AUDACITY: A comprehensive approach for the detection and classification of Runs of Homozygosity in medical and population genomics

Alberto Magia,⇑ Tania Giangregorio, Roberto Semeraro, Giulia Carangelo, Flavia Palombo, Giovanni Romeo, Marco Seri, Tommaso Pippucci

✉  Department of Information Engineering, University of Florence, Florence, Italy
✉  Department of Medical and Surgical Sciences, University of Bologna, Bologna, Italy
✉  Department of Experimental and Clinical Medicine, University of Florence, Florence, Italy
✉  Medical Genetics Unit, Sant’Orsola-Malpighi University Hospital, Bologna, Italy

A R T I C L E   I N F O

Article history:
Received 31 March 2020
Received in revised form 30 June 2020
Accepted 2 July 2020
Available online 14 July 2020

A B S T R A C T

Runs of Homozygosity (RoHs) are popular among geneticists as the footprint of demographic processes, evolutionary forces and inbreeding in shaping our genome, and are known to confer risk of Mendelian and complex diseases. Notwithstanding growing interest in their study, there is unmet need for reliable and rapid methods for genomic analyses in large data sets. AUDACITY is a tool integrating novel RoH detection algorithm and autozygosity prediction score for prioritization of mutation-surrounding regions. It processes data in VCF file format, and outperforms existing methods in identifying RoHs of any size. Simulations and analysis of real exomes/genomes show its potential to foster future RoH studies in medical and population genomics.

© 2020 The Authors. Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Runs of Homozygosity (RoH) are sizeable stretches of consecutive homozygous genotypes that arise in the genome of an individual who receives copies of an identical ancestral haplotype, a situation known as autozygosity [6]. RoHs are present in any human genome, but their size generally reflects the number of generations over which recombination had the chance to operate in breaking up haplotypes descending from a parental common ancestor.

As a consequence, the RoH burden increases in the offspring of consanguineous matings, as well as within isolated populations as a result of elevated levels of population background relatedness [25]. Within such populations, apparently unrelated parents often result to be connected as closely as third cousins or more when analyzed at the genome-wide level [12]. Conversely, RoH number shows distinctive population patterns which seem to follow the “out of Africa” serial-migration model, being less present in Africans while spreading in the other continental groups because of successive migrations that decreased the effective population size, reducing haplotype diversity and thus favoring the occurrence of RoHs [25].

Due to the fact that RoHs are enriched for rare deleterious variants [34], autozygosity is associated with an increased risk of autosomal recessive diseases [4]. In patients born to consanguineous parents the homozygous disease-causing variant usually resides within long (several to tens of megabases) tracts of autozygosity.

Exploiting this, homozygosity mapping has successfully identified during last decades genes underlying many hundreds of rare recessive diseases [2]. While a few, randomly distributed long RoHs stand out in the genome of inbred individuals, shorter RoHs are frequent also in outbred populations and tend to be relatively concentrated in genomic regions mainly governed by Linkage Disequilibrium (LD). Nonetheless, short RoHs may represent true autozygosity [25], and may surround autosomal recessive genes likely as a result of founder effects [14,24]. Beyond Mendelian genetics, RoHs have been recently investigated in complex conditions and quantitative traits, and have been shown to be indicative of selection signals [8].

The gold-standard technology for RoH detection is still considered to be array Single Nucleotide Polymorphisms (aSNPs), although following the advent of Next Generation Sequencing (NGS) a number of methods have been either adapted from aSNP or originally tailored to Whole Exome Sequencing (WES) data [26]. Arrays have lower genotyping error rates than NGS [36],

https://doi.org/10.1016/j.csbj.2020.07.003
2001-0370© 2020 The Authors. Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

* Corresponding authors.
E-mail addresses: albertomagi@gmail.com (A. Magi), tommaso.pippucci@gmail.com (T. Pippucci).
but give only access to a fixed set of about 1 million common SNPs. As WES and now Whole Genome Sequencing (WGS) are becoming at hand for research and diagnostic laboratories world-wide, RoH studies are making more and more extensive use of NGS data in medical as well as in population genomics [2,3,5,29,8].

However, an approach is lacking that comprehensively address the problem to reliably identify autozygosity by detecting RoH of any size, being as sensitive to the different characteristics of data underlying WES and WGS, as robust to undergo computationally intensive tasks in ever larger data sets.

