

# Detection of Runs of Homozygosity from Whole Exome Sequencing Data: State of the Art and Perspectives for Clinical, Population and Epidemiological Studies

Tommaso Pippucci<sup>a, b</sup> Alberto Magi<sup>c</sup> Alessandro Gialluisi<sup>d</sup>  
Giovanni Romeo<sup>a, b</sup>

<sup>a</sup>Medical Genetics Unit, Polyclinic Sant'Orsola-Malpighi and <sup>b</sup>Department of Medical and Surgical Sciences, University of Bologna, Bologna, and <sup>c</sup>Department of Experimental and Clinical Medicine, University of Florence, Florence, Italy; <sup>d</sup>Language and Genetics Department, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

## Key Words

Exome sequencing · Runs of homozygosity · Homozygosity mapping

## Abstract

Runs of homozygosity (ROH) are sizeable stretches of homozygous genotypes at consecutive polymorphic DNA marker positions, traditionally captured by means of genome-wide single nucleotide polymorphism (SNP) genotyping. With the advent of next-generation sequencing (NGS) technologies, a number of methods initially devised for the analysis of SNP array data (those based on sliding-window algorithms such as PLINK or GERMLINE and graphical tools like HomozygosityMapper) or specifically conceived for NGS data have been adopted for the detection of ROH from whole exome sequencing (WES) data. In the latter group, algorithms for both graphical representation (AgileVariantMapper, HomSI) and computational detection ( $H^3M^2$ ) of WES-derived ROH have been proposed. Here we examine these different approaches and discuss available strategies to implement ROH detection in WES analysis. Among sliding-window algorithms,

PLINK appears to be well-suited for the detection of ROH, especially of the long ones. As a method specifically tailored for WES data,  $H^3M^2$  outperforms existing algorithms especially on short and medium ROH. We conclude that, notwithstanding the irregular distribution of exons, WES data can be used with some approximation for unbiased genome-wide analysis of ROH features, with promising applications to homozygosity mapping of disease genes, comparative analysis of populations and epidemiological studies based on consanguinity.

© 2014 S. Karger AG, Basel

## Introduction

In a single diploid genome, runs of homozygosity (ROH) can be defined as sizeable stretches of homozygous genotypes at consecutive polymorphic DNA marker positions [1]. Several ROH definitions have been proposed to date, which vary from the minimal genomic/genetic length (measured in base pairs, units of genetic recombination or cM, respectively) to the minimal num-

ber of markers required to label a given region of the genome as ROH [2–6]. In general, consensus criteria for defining and classifying ROH are lacking, leading to discrepancies between the approaches used and consequently to difficulties in the interpretation and comparison among different studies [7].

At least in part this variability of definitions reflects the technical constraints that influence the resolution of the genetic map used, and consequently the reliability of the identified regions. When the most densely spaced polymorphic markers available were microsatellites, the map density of typical genome-wide scans was approximately 1 marker every 10 cM, allowing the detection of only very long homozygous segments (40 cM) [8]. With the advent of denser genome-wide single nucleotide polymorphism (SNP) array platforms, more than 1 million SNP per individual genome can be simultaneously genotyped providing average map densities of 1 SNP every 3 kb or less. It follows that there is now virtually no limit to the minimal ROH size that can be theoretically detected.

However, if, on one hand, the availability of highly informative SNP maps has enormously increased the sensitivity in mapping ROH of every size, on the other hand, it has introduced a number of factors that potentially affect the specificity with which genomic regions are correctly recognized as ROH. First, genotyping errors of SNP array platforms may introduce spurious heterozygous genotypes in an otherwise continuous stretch of homozygous SNP, causing the ROH to be incorrectly broken. The error rates of the most widely adopted genotyping platforms are usually around 0.1% [9], and such events should therefore be taken into account in the ROH definition. One can refer to Howrigan et al. [7] and Ferencakovic et al. [10] for a deeper treatment of this matter. Second, although the SNP included in genotyping platforms are in principle evenly distributed all along the genome, it is always possible that poorly represented chromosomal regions occur for which the homozygous/heterozygous state cannot be accurately inferred. This may lead to the inclusion of DNA segments in ROH that would otherwise be classified as heterozygous if investigated over a larger number of polymorphic loci. Finally, the tremendous increment in the number of interrogated markers, and consequently in the resolution of the available platforms, magnifies the inadequacy of a ROH definition that is expressed solely in terms of size and/or SNP numbers.

With the advent of next-generation sequencing (NGS) technologies, the map of the human genome is virtually complete. All accessible bases can now be genotyped and

millions of polymorphisms, from the most common to the most private ones, can be obtained for each individual genome. This has a predictable impact on the extent to which homozygosity can be identified, providing the highest possible map resolution and marker information. Nonetheless, the challenges in the analysis of high density SNP arrays may even increase in this novel context.

The higher the SNP density, the higher the probability to detect stretches of many consecutive SNP that occur as homozygous due to linkage disequilibrium (LD). Strong LD typically extends up to a few hundred kb. Consequently, especially short ROH are distributed throughout the genome and occur commonly in individuals of all populations. Different approaches have been proposed for accounting for LD, which have been adopted in ROH studies: imposing a cutoff for the length of ROH at 500 kb [11]; weighting individual ROH by the internal level of LD approximated by the squared genotypic correlation coefficient  $r^2$  [1]. Conversely, Pemberton et al. [12] applied the autozygosity LOD score method proposed by Broman and Weber [8] and later by Wang et al. [3]. However, it is important to notice that LD properties across the whole exome sequence (WES) target are poorly studied (at least to the best of our knowledge) and may not reflect those known for the genome.

