Haplotype frequencies of three populations in a ROH in Chromosome 16 that overlaps with VKORC1

	Haplotype A	Haplotype B	Haplotype C
Chinese	0.31	0.0052	0.099
Malay	0.28	0.045	0.10
Indian	0.0060	0.34	0

Abbreviation: ROH, region of homozygosity.

Figure 8 Results of PCA on haplotype frequencies of ROH regions. A full color version of this figure is available at the *Journal of Human Genetics* journal online.

of ROH). With that, we are able to obtain observed frequencies for the (A, A), (B, B) and (C, C) genotypes. We use the χ^2 test with three degrees of freedom to test if there is departure of the observed from the expected. A large majority of ROH loci (>92%) adhere to HWE, suggesting that assumptions of autozygosity and random mating are true for most ROH loci. Of the regions that show departure from HWE (false discovery rate <0.01), majority show excess homozygosity than would be expected. The reasons for departure from HWE are not immediately clear, and could be due to various reasons such as positive selection (see section ‘Comparison with regions associated with positive selection’) or nonrandom mating.

Comparison with varLD

As described in the section ‘Identification of regions with differential LD between populations’, we identify 16, 10 and 13 regions with differential LD variation between the Chinese and Indian populations, Malay and Indian populations, and Chinese and Malay populations, respectively. Of the 16 regions, 14 overlap with a common ROH and 10 out of 14 show significant differences in haplotype frequency between the Chinese and Indian populations. Of the 10 regions, 7 overlap with a common ROH and 7 out of 7 show significant differences in haplotype frequency between the Malay and Indian populations. Of the 13 regions, 8 overlap with a common ROH and 8 out of 8 show significant differences in haplotype frequency between the Chinese and Malay populations.

We observe that the majority of regions (74%) that show LD differences between populations correspond to regions where ROHs are observed, and furthermore, the haplotype frequencies in these regions differ between the populations. These results indicate that ROH patterns explain a large proportion of LD variations.

Comparison with regions associated with positive selection

We investigate if the regions detected for recent positive natural selection overlap with ROHs. We consider the top 10 candidate regions for recent positive selection in each of the populations, as published in a previous study.¹⁴ These regions were detected based on the clustering of SNPs with high integrated haplotype score.³⁰ Out of the 30 regions considered, 28 regions overlap with a common ROH defined in this study, with 20 regions completely within an ROH and the other 8 regions with a high percentage of overlap (at least 60%). This suggests the occurrence of ROHs as a possible consequence of positive selection, where the positively selected haplotypes rise to a high frequency, resulting in a high possibility of ROHs due to autozygosity.

Out of the 28 regions, 10 of them overlap with an ROH that failed HWE. Performing Fisher’s exact test on a 2 by 2 table with indicators for departure from HWE and indicators for positive selected regions as rows and columns, we obtain an odds ratio of 1.89 (P -value=0.05). The departure from HWE may be a consequence of positive selection. An ROH that has a higher frequency than would be expected for its length may also be an evidence of positive selection (see Supplementary Methods Figure S8).

Effect of heterozygosity on the relationship between ROH and LD

When we filter the individual regions using a stricter confidence threshold of the 75th percentile (that is, allowing less heterozygosity), we identify 414 common regions, but the relationship of these regions with haplotype frequency, regional LD and positive selection is weak (see Supplementary Methods Figures S4 and S5 and section Comparison with VarLD (results based on these 414 common regions)). We also see poorer separation of the populations by PCA, but this is likely due to the fewer number of common regions identified. At the 25th percentile threshold, the percentage of heterozygosity is still kept low at <5% for a large majority of the regions (See Supplementary Methods Figure S9). With an overly strict confidence score threshold, many regions are omitted and this decreases the number of common regions formed from 1256 to 414. Allowing for some heterozygosity within the regions allows detection of older ROH loci (heterozygosity caused by recent recombination or mutation), which have a stronger relationship with LD and positive selection (see Simulation section in Supplementary Methods).