We thus aimed to develop a rapid and accurate approach for RoH detection and characterization that exploits genotypes in Variant Calling Format (VCF) originated from either WES or WGS. To this end, we modeled NGS genotype calls by means of DIDOH\textsuperscript{3}M\textsuperscript{2}, a discrete-input, discrete-output Hidden Markov Model (HMM) obtained as a modification of our previous algorithm H\textsuperscript{3}M\textsuperscript{2} [19], and calculated for each of the so identified RoHs a logarithm of the odds (\textit{RLOD}) score reflecting its probability to be autozygous (Fig. 1).

We packaged DIDOH\textsuperscript{3}M\textsuperscript{2} and RLOD score in the AUDACITY (AUtozygosity iDentification And ClassIfication Tool) software tool and we show how such an approach outperforms current strategies to characterize RoHs that, irrespective of their size, are relevant for population studies exploiting WES/WGS data as well as for the identification of genes underlying recessive diseases.

2. Methods

2.1. Ethical considerations

Written informed consents were obtained from all patients or their parents/legal guardians who underwent WES for diagnostic or research purposes at the Medical Genetics Unit, Sant’Orsola-Malpighi University Hospital, and analysis of their WES was approved by the local Medical Ethics Committees.

2.2. DIDOH\textsuperscript{3}M\textsuperscript{2} algorithm

The HMM underlying DIDOH\textsuperscript{3}M\textsuperscript{2} (discrete-input, discrete-output homozygosity heterogeneous hidden Markov model) is a two-state HMM where the hidden states are the non-homozygous (\textit{S}\textsubscript{1}) and homozygous state (\textit{S}\textsubscript{2}) and the observations are the genotypes \textit{G}\textsubscript{i} assigned to each interrogated SNV \textit{i} (\textit{G}\textsubscript{i} \in \{\textit{G}\textsubscript{Ref}, \textit{G}\textsubscript{Het}, \textit{G}\textsubscript{Hom}\}), where \textit{G}\textsubscript{Hom} is homozygous reference, \textit{G}\textsubscript{Het} is heterozygous, while \textit{G}\textsubscript{Ref} is homozygous alternative) along the length of the genome.

The emission matrix, \textit{B}, has the following form:

\[
B = \begin{pmatrix}
1 - R_1 & R_1 \\
1 - R_2 & R_2
\end{pmatrix}
\]

where \textit{R}\textsubscript{1} and \textit{R}\textsubscript{2} are the probabilities of finding a heterozygous SNV in non-homozygous and homozygous genomic regions, respectively. In practice, \textit{R}\textsubscript{1} models the proportion of heterozygous SNVs in non-homozygous regions, while \textit{R}\textsubscript{2} the presence of heterozygous SNVs in homozygous regions which results mainly from sequencing and alignment errors.

We incorporated the distance between adjacent SNVs (\textit{d}\textsubscript{i}) and the likelihood of each observed genotype (\textit{P}\textsubscript{i}) into the transition probabilities of the HMM by considering a modified transition matrix defined for \textit{i} = 1, ..., \textit{n} - 1 where \textit{n} is the number of genomic markers:

\[
A_i = \begin{pmatrix}
1 - p_1(1 - e^{-f_i}) & p_1(1 - e^{-f_i}) \\
p_2(1 - e^{-f_i}) & 1 - p_2(1 - e^{-f_i})
\end{pmatrix}
\]

Fig. 1. AUDACITY workflow. Panel a: AUDACITY takes as input genotype data in VCF file format and follows 4 analysis steps for RoH identification by DIDOH\textsuperscript{3}M\textsuperscript{2} (Genotype Extraction, Hidden Markov Model, Viterbi algorithm) and classification by RLOD (Score calculation). For this latter step, AUDACITY can exploit allele frequencies by 1 K Genome Project Phase 3 as well as calculate custom allele frequencies from genotypes in the input VCF file. Panel b: The Hidden Markov Model is the core of the DIDOH\textsuperscript{3}M\textsuperscript{2} algorithm leading to RoH identification. Panel c: RLOD allows to classify ROH based on the allele frequencies of SNP genotypes forming the RoH diplotype.
where $p_i$ ($p_j$) is the probability to shift from $S_i$ to $S_j$ ($S_j$ to $S_i$) in a homogeneous HMM, $f_1 = (d_i/d_{\text{norm}}) + (1 - P_i)/P_{\text{norm}}^i$, $(d_{\text{norm}}, d_{\text{norm}})$ is the distance normalization parameter and $P_{\text{norm}}$ is the genotype likelihood normalization parameter.