Autozygosity is that particular type of homozygosity that reflects ‘identity by descent’ (IBD) and results from the co-occurrence at a given locus of the same allele deriving from a common ancestor, by way of nonrandom mating (inbreeding). When the two alleles derive from non-common ancestors (random mating), the genotype is said to be allozygous and the two alleles are ‘identical by state’. Autozygosity is a crucial concept in homozygosity mapping [13]. Pemberton et al. [12] noticed that, despite their small size, short ROH can reflect autozygosity.

The above-mentioned analytical advancements introduced by NGS technologies improve the detection of small ROH, paving the way to the effective adoption of homozygosity mapping in contexts characterized by low inbreeding levels. The application of methods such as the autozygosity LOD score enhance the ability to discriminate whether the identified ROH are likely disease-associated or not [12], and it can be generally applied as a measure of significance of the identified ROH.

To date, a number of methods initially devised for the analysis of SNP array data or specifically conceived for NGS data have been adopted for the detection of ROH from WES data. The former ones are reknown algorithms for SNP array genotyping data analysis, such as PLINK [14] or GERMLINE [15], while the latter group includes

original methods that exploit B allele frequency (BAF) or NGS genotype data measures to infer homozygosity from WES-aligned reads. In a NGS alignment output, BAF can be calculated as the ratio between reads carrying the B allele and total reads at a given polymorphic position. Two of these approaches [16, 17] basically provide graphical outputs that highlight regions of extended WES homozygosity along the chromosomes. A third approach presents an original algorithm that relies on a heterogeneous hidden Markov model (HMM) to incorporate SNP-to-SNP distances in the detection of ROH of every genomic size [18].

The purpose of this review is to examine all the computational approaches tested to date for the ROH detection from WES data, discuss the various approaches proposed to implement this strategy and outline the future perspectives they open for homozygosity mapping of disease genes, population analysis and epidemiological studies.

### Approaches for the Graphical Visualization of ROH

An initial demonstration that homozygosity information could be retrieved directly from WES data was given by Becker et al. [16] examining individuals affected with autosomal recessive osteogenesis imperfecta (OMIM 613982). They calculated the BAF of hundreds of thousands of dbSNP (<http://ncbi.nlm.nih.gov/SNP>) positions scattered across the WES-targeted regions, and plotted the obtained ratios against the chromosomal coordinates. Without performing SNP arrays, they were able to isolate the regions corresponding to extended ROH by visualizing chromosomal BAF profiles. In an affected individual, one of these regions harbored a homozygous truncating mutation of *SERPINF1* (OMIM 172860), which was subsequently demonstrated to be the cause of the disorder.

The web-based tool HomozygosityMapper, originally tailored for an assortment of SNP array platforms, has been extended to integrate variant call format files (the standard file format for storing NGS genotype variant calls) as input, and bed extensible data files (the standard file format for storing genomic features by coordinates) as output containing placements of the identified ROH [19]. HomozygosityMapper offers an online intuitive graphical interface allowing users to interactively analyze NGS data for homozygosity mapping in humans and a number of species.

Two downloadable softwares have been designed for the graphical representation of autozygous regions de-

tectable from WES data. Following the assumption that WES data can simultaneously catalogue the autozygous intervals and all candidate deleterious variants within these intervals, Carr et al. [17] developed a suite of two computer programs, namely AgileGenotyper and AgileVariantMapper. These two programs differ in the two alternative ways by which they identify autozygosity: AgileVariantMapper uses genotype call data of only variant positions originally detected by WES variant calling analysis, whereas AgileGenotyper deduces genotypes from BAF measure at >0.5 million a priori fixed exonic positions corresponding to polymorphisms previously discovered by the 1000 Genomes Project (<http://www.1000genomesproject.org>).

With the former approach, some tens of thousands of variant genotypes will be included in the analysis, with a predictable high proportion of false positive heterozygous genotypes; with the latter, the risk of false positives is lower but for the vast majority (approximately 95%) of the tested polymorphic positions, the minor allele frequency will be very low. This is expected, since these markers were identified by the 1000 Genomes Project, where SNPs can be rare variants found in few individuals or even private ones observed only once. There are about 4 million 1000 Genomes Project SNPs within or surrounding coding regions. On one hand, the inclusion of these SNPs is useful to enhance map density at an unprecedented resolution, but on the other hand, it increases the probability to identify stretches of nearly unvariable loci, at which all individuals can display homozygous reference genotypes. The identification of ROH formed by these stretches of consecutive homozygous reference alleles can be uninformative.

By comparing the graphical outputs of each of the two programs with that obtained by SNP array genotype analysis on a small set of offspring of consanguineous unions affected with rare recessive disorders, Carr et al. [17] noticed that regions derived from WES data (with any of the two different approaches) generally result noticeably more fragmented and consequently less clearly delineated with respect to those obtained with SNP microarray data. This situation occurs regularly with very small ROH, but it is a common feature also of substantially long (several Mb) ROH that surround the mutated gene. The authors concluded that autozygosity detection relying solely on WES data can replace SNP array data but can require a massive use of subjective data interpretation.

Recently, another graphical software, namely HomSI [20] has been presented, exploiting variant call format files to identify ROH with a sliding-window approach.

Similarly to AgileVariantMapper, it has been conceived for the concurrent identification of ROH and mutations in consanguineous families.

### Sliding-Window Algorithms

Methods such as PLINK or GERMLINE have originally been developed for the analysis of SNP array data. These are known as sliding-window algorithms, since they scan the chromosomes by moving a window of a fixed size along their whole length in search of stretches of consecutive homozygous SNP.