DISCUSSION

In summary, this study identifies and investigates the population characteristics of ROHs in three Singapore populations, Chinese, Malay and Indian. We report an abundance of ROHs, with an average of >100 ROHs per individual. On average, the Indians have lower numbers and total length of ROHs per individual than the Chinese and Malays, possibly indicative of a larger founder population. However, there are several Indians with multiple large ROHs, suggesting that they may be offsprings of parents who are close relatives. In India, consanguineous marriages are more prevalent in the South, especially in Tamil Nadu, from where many Singapore Indians descended. From the Consanguinity/Endogamy Resource (http://www.consang.net/index.php/Main_Page), data from a 1982 study have shown the prevalence of consanguineous marriages among Singapore Indians to be 4% compared with only 0.3% in Singapore Chinese. Published data have shown that the number of ROHs of several megabases increase markedly in the offsprings of consanguineous marriages,^{3,24} with an average of 6.25% homozygosity expected in the genome of the offsprings of first cousin marriages.⁷ Li *et al.*³ have shown that in a family with four children from first cousin

marriages, multiple ROHs ranging from 3.06 to 53.17 Mb were observed in all the children. Woods *et al.*²⁴ have also shown a marked increase in homozygosity levels in individuals with a recessive disease whose parents were first cousins, where, on average, 11% of their genomes were homozygous.

In addition, we identify 1256 common ROH loci, and investigate the occurrence of ROHs and haplotype frequency, regional LD and positive selection. Based on the results for this data set, we find that the frequency of occurrence of ROHs is positively associated with haplotype frequency and regional LD. The preferential occurrence of ROHs in regions of high LD and low recombination has also been observed in other studies.¹⁰ The majority of regions detected for recent positive selection and regions with differential LD between populations overlap with ROH loci. By considering both the location of the ROH and the allelic form of the ROH, we are able to separate the populations by PCA, demonstrating that ROHs contain information on population structure and the evolutionary and demographic history of a population.

The ability of genome-wide SNP markers for population structure analysis has been widely acknowledged. Here, we are not proposing the superiority of ROHs in population structure analysis. It is expected that using genome-wide SNP data allows very good separation of populations through PCA because of the amount of information it contains (see 14 for PCA analysis using SNPs on the same population samples). In this paper, we have shown that it is possible to distinguish populations using just ~1000 segments of the genome. Comparatively, if we were to choose 1000 random segments of the genome and perform a similar analysis, we would not obtain as good a separation as with ROHs (see Supplementary methods Figure S7). The unique characteristics of ROHs allow us to study common haplotypes conveniently; it is complementary to SNP-based analysis. In SNP-based analysis, we simply compare SNP-level frequencies between populations but in ROH-based analysis, we are able to capture differences in LD or haplotype structures.

Majority of the ROH loci overlap with known genes but their association with complex phenotypes is still rudimentary. This warrants further characterization of ROHs in different populations, investigation of their roles in the genetics of complex phenotypes and further studies of population evolutionary genetics. These future studies will be of importance given the abundance of ROHs in the human genome and the differences of ROHs between populations.

A sufficiently large number of SNPs is required to accurately detect ROHs.^{1,2} To this end, we have used two highly dense SNP arrays (Illumina 1 M and Affymetrix 6.0) with > 1.58 million unique SNPs. Using a confidence score metric that takes into account percentage of heterozygosity as well as the number of SNPs in the region, we discard individual regions whose confidence scores are below the 25th percentile of the confidence scores. We use the PennCNV algorithm that relies on signal intensity data to detect putative ROHs. We then filter out false positives by checking SNP genotypes within the ROH. To our knowledge, most studies on ROHs use only SNP genotypes, but this approach may produce false positives caused by hemizygous deletions. On the other hand, due to the noise in signal intensity data, the regions called by PennCNV could also result in false-positive regions. We feel it is important to use a combination of the methods (that is, signal intensity data and genotype data) to minimize false-positive rates.