As a result, we obtained a heterogeneous HMM, where the larger (smaller) di ($P_i$) and the greater the probability to shift from one state to another, $d_{\text{norm}}$ (the distance normalization parameter) and $P_{\text{norm}}$ (genotype probability normalization parameter) modulate the impact of $d_i$ and $P_i$, respectively, on the transition probability between the two hidden states ($S_i$, $S_j$), $p_1$, $p_2$, $R_1$, $R_2$, together with $d_{\text{norm}}$ and $P_{\text{norm}}$, are all set as parameters for DIDOH$^M^2$ instead of being estimated by an expectation-maximization algorithm since they result to be useful to set the resolution of RoH instead of being estimated by an expectation-maximization algorithm.

We constructed genealogies with consanguinity loops formed by unions between 1st, 2nd or 3rd cousins unions (1C/2C/3C) and with one single offspring to the consanguinuous parents (index offspring). To condition recombination patterns on the presence of a recessive disease-linked locus, we forced a specific locus (chr1: 212677319) to be inherited by the index offspring as 2 copies of an ancestral allele dropping from one of the 2 common ancestors of the parents in the pedigree.

Since Markerdrop associates a founder-tracking label with each dropping haplotype, we were able to trace pairs of adjacent SNPs between which the simulated recombinations took place. To locate each recombination spot to an exact genomic position, we randomly drew a single base pair coordinate according to the hg19 reference genome between every 2 adjacent SNPs with different founder-tracking labels. We assigned to each of the 2 SNP backbone haplotypes in the family founders one among 160 phased 1KGP Phase1 SNP haplotypes of European unrelated individuals [1]. For WGS, we used all SNVs called in the European 1KGP samples (about 16 M SNVs). For WES, to obtain a set of SNV sites that could be representative of most widely adopted exome target enrichment kits, we used the subset of SNV sites (about 470 K SNPs) that had median depth of at least 20X calculated across 5 of our WES performed in-house with each of the following kits: Agilent SureSelect Human All Exon v6 (Agilent Technologies Inc., La Jolla, CA, USA), SeqCap EZ v2.0/v3.0, (Roche NimbleGen, Basel, Switzerland) BGI (BGI, Shenzhen, China) and Nextera Rapid Capture Exome (Illumina Inc., San Diego, CA, USA).

WGS/WES variant call sets were eventually created by superimposing on the SNP backbone haplotypes in the index offspring the corresponding 1KGP haplotypes according to the recombination patterns traced by the founder-tracking labels.

We constructed genealogies with consanguinity loops formed by unions between 1st, 2nd or 3rd cousins unions (1C/2C/3C) and with one single offspring to the consanguinuous parents (index offspring). To condition recombination patterns on the presence of a recessive disease-linked locus, we forced a specific locus (chr1: 212677319) to be inherited by the index offspring as 2 copies of an ancestral allele dropping from one of the 2 common ancestors of the parents in the pedigree.

Since Markerdrop associates a founder-tracking label with each dropping haplotype, we were able to trace pairs of adjacent SNPs between which the simulated recombinations took place. To locate each recombination spot to an exact genomic position, we randomly drew a single base pair coordinate according to the hg19 reference genome between every 2 adjacent SNPs with different founder-tracking labels. We assigned to each of the 2 SNP backbone haplotypes in the family founders one among 160 phased 1KGP Phase1 SNP haplotypes of European unrelated individuals [1]. For WGS, we used all SNVs called in the European 1KGP samples (about 16 M SNVs). For WES, to obtain a set of SNV sites that could be representative of most widely adopted exome target enrichment kits, we used the subset of SNV sites (about 470 K SNPs) that had median depth of at least 20X calculated across 5 of our WES performed in-house with each of the following kits: Agilent SureSelect Human All Exon v6 (Agilent Technologies Inc., La Jolla, CA, USA), SeqCap EZ v2.0/v3.0, (Roche NimbleGen, Basel, Switzerland) BGI (BGI, Shenzhen, China) and Nextera Rapid Capture Exome (Illumina Inc., San Diego, CA, USA).

WGS/WES variant call sets were eventually created by superimposing on the SNP backbone haplotypes in the index offspring the corresponding 1KGP haplotypes according to the recombination patterns traced by the founder-tracking labels.

To generate sets of recombination patterns large enough to reproduce a representative spectrum of inbreeding coefficient (F) in the index offspring, we run 100 Markerdrop simulations for each pedigree with 1C, 2C and 3C genealogy loop. To account for the variability ascribable to different SNV sets among the selected 1KGP subjects, we assigned 10 and 5 different combinations of 1KGP founders’ haplotype pairs for WGS and WES, respectively.

2.5. Definition of true autozygous/non-autozygous RoH

To define true RoH, we wanted to find the minimum number of consecutive SNVs that were not homozygous by chance. To this end, we randomly picked up 100 K stretches of n consecutive SNVs from our WGS/WES call sets and used Hardy-Weinberg’s law to calculate the probability that all the n SNVs were homozygous. We included increasing n SNVs and chose the minimum n for which, on average over the 100 K stretches, the probability of finding n homozygous SNVs was $\leq 0.01$.