The PLINK software package [21] implements such a sliding-window algorithm that can be invoked with the option `-homozyg` and that allows the user to set the desired size (in terms of SNP number or physical genomic length) for the window that slides along the chromosome. The window moves forward from the 5' to the 3' extremity of a chromosome on a SNP-per-SNP basis; at each given SNP, the proportion of overlapping windows that appear homozygous is calculated in order to classify the SNP as being in a homozygous segment or not. If this proportion is higher than a defined threshold (default value: 0.05), the SNP is designated as being in a homozygous segment. A variable number of heterozygous (default value = 1) or missing (default value = 5) SNP per window can be specified by the user in order to tolerate genotyping errors and failures. Subsequently, PLINK calls ROH if: (a) the amount of consecutive SNP in a homozygous segment exceeds a predefined threshold in terms of SNP number (default value = 100) and/or covered chromosomal length (default value = 1,000 kb); (b) the SNP density is higher than a user-specified threshold (default value = 1 every 50 kb), or (c) the SNP-to-SNP distance is never greater than a user-specified threshold (default value = 1,000 kb).

GERMLINE [22] is a tool mainly designed for IBD calculation. For ROH detection, it adopts a sliding-window algorithm that can be invoked with the option `-homoz` or `-homoz-only` and that, like PLINK, is flexible with respect to several parameters such as window size, minimum length of the ROH and tolerance for heterozygous mismatches. Unlike PLINK, GERMLINE breaks up SNP stretches into non-overlapping windows of a user-specified length in terms of SNP, and only if several consecutive windows tagged as homozygous exceed a threshold in terms of physical or genetic distance, the region is labeled as homozygous.

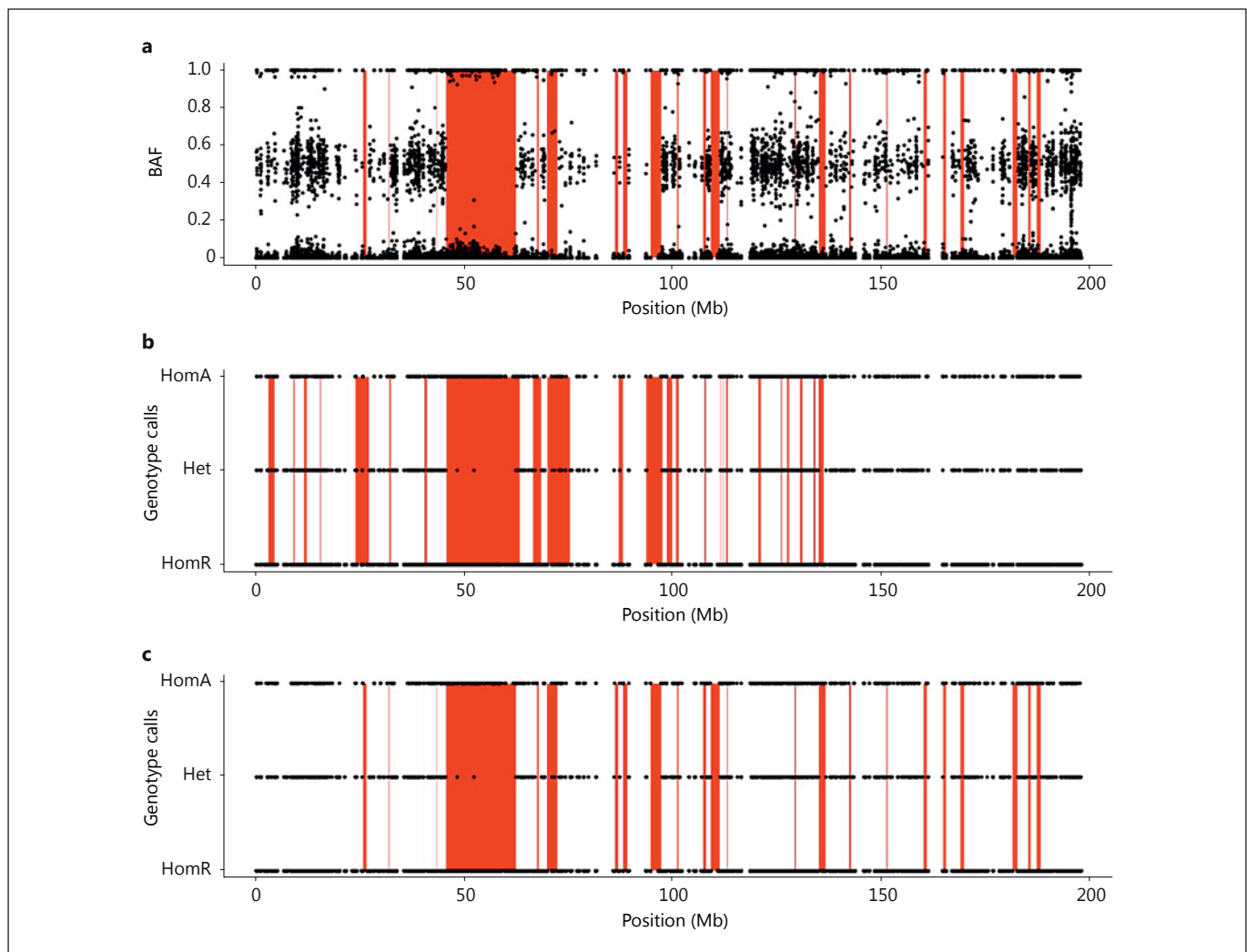
In a proof of principle paper, Pippucci et al. [14] introduced the use of sliding-window algorithms for WES-based ROH detection by applying PLINK to the WES data of two siblings born to first-cousin parents who were affected with dysmyelinating leukodystrophy and spastic paraparesis (OMIM 612319) caused by an already known splice-site homozygous mutation in FA2H (OMIM 611026) [23]. They named their approach EX-HOM (EXome-HOMOzygosity) – an approach that made use of all exonic or perieyonic autosomal dbSNP sites and of all novel SNV regions retrieved in at least one of the two siblings to create a genetic map consisting of more than 100,000 markers. By adapting PLINK parameters to the map in use, they sought ROH shared by the two siblings. EX-HOM regions >1 Mb in length showed a substantial overlap with those identified as LOD score peaks by linkage analysis, leading to the conclusion that the EX-HOM approach can correctly identify disease-related long homozygous regions.

Based on a similar assumption, Zhuang et al. [15] compared the performances of GERMLINE in the detection of IBD regions from WES and high-density SNP array data of a dataset of Ashkenazi Jewish individuals, in order to empirically measure the accuracy of this algorithm on WES data. They reported a poor overlap of IBD discovered by applying GERMLINE to genotyping data of the same dataset, even when focusing on IBD segments as extended as tens of Mb. Due to this lack of concordance between WES and SNP array-based analyses, Zhuang et al. [15] argued that WES data are of limited utility for IBD detection and homozygosity mapping. They ascribed these limitations mainly to the scattered distribution of WES data that generate stretches of irregularly spaced SNP, and commented on the need to adapt detection algorithms to the density variations and decreased allelic diversity typical of WES data.

### HMM Algorithms

State-of-the-art ROH detection methods have been conceived to be used with equally spaced SNP maps. The SNP distribution in exome targets is much more irregular and sparse. This represents an intrinsic limitation to the application of such approaches to WES data, particularly in the detection of medium and short ROH. To meet the requirement for a computational approach that could more efficiently identify ROH of any size in WES, Magi et al. [18] developed a method exploiting a heterogeneous HMM. Such an algorithm, named  $H^3M^2$  (homozygosity





**Fig. 1.** Chromosomal ROH detection by  $H^3M^2$  (a), GERMLINE (b) and PLINK (c). Distribution of BAF and genotype calls along physical coordinates of chromosome 3. At each interrogated marker locus, BAF is expressed as the ratio between the number of reads carrying the alternate allele and the number of total reads covering that site. The corresponding genotype call is expressed as

0 (homozygous reference), 0.5 (heterozygous) or 1 (homozygous alternate). Of note, the largest ROH, spanning about 18 Mb, is successfully identified by any of the 3 methods. The overall view of chromosome 3 shows the highest overlap between ROH detected by  $H^3M^2$  and PLINK.

heterogeneous HMM), takes BAF measures over all 1000 Genomes Project exome target variant positions rather than genotype calls as input.

The originality of this algorithm consists in its capability to incorporate distances between consecutive SNPs to discriminate between the homozygosity and the heterozygosity states. This feature of  $H^3M^2$  offers advantage in the detection of small-sized (<500 kb) and medium-sized (between 500 and 1,500 kb) ROH [18]. Differently,  $H^3M^2$  and PLINK show similar performances in the WES-based detection of long ROH (>1,500 kb).

As AgileGenotyper,  $H^3M^2$  exploits all 1000 Genomes Project SNPs in the targeted exome, thus possibly incurring in the identification of uninformative ROH. In addition, SNPs in the targeted exome can be extremely close to each other, and consequently LD can be extremely high. Similarly to other ROH detection methods, at present,  $H^3M^2$  does not account for LD, nor does it weight detected ROH according to their haplotype frequency. The implementation of such measures in the method would be particularly important for ROH detection in the context of WES data.

Other methods based on HMMs exist that are designed to detect autozygosity of IBD regions: IBDSeq [24] and GIBDLD [25]. The former has been conceived for NGS whole genome data, the latter for SNP arrays, but neither of the two has been tested on WES data yet.

## Homozygosity Mapping

The results so far summarized emphasize that homozygosity mapping of rare autosomal recessive disorders can be conducted with no need of SNP array data. A number of available tools are able to simultaneously perform homozygosity mapping and the identification of candidate mutations using WES as a single source of data, as previously discussed [14, 16]. Figure 1 shows that all applied method (PLINK, GERMLINE,  $H^3M^2$ ) successfully identified a 18-Mb autozygous region harboring the CACNA2D2 (OMIM 607082) homozygous mutation responsible of epileptic encephalopathy in a single offspring to consanguineous parents [26].

In general, exceptionally large regions, such as those commonly observed in the offspring of consanguineous unions, are easily identified by any of the available algorithms. The methods, however, can greatly differ in the overall accuracy. Figure 1 shows the higher concordance between  $H^3M^2$  and PLINK compared to GERMLINE. The choice of the more accurate approaches in the detection of small and medium ROH is particularly relevant when the expected size of the disease-related ROH is smaller than that resulting from consanguineous unions [27]. The use of such approaches open promising perspectives for WES-based homozygosity mapping also in non-inbred individuals.

## Population Analysis

Up to now, an analysis of WES data has never been used in the study of ROH in populations. We evaluated the potentiality of WES data to explore genomic patterns of homozygosity in human populations by applying  $H^3M^2$  to a dataset of samples from 100 individuals of 5 populations sequenced by the 1000 Genomes Project (see Appendix).

Following the model established by Pemberton et al. [12], we analyzed the ROH sizes as a mixture of three normal distributions representing three distinct ROH classes: class A (<500 kb), class B (between 500 and 1,500 kb) and class C (>1,500 kb).

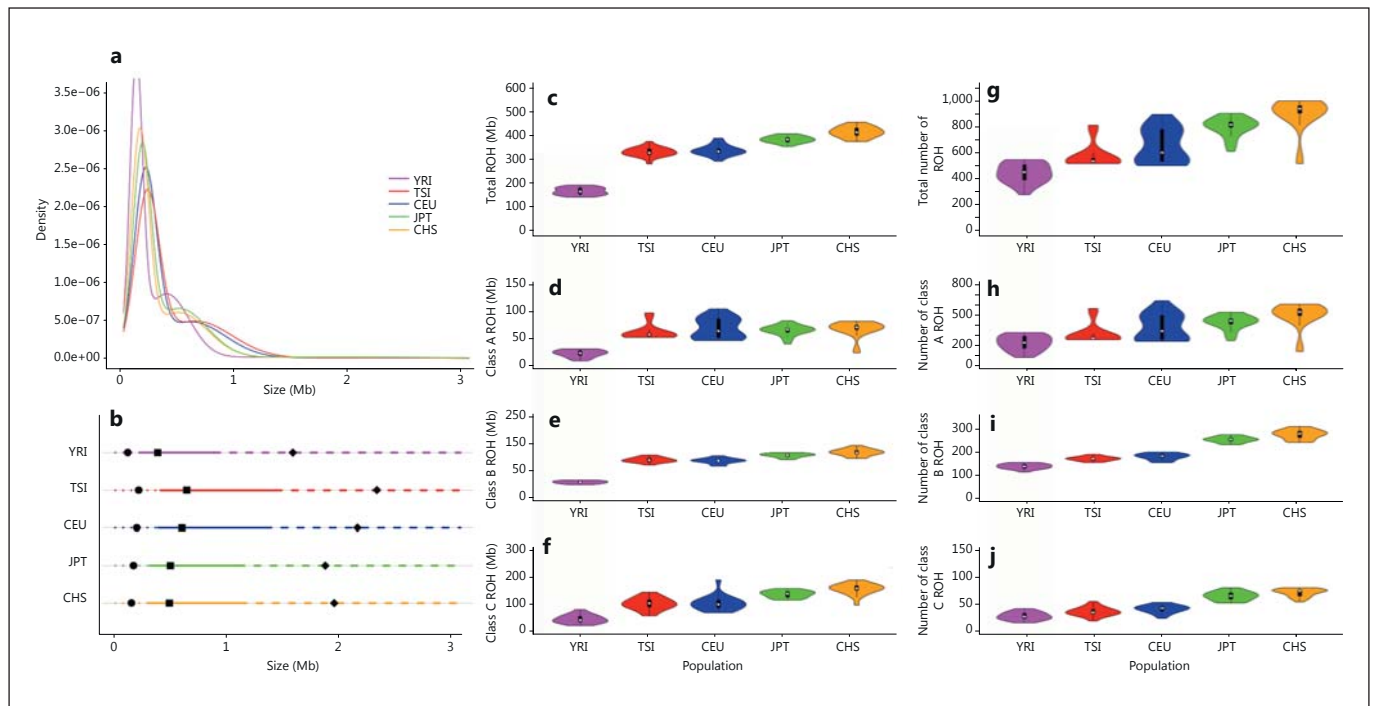
**Table 1.** Summary statistics for the WES data generated by the 1000 Genomes Project consortium

	Ethnic group				
	YRI	TSI	CEU	JPT	CHS
<i>Mean lengths of homozygous segments</i>					
Class A, kb	118	233	208	179	163
Class B, kb	407	731	661	561	571
Class C, Mb	1.67	2.58	2.3	2.0	2.1
<i>Boundary sizes between</i>					
Class A and B, kb	226	450	407	346	334
Class B and C, Mb	0.94	1.6	1.47	1.23	1.28
<i>Length of homozygous segments for each population, Mb</i>					
Total	203	418	415	435	481
Class A	26.6	73.5	75.7	78.8	79.3
Class B	81.2	171.8	168.8	174.4	192.3
Class C	95.4	172.8	170.4	182.6	209.4
<i>Number of homozygous segments for each population</i>					
Total	458	586	658	800	871
Class A	231	321	370	447	491
Class B	178	208	226	275	297
Class C	50	57	63	78	84

The results of the clustering analysis reported in figure 2a, b show that the mean of each class and the boundaries between the different classes vary across the 5 populations. Total lengths and total numbers of ROH per individual across the 5 populations are represented as violin plots in figure 2c–j. Several patterns emerge from the comparison of the total lengths and total numbers of ROH per individual across the populations (see table 1).

First, the total length of ROH increases with the distance from Africa (fig. 2c). Second, also the total length of each of the three classes increases with the distance from Africa and in all populations: class A contributes 15% of the total amount of ROH, class B 40% and class C 45% (fig. 2d–f). Third, the total numbers of ROH per individual (fig. 2g) show similar patterns to those observed for total lengths. For each of the populations, class A accounts for 55% of the total numbers of ROH, class B for 35% and class C for 10% (fig. 2h–j).

These results are in general agreement with previously published work that examined SNP array-based homozygosity at a population level [5, 12]. However, the total length and total number of class C ROH are surprisingly high for outbred populations. These values seem to be inflated, likely reflecting an excess of regions characterized by low marker density. These regions can be present in an exome-targeted map, where they correspond to re-



**Fig. 2.** Genomic patterns of homozygosity in 5 human populations. **a** Gaussian kernel density estimates of the ROH size distribution in each of the 5 populations (YRI, TSI, CEU, JPT and CHS). **b** Mean of each ROH class (A, B and C) and the boundaries between classes A and B and classes B and C for each of the 5 populations. Dotted lines and ● define the limits and mean size of class A (small) ROH; continuous lines and ■ define the limits and mean size of class B (medium) ROH; dashed lines and ◆ define the lim-

its and mean size of class C (large) ROH. **c–f** Distribution of total ROH lengths over all individuals in each of the 5 populations, for all three classes combined (**c**), for class A (**d**), for class B (**e**) and for class C (**f**). **g–j** Distribution of the numbers of ROH over all individuals in each of the 5 populations, for all three classes combined (**g**), for class A (**h**), for class B (**i**) and for class C (**j**). **c–j** Data are shown as violin plots: each violin contains a vertical black line (25–75% range) and a horizontal white line (median).

gions void of genes (exons). Probably, these regions would be otherwise identified as heterozygous if interrogated over an adequate number of polymorphic positions. This observation highlights how the sensitivity of  $H^3M^2$  leads to the identification of regions with few sparse SNPs, which are likely spurious ROH calls. Therefore, further research is required to assess the full applicability of WES-based methods such as  $H^3M^2$  to population data.

### Potential Impact on the Estimation of the Inbreeding Coefficient $F$

In addition to the usefulness of WES-derived ROH in medical and population genetics, a comprehensive catalogue of individual and population genomic patterns of homozygosity may be helpful in the field of genetic epidemiology, especially of rare recessive monogenic disor-

ders. Here, we limit the discussion of these possible applications to the contribution that WES data can give to the calculation of the homozygosity index (HI). Originally proposed by Ten Kate et al. [28] and by Gialluisi et al. [29], HI was to infer allelic frequencies ( $q$ ) of autosomal recessive disorders based on the relative proportion of homozygous patients in a sample population, the inbreeding coefficient ( $F$ ) and the mutational spectrum of even a relatively small sample ( $n = 25$ ) of affected individuals [29, 30].

Intuitively, for any autosomal recessive disorders, the ratio of homozygotes versus compound heterozygotes for any given disease allele reflects the frequency of that disorder in the population. One can roughly say that the higher the proportion of homozygotes, the rarer the disorder in a given population. This method (see Appendix) has been demonstrated to accurately estimate  $q$ s for phenylketonuria, familial Mediterranean fever [29] and Wil-

**Table 2.** Summary of the tools for homozygosity mapping

Tool	Algorithm	Input data	Accurately detected ROH	Output
PLINK	sliding window	genotype calls	B and C classes	PLINK output file
GERMLINE	sliding window	genotype calls	C class	GERMLINE output file
AgileGenotyper	frequentist genotype assignment	B allele frequency	C class	colour-based visualization
$H^3M^2$	HMM	B allele frequency	A, B, C classes	bed extensible data file
HomozygosityMapper	sliding blocks	genotype calls	C class	colour-based visualization (and bed extensible data file)
HomSI	sliding window	genotype calls	C class	colour-based visualization

Class A = ROH <500 kb; class B = 500 kb < ROH < 1.5 Mb; class C = ROH >1.5 Mb.

son disease [30] in highly consanguineous or endogamous populations. This approach has the advantage over traditional descriptive epidemiology to generate estimates that are not affected by biases related to underdiagnosis and incomplete epidemiological records of vast and remote areas [29, 30]. Moreover the HI approach does not require the collection of additional mutation data from the general control population but relies only on the analysis of mutations identified in a small group of patients.

Difficulties in obtaining accurate  $F$  estimates represent the major limitation for the correct estimates of  $q$  by this approach. This is particularly problematic when analyzing samples of affected individuals born to apparently unrelated parents, for which the estimates of  $F$ , usually based on demographic data, are often unreliable and outdated. Even for samples of patients born to consanguineous matings, classical estimates based on pedigree reconstruction seem to underestimate the factual levels of inbreeding [31]. Therefore, estimates based on genomic data are generally better to retrieve reliable  $F$  values. To this end several experimental methods have been proposed to accurately estimate  $F$  from the interrogation of hundreds of thousands or millions of polymorphisms that can be obtained with SNP array platforms. The majority of these methods compute  $F$  relative to the proportion of ROH longer than a given minimal threshold in the individual genome [11, 32]. A statistically based estimation of  $F$  from lengths of ROH exploiting a maximum-likelihood approach has also been formulated [33].

With WES data, a likewise large number of polymorphisms can be interrogated, covering a much broader spectrum of allele frequencies. The opportunity to access alleles rarer than those offered by SNP array platforms,

including those alleles that are represented only or mainly in the examined population(s), may be beneficial to the more accurate discrimination between autozygous and simply homozygous regions, in turn leading to the more accurate calculation of  $F$  based on genomic estimates. An additional benefit that can be contributed by WES data is a higher sensitivity in detecting rare disease-related alleles compared to diagnostic screening panels, thus increasing the accuracy in the calculation of HI [29]. Further research is needed to evaluate which of the presently available methods for ROH detection based on WES data are more suitable to the scope of obtaining an accurate  $F$  estimate. Implementation of such an approach along with the classical application of WES to variant calling could globally improve performance of the HI method.

In conclusion, due to the central role of consanguineous marriages in a large proportion of the global population [34], the HI method represents a combined approach of molecular and genetic epidemiology based on consanguinity which is useful to assess the need of genetic screening for autosomal recessive disorders and to establish priorities for genetic testing at the population level. At the same time, the collection of data related to WES-based homozygosity and rare polymorphisms in appropriate databases can improve epidemiological estimates based on the HI method.

### Conclusive Remarks

In this review, we tried to outline the state of the art of currently available methods for ROH detection from WES data. A handful of algorithms, initially designed for SNP array data or originally conceived for WES data,



have been proposed to date (see table 2). Sliding-window approaches, exploited by methods designed for SNP arrays, can generally detect long ROH even when applied to WES data, with PLINK ensuring higher sensitivity/specificity trade-offs.

Of the graphical visualization tools, Homozygosity-Mapper has recently been extended to process WES data in addition to SNP array data. AgileGenotyper and Hom-SI are algorithms for the visualization of ROHs that can be useful, easy-to-use and simple tools to obtain a qualitative measure of homozygosity surrounding the locus of interest.  $H^3M^2$  is a HMM algorithm which among the present methods is the most extensively tested for the detection of ROH of any size.

For the purpose of homozygosity mapping, all these methods seem to be suited for the identification of large ROHs, corroborating the idea that homozygosity mapping could be successfully performed without support of SNP array data. The higher accuracy in the detection of short and medium ROHs by  $H^3M^2$  is promising for the application of WES-based homozygosity mapping in outbred individuals.

As a first attempt to exploit WES data in population analysis, we tested  $H^3M^2$  over 100 samples of 5 different human populations, finding general agreement with previously reported trends for patterns of homozygosity. A major concern regarding the application of WES data to the quantitative estimation of genomic homozygosity is the inadequate coverage of exon-devoid regions, which is likely a source of both false positive and false negative calls. Further research is needed to assess the reliability of WES-based measures in this specific field.

Finally, mutations causing autosomal recessive disorders identified by WES or by any other molecular genetic approach are a valuable source of data for genetic epidemiology, as exemplified by the HI method, which offers an inexpensive way of assessing the need of genetic screening for specific autosomal recessive disorders in specific populations.

In conclusion, notwithstanding the sparse and irregular distribution of exons in the genome, WES data can be used with some approximation for unbiased genome-wide analyses of genomic features, as already widely demonstrated for structural variations [35].

Undoubtedly, whole genome sequencing will allow more accurate and regular mapping of every type of genomic variations and features. However, the analysis of whole genome sequencing is still not free of challenges, starting from data interpretation. The continuously decreasing cost and increasing accessibility of WES, along

with its ability to provide a virtually exhaustive catalogue of functionally relevant variations such as coding mutations, have opened the way to the exploitation of WES data for different uses. With regard to ROH, in this review we tried to show what is currently feasible with WES data and what can be improved in the future.

## Appendix: Methods

### Population Analysis

Our sample consisted of 100 individuals from 5 different human populations: 20 YRI (Yoruba from Ibadan, Nigeria), 20 CEU (Utah residents with Northern and Western European ancestry from the CEPH collection), 20 TSI (Tuscans in Italy), 20 CHS (Han Chinese, South China) and 20 JPT (Japanese in Tokyo, Japan).

Boundary sizes between classes A and B and between classes B and C were estimated using the following formulas:

$$CB_{AB}^i = \frac{A_{max}^i + B_{min}^i}{2}$$

$$CB_{BC}^i = \frac{B_{max}^i + C_{min}^i}{2}$$

where  $A_{max}^i$ ,  $B_{min}^i$ ,  $B_{max}^i$ , and  $C_{min}^i$  are the minimum and maximum ROH sizes for the three classes for population  $i$ , respectively. The total lengths and total numbers of ROH per individual across the three classes were calculated, and their distributions for each population were subsequently studied.

### HI Method

The HI approach consists in computing  $q$  of a given autosomal recessive disorder as:

$$q = \frac{F(1 - HI)}{(HI - \sum q_i^2)(1 - F)},$$

where HI is of the subset (i.e. the frequency of homozygotes relative to the total of homozygotes and compound heterozygotes among patients) and  $q_i$  is the relative frequency of the  $i$ -th disease allele (with  $i = 0, 1, 2, \dots, n - 1, n$ ).

## Acknowledgements

Tommaso Pippucci is supported by the Italian Ministry of Health, Young Investigators Award, project GR-2009-1574072.

## References

- Nothnagel M, Lu TT, Kayser M, Krawczak M: Genomic and geographic distribution of SNP-defined runs of homozygosity in Europeans. *Hum Mol Genet* 2010;19:2927–2935.
- Gibson J, Morton NE, Collins A: Extended tracts of homozygosity in outbred human populations. *Hum Mol Genet* 2006;15:789–795.
- Wang S, Haynes C, Barany F, Ott J: Genome-wide autozygosity mapping in human populations. *Genet Epidemiol* 2009;33:172–180.
- Curtis D, Vine AE, Knight J: Study of regions of extended homozygosity provides a powerful method to explore haplotype structure of human populations. *Ann Hum Genet* 2008;72:261–278.
- Auton A, Bryc K, Boyko AR, Lohmueller KE, Novembre J, Reynolds A, Indap A, Wright MH, Degenhardt JD, Gutenkunst RN, King KS, Nelson MR, Bustamante CD: Global distribution of genomic diversity underscores rich complex history of continental human populations. *Genome Res* 2009;19:795–803.
- MacLeod IM, Meuwissen TH, Hayes BJ, Goddard ME: A novel predictor of multilocus haplotype homozygosity: comparison with existing predictors. *Genet Res (Camb)* 2009;91:413–426.
- Howrigan DP, Simonson MA, Keller MC: Detecting autozygosity through runs of homozygosity: a comparison of three autozygosity detection algorithms. *BMC Genomics* 2011;2:460.
- Broman KW, Weber JL: Long homozygous chromosomal segments in reference families from the centre d'étude du polymorphisme humain. *Am J Hum Genet* 1999;65:1493–1500.
- Hong H, Xu L, Liu J, Jones WD, Su Z, Ning B, Perkins R, Ge W, Miclaus K, Zhang L, Park K, Green B, Han T, Fang H, Lambert CG, Vega SC, Lin SM, Jafari N, Czika W, Wolfinger RD, Goodsaid F, Tong W, Shi L: Technical reproducibility of genotyping SNP arrays used in genome-wide association studies. *PLoS One* 2012;7:e44483.
- Ferenkovic M, Solkner J, Curik I: Estimating autozygosity from high-throughput information: effects of SNP density and genotyping errors. *Genet Sel Evol* 2013;45:42.
- 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: Runs of homozygosity in European populations. *Am J Hum Genet* 2008;83:359–372.
- Pemberton TJ, Absher D, Feldman MW, Myers RM, Rosenberg NA, Li JZ: Genomic patterns of homozygosity in worldwide human populations. *Am J Hum Genet* 2012;10:275–292.
- Lander ES, Botstein D: Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science* 1987;236:1567–1570.
- Pippucci T, Benelli M, Magi A, Martelli PL, Magini P, Torricelli F, Casadio R, Seri M, Romeo G: EX-HOM (EXome HOMozygosity): a proof of principle. *Hum Hered* 2011;72:45–53.
- Zhuang Z, Gusev A, Cho J, Pe'er I: Detecting identity by descent and homozygosity mapping in whole-exome sequencing data. *PLoS One* 2012;7:e47618.
- Becker J, Semler O, Gilissen C, Li Y, Bolz HJ, Giunta C, Bergmann C, Rohrbach M, Koerber F, Zimmermann K, de Vries P, Wirth B, Schoenau E, Wollnik B, Veltman JA, Hoischen A, Netzer C: Exome sequencing identifies truncating mutations in human SERPINF1 in autosomal-recessive osteogenesis imperfecta. *Am J Hum Genet* 2011;88:362–371.
- Carr IM, Bhaskar S, O'Sullivan J, Aldahmesh MA, Shamseldin HE, Markham AF, Bonthron DT, Black G, Alkuray FS: Autozygosity mapping with exome sequence data. *Hum Mutat* 2013;34:50–56.
- Magi A, Tattini L, Palombo F, Benelli M, Gialluisi A, Giusti B, Abbate R, Seri M, Gensini GF, Romeo G, Pippucci T:  $H^3M^2$ : detection of runs of homozygosity from whole-exome sequencing data. *Bioinformatics* 2014, DOI: 10.1093/bioinformatics/btu401.
- Seelow D, Schuelke M: HomozygosityMapper2012 – bridging the gap between homozygosity mapping and deep sequencing. *Nucleic Acids Res* 2012;40(Web Server issue):W516–W520.
- Gormez Z, Bakir-Gungor B, Sagioglu MS: HomSI: a homozygous stretch identifier from next-generation sequencing data. *Bioinformatics* 2014;30:445–447.
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC: PLINK: a toolset for whole genome association and population based linkage analyses. *Am J Hum Genet* 2007;81:559–575.
- Gusev A, Lowe JK, Stoffel M, Daly MJ, Altshuler D, Breslow JL, Friedman JM, Pe'er I: Whole population, genome-wide mapping of hidden relatedness. *Genome Res* 2009;19:318–326.
- Garone C, Pippucci T, Cordelli DM, Zuntini R, Castegnaro G, Marconi C, Graziano C, Marchiani V, Verrotti A, Seri M, Franzoni E: FA2H-related disorders: a novel c.270 + 3A>T splice-site mutation leads to a complex neurodegenerative phenotype. *Dev Med Child Neurol* 2011;53:958–961.
- Browning BL, Browning SR: Detecting identity by descent and estimating genotype error rates in sequence data. *Am J Hum Genet* 2013;93:840–851.
- Han L, Abney M: Using identity by descent estimation with dense genotype data to detect positive selection. *Eur J Hum Genet* 2013;21:205–211.
- Pippucci T, Parmeggiani A, Palombo F, Maresca A, Angius A, Crisponi L, Cucca F, Li-guori R, Valentino ML, Seri M, Carelli V: A novel null homozygous mutation confirms CACNA2D2 as a gene mutated in epileptic encephalopathy. *PLoS One* 2013;8:e82154.
- Hildebrandt F, Heeringa SF, Rüschemdorf F, et al: A systematic approach to mapping recessive disease genes in individuals from outbred populations. *PLoS Genet* 2009;5:e1000353.
- Ten Kate LP, Teeuw M, Henneman L, Cornel MC: Autosomal recessive disease in children of consanguineous parents: inferences from the proportion of compound heterozygotes. *J Community Genet* 2010;1:37–40.
- Gialluisi A, Pippucci T, Anikster Y, Ozbek U, Medlej-Hashim M, Megarbane A, Romeo G: Estimating the allele frequency of autosomal recessive disorders through mutational records and consanguinity: the homozygosity index (HI). *Ann Hum Genet* 2012;76:159–167.
- Gialluisi A, Incollu S, Pippucci T, Lepori MB, Zappu A, Loudianos G, Romeo G: The homozygosity index (HI) approach reveals high allele frequency for Wilson disease in the Sardinian population. *Eur J Hum Genet* 2013;21:1308–1311.
- 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: Quantification of homozygosity in consanguineous individuals with autosomal recessive disease. *Am J Hum Genet* 2006;78:889–896.
- Carothers AD, Rudan I, Kolcic I, Hayward C, Wright AF, Campbell H, Teague P, Hastie ND, Weber JL: Estimating human inbreeding coefficients: comparison of genealogical and marker heterozygosity approaches. *Ann Hum Genet* 2006;70:666–676.
- Leutenegger AL, Prum B, Gnani E, Verny C, Lemaître A, Clerget-Darpoux F, Thomson EA: Estimation of the inbreeding coefficient through the use of genomic data. *Am J Hum Genet* 2003;73:516–523.
- Bittles AH: A community genetics perspective on consanguineous marriage. *Community Genet* 2008;11:324–330.
- Magi A, Tattini L, Cifola I, D'Aurizio R, Benelli M, Mangano E, Battaglia C, Bonora E, Kurg A, Seri M, Magini P, Giusti B, Romeo G, Pippucci T, De Bellis G, Abbate R, Gensini GF: EXCAVATOR: detecting copy number variants from whole-exome sequencing data. *Genome Biol* 2013;14:R120.