We also use PLINK, a widely used software for ROH detection, on genotypes from both platforms using the following parameters: 500 kb window with two heterozygous SNPs allowed, minimum length of 500 kb, 50 SNPs as minimum number of SNPs and minimum density

of 1 SNP per 10 kb. We find that 75% of the regions found by PennCNV are detected by PLINK, suggesting that the results of the analysis will likely give similar conclusions using PLINK. A formal and systematic comparison of multiple algorithms for ROH detection will be interesting.

Potential biases in the detection of ROHs include false-negative regions due to ascertainment bias in SNP selection for the SNP arrays and false-positive regions due to the lack of minor allele frequency (MAF) criterion applied before the identification of ROHs. With regards to the former, SNPs from genotyping platforms are mostly tagged SNPs from the HapMap project, so populations that were not analyzed in the HapMap project will have less chance of their population-specific SNPs being included in the array. However, both the Illumina 1 M and Affymetrix 6.0 arrays have a high marker density and uniformity. With regards to the later, we do not expect our results to be affected considerably by not filtering SNPs with low MAF, for several reasons. First, we have very dense SNP genotyping data of > 1.58 million SNPs, and as an ROH is defined as a region of consecutive homozygosity of > 500 kb, it is unlikely that there exists a large number of consecutive low-MAF SNPs that cause a false-positive identification. In any case, these monomorphic/near monomorphic SNPs are uninformative and would not affect the haplotype analyses. It is of concern if the region is detected because the monomorphic/low-MAF SNPs are genotyped, whereas other SNPs present in the region are missed (due to ascertainment bias). However, as ROH detection is not reliant on a single SNP, but on many consecutive homozygous SNPs in a 500 kb region, we do not expect either issue to be of serious concern.

Some studies²³ have adopted the strategy of removing SNPs in high LD before defining an ROH (that is, thinning the data set but requiring a lower number of SNPs for the definition of ROH). However, we found poor correlation between the frequency of the ROHs we identified and the mean or median pairwise D' or r^2 statistics (for SNPs within the ROH, up to 250 kb apart, see Supplementary Methods Figure S6), meaning that a SNP being in high LD in the vicinity is not sufficient for its inclusion in an ROH, and a SNP in low LD is not sufficient for its exclusion in an ROH.

In conclusion, our study is one of the first to describe the population characteristics of ROHs in the three Singapore populations (Chinese, Malay and Indian). Our results are in support that ROHs contain population demographic and ancestral haplotype information.

ACKNOWLEDGEMENTS

We thank Dr Teo Yik Ying for helpful discussions related to this work and Rick Ong for identifying regions with differential LD between populations using the VarLD program. TSM acknowledges support from the National University of Singapore Graduate School for Integrative Sciences and Engineering (NGS) Scholarship.

- 1 Ku, C. S., Naidoo, N., Teo, S. M. & Pawitan, Y. Regions of homozygosity and their impact on complex diseases and traits. *Hum. Genet.* **129**, 1–15 (2011).
- 2 Gibson, J., Morton, N. E. & Collins, A. Extended tracts of homozygosity in outbred human populations. *Hum. Mol. Genet.* **15**, 789–795 (2006).
- 3 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).
- 4 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).
- 5 Simon-Sanchez, J., Scholz, S., Fung, H. C., Matarin, M., Hernandez, D., Gibbs, J. R. *et al.* Genome-wide SNP assay reveals structural genomic variation, extended homo-