We found that the minimum n was 50 and 60 for WGS and WES, respectively, and defined each of these stretches spanning $\geq 100$kb, as true RoH. We made use of the founder-tracking labels associated with each of the 2 founders’ haplotypes to discriminate between autozygosity (same label) and non-autozygosity (different labels).
2.6. Estimation of inbreeding coefficient F

We estimated the inbreeding coefficient (F) of the simulated 1C, 2C, 3C index offspring using FSuite [11] with default options, creating 100 random submaps with one marker every 0.5 cM using SNVs in common between WES and WGS and with minor allele frequency > 0.05.

2.7. Linkage disequilibrium based SNV pruning

The pruned subset of SNVs was generated by using PLINK [28] with the –indep-pairwise option and the following parameter settings: window size in SNPs (50), number of SNPs to shift the window at each step (5) and r² threshold (0.5).

2.8. Incorporation of different allele frequency sets into the calculation of RLOD

To calculate RLOD, DIDOH²M² allows either to derive allele frequencies directly from the batch of samples under analysis or to use global allele frequencies pre-calculated by the 1KGP project. If a batch is relatively small, frequencies based on few subjects may affect RLOD calculation. However, users may be interested in using allele frequencies from the sample under study because they retain that global allele frequencies may not properly reflect the samples’ population. We therefore calculated RLOD using sets of allele frequencies derived from subsets of 10, 50, 100 individuals in the 1KGP Phase 1 European population and from the global 1KGP Phase 1 population and evaluated the performance of RLOD to predict autozygosity under the different sets.

2.9. RoH clustering

RoH clustering was performed with a three-component Gaussian mixture models by using the Mclust function from the mclust package (v.3) in R allowing component magnitudes, means, and variances to be free parameters. RoHs were then partitioned in the three classes and boundaries sizes between classes A and B and between classes B and C were estimated using the following formulas:

\[ C_{AB} = \frac{C_{\text{min}} - C_{\text{max}}}{2} \]

\[ C_{AC} = \frac{C_{\text{min}} + C_{\text{max}}}{2} \]

where \( C_{\text{min}} \) and \( C_{\text{max}} \) are the minimum and maximum RoH sizes for the three classes and boundaries sizes between classes A and B.

3. Results

3.1. RoH identification

To evaluate the performance of DIDOH²M² for different parameter settings we performed several analyses based on synthetic data (see Methods and Supplemental materials), and we found that when we want to study only large homozygous segments we should set large values of \( d_{\text{norm}} \) (10⁵, 10⁶) and small values of \( p_1 \) (0.1, \( p_2 \) must be set to 0.1). On the other hand, to increase the resolution of the algorithm and detect small RoHs, small \( d_{\text{norm}} \) (10³, 10⁴) and larger \( p_1 \) values (0.2, 0.3) are recommended.

As a further step, to test DIDOH²M² in the analysis of real genotypes data for different parameter settings, we leveraged WES and WGS genotype calls of 200 subjects from 1000 Genomes Project (see Methods and Supplemental materials) and we studied the RoHs identified by our method in terms of their cumulative global size and number. We found that while using higher values of \( R_1 \) increase both size and number of homozygous segments, the use of smaller values of \( R_2 \) increase the number but decrease the cumulative size of RoHs (Supplemental Fig. 4).

These results are a direct consequence of the role of \( R_1 \) and \( R_2 \) parameters in our heterogeneous HMM. \( R_1 \) represents the proportion of heterozygous markers that defines non-homozygous segments and all the segments that have a heterozygous proportion smaller than \( R_1 \) are identified as homozygous. For this reason, the larger \( R_1 \) and the larger the total size and number of homozygous segments identified by our model. On the other hand, \( R_2 \) represents the proportion of heterozygous markers that our HMM tolerates in a homozygous region. Larger values of \( R_2 \) allow to identify as homozygous regions with a higher number of heterozygous markers, while for small values of \( R_2 \) homozygous regions are called only if they contain a smaller fraction of heterozygous markers.

Hence, increasing the value of \( R_2 \) impose the algorithm to split large homozygous regions (with a fraction of heterozygous markers larger than \( R_1 \)) in small segments (with a fraction of heterozygous markers smaller than \( R_1 \)) thus increasing the total number of detected RoHs and decreasing their cumulative size.

By setting the most conservative set of parameters (\( R_1 = 1/100 \) and \( R_2 = 1/100000 \)), DIDOH²M² detected an average of around 90 Mb for WGS (around 1000 RoHs) and 30 Mb (around 20 RoHs) for WES data, while using more inclusive parameters (\( R_1 = 5/100 \) and \( R_2 = 1/1000 \)) it detected around 800 Mb of homozygous segments for WGS (around 20000 RoHs) and 450 Mb (around 1200 RoHs) for WES data.

To make a comparison with existing tools to identify RoH from VCF data, we applied PLINK [28], BCFTools [23] and VCFTools [10] to the WES and WGS genotype calls of the 200 aforementioned subjects from 1000 Genomes Project. To allow comprehensive evaluation of performance, we tested different combinations of parameters for each of those tools (DIDOH²M², PLINK, BCFTools) which allow the user to tune parameter settings (See Supplemental materials).

To estimate the true cumulative individual RoH length and size, we created a gold standard dataset generated using the genotype calls released by the 1KGP consortium for the afore-mentioned 200 subjects. We took as true RoH every region larger than 100 Kb and made by at least 200 consecutive homozygous single nucleotide variants (SNVs) across around 38 million and 1.5 million SNV genotypes called by the 1KGP consortium in WGS and WES, respectively (Supplemental materials).

VCFTools was always characterized by substantial over-calling while the average RoH length/number identified by DIDOH²M² and PLINK Across different parameter settings varied below and above the true value. BCFTools had a contrasting behavior, displaying excess or lack of true RoH in its results of both WGS and WES analyses (Supplemental Fig. 4). RoHs identified by DIDOH²M² had the lowest fractions of heterozygous SNVs (Fig. 2, Panels b,d and Supplemental Fig. 3), suggesting less spurious calls. In particular, by setting \( R_1 = 2/100 \) (for WGS data) or \( R_1 = 4/100 \) (for WES data) DIDOH²M² outperformed the three existing tools achieving the best trade-off between precision and recall over true RoH in both WGS and WES (Fig. 2, Panels a, c and Supplemental Fig. 3).

3.2. Prediction of autozygosity by RLOD

To estimate RoH probability, for each homozygous segment identified by DIDOH²M², we then computed a RoH LOD score (RLOD) comparing the probability of the most likely genotype with that of the observed genotype at each of the N homozygous SNVs within the RoH:
During simulations, a pair of 1KGP phased haplotypes for WES SNVs was assigned to each of the founder subjects of the synthetic pedigrees and let to drop along the genealogy. A founder-tracking label was linked to the flowing alleles, so that each was inherited by the offspring could be traced back to its founder of origin. By checking the pairs of founder-tracking labels linked to the RoH detected in the index offspring (Supplementary materials) we were able to unambiguously split RoH of the subject into autozygous (same label for both alleles) and non-autozygous (different label for each allele) segments.

We calculated the genomic inbreeding coefficient ($gF$) of all the 100 simulated offsprings for each 1C, 2C and 3C genealogy loop by FSuite [11] setting parameters as specified in Supplementary materials.

As recognized in the literature [37,20], $gF$ values are dispersed around the mean across pedigrees with identical genealogy loops. According to our simulations, for the 1C, 2C and 3C genealogy loops $gF$ values were distributed as follows: mean = 0.0675 and sd = 0.0231 (1C), mean = 0.0245 and sd = 0.0126 (2C), and mean = 0.0104 and sd = 0.00749 (3C), and show substantial overlap between pedigrees with the different genealogy loops. As a result, knowledge of the pedigree is not particularly useful to predict the inbreeding level of the offspring. To group together simulations with similar inbreeding levels, we therefore calculated the median $gF$ value for each genealogy (0.066, 0.023 and 0.0105 for 1C, 2C and 3C, respectively)
respective) and created four \( g^f \) ranges (from high to low inbreeding levels: F1: 0.066–1; F2: 0.023–0.066; F3: 0.00105–0.023; F4: 0–0.0105) into which we classified the simulated offsprings based on their \( g^f \) value rather than pedigree.

In published population as well as case-control RoH studies [9,13,16,17,20–22], either minimum size threshold or linkage disequilibrium (LD)-based SNP pruning is applied to avoid calling short RoHs that are very common or that are homozygous by equilibrium (LD)-based SNP pruning is applied to avoid calling [9,13,16,17,20–22], either minimum size threshold or linkage disequilibrium. This resulted in 6 RoH-call sets generated by applying any of the 3 size thresholds, alone or in combination with the LD cut-off (100 Kb, 500 Kb, 1500 Kb, 100 Kb + LD, 500 Kb + LD, 1500 Kb + LD). On these we then computed \( RLOD \) using SNP allele frequencies of the 1KGP Phase 1 European population. Subsequently we simulated WGS data following the same steps as for WES simulations but using a marker map extended using SNV allele frequencies of the 1KGP Phase 1 European population. We then simulated WES data following the same steps as for WES simulations but using a marker map extended using SNV allele frequencies of the 1KGP Phase 1 European population. We then calculated precision and recall for each RoH-call set as follows:

- As precision, we calculated the fraction of RoH calls by \( DIDOH^M \) that have any overlap with true autozygous RoH;
- As recall, we calculated the fraction of true autozygous RoH called by \( DIDOH^M \).

We then compared the capability of \( RLOD \) to identify autozygosity with that of RoH size, because long RoHs are commonly considered to be truly autozygous regions. We were able to demonstrate that \( RLOD \) largely outperforms size in discriminating between autozygosity and non-autozygosity when applied to both WES and WGS data, with more evident gain in performance as \( g^f \) range decreases (Fig. 3).

For any \( g^f \) range, the best trade-off between precision and recall is obtained by the combination of \( RLOD \) with the most stringent threshold for size (1500 Kb) and \( LD \geq 0.5 \). In WES data, progressive regression of the trade-off point is observed from higher to lower \( g^f \) ranges (Fig. 3a–d), while in WGS \( RLOD \) improved performance even more markedly than in WES without apparent loss in accuracy in lower \( g^f \) ranges (Fig. 3e–h).

\( DIDOH^M \) allow users to use allele frequencies (AFs) retrieved from 1KGP or custom AFs calculated directly from genotypes in the VCF file under analysis. Importantly, we did not notice any major change in \( RLOD \) performance when using increasing numbers of samples to calculate AFs from 50 samples upwards. As shown in Supplemental Fig. 6, \( RLOD \) provided comparable accuracy in identifying autozygosity from WES as well WGS simulated data using 1KGP Phase 1 global or European AFs, as it did using custom AFs calculated from a number of samples of 50 or more.

### 3.3 Prioritization of mutation-surrounding RoH by \( RLOD \)

Most approaches for the prioritization of candidate variants for autosomal recessive diseases rely on the size of the surrounding RoH [37,30,2]. However size is not always optimal to predict which, among many long RoH in the patient’s genome, is the one containing the causative variant [38], because also short RoH can happen to be autozygous and surround autosomal recessive genes [25,15,24]. Of the few available alternative strategies that use statistical methods and exploit haplotype frequencies [38,15], none is tailored for NGS data. To compare the capability of \( RLOD \) to prioritize the mutation-surrounding RoH (msRoH) with that of RoH size, while performing the simulations described above we forced a specific disease-linked locus to be inherited by the offspring as two copies of an ancestral allele (Supplementary materials). We then ranked RoHs of the index offspring’s WES simulations per genealogy loop by both \( RLOD \) and size, and evaluated which of the two measures was the most efficient in prioritizing the msRoH as follows:

- we calculated how many times the disease-linked RoH ranked as 1st by any of the two measures;
- we calculated how many times the disease-linked RoH rank by one measure was higher or equal to that by the other;

We found that the msRoH ranks as 1st significantly more times by \( RLOD \) than by size, and ranks higher or equal by \( RLOD \) than by size both in WES and WGS (Tables 1 and 2, Fig. 4 and Supplemental Fig. 7).

Overall, these results show that \( RLOD \) outperforms size in prioritizing the msRoH in a patient’s genome, proving to be useful as part of the toolkit for prioritization of candidate variants with recessive effect. The lower is the \( g^f \) range (Fig. 4), as well as the degree of parental consanguinity (Supplemental Fig. 7), the more significant are these differences (Table 1 and Table 2).

To replicate the conclusions of the simulation analysis in the WES of 15 real patients, we used data where the homozygous disease-causing variant was found to be surrounded by RoH as identified using \( DIDOH^M \). This data set included 13 unrelated patients whose parents were closely inbred (1st/2nd cousins), and 2 for whom parental consanguinity was not reported, all undergoing WES for research [27,14,24] or diagnostics.

WES data were processed as in the simulations to identify patients’ RoHs, and the ranking position by \( RLOD \) and size of the mutation-surrounding RoH was retrieved from \( DIDOH^M \) results. \( RLOD \) conferred the msRoH higher or equal rank than size in 9 out of the 15 WES (60%). The overall distance between ranking positions by size and \( RLOD \) across the 15 WES, calculated as the sum of the distances between the ranking positions by size and \( RLOD \) (totaling 54), is in favor of this latter.

As shown in Fig. 5, when the ranking position by \( RLOD \) is higher than that by size, the distance between them is larger (mean = 8.25 ranking positions) than in the opposite situation (mean = 2 ranking positions), demonstrating that \( RLOD \) outdistant size in the majority of cases, while in instances where the size confers the msRoH higher rank than \( RLOD \), the 2 measures achieve comparable ranking positions. Notably, the higher the distance between the ranking positions by size and \( RLOD \), the shorter the size of the mutation-surrounding RoH (\( r = -0.69 \)) (Fig. 5).

This underscores the capability of \( RLOD \) to pick out small msRoHs among the many regions of similar or larger size found throughout the genome, reflecting the simulation analysis showing that \( RLOD \) outperformed size especially for low \( g^f \) ranges. \( RLOD \) was able to outdistance size in prioritizing small msRoHs in the WES of patients with autosomal recessive disorders, as for RoHs surrounding disease-causing MYO15A and ATAD3A variants, both smaller than 2 Mb [14,24].

### 3.4 Characterization of RoH across worldwide populations by \( DIDOH^M \)

To show the potentiality of our computational approach to explore genomic patterns of homozygosity in human populations
we used $DIDOH^3M^2$ to analyze genotypes of 600 individuals from six populations sequenced by the 1000 Genomes Project (100 YRI, Yoruba from Ibadan, Nigeria; 100 BEB, Bengali from Bangladesh; 100 CEU, Utah residents with Northern and Western European ancestry from the CEPH collection; 100 JPT, Japanese in Tokyo, Japan; 100 CLM, Colombians from Medellin, Colombia; 100 FIN, Finnish in Finland).

We performed $DIDOH^3M^2$ analysis with $p_2 = 0.1, p_1 = 0.1, d_{num} = 10^3, R_1=4/100$ and $R_2=1/1000$ and, following the model proposed by [25], for each population separately we analyzed RoH sizes as a mixture of three normal distributions representing three distinct RoH classes: short RoHs ranging tens of Kb (class A), medium RoHs ranging hundreds of Kb (class B), large RoHs ranging up to tens of Mb (class C) (Fig. 6, Panel a). Class A reflect homozygosity for ancient haplotypes that contribute to local LD patterns, class B result from background relatedness owing to limited population size while class C result from recent parental relatedness.

The mean size of each class and the boundaries between different classes vary across the 6 populations, in particular class B and C RoHs (Fig. 6 Panel b). For each class, we observed the smallest mean size for YRI and the largest mean size for JPT and CLM populations. As a further step we calculated the overall RoH length per individual across the three classes and we studied its distributions within each population and compared the populations with each other.

As shown in Fig. 6, Panels c–f, the total lengths of RoH (Fig. 6, Panels f) generally increase with increasing distance from Africa of the geographical location of the population, where an isolated population such as the Finnish are comparable with the European CEU population and an admixed population such as the Colombian shows greater variability. Class A (Fig. 6, Panels f) and class B (Fig. 6, Panels e) RoH generally follow this pattern, only with decreasing variability in the admixed Colombian population from class B to class A RoH. Total lengths of class C RoH (Fig. 6, Panel d) are not characterized by the same stepwise increase. Instead, they are higher in the Finnish and are more variable in the Colombian population.

Finally, for each individual, we performed pairwise comparisons between the total lengths of class A, class B, and class C RoHs. In agreement with results reported by [25] (Fig. 5, Panels g–i), the total length of class A and B RoH are highly correlated ($R = 0.91$), while the correlation with class C is much smaller ($C-A R = 0.34$, $C-B R = 0.99$).
C-B $R = 0.37$), suggesting that class A and B RoHs, as expected, may have arisen via a different process of class C.

3.5. AUDACITY tool

The DIDOH$^3$M$^2$ algorithm and RLOD score calculation described and tested in previous sections have been packaged in the AUDACITY software tool. AUDACITY is a collection of perl, R and fortran codes. A schematic representation of its workflow is reported in Fig. 1. AUDACITY takes as input the genotype data of multiple samples in VCF file format, selects all the biallelic SNVs, applies the DIDOH$^3$M$^2$ algorithm, calculates the RLOD and gives as output a bed file containing coordinates (Chr, Start, End), the RoH length, the number of SNPs and the RLOD score for each detected RoH.

In default setting, the AUDACITY tool calculates the RLOD of each RoH by exploiting the allele frequencies of all the biallelic SNVs discovered by the 1 K genome project Phase 3 dataset. In alternative, the AUDACITY tool allows to calculate custom allele frequency of the individuals from the input VCF file.

On a desktop computer with a 2.5 GHz cpu and 8 GB of ram, it takes two hours to perform the analysis of a VCF file with the genotype calls of ten WGS experiments. AUDACITY is publicly available at https://sourceforge.net/projects/audacity-tool/.

4. Discussion

In this study, we created and tested a novel method, AUDACITY, for simultaneous identification and prioritization of RoHs from WES and WGS data. Prioritization was done by computing RLOD, a LOD score that inversely reflects the cumulative frequency of the homozygous alleles forming the RoH diplotype, and then by ranking the RoHs according to their RLOD.

The idea of using allele frequencies to assess the probability of a RoH to be autozygous was firstly postulated by Broman and Weber [6], who proposed a sliding-window method to identify RoHs cou-
Fig. 5. Mutation-surrounding (ms) ROH prioritization by RLOD and RoH size in real WES. The difference in msRoH ranking position by RLOD and RoH size is shown for WES data of 15 patients in whom the homozygous disease-causing variant had been identified as part of a research project or of a diagnostic process. Colors represent different size ranges of the msRoH.

Fig. 6. Genomic patterns of homozygosity in 6 human populations. a Gaussian kernel density estimates of the ROH size distribution in each of the 6 populations (YRI, BEB, CEU, JPT, CLM and FIN); 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 6 populations. Dotted lines define the limits of class A (small) ROHs, continuous lines define the limits of class B (medium) ROH and dashed lines define the limits of class C (large) ROH. Filled circles represents the mean size of class A ROH, filled squares the mean size of Class B ROH and filled rhombus the mean size of Class C ROH. c–f Distribution of total ROH lengths over all individuals in each of the 6 populations, for all three classes combined (c), for class A (d), for class B (e) and for class C (f). Data are shown as violin plots: each violin contains a vertical black line (25–75% range) and a horizontal white line (median). g–j report the Pairwise comparison of per-Individual Total ROH Lengths across Size Classes. A vs B (g), A vs C (h) and B vs C (i).
pling detection and inference about autozygosity. When this method was proposed, genome-wide scans were performed with hundreds of microsatellite markers, far fewer than the million SNVs that can be interrogated in WES and WGS data.

We decided to uncouple the processes of RoH detection and autozygosity prediction, first running the new HMM algorithm to single out the regions of homozygosity, and only after applying the RLOD calculation limited to the so-identified regions. In this way, we avoided the computationally intensive task of doing iterative RLOD calculations over million markers by overlapping windows along the genome, resulting in faster computation at no expense in performance.

By taking multi-sample VCF as input file format, DIDOH\textsuperscript{H}M\textsuperscript{2} overcomes the major limitation of our former tool H\textsuperscript{2}M\textsuperscript{2} [19], that was able to deal only with the huge, and thus less manageable for most end-users, single-sample BAM files. Indeed, VCF is by far the most accessible NGS data format for laboratories worldwide, and since it can be populated with genotypes of many samples, it is suitable for DIDOH\textsuperscript{H}M\textsuperscript{2} analysis of diagnostic series, case-control cohorts as well as for population studies.

By an extensive work of performance comparison, we demonstrated how DIDOH\textsuperscript{H}M\textsuperscript{2} outperforms popular existing tools in the accuracy to detect RoH from both WES and WGS. In particular, as the previous H\textsuperscript{2}M\textsuperscript{2}, our algorithm proves to be as more accurate than the other tools as RoH size is smaller.

While researches have long been focused on long RoHs unveiling the presence of homozygous recessive alleles in patients from consanguineous families, the increasing availability of WGS will allow to disclose the effect of short RoH on complex disease risk and on the demographic history of human populations [8]. In this perspective, it is of importance to classify RoH based on the allelic composition of their diplotype rather than on their size, because the former has the potential to shed light on the RoH origin and relevance with respect to genomic variables such as recombination rate, positive selection and recessive effect of alleles modulating human traits and diseases. RLOD proved to be able to reliably predict autozygosity from WES and WGS data, and we demonstrated by simulation analysis and application to real WES data that this property can be used to prioritize mRoH implicated in recessive disorders more efficiently than size.

As noted already by other authors [25], using fixed size cut-offs to RoH length such as 500 Kb, 1 Mb or 2 Mb [18,17,32] based on the assumption that RoH below these thresholds are chance homozygosity mainly governed by LD patterns and therefore not biologically relevant, is at risk of overlooking true autozygosity. In view of an increasingly adoption of WGS by clinical and research laboratories, RLOD will be useful in prioritizing RoHs where disease-causing variants may be not easily tackled, i.e. because non-coding and therefore difficult to single out from the wealth of homozygous candidate variants dispