Hum Genet (2011) 129:1–15
DOI 10.1007/s00439-010-0920-6

REVIEW PAPER

Regions of homozygosity and their impact on complex diseases and traits

Chee Seng Ku · Nasheen Naidoo · Shu Mei Teo · Yudi Pawitan

Received: 8 August 2010 / Accepted: 4 November 2010 / Published online: 23 November 2010

Abstract Regions of homozygosity (ROHs) are more abundant in the human genome than previously thought. These regions are without heterozygosity, i.e. all the genetic variations within the regions have two identical alleles. At present there are no standardized criteria for defining the ROHs resulting in the different studies using their own criteria in the analysis of homozygosity. Compared to the era of genotyping microsatellite markers, the advent of high-density single nucleotide polymorphism genotyping arrays has provided an unparalleled opportunity to comprehensively detect these regions in the whole genome in different populations. Several studies have identified ROHs which were associated with complex phenotypes such as schizophrenia, late-onset of Alzheimer's disease and height. Collectively, these studies have conclusively shown the abundance of ROHs larger than 1 Mb in outbred populations. The homozygosity association approach holds great promise in identifying genetic susceptibility loci harboring recessive variants for complex diseases and traits.

Introduction

Human genetic variations are the differences in DNA sequences within the genome of individuals within populations. These variations can take many forms, including single nucleotide variants or substitutions, tandem repeats (short tandem repeats and variable number of tandem repeats), small indels (insertions and deletions of a short DNA sequence), duplications or deletions that change the copy number of a larger segment of a DNA sequence ($\geq$1 kb) i.e. copy number variations (CNVs), and other chromosomal rearrangements such as inversions and translocations (also known as copy-neutral variations) (Nakamura 2009; Frazer et al. 2009; Ku et al. 2010a). The amount of genetic variation in the human genome is more abundant than previously thought, and this has been further corroborated with the findings from whole genome resequencing studies where several million single nucleotide polymorphisms (SNPs) and several hundred thousand indels and structural variants were identified (Wheeler et al. 2008; Bentley et al. 2008; Wang et al. 2008; Kim et al. 2009). In addition to SNPs (Altshuler et al. 2008; Hindorff et al. 2009), other genetic variations have also been found to be associated with various complex diseases and traits (Haberman et al. 2008; Hannan 2010; Wain et al. 2009; Stankiewicz and Lupski 2010).

By comparison, the region of homozygosity (ROH) is not currently classified as a type of genetic variation as there is no consensus on whether it should be classified as one type of 'structural' genetic variation. The reasons for this are two fold: (a) the ROH is not a 'genetic alteration' of the DNA sequence like other genetic variations and, (b) the research on their genome-wide mapping is still relatively new. However, the extent of ROHs varies among individuals and between different populations. In comparison to other types of genetic variations where the inter-population differences have been well documented (International HapMap Consortium 2005, 2007; Jakobsson et al. 2008; Teo et al. 2009), published data has increasingly shown the

C. S. Ku (✉) · N. Naidoo · S. M. Teo
Department of Epidemiology and Public Health,
Centre for Molecular Epidemiology,
Yong Loo Lin School of Medicine,
National University of Singapore, Singapore, Singapore
e-mail: g0700040@nus.edu.sg

Y. Pawitan (✉)
Department of Medical Epidemiology and Biostatistics,
Karolinska Institutet, Stockholm, Sweden
e-mail: yudi.pawitan@ki.se

inter-individual and inter-population variations in the profiles of homozygosity (Gibson et al. 2006; McQuillan et al. 2008; Nothnagel et al. 2010; O'Dushlaine et al. 2010).

Research on ROHs has started to gain impetus, as is evidenced by the increasing numbers of publications after the first study by Gibson et al. (2006) reporting its abundance in the human genomes of outbred populations. Further studies have investigated the population genetics aspects of ROHs in healthy individuals (Li et al. 2006; McQuillan et al. 2008; Nothnagel et al. 2010; Nalls et al. 2009b), and also performed association analyses to identify ROHs that are associated with complex diseases and traits in a case–control study design (Lencz et al. 2007; Nalls et al. 2009a; Vine et al. 2009; Yang et al. 2010b).

The aim of this paper is to review the recent progress and to elaborate on the issues and challenges in genome-wide mapping of ROHs in the human genome using high-density SNPs genotyping arrays in normal populations and in disease association studies. We also highlight the findings showing associations between ROHs and complex phenotypes. Finally, we discuss the future directions and the potential applications of ROHs as surrogate markers in identifying recessive loci for complex phenotypes. This approach is also known as 'genome-wide homozygosity association' and could be a promising alternative to finding the 'missing heritability' for complex phenotypes (Manolio et al. 2009). Population genetics and selection pressure on ROHs are briefly discussed, as these topics are beyond the scope of this review paper. Other interesting areas of ROHs research such as studies of homozygosity in inbreeding and isolated populations and findings from animal and plant genetics deserve to be reviewed in a separate paper.

What is a region of homozygosity?

A ROH defines a continuous or uninterrupted stretch of a DNA sequence without heterozygosity in the diploid state, that is in the presence of both copies of the homologous DNA segment. Thus, all the genetic variations, such as SNPs (biallelic marker) or microsatellites (multiallelic marker) within the homologous DNA segments have two identical alleles that create homozygosity (Gibson et al. 2006). The ROH is different from one-copy deletion (or hemizygous deletion), which could also lead to the homozygosity, e.g. in genome-wide SNPs genotyping data. However this is considered as a 'spurious homozygosity' because only one allele of the SNPs is present in the deleted region for one-copy deletions. Thus, the DNA fragments with only the single allele are hybridized on the genotyping array. As a result, the signal intensity of only one allele is measured and subsequently used in genotype calling, and hence it would be incorrectly labeled a homozygote genotype. Therefore, the result of 'homozygosity' is due to the absence of the other allele, instead of 'true homozygosity' where two identical alleles are present (Peiffer et al. 2006). The distinction between 'true homozygosity' as opposed to 'spurious homozygosity' due to one-copy deletion is difficult to determine just by inspection of the genotype data alone. The allelic signal intensity ratio (the relative ratio of the fluorescent signals between two probes/alleles at each SNP) is needed to differentiate between the two types of homozygosity (Peiffer et al. 2006; Wang et al. 2007). Therefore, for studies that used only SNPs genotype data to identify the ROHs, i.e. to screen regions with a minimum consecutive homozygote SNPs, the possibility that some regions are caused by one-copy deletion cannot be firmly excluded, because deletions are also widespread in the human genome (McCarroll et al. 2008; Conrad et al. 2010).

Cytogenetic abnormalities such as uniparental isodisomy can also result in homozygosity where two copies of a single parental homologous DNA segment are inherited from one parent. As such it cannot be distinguished from homozygosity resulting from other factors such as parental consanguinity using the allelic signal intensity ratio as in the case of one-copy deletion. Thus for studies that involved unrelated samples where checking the Mendelian transmission errors in the ROHs is not possible, the possibility of uniparental isodisomy leading to homozygosity cannot be definitively ruled out. Assessing the transmission errors requires data from trios or families. However, the likelihood that a considerable fraction of ROHs will be accounted for by uniparental isodisomy is low given that this cytogenetic abnormality is rare (Curtis 2007).

Currently, there is no consensus or standardized criteria used to define the ROH. However, previous studies have focused on regions $\geq$1 Mb, and thus the true extent of homozygosity in the human genome could be underestimated (Gibson et al. 2006; Li et al. 2006). More recent studies have defined a ROH at a minimum length of 500 kb (Yang et al. 2010b) with the intention of avoiding underestimation of the numbers of regions in the human genome. This is because shorter ROHs are now also thought to be associated with complex phenotypes. However, setting a shorter length for definition will increase the number of false positive signals i.e. increase the sensitivity at the expense of specificity. Therefore, in discovery studies, balancing both the sensitivity and specificity when setting the criteria to identify ROHs is critical.

By focusing only on regions $\geq$500 kb or 1 Mb, the 'noise' introduced by one-copy deletions is likely to be minimal, thus reducing the potential to cause spurious homozygosity. This is because large deletions of $\geq$500 kb are relatively rare in the human genome—as supported by data from high-resolution genome-wide mapping of CNVs

studies (McCarroll et al. 2008; Conrad et al. 2010; Ku et al. 2010b; Park et al. 2010a; Yim et al. 2010). Therefore, a critical issue to be addressed in future homozygosity mapping studies is determining the optimal cutoff of the length of the ROH to be adopted, as this will avoid over-estimating the homozygosity when the length is set too low and which can then be easily confounded by one-copy deletion of hundreds of kilobases or smaller. Although some studies have reduced the cutoff length to 500 kb (Yang et al. 2010b), it is still uncertain whether this new cutoff can readily reflect the true extent of homozygosity in the human genome.

Defining criteria and terminologies

Before the term 'copy number variation (CNV)' was first introduced in 2006 (Freeman et al. 2006), various different terms were used to describe these copy number variable regions such as 'large-scale copy number variants' and 'intermediate-sized variants' (Sebat et al. 2004; Iafrate et al. 2004). To date, various terminologies have also been used to describe the ROHs such as 'extended tracts of homozygosity' (Gibson et al. 2006), 'long contiguous stretches of homozygosity' (Li et al. 2006), 'runs of homozygosity' (Nothnagel et al. 2010; McQuillan et al. 2008), 'autozygosity regions' (Nalls et al. 2009b) and 'homozygosity-by-descent' (Polasek et al. 2010). Different studies have used their own criteria in identifying ROHs with some studies employing more stringent criteria compared to others applying a more liberal definition (Gibson et al. 2006; Li et al. 2006; Nothnagel et al. 2010; McQuillan et al. 2008; Nalls et al. 2009b; Curtis et al. 2008). For example, Curtis et al. (2008) used their own developed software and the criteria of a minimum of 10 consecutive, homozygous SNPs extending over 1 Mb. In comparison, other studies employed the default definition implemented in the 'Runs of homozygosity' function in the PLINK software (http://pngu.mgh.harvard.edu/~purcell/plink/). These criteria are (a) the length of the ROH $\geq$1 Mb, (b) a minimum of 100 SNPs per ROH, and (c) a density of at least 1 SNP per 50 kb (Nothnagel et al. 2010). As all the studies are referring to the same type of 'DNA sequence feature' it is essential to standardize the terminology to be used in describing these regions to avoid confusion.

Polymorphic markers used to detect ROHs

Although long continuous ROHs have been documented a decade ago in reference families from the Centre D'etude Du Polymorphisme Humain (CEPH) (Broman and Weber 1999), no large-scale population-based study had been performed to interrogate the extent of ROHs in the human genome until the first study by Gibson et al. (2006). The recent advances in genome-wide mapping or detection of ROHs have been driven mainly by the availability of highly accurate SNPs databases such as the International HapMap Project, and the technology to genotype several hundred thousand to several million SNPs throughout the human genome (International HapMap Consortium 2005, 2007; Gibbs and Singleton 2006; Ragoussis 2009). The early study in the CEPH families used approximately 8,000 short tandem repeat markers and detected long continuous ROHs. In contrast, subsequent studies have applied SNPs as the polymorphic markers to detect the ROHs (Gibson et al. 2006; Li et al. 2006; McQuillan et al. 2008; Nothnagel et al. 2010; Nalls et al. 2009b). At the single marker level, short tandem repeats are more informative than SNPs because they are multiallelic markers. However, SNPs are more numerous and collectively can yield more information than short tandem repeats and offer a higher resolution compared to other genetic markers—both of which are important to accurately identify the numbers and sizes of ROHs.

Genotyping a large number of SNPs in a microarray platform presents a powerful tool to detect ROHs comprehensively across the whole genome (Gibbs and Singleton 2006; Ragoussis 2009). This also enables investigation into the number, length or size, and location or distribution of the ROHs in the human genome in a more unbiased manner compared to microsatellite markers (Gibson et al. 2006; Li et al. 2006; McQuillan et al. 2008; Nothnagel et al. 2010; Nalls et al. 2009b). The SNPs genotyping platforms also allow studies of the relationship between ROHs and recombination or linkage disequilibrium (LD) patterns, as the SNPs data can be used for haplotype analyses and to calculate the recombination rates (Curtis et al. 2008). The ability to investigate the co-occurrence of ROHs in the areas with extensive LD or low recombination is important in investigating the mechanisms contributing towards the high frequency of ROHs in the human genome.

Genotyping of a sufficiently large number of SNPs is required to accurately detect the ROHs. The study by Gibson et al. (2006) used data from the International HapMap Phase I Project comprising of approximately 1 million SNPs (International HapMap Consortium 2005), whilst other studies have used lower density genotyping arrays ranging from 300,000 to 550,000 SNPs. The importance of having high-density polymorphic markers was shown by Gibson et al. (2006) who found the largest ROH of 17.9 Mb containing 3,922 SNPs from the SNPs data from HapMap Phase I. However, using the data from HapMap Phase II comprising of >3 million SNPs (International HapMap Consortium 2007), a total of 12,778 SNPs were found in the region with 11 heterozygotes. These heterozygotes interrupted the ROH and have divided it into 12 smaller

segments (Gibson et al. 2006). However, it is unclear whether these 11 heterozygotes are genotyping errors or true heterozygotes occurring as a result of recent mutations. Thus, to account for genotyping errors, studies have allowed some missing genotypes and heterozygotes for each ROH to avoid artificially splitting the region (Table 1).

This hints that the sizes of ROHs may be over-estimated in previous studies when using lower density SNPs genotyping arrays. Therefore, the numbers and sizes of ROHs identified by previous studies are likely to be different or altered when higher density SNPs data is available for analysis on the same samples. This also implies that a cautious interpretation should be imposed for ROHs of several megabases for studies using lower resolution SNPs data. A higher density of SNPs is needed for a definitive assessment of ROHs. Although the SNPs genotyping array is an invaluable tool to detect ROHs, it is not without limitations. Similar to CNV detection using SNPs genotyping platforms, the boundaries of the ROHs cannot be determined accurately at a single nucleotide resolution, as accuracy depends on the SNPs resolution. Therefore, like CNVs, the sizes of ROHs could be inflated, i.e. the ROHs detected in previous studies could be smaller than currently estimated. However, there is currently no data supporting this speculation for ROHs as compared to CNVs (McCarroll et al. 2008; Perry et al. 2008).

Methods of detecting ROHs

Several targeted and genome-wide molecular methods are available to detect structural variations such as CNVs (deletions and duplications) and copy-neutral variations (translocations and inversions). However, unlike with structural variations, ROHs cannot be detected with technologies used in molecular genetics such as fluorescence in situ hybridization (FISH) and bacterial artificial chromosome (BAC) clone or oligonucleotide-based comparative genomic hybridization (CGH) arrays (Carson et al. 2006; Feuk et al. 2006; Carter 2007). Furthermore, several new sequencing-based approaches for detecting structural variations such as paired-end sequencing mapping and depth-of-coverage of the sequence read are also unfit to detect ROHs (Korbel et al. 2007; Kidd et al. 2008; Yoon et al. 2009).

The genome-wide mapping of ROHs can only be done using SNPs genotyping arrays or direct sequencing. The whole-genome resequencing or de novo genome assembly using the next or third generation sequencing technologies will offer an almost complete solution to detecting most of the genetic variations including ROHs within the human genome. However, these high-throughput sequencing technologies were not readily available until recently, and the cost is still prohibitively expensive to sequence the whole human genome in a population-based study (Mardis 2008; Metzker 2010). As a result, SNPs genotyping arrays are the main tools for ROH mapping. The SNPs data can be used in two different ways to detect the ROHs. The first approach is to screen the whole genome in a sliding window manner for consecutive SNPs showing homozygotes over a certain length such as 1 Mb, as implemented in PLINK (Purcell et al. 2007). Since this approach only uses genotype data, it is unable to distinguish between true homozygosity and the spurious homozygosity caused by one-copy deletion without further investigations of CNVs in the samples.

This limitation has been overcome by the second approach which relies on the signal intensity data. Two types of signal intensity data are generated by the SNPs genotyping array: (a) the total signal intensity or log R ratio (LRR) and (b) the allelic intensity ratio or B allele frequency (BAF). The combination of LRR and BAF can be used to determine several different states of copy numbers such as homozygous and hemizygous deletions, and one-copy and two-copy duplications, and ROHs as implemented in the PennCNV algorithm. The BAF is needed to differentiate between ROH from normal diploid copies and one-copy deletion (Wang et al. 2007). Figure 1 illustrates the difference in LRR and BAF patterns between ROH and one-copy deletion. For the one-copy deletion, there is a decrease in LRR in addition to the absence of heterozygosity as shown in the BAF panel. Conversely, no reduction in LRR will be seen for ROH, but the absence of heterozygosity is observed. Most of the genome-wide studies of ROHs have used SNPs genotyping arrays. In comparison, the commonly used oligonucleotide-based CGH arrays in detecting CNVs produced only total signal intensity data. This renders them unable to be used for identifying ROHs.

In addition to the most commonly used PLINK software for detecting and analyzing ROHs (Table 1), other methods have also been recently developed for these purposes (Seelow et al. 2009; Browning and Browning 2010; Polasek et al. 2010). The development of powerful and accurate tools or methods for the detection and analysis is a prerequisite for the success of research into ROHs. Furthermore, new algorithms to identify disease-related segments based on homozygosity using case–control data have also been developed. This will enhance studies to identify ROHs that differ between cases and controls, as these regions may contain recessive variants underlying the diseases (Wang et al. 2009). All the ROHs detection methods have their own strengths and limitations with varying rates of false-positive and false-negative results and as such, a combination of methods would be more ideal to minimize these limitations.

Table 1 Summaries of genome-wide association studies of ROHs and complex phenotypes using high-density SNP genotyping arrays

Phenotype and study	Sample size and genotyping platform	Software, criteria of ROHs, association analysis	Major results
Schizophrenia (Lencz et al. 2007)	178 cases and 144 controls Affymetrix 500K	Software • Whole-genome homozygosity analysis (WGHA) performed with customized python scripting in the HelixTree environment Criteria • ROH—any window of 100 or more consecutive SNPs that are homozygous, not receiving a heterozygous call • Common ROHs—only those ROHs in which 10 or more subjects share ≥100 identical homozygous calls were retained for further analysis Association analysis • Case–control comparisons of frequency of presence for each common ROH were examined by using χ^2 test	• A total of 339 common ROHs were identified • Schizophrenia cases demonstrated a significantly greater number of common ROHs than controls • 9 ROHs significantly differed in frequency between cases and controls
Bipolar disorder (Vine et al. 2009)	553 cases and 547 controls Affymetrix 500K	This study applied the WGHA approach as demonstrated in the Lencz et al. (2007) study	• A total of 239 common ROHs were identified • The total number of common ROHs did not differ between cases and controls • 7 common ROHs were significant at $p < 0.05$
Late-onset Alzheimer's disease (Nalls et al. 2009a)	837 cases and 550 neurological normal controls Affymetrix 500K	Software • PLINKv1.02 • A sliding window of 50 SNPs, allowing at most 2 missing genotypes and 1 heterozygote call per ROH Criteria • ROH—at least 1 Mb of consecutive homozygous genotypic calls •Minimum SNP density coverage—at least 50 SNPs per megabase Association analysis • 1,090 consensus regions from overlapping ROHs were defined •Each consensus region was found in no less than 10 participants • The consensus ROHs were analyzed using the maxT permutation test algorithm for case/control studies in PLINKv1.02	• One homozygous consensus region in chromosome 8 was found to be significantly overrepresented in cases when compared to controls • The cases presented a slightly higher degree of extended homozygosity when compared with the control group

Table 1 continued

Phenotype and study	Sample size and genotyping platform	Software, criteria of ROHs, association analysis	Major results
Height (Yang et al. 2010b)	Discovery study 998 US Caucasian subjects Affymetrix 500K Replication study 8,385 Caucasian subjects from the Framingham Heart Study Affymetrix 500K plus 50K supplemental array	Software • PLINKv1.01 • A sliding window of 5 Mb (minimum 50 SNPs), allowing 5 missing SNPs and 1 heterozygous site per window Criteria • A minimum of 100 consecutive SNPs in a ROH • Minimum length for a ROH, 500 kb • Minimum density in a ROH, 50 kb per SNP • Maximum gap between 2consecutive homozygous SNPs—100 kb Association analysis • Individual ROHs were divided into different ROH groups using the homozyg-group command in the Runs of Homozygosity program • For each ROH group containing >50 subjects—Student's *t* test to compare the adult height of subjects with this ROH group to the height of subjects without this ROH group	Discovery study • 113,910 individual ROHs in 998 subjects • For the association analyses between human adult height and ROHs, 3,322 ROH groups containing more than 50 individual ROHs • 80 ROH groups overlapped with copy number polymorphisms and were excluded from the subsequent association analyses. • One ROH group (ROH 12q21.31) was significantly associated with adult height even after Bonferroni correction Replication study • A significant association with adult height was successfully replicated for the ROH group by FBAT analysis
Colorectal cancer (Spain et al. 2009)	921 cases and 929controls Illumina Infinium Human Hap550 BeadChips	Software • PLINKv1.05 • A sliding window of 50 SNPs, allowing 2% heterozygous SNPs and 5 missing calls in each window Criteria • This study initially analyzed ROHs that were ≥50 SNPs in length • Repeated the analysis using a number of different criteria to define a ROH (≥30 SNPs, ≥40 SNPs, ≥60 SNPs, ≥2 Mb, ≥4 Mb, and ≥10 Mb) Association analysis • Statistical analyses were performed using packages available in R	• No evidence was found for an association between total size of the ROHs in each individual and colorectal cancer • This study calculated the frequencies of cases and controls in which one or more ROHs of ≥4 Mb were detected • 159 of 921 (17%) cases and 142 of 929 (15%) controls had ROHs ($p = 0.14$, Fisher's exact test)
Childhood acute lymphoblastic leukemia (Hosking et al. 2010)	824 cases and 2,398 controls Illumina Infinium Human370 Duo BeadChips	Software • PLINKv1.06 • A sliding window of SNPs across the entire genome, 2% heterozygous SNPs were allowed in each window, 5 missing calls per window Criteria • ROH, ≥75 consecutive SNPs • Only ROHs which occurred in ≥10 persons were retained for analysis Association analysis • Subsequent statistical analyses were performed using packages available in R • Comparison of the distribution of categorical variables was performed using the χ^2 test	• A total of 396 ROHs were identified • Patients and controls showed no significant difference in the average number of ROH • 4 ROHs differed significantly ($p < 0.01$) between cases and controls

Table 1 continued

Phenotype and study	Sample size and genotyping platform	Software, criteria of ROHs, association analysis	Major results
Breast and prostate cancer (Enciso-Mora et al. 2010)	Breast cancer 1,183 cases and 1,185 controls Illumina Infinium Human550 Duo BeadChips Prostate cancer 1,177 cases and 1,149 controls Illumina Infinium Human217 and Human 317 BeadChips	Software • PLINK v1.06 • A sliding window of SNPs across the genome, 2% heterozygous SNPs were permitted in each window, 5 missing calls per window Criteria • ROH, ≥80 consecutive SNPs • Only considered ROH that occurred in ≥ 10 individuals Association analysis • Subsequent statistical analyses were performed using packages available in R • Comparison of the distribution of categorical variables was performed using the χ^2 test	• A total of 415 and 426 ROHs were identified in breast cancer and prostate cancer series, respectively • 6 ROHs differed significantly ($p < 0.01$) between breast cancer cases and controls. • 4 ROHs differed significantly ($p < 0.01$) between prostate cancer cases and controls

Different studies have applied different filtering or quality control criteria of the genome-wide SNP data and samples before the data was used for ROH analysis and association studies

Mechanisms generating ROHs

Several mechanisms and factors have been postulated to explain the high frequency of ROHs in the human genome namely, parental consanguinity, uniparental isodisomy and the presence of 'common extended haplotypes'. One of the most common and well established mechanisms leading to ROHs of several megabases is parental consanguinity, in which the offspring inherits chromosomal segments that are identical-by-descent from each parent. Published data has shown that the number of ROHs of several megabases increased markedly in the offspring of consanguineous marriages (Li et al. 2006; Woods et al. 2006) with up to 6% of homozygosity anticipated in the genome of the offspring of first cousin marriages (Broman and Weber 1999). Li et al. (2006) showed that in a family with 4 children from first cousin marriages, multiple ROHs ranging from 3.06 to 53.17 Mb were observed in all the children. Woods et al. (2006) also showed a marked increase in homozygosity levels in individuals with a recessive disease whose parents were first cousins, where 11% of their genomes were homozygous on average. Additionally, the cumulative length of ROHs per genome was found to be larger in two isolated rather than in two more cosmopolitan (non-isolated) European populations (McQuillan et al. 2008). Therefore, when compared to outbred populations, there is an expected increase in the level of homozygosity or number of ROHs in populations where consanguineous marriages are prevalent, as well as in isolated populations where limited random mating or a restricted mate choice has taken place. However, this is unlikely to be the main factor responsible for the high frequency of ROHs in outbred populations in which parental consanguinity is uncommon.

Another widely discussed mechanism is cytogenetic abnormalities such as uniparental disomy, which can be divided into uniparental isodisomy and uniparental heterodisomy. Only uniparental isodisomy can cause homozygosity as the offspring inherits two identical copies of a homologous chromosomal segment from only one parent. As a result, no heterozygosity would be observed in that particular homologous chromosomal segment (Ting et al. 2007). Similarly, this is also an unlikely explanation for the abundance of ROHs reported in the literature; given that uniparental disomies are rare genetic abnormalities that can cause severe and rare genomic disorders when their locations affect imprinted genes. Examples of these disorders are Prader–Willi Syndrome, Angelman Syndrome and Silver–Russell syndrome (Gurrieri and Accadia 2009; Van Buggenhout and Fryns 2009; Abu-Amero et al. 2008). This is further supported by previous studies concluding that the ROHs are not due to genetic abnormalities as no excess apparent deviation from Mendelian transmission was observed. More specifically, transmis-

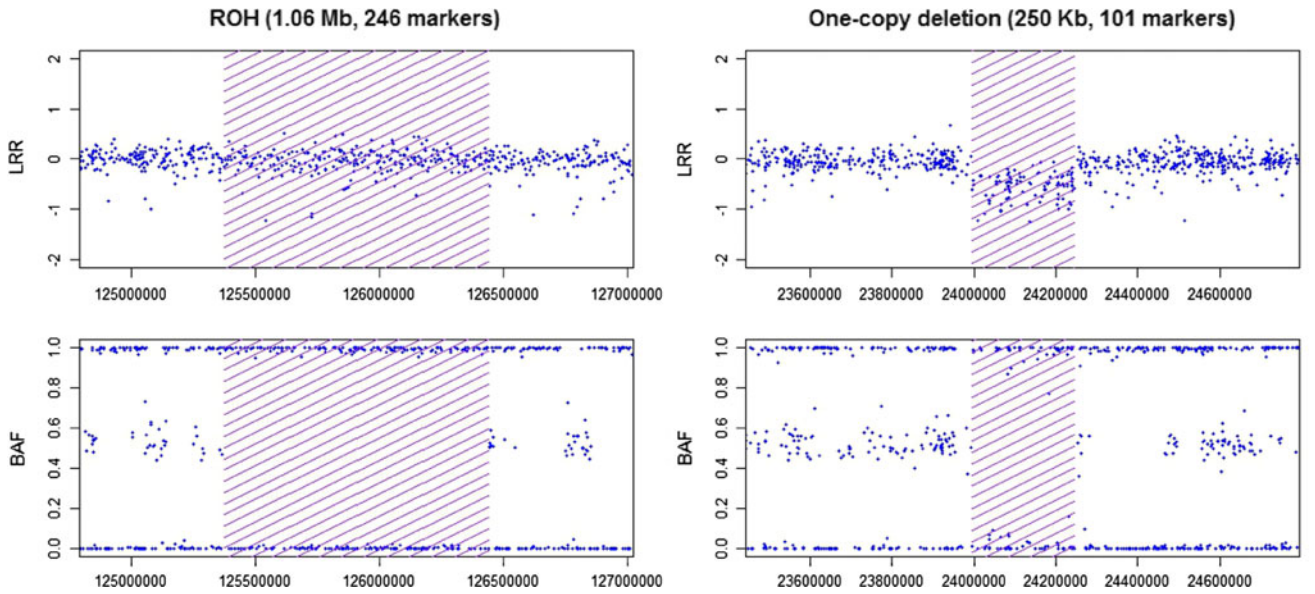

Fig. 1 Plots of the differences in the LRR (Log R Ratio) and BAF (B Allele Frequency) patterns for the ROH (*left panels*) and one-copy deletion (*right panels*) generated from a sample derived from our previous study (Ku et al. 2010b) and genotyped by the Illumina 1 M Beadchip. The ROH and one-copy deletion were detected using the LRR and BAF information by PennCNV algorithm (LRR: total fluorescent intensity signals from both sets of probes/alleles at each SNP, BAF: the relative ratio of the fluorescent signals between two probes/alleles at each SNP) (Wang et al. 2007). The size of the ROH is approximately 1.06 Mb (1,064,933 bases) spanning from 125374832 to 126439764 in chromsome 2. This region contains 246 markers. The size of the one-copy deletion is approximately 250 kb (250,186 bases) spanning from 23994408 to 24244593 in Chromosome 22. This region contains 101 markers. The regions affected by the ROH and one-copy deletion were shaded and the blue dots represent markers in the genotyping array

sion errors occur more rarely in ROHs than would be expected by chance as shown by the observed number of Mendelian transmission errors within a ROH which is less than the expected number (Curtis 2007). Since this study has clearly demonstrated that the ROHs are not usually due to cytogenetic abnormalities, it then indirectly supports the presence of common extended haplotypes as the mechanism contributing toward the high frequency of ROHs in human genomes.

The presence of common extended haplotypes therefore becomes the most likely factor responsible for the high frequency of ROHs which are passed on from both parents to the offspring in the genomes of outbred populations. Data demonstrating the co-occurrence of ROHs in regions with extensive LD and low recombination rates also support the hypothesis of common extended haplotypes in generating homozygosity in the genomes of outbred populations (Gibson et al. 2006; Curtis et al. 2008). A further process believed to be driving the increasing frequency of common extended haplotypes is positive selection. ROHs resulting from common extended haplotypes may be indicative of positive selection pressure of functional importance of these regions. Several methods have been used to quantify the positive selection pressure on ROHs namely, the integrated haplotype score (iHS), Tajima's *D* test and the Fixation index (F_{ST}). Numerous large (several megabases) and common (>25%) ROHs were found to have high values for these metrics indicating the signal for positive selection (Enciso-Mora et al. 2010; Hosking et al. 2010).

Genome-wide mapping of ROHs in the human genome

It was not previously expected that the genomes of outbred populations contain ROHs of several megabases until the first few early reports in 2006 and 2007 (Gibson et al. 2006; Li et al. 2006; Simon-Sanchez et al. 2007). One study found ROHs of >5 Mb in 26 of the 272 unrelated samples assessed (Simon-Sanchez et al. 2007). Similarly, another study performed in Han Chinese also observed the high frequency of ROHs, where 34 of the 515 unrelated individuals contained ROHs ranging from 2.94 to 26.27 Mb (Li et al. 2006). While Gibson et al. (2006) studied the samples from the International HapMap Projects and identified 1,393 ROHs exceeding 1 Mb in 209 unrelated HapMap individuals. Several hundreds of ROHs were found in each of the HapMap populations, and the average number of ROHs (>1 Mb) per individual was found to be lowest in the Yoruba Ibadan Nigerian (YRI) population compared to other populations within the HapMap Phase I Project (Gibson et al. 2006). In addition to demonstrating that ROHs are remarkably common, even in the unrelated individuals from the apparently outbred populations, Gibson et al. (2006) also demonstrated the value of including diverse

populations to examine the differences in ROHs. In the YRI population, the samples have the least number of ROHs per individual. This finding is expected, because the populations of African ancestry are older in human history and hence have more generations and a higher number of recombination events than other populations (recombination occurs during meiosis in each generation). Recombination is one of the important processes to interrupt the long continuous ROHs into smaller segments over the generations. Population differences in ROHs have also been well documented in other studies (Nothnagel et al. 2010).

Each of the previous studies identified a different number of ROHs per individual (Li et al. 2006; Nothnagel et al. 2010; McQuillan et al. 2008; Nalls et al. 2009b; Curtis et al. 2008). These differences are likely reflective of technical and methodological variations such as differing genotyping platforms or SNPs data, differing defining criteria and differing analytical techniques used. Both the genotyping platform and defining criteria can significantly influence the profile of ROHs by way of number, size, cumulative length and genomic distributions. Slight alterations in defining criteria can substantially affect the number of ROHs detected and as a result comparisons between studies are difficult. Therefore, it is critical to develop a set of standardized criteria in identifying ROHs and to establish a database to catalog these regions in the human genome from published studies, similar to other databases developed for SNPs and structural variants (CNVs) such as the dbSNP and Database of Genomic Variants, respectively (Day 2010; Iafrate et al. 2004). This database will enable researchers to quickly compare their results with published data. Consensus on defining the ROHs and the construction of a database to serve as a reference will help in expediting research in ROHs.

LD-pruning of SNPs in mapping of ROHs

The SNPs genotyping data is undoubtedly invaluable for identifying ROHs. However, there is an issue of whether pruning the list of SNPs to remove local LD (i.e. to remove SNPs that are in strong LD) should be done before the data can be used for ROHs. The idea of LD-pruning of SNP data is that the LD between the SNPs can inflate the chance of occurrence of biologically meaningless ROHs. However, there are still uncertainties with regards to the LD-pruning step such as the optimal cutoff of LD (measured by r^2) to be used, although some studies have used the conventional and arbitrary cutoff of $r^2 > 0.8$. More importantly, it is unclear about the quality and performance in terms of sensitivity and specificity for mapping ROHs using LD-pruning SNPs data compared to data without the LD-pruning step. This is an interesting research subject worth pursuing and studies should be done to assess the importance of this LD-pruning step. However, unless significant differences in the sensitivity and specificity are shown using LD-pruning SNP data, the LD-pruning step may not necessarily be needed.

Some of the studies using whole-genome SNPs genotyping arrays have omitted the LD-pruning step before the data was used for mapping ROHs, even though Gibson et al. (2006) used the SNP data from the International HapMap Project where the LD information is readily available. However, others have taken the LD between SNPs into account and used the pairwise LD SNP pruning function in PLINK, with a default value of $r^2 > 0.8$ (Enciso-Mora et al. 2010; Hosking et al. 2010). For example, one study found 370,611 separate tag groups which is a 27.6% reduction of information compared with the original number of SNPs. To account for this, the study adopted a more stringent cutoff of a minimum of 80 consecutive SNPs (instead of 58) to identify ROHs (Enciso-Mora et al. 2010). Similarly Lencz et al. (2007) also took into consideration the LD between the SNPs through setting a more stringent threshold of 100 consecutive SNPs that are homozygous. In comparison, another study removed SNPs in LD with $r^2 < 0.1$ leaving only 30,307 SNPs to form the 'low-LD panel' for some analyses (Spain et al. 2009). Although these studies have taken LD between SNPs into account, it is unclear whether an improvement in sensitivity and specificity was achieved by implementing this LD-pruning step since no evaluation was done to directly compare the differences between the ROHs profile with and without the LD-pruning step. Therefore, the LD-pruning step is conceptually correct; however to warrant this step to be performed in future genome-wide mapping of ROHs, more published data demonstrating its advantages is needed.

Implications on complex diseases and traits

Many novel pathogenic genes or mutations underlying autosomal recessive disorders have been identified through homozygosity mapping. This approach has been shown to be powerful and is particularly useful in investigating autosomal recessive disorders especially in populations with a high prevalence of consanguinity. This is evident from the enormous number of studies identifying causal mutations for autosomal recessive disorders in consanguineous families (Abu Safieh et al. 2010; Harville et al. 2010; Walsh et al. 2010; Pang et al. 2010; Lapunzina et al. 2010; Nicolas et al. 2010; Uz et al. 2010; Iseri et al. 2010; Collin et al. 2010). However, the first study applying the homozygosity association approach at the genome-wide scale for complex diseases only appeared in 2007 (Lencz et al. 2007). Table 1 summarizes the

genome-wide ROH association studies of complex phenotypes using high-density genotyping arrays.

The 'homozygosity analysis' has been shown to be useful for the identification of disease susceptibility genes in both monogenic and complex diseases (Miyazawa et al. 2007; Jiang et al. 2009). The effects of inbreeding or consanguinity and recessive variants or heterozygosity levels on the risk of complex phenotypes (diseases and quantitative traits) have been previously well established (Rudan et al. 2003a, 2003b, 2006; Campbell et al. 2007). A strong linear relationship between the inbreeding coefficient and blood pressure was found and several hundred recessive loci were predicted as contributing to blood pressure variability. Recessive or partially recessive genetic variants account for 10–15% of the total variation in blood pressure (Rudan et al. 2003a). Higher levels of relative heterozygosity were shown to be associated with lower blood pressure and total and low-density lipoprotein cholesterol by measuring genome-wide heterozygosity (Campbell et al. 2007). In addition to quantitative traits, inbreeding was also found to be a significant positive predictor for a number of late-onset complex diseases such as coronary heart diseases, stroke, cancer and asthma (Rudan et al. 2003b). These studies have strongly supported the hypothesis that the genetics of complex phenotypes include a component of recessively acting variants; however, these studies did not directly investigate the associations of complex phenotypes with ROHs detected using polymorphic markers.

Although the information regarding the extent of ROHs in the human genome is still limited compared with SNPs, indels and CNVs, their potential impact on complex diseases and traits could also be significant as other genetic variations. The importance of ROHs to complex phenotypes remains largely unexplored; however, several studies have shown significant differences in ROHs between cases and controls in a genome-wide investigation for schizophrenia (Lencz et al. 2007), late-onset Alzheimer's disease (Nalls et al. 2009a) and height (Yang et al. 2010b). The idea underlying the homozygosity association approach is to uncover recessive variants contributing to complex phenotypes. The success of this approach has been demonstrated in several studies. Nine common ROHs significantly differentiated schizophrenia cases from controls. More interestingly, four of the regions contained or were located near to the genes that are known to be associated with schizophrenia such as *NOS1AP*, *ATF2*, *NSF*, and *PIK3C3* (Lencz et al. 2007). This proof-of-principle study has demonstrated the applications of the whole-genome homozygosity association approach in identifying genetic risk loci for complex phenotypes and it represents an alternative and new avenue in addition to SNPs analysis.

Similarly in a large-scale association study involving 837 late-onset Alzheimer's disease cases and 550 controls, one ROH on chromosome 8 was identified, and three of the genes (*STAR, EIF4EBP1* and *ADRB3*) in the region are biologically plausible candidates (Nalls et al. 2009a). Success was also achieved for complex quantitative traits such as height (Yang et al. 2010b), where strong statistical evidence showing association of one ROH with height was obtained in a total sample size of >10,000 in both the genome-wide discovery and replication studies. The height of individuals with the particular ROH was significantly higher (increased by 3.5 cm) than the individuals without the region. The identification of this ROH added further support to the contribution of recessive loci to adult height variation (Kimura et al. 2008; Xu et al. 2002). Nonetheless, other studies produced negative results, as no evidence of homozygosity was found for bipolar disorder (Vine et al. 2009).

To date, the results showing the association between homozygosity with various cancers are also controversial (Hosking et al. 2010; Assié et al. 2008; Enciso-Mora et al. 2010). For example, two studies investigating the homozygosity in colorectal cancers derived an opposing conclusion which is likely due to the differences between the two studies such as the sample sizes, the density of genotyping platforms and the analysis (Bacolod et al. 2008; Spain et al. 2009). Although studies have found statistically negative results after imposing the stringent Bonferroni correction for multiple-testing, a number of ROHs warrant further investigation as these regions overlapped with biologically plausible genes for the phenotypes. One ROH was found to encompass the gene encoding erythropoietin receptor (*EPOR*) protein. Over-expression of this protein has been documented in acute lymphoblastic leukemia (Hosking et al. 2010).

Many reasons can be speculated for the inconsistencies as to why associations of ROHs were only found in some diseases or studies but not others. This could also indicate that the effects of homozygosity on the risk of complex phenotypes may be disease or trait-dependent, for example some quantitative traits have shown significant variance due to recessive alleles such as systolic blood pressure, total cholesterol and low-density lipoprotein cholesterol. This implies that the effects of homozygosity may be greater in influencing the variation of these traits than others (Campbell et al. 2009). On the other hand, it could also be population-dependent since differences in homozygosity between populations have been documented. Although a number of genome-wide homozygosity association studies have been performed, the optimum study design or analysis methods for assessing the associations or effects of ROHs on the disease risk has not yet been well established. This is, however, vital before breakthrough discoveries can be made in this research area.

The idea for using the homozygosity association approach to dissect the genetics of complex phenotypes is

to reveal the recessive loci that only express their effects (or increase the risk of complex diseases) in the presence of two deleterious recessive alleles, in a recessive disease model. In addition to autosomal recessive disorders, complex diseases can also be affected by recessive variants. The conventional single-SNP analysis approach applied in GWAS may not be statistically powerful enough to identify recessive alleles with small effect sizes and moreover, the recessive model is not usually tested. Until the effect of homozygosity on complex phenotypes is better understood, it is premature to make any conclusions, as the field is still in its infancy compared to association studies between SNPs and CNVs for complex diseases and traits. However, collectively these studies have demonstrated the feasibility of using the homozygosity association approach to identify susceptibility loci for complex phenotypes and have produced encouraging results. This also further underscores the need to further investigate and catalog the extent of ROHs in different populations. Similar to the other genetic variations, ROHs have the potential of becoming the genetic markers in GWAS. In fact, homozygosity mapping has been commonly used to identify the loci for recessive diseases in consanguineous families.

Strengths and shortcomings of genome-wide homozygosity association studies

From the statistical analysis point of view, the advantage of the genome-wide homozygosity association approach is that it suffers lesser penalty from Bonferroni correction for multiple-testing as significantly fewer ROHs are involved compared to the number of SNPs tested in GWAS. Thus, it needs a less stringent *p* value cutoff to declare genome-wide significance. Thus, the genome-wide ROHs association approach has a higher statistical power or requires a fewer number of samples in the studies than the 'conventional GWAS'.

GWAS is an indirect approach that relies on LD to identify the causal variants, thus the results from GWAS are pinpointing genetic loci rather than revealing the causal variants directly (Wang et al. 2005; Hirschhorn and Daly 2005). Similarly in genome-wide homozygosity association studies, one or more ROHs are identified as susceptibility risk loci rather than revealing the actual recessive variants causing the disease. For example, the homozygous consensus region in chromosome 8 was found to be associated with late-onset Alzheimer disease contains seven genes. However, the number of recessive variants within these genes or this region responsible for this 'statistical association signal' and which are functionally important in causing the diseases is unknown (Nalls et al. 2009a). The approaches to be taken from identifying the disease or trait-associated ROHs to locating the functional recessive variants is also unclear. Moreover, the sizes of ROHs are many folds larger than the LD blocks detected by conventional SNP analysis in GWAS, thus making the fine mapping of recessive variants harder. Therefore, the genome-wide association of ROHs, at best, can only pinpoint to a relatively large region harboring as yet to be identified recessive variants.

One common issue and problem in case–control association studies of CNVs and ROHs is how to construct the common CNV and ROH regions in the first place. This step is required to group the individual CNVs or ROHs into a common and discrete region. Similar to CNVs, it is unclear how to partition the individual ROHs into ROH groups so that the frequencies can be used for association analysis. This represents an important analytical challenge in these studies. Genome-wide studies investigating the association of common CNVs with complex phenotypes have so far yielded limited successes (Wellcome Trust Case Control Consortium 2010). As for ROHs, different studies have used their own methods to define ROH groups as no standardized criteria are available. Alternatively this step can be easily performed as the individual ROHs can be divided into different ROH groups by using the 'homozyg-group' command in the 'Runs of Homozygosity' program in PLINK. As a result, each ROH group is actually the overlapping region among all the individual ROHs in the group i.e. the consensus region (the region shared by all overlapping ROHs) (Fig. 2). Using this approach, Yang et al. (2010b) identified 3,322 ROH groups containing more than 50 individual ROHs. While Nalls et al. (2009a) identified 1,090 consensus regions from overlapping ROHs, but each consensus region was found in 10 or more individuals.

Besides identifying the ROH groups for association analysis, attempts were also made to compute other parameters such as the total length of the genome comprised by ROHs (the sum of the length of all ROHs), average length of each ROH (the total length divided by the number of ROHs) and the number of ROHs per individual and

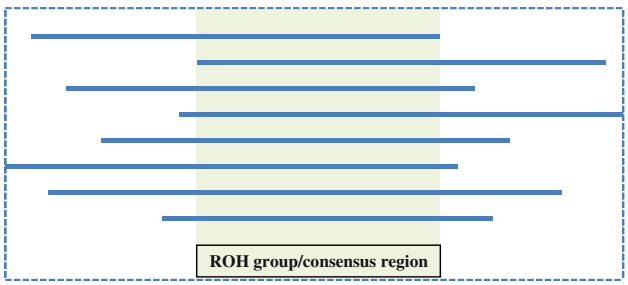

Fig. 2 Schematic diagram illustrating the ROH group or consensus region (*shadowed rectangle*) of several individual ROHs (*blue line*). Only 8 individual ROHs are shown for illustrative purposes with each individual ROH extending in both directions from the consensus region

compare these parameters between cases and controls. Nonetheless, no significant result was observed for late-onset Alzheimer disease (Nalls et al. 2009a). Likewise, no significant difference was found in the average number of ROHs between acute lymphoblastic leukemia, breast and prostate cancers with their controls (Hosking et al. 2010; Enciso-Mora et al. 2010).These analyses may not be very fruitful and have a limited interpretation. Even though significant results were obtained for all the three parameters, the findings are not informative in pointing to specific ROHs that are important to the disease. It can only be concluded that the overall extent of homozygosity is significantly greater in cases compared to controls and thus some recessive variants may be predisposed to the disease risk.

Conclusions

Published data have conclusively demonstrated the high frequency of ROHs in the genomes of outbred populations, and previous studies have also successfully unraveled the associations between ROHs and several complex phenotypes such as schizophrenia, late onset Alzheimer's diseases and height. These studies have shown the promise of the homozygosity association approach in identifying recessive loci for complex phenotypes. However, to what extent this approach contributes toward dissecting the genetics of complex phenotypes is yet to be determined. The analysis of ROHs is now feasible and convenient given the readily available high-density SNPs genotype data and the powerful detection tools such as the PLINK and PennCNV algorithms. Cataloging ROHs in different populations is important, as it lays the foundation for exploring the recessive variants for complex phenotypes.

Currently, the results from GWAS focusing on SNPs analysis alone, explains only a small fraction of the heritability of complex phenotypes (Manolio et al. 2009). Several reasons accounting for the missing heritability have been postulated (Eichler et al. 2010). The missing heritability has challenged the validity of the common-disease common variant (CD/CV) hypothesis (Schork et al. 2009), and has also diverted the research focus to rare variants (Bodmer and Bonilla 2008; Gorlov et al. 2008; Dickson et al. 2010). However, more recent studies have shown that common variants, or more specifically common SNPs, can explain a greater proportion of the heritability than what has been accounted for by GWAS done to date. These SNPs, however, are hidden within the GWAS data, and require larger sample sizes to be discovered (Yang et al. 2010a; Park et al. 2010b). The homozygosity association approach will offer an additional avenue to discovering genetic risk loci that may be missed by the conventional SNPs analysis in GWAS. The homozygosity analysis can be 'easily' performed using the SNPs genotype data and the available detection algorithms, and this is also in line with the ethos of maximizing the information from the GWAS dataset. However several issues and problems still remain as has been discussed.

The power of the homozygosity mapping approach in identifying genes and mutations for autosomal recessive disorders has been previously shown, but currently available data is limited in order to evaluate the success of this approach when applied to complex phenotypes. Hence more studies are needed in the future. Finally we advocate the use of the homozygosity association approach as an additional method of identifying loci harboring recessive variants for complex diseases and traits, which may have been undetected when conventional SNPs analysis was performed alone. The success of this approach has been demonstrated in several complex phenotypes applying the approach. The results so far are encouraging enough to warrant further studies on ROHs to investigate their impacts on complex phenotypes.

Cataloging the ROHs in human genomes and investigating their associations with complex phenotypes should build on the existing GWAS data and these are important areas to pursue in future. The contribution and the role of ROHs in complex phenotypes have been considerably neglected in GWAS; therefore we encourage researchers to explore the associations of ROHs with various phenotypes using their existing SNP data. As the high-density SNPs genotype data have already been generated by several hundred GWAS, the studies of ROHs should be relatively uncomplicated. The availability of these SNP datasets will facilitate the assessment of the roles that ROHs have in complex phenotypes.

References

Abu Safieh L, Aldahmesh MA, Shamseldin H, Hashem M, Shaheen R, Alkuraya H, Al Hazzaa SA, Al-Rajhi A, Alkuraya FS (2010) Clinical and molecular characterisation of Bardet-Biedl syndrome in consanguineous populations: the power of homozygosity mapping. J Med Genet 47:236–241

Abu-Amero S, Monk D, Frost J, Preece M, Stanier P, Moore GE (2008) The genetic aetiology of Silver–Russell syndrome. J Med Genet 45:193–199

Altshuler D, Daly MJ, Lander ES (2008) Genetic mapping in human disease. Science 322:881–888

Assié G, LaFramboise T, Platzer P, Eng C (2008) Frequency of germline genomic homozygosity associated with cancer cases. JAMA 299:1437–1445

Bacolod MD, Schemmann GS, Wang S, Shattock R, Giardina SF, Zeng Z, Shia J, Stengel RF, Gerry N, Hoh J, Kirchhoff T, Gold B, Christman MF, Offit K, Gerald WL, Notterman DA, Ott J, Paty PB, Barany F (2008) The signatures of autozygosity among patients with colorectal cancer. Cancer Res 68:2610–2621

Bentley DR, Balasubramanian S, Swerdlow HP et al (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53–59

Bodmer W, Bonilla C (2008) Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet 40:695–701

Broman KW, Weber JL (1999) Long homozygous chromosomal segments in reference families from the centre d'Etude du polymorphisme humain. Am J Hum Genet 65:1493–1500

Browning SR, Browning BL (2010) High-resolution detection of identity by descent in unrelated individuals. Am J Hum Genet 86:526–539

Campbell H, Carothers AD, Rudan I, Hayward C, Biloglav Z, Barac L, Pericic M, Janicijevic B, Smolej-Narancic N, Polasek O, Kolcic I, Weber JL, Hastie ND, Rudan P, Wright AF (2007) Effects of genome-wide heterozygosity on a range of biomedically relevant human quantitative traits. Hum Mol Genet 16:233–241

Campbell H, Rudan I, Bittles AH, Wright AF (2009) Human population structure, genome autozygosity and human health. Genome Med 1:91

Carson AR, Feuk L, Mohammed M, Scherer SW (2006) Strategies for the detection of copy number and other structural variants in the human genome. Hum Genomics 2:403–414

Carter NP (2007) Methods and strategies for analyzing copy number variation using DNA microarrays. Nat Genet 39:S16–S21

Collin RW, Safieh C, Littink KW, Shalev SA, Garzozi HJ, Rizel L, Abbasi AH, Cremers FP, den Hollander AI, Klevering BJ, Ben-Yosef T (2010) Mutations in C2ORF71 cause autosomal-recessive retinitis pigmentosa. Am J Hum Genet 86:783–788

Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, Fitzgerald T, Hu M, Ihm CH, Kristiansson K, Macarthur DG, Macdonald JR, Onyiah I, Pang AW, Robson S, Stirrups K, Valsesia A, Walter K, Wei J, Wellcome Trust Case Control Consortium, Tyler-Smith C, Carter NP, Lee C, Scherer SW, Hurles ME (2010) Origins and functional impact of copy number variation in the human genome. Nature 464:704–712

Curtis D (2007) Extended homozygosity is not usually due to cytogenetic abnormality. BMC Genet 8:67

Curtis D, Vine AE, Knight J (2008) Study of regions of extended homozygosity provides a powerful method to explore haplotype structure of human populations. Ann Hum Genet 72:261–278

Day IN (2010) dbSNP in the detail and copy number complexities. Hum Mutat 31:2–4

Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB (2010) Rare variants create synthetic genome-wide associations. PLoS Biol 8:e1000294

Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH (2010) Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet 11:446–450

Enciso-Mora V, Hosking FJ, Houlston RS (2010) Risk of breast and prostate cancer is not associated with increased homozygosity in outbred populations. Eur J Hum Genet 18:909–914

Feuk L, Carson AR, Scherer SW (2006) Structural variation in the human genome. Nat Rev Genet 7:85–97

Frazer KA, Murray SS, Schork NJ, Topol EJ (2009) Human genetic variation and its contribution to complex traits. Nat Rev Genet 10:241–251

Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, Aburatani H, Jones KW, Tyler-Smith C, Hurles ME, Carter NP, Scherer SW, Lee C (2006) Copy number variation: new insights in genome diversity. Genome Res 16:949–961

Gibbs JR, Singleton A (2006) Application of genome-wide single nucleotide polymorphism typing: simple association and beyond. PLoS Genet 2:e150

Gibson J, Morton NE, Collins A (2006) Extended tracts of homozygosity in outbred human populations. Hum Mol Genet 15:789–795

Gorlov IP, Gorlova OY, Sunyaev SR, Spitz MR, Amos CI (2008) Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. Am J Hum Genet 82:100–112

Gurrieri F, Accadia M (2009) Genetic imprinting: the paradigm of Prader–Willi and Angelman syndromes. Endocr Dev 14:20–28

Haberman Y, Amariglio N, Rechavi G, Eisenberg E (2008) Trinucleotide repeats are prevalent among cancer-related genes. Trends Genet 24:14–18

Hannan AJ (2010) Tandem repeat polymorphisms: modulators of disease susceptibility and candidates for 'missing heritability'. Trends Genet 26:59–65

Harville HM, Held S, Diaz-Font A, Davis EE, Diplas BH, Lewis RA, Borochowitz ZU, Zhou W, Chaki M, MacDonald J, Kayserili H, Beales PL, Katsanis N, Otto E, Hildebrandt F (2010) Identification of 11 novel mutations in eight BBS genes by high-resolution homozygosity mapping. J Med Genet 47:262–267

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci USA 106:9362–9367

Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6:95–108

Hosking FJ, Papaemmanuil E, Sheridan E, Kinsey SE, Lightfoot T, Roman E, Irving JA, Allan JM, Taylor M, Tomlinson IP, Greaves M, Houlston RS (2010) Genome-wide homozygosity signatures and childhood acute lymphoblastic leukemia risk. Blood 115:4472–4477

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C (2004) Detection of large-scale variation in the human genome. Nat Genet 36:949–951

International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299–1320

International HapMap Consortium, Frazer KA, Ballinger DG et al (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851–861

Iseri SU, Wyatt AW, Nürnberg G, Kluck C, Nürnberg P, Holder GE, Blair E, Salt A, Ragge NK (2010) Use of genome-wide SNP homozygosity mapping in small pedigrees to identify new mutations in VSX2 causing recessive microphthalmia and a semidominant inner retinal dystrophy. Hum Genet 128:51–60

Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung HC, Szpiech ZA, Degnan JH, Wang K, Guerreiro R, Bras JM, Schymick JC, Hernandez DG, Traynor BJ, Simon-Sanchez J, Matarin M, Britton A, van de Leemput J, Rafferty I, Bucan M, Cann HM, Hardy JA, Rosenberg NA, Singleton AB (2008) Genotype, haplotype and copy-number variation in worldwide human populations. Nature 451:998–1003

Jiang H, Orr A, Guernsey DL, Robitaille J, Asselin G, Samuels ME, Dubé MP (2009) Application of homozygosity haplotype analysis to genetic mapping with high-density SNP genotype data. PLoS One 4:e5280

Kidd JM, Cooper GM, Donahue WF et al (2008) Mapping and sequencing of structural variation from eight human genomes. Nature 453:56–64

Kim JI, Ju YS, Park H et al (2009) A highly annotated whole genome sequence of a Korean individual. Nature 460:1011–1015

Kimura T, Kobayashi T, Munkhbat B, Oyungerel G, Bilegtsaikhan T, Anar D, Jambaldorj J, Munkhsaikhan S, Munkhtuvshin N, Hayashi H, Oka A, Inoue I, Inoko H (2008) Genome-wide association analysis with selective genotyping identifies candidate loci for adult height at 8q21.13 and 15q22.33–q23 in Mongolians. Hum Genet 123:655–660

Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, Taillon BE, Chen Z, Tanzer A, Saunders AC, Chi J, Yang F, Carter NP, Hurles ME, Weissman SM, Harkins TT, Gerstein MB, Egholm M, Snyder M (2007) Paired-end mapping reveals extensive structural variation in the human genome. Science 318:420–426

Ku CS, Loy EY, Salim A, Pawitan Y, Chia KS (2010a) The discovery of human genetic variations and their use as disease markers: past, present and future. J Hum Genet 55:403–415

Ku CS, Pawitan Y, Sim X, Ong RT, Seielstad M, Lee EJ, Teo YY, Chia KS, Salim A (2010b) Genomic copy number variations in three Southeast Asian populations. Hum Mutat 31:851–857

Lapunzina P, Aglan M, Temtamy S, Caparrós-Martín JA, Valencia M, Letón R, Martínez-Glez V, Elhossini R, Amr K, Vilaboa N, Ruiz-Perez VL (2010) Identification of a frameshift mutation in Osterix in a patient with recessive osteogenesis imperfecta. Am J Hum Genet 87:110–114

Lencz T, Lambert C, DeRosse P, Burdick KE, Morgan TV, Kane JM, Kucherlapati R, Malhotra AK (2007) Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. Proc Natl Acad Sci USA 104:19942–19947

Li LH, Ho SF, Chen CH, Wei CY, Wong WC, Li LY, Hung SI, Chung WH, Pan WH, Lee MT, Tsai FJ, Chang CF, Wu JY, Chen YT (2006) Long contiguous stretches of homozygosity in the human genome. Hum Mutat 27:1115–1121

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM (2009) Finding the missing heritability of complex diseases. Nature 461:747–753

Mardis ER (2008) Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet 9:387–402

McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A, Shapero MH, de Bakker PI, Maller JB, Kirby A, Elliott AL, Parkin M, Hubbell E, Webster T, Mei R, Veitch J, Collins PJ, Handsaker R, Lincoln S, Nizzari M, Blume J, Jones KW, Rava R, Daly MJ, Gabriel SB, Altshuler D (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat Genet 40:1166–1174

McQuillan R, Leutenegger AL, Abdel-Rahman R, Franklin CS, Pericic M, Barac-Lauc L, Smolej-Narancic N, Janicijevic B, Polasek O, Tenesa A, Macleod AK, Farrington SM, Rudan P, Hayward C, Vitart V, Rudan I, Wild SH, Dunlop MG, Wright AF, Campbell H, Wilson JF (2008) Runs of homozygosity in European populations. Am J Hum Genet 83:359–372

Metzker ML (2010) Sequencing technologies—the next generation. Nat Rev Genet 11:31–46

Miyazawa H, Kato M, Awata T, Kohda M, Iwasa H, Koyama N, Tanaka T, Huqun, Kyo S, Okazaki Y, Hagiwara K (2007) Homozygosity haplotype allows a genomewide search for the autosomal segments shared among patients. Am J Hum Genet 80:1090–1102

Nakamura Y (2009) DNA variations in human and medical genetics: 25 years of my experience. J Hum Genet 54:1–8

Nalls MA, Guerreiro RJ, Simon-Sanchez J, Bras JT, Traynor BJ, Gibbs JR, Launer L, Hardy J, Singleton AB (2009a) Extended tracts of homozygosity identify novel candidate genes associated with late-onset Alzheimer's disease. Neurogenetics 10:183–190

Nalls MA, Simon-Sanchez J, Gibbs JR, Paisan-Ruiz C, Bras JT, Tanaka T, Matarin M, Scholz S, Weitz C, Harris TB, Ferrucci L, Hardy J, Singleton AB (2009b) Measures of autozygosity in decline: globalization, urbanization, and its implications for medical genetics. PLoS Genet 5:e1000415

Nicolas E, Poitelon Y, Chouery E, Salem N, Levy N, Mégarbané A, Delague V (2010) CAMOS, a nonprogressive, autosomal recessive, congenital cerebellar ataxia, is caused by a mutant zinc-finger protein, ZNF592. Eur J Hum Genet 18:1107–1113

Nothnagel M, Lu TT, Kayser M, Krawczak M (2010) Genomic and geographic distribution of SNP-defined runs of homozygosity in Europeans. Hum Mol Genet 19:2927–2935

O'Dushlaine CT, Morris D, Moskvina V, Kirov G, Consortium IS, Gill M, Corvin A, Wilson JF, Cavalleri GL (2010) Population structure and genome-wide patterns of variation in Ireland and Britain. Eur J Hum Genet 18:1248–1254

Pang J, Zhang S, Yang P, Hawkins-Lee B, Zhong J, Zhang Y, Ochoa B, Agundez JA, Voelckel MA, Fisher RB, Gu W, Xiong WC, Mei L, She JX, Wang CY (2010) Loss-of-function mutations in HPSE2 cause the autosomal recessive urofacial syndrome. Am J Hum Genet 86:957–962

Park H, Kim JI, Ju YS, Gokcumen O, Mills RE, Kim S, Lee S, Suh D, Hong D, Kang HP, Yoo YJ, Shin JY, Kim HJ, Yavartanoo M, Chang YW, Ha JS, Chong W, Hwang GR, Darvishi K, Kim H, Yang SJ, Yang KS, Kim H, Hurles ME, Scherer SW, Carter NP, Tyler-Smith C, Lee C, Seo JS (2010a) Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. Nat Genet 42:400–405

Park JH, Wacholder S, Gail MH, Peters U, Jacobs KB, Chanock SJ, Chatterjee N (2010b) Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nat Genet 42:570–575

Peiffer DA, Le JM, Steemers FJ, Chang W, Jenniges T, Garcia F, Haden K, Li J, Shaw CA, Belmont J, Cheung SW, Shen RM, Barker DL, Gunderson KL (2006) High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. Genome Res 16:1136–1148

Perry GH, Ben-Dor A, Tsalenko A, Sampas N, Rodriguez-Revenga L, Tran CW, Scheffer A, Steinfeld I, Tsang P, Yamada NA, Park HS, Kim JI, Seo JS, Yakhini Z, Laderman S, Bruhn L, Lee C (2008) The fine-scale and complex architecture of human copy-number variation. Am J Hum Genet 82:685–695

Polasek O, Hayward C, Bellenguez C, Vitart V, Kolcić I, McQuillan R, Saftić V, Gyllensten U, Wilson JF, Rudan I, Wright AF, Campbell H, Leutenegger AL (2010) Comparative assessment of methods for estimating individual genome-wide homozygosity-by-descent from human genomic data. BMC Genomics 11:139

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC (2007) PLINK: a toolset for whole genome association and population based linkage analyses. Am J Hum Genet 81:559–575

Ragoussis J (2009) Genotyping technologies for genetic research. Annu Rev Genomics Hum Genet 10:117–133

Rudan I, Rudan D, Campbell H, Carothers A, Wright A, Smolej-Narancic N, Janicijevic B, Jin L, Chakraborty R, Deka R, Rudan P (2003a) Inbreeding and risk of late onset complex disease. J Med Genet 40:925–932

Rudan I, Smolej-Narancic N, Campbell H, Carothers A, Wright A, Janicijevic B, Rudan P (2003b) Inbreeding and the genetic complexity of human hypertension. Genetics 163:1011–1021

Rudan I, Campbell H, Carothers AD, Hastie ND, Wright AF (2006) Contribution of consanguinuity to polygenic and multifactorial diseases. Nat Genet 38:1224–1225

Schork NJ, Murray SS, Frazer KA, Topol EJ (2009) Common vs rare allele hypotheses for complex diseases. Curr Opin Genet Dev 19:212–219

Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Månér S, Massa H, Walker M, Chi M, Navin N, Lucito R, Healy J, Hicks J, Ye K, Reiner A, Gilliam TC, Trask B, Patterson N, Zetterberg A, Wigler M (2004) Large-scale copy number polymorphism in the human genome. Science 305:525–528

Seelow D, Schuelke M, Hildebrandt F, Nürnberg P (2009) HomozygosityMapper—an interactive approach to homozygosity mapping. Nucleic Acids Res 37:593–599

Simon-Sanchez J, Scholz S, Fung HC, Matarin M, Hernandez D, Gibbs JR, Britton A, de Vrieze FW, Peckham E, Gwinn-Hardy K, Craw-

ley A, Keen JC, Nash J, Borgaonkar D, Hardy J, Singleton A (2007) Genome-wide SNP assay reveals structural genomic variation, extended homozygosity and cell-line induced alterations in normal individuals. Hum Mol Genet 16:1–14

Spain SL, Cazier JB, CORGI Consortium, Houlston R, Carvajal-Carmona L, Tomlinson I (2009) Colorectal cancer risk is not associated with increased levels of homozygosity in a population from the United Kingdom. Cancer Res 69:7422–7429

Stankiewicz P, Lupski JR (2010) Structural variation in the human genome and its role in disease. Annu Rev Med 61:437–455

Teo YY, Sim X, Ong RT, Tan AK, Chen J, Tantoso E, Small KS, Ku CS, Lee EJ, Seielstad M, Chia KS (2009) Singapore Genome Variation Project: a haplotype map of three Southeast Asian populations. Genome Res 19:2154–2162

Ting JC, Roberson ED, Miller ND, Lysholm-Bernacchi A, Stephan DA, Capone GT, Ruczinski I, Thomas GH, Pevsner J (2007) Visualization of uniparental inheritance, Mendelian inconsistencies, deletions, and parent of origin effects in single nucleotide polymorphism trio data with SNPtrio. Hum Mutat 28:1225–1235

Uz E, Alanay Y, Aktas D, Vargel I, Gucer S, Tuncbilek G, von Eggeling F, Yilmaz E, Deren O, Posorski N, Ozdag H, Liehr T, Balci S, Alikasifoglu M, Wollnik B, Akarsu NA (2010) Disruption of ALX1 causes extreme microphthalmia and severe facial clefting: expanding the spectrum of autosomal-recessive ALX-related frontonasal dysplasia. Am J Hum Genet 86:789–796

Van Buggenhout G, Fryns JP (2009) Angelman syndrome (AS, MIM 105830). Eur J Hum Genet 17:1367–1373

Vine AE, McQuillin A, Bass NJ, Pereira A, Kandaswamy R, Robinson M, Lawrence J, Anjorin A, Sklar P, Gurling HM, Curtis D (2009) No evidence for excess runs of homozygosity in bipolar disorder. Psychiatr Genet 19:165–170

Wain LV, Armour JA, Tobin MD (2009) Genomic copy number variation, human health, and disease. Lancet 374:340–350

Walsh T, Shahin H, Elkan-Miller T, Lee MK, Thornton AM, Roeb W, Abu Rayyan A, Loulus S, Avraham KB, King MC, Kanaan M (2010) Whole exome sequencing and homozygosity mapping identify mutation in the cell polarity protein GPSM2 as the cause of nonsyndromic hearing loss DFNB82. Am J Hum Genet 87:90–94

Wang WY, Barratt BJ, Clayton DG, Todd JA (2005) Genome-wide association studies: theoretical and practical concerns. Nat Rev Genet 6:109–118

Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, Hakonarson H, Bucan M (2007) 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

Wang J, Wang W, Li R et al (2008) The diploid genome sequence of an Asian individual. Nature 456:60–65

Wang S, Haynes C, Barany F, Ott J (2009) Genome-wide autozygosity mapping in human populations. Genet Epidemiol 33:172–180

Wellcome Trust Case Control Consortium, Craddock N, Hurles ME et al (2010) Genome-wide association study of CNVs in 16, 000 cases of eight common diseases and 3, 000 shared controls. Nature 464:713–720

Wheeler DA, Srinivasan M, Egholm M et al (2008) The complete genome of an individual by massively parallel DNA sequencing. Nature 452:872–876

Woods CG, Cox J, Springell K, Hampshire DJ, Mohamed MD, McKibbin M, Stern R, Raymond FL, Sandford R, Malik Sharif S, Karbani G, Ahmed M, Bond J, Clayton D, Inglehearn CF (2006) Quantification of homozygosity in consanguineous individuals with autosomal recessive disease. Am J Hum Genet 78:889–896

Xu J, Bleecker ER, Jongepier H, Howard TD, Koppelman GH, Postma DS, Meyers DA (2002) Major recessive gene(s) with considerable residual polygenic effect regulating adult height: confirmation of genomewide scan results for chromosomes 6, 9, and 12. Am J Hum Genet 71:646–650

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, Goddard ME, Visscher PM (2010a) Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42:565–569

Yang TL, Guo Y, Zhang LS, Tian Q, Yan H, Papasian CJ, Recker RR, Deng HW (2010b) Runs of homozygosity identify a recessive locus 12q21.31 for human adult height. J Clin Endocrinol Metab 95:3777–3782

Yim SH, Kim TM, Hu HJ, Kim JH, Kim BJ, Lee JY, Han BG, Shin SH, Jung SH, Chung YJ (2010) Copy number variations in East-Asian population and their evolutionary and functional implications. Hum Mol Genet 19:1001–1008

Yoon S, Xuan Z, Makarov V, Ye K, Sebat J (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res 19:1586–1592