- zygosity and cell-line induced alterations in normal individuals. *Hum. Mol. Genet.* **16**, 1–14 (2007).
- 6 Curtis, D., Vine, A. E. & Knight, J. Study of regions of extended homozygosity provides a powerful method to explore haplotype structure of human populations. *Ann. Hum. Genet.* **72**, 261–278 (2008).
 - 7 Broman, K. W. & Weber, J. L. Long homozygous chromosomal segments in reference families from the centre d'Etude du polymorphisme humain. *Am. J. Hum. Genet.* **65**, 1493–1500 (1999).
 - 8 Curtis, D. Extended homozygosity is not usually due to cytogenetic abnormality. *BMC Genet.* **8**, 67 (2007).
 - 9 McQuillan, R., Leutenegger, A. L., Abdel-Rahman, R., Franklin, C. S., Pericic, M., Barac-Lauc, L. *et al.* Runs of homozygosity in European populations. *Am. J. Hum. Genet.* **83**, 359–372 (2008).
 - 10 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).
 - 11 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).
 - 12 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).
 - 13 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).
 - 14 Teo, Y. Y., Sim, X., Ong, R. T., Tan, A. K., Chen, J., Tantoso, E. *et al.* Singapore genome variation project: a haplotype map of three Southeast Asian populations. *Genome Res.* **19**, 2154–2162 (2009).
 - 15 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).
 - 16 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).
 - 17 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).
 - 18 Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300 (1995).
 - 19 Lewontin, R. C. & Kojima, K. The evolutionary dynamics of complex polymorphisms. *Evolution* **14**, 458–472 (1960).
 - 20 Ong, R. T. H. & Teo, Y. Y. varLD: A program for quantifying variation in linkage disequilibrium patterns between populations. *Bioinformatics* **26**, 1269–1270 (2010).
 - 21 Teo, Y. Y., Fry, A. E., Bhattacharya, K., Small, K. S., Kwiatkowski, D. P. & Clark, T. G. Genome-wide comparisons of variation in linkage disequilibrium. *Genome Res.* **19**, 1849–1860 (2009).
 - 22 Kirin, M., McQuillan, R., Franklin, C. S., Campbell, H., McKeigue, P. M. & Wilson, J. F. Genomic runs of homozygosity record population history and consanguinity. *PLoS ONE* **5**, e13996 (2010).
 - 23 Auton, A., Bryc, K., Boyko, A. R., Lohmueller, K. E., Novembre, J., Reynolds, A. *et al.* Global distribution of genomic diversity underscores rich complex history of continental human populations. *Genome Res.* **19**, 795–803 (2009).
 - 24 Woods, C. G., Cox, J., Springell, K., Hampshire, D. J., Mohamed, M. D., McKibbin, M. *et al.* Quantification of homozygosity in consanguineous individuals with autosomal recessive disease. *Am. J. Hum. Genet.* **78**, 889–896 (2006).
 - 25 Aquilante, C. L., Langae, T. Y., Lopez, L. M., Yarandi, H. N., Tromberg, J. S., Mohuczy, D. *et al.* Influence of coagulation factor, vitamin K epoxide reductase complex subunit 1, and cytochrome P450 2C9 gene polymorphisms on warfarin dose requirements. *Clin. Pharmacol. Ther.* **79**, 291–302 (2006).
 - 26 Harrington, D. J., Underwood, S., Morse, C., Shearer, M. J., Tuddenham, E. G. D. & Mumford, A. D. Pharmacodynamic resistance to warfarin associated with a Val66Met substitution in vitamin K epoxide reductase complex subunit 1. *Thromb. Haemost.* **93**, 23–26 (2005).
 - 27 Zhu, Y., Shennan, M., Reynolds, K. K., Johnson, N. A., Herrnberger, M. R., Valdes, R. Jr. *et al.* Estimation of warfarin maintenance dose based on VKORC1 (–1639 G>A) and CYP2C9 genotypes. *Clin. Chem.* **53**, 1199–1205 (2007).
 - 28 Yuen, E., Gueorgieva, I., Wise, S., Soon, D. & Aarons, L. Ethnic differences in population pharmacokinetics and pharmacodynamics of warfarin. *J. Pharmacokinet. Pharmacodyn.* **37**, 3–24 (2009).
 - 29 Lee, S. C. Inter-ethnic variability in warfarin requirement is explained by VKORC1 genotype in an Asian population. *Clin. Pharmacol. Ther.* **79**, 197–205 (2006).
 - 30 Voight, B. F., Kudavalli, S., Wen, X. & Pritchard, J. K. A map of positive selection in the human genome. *PLoS Biol.* **4**, e72 (2006).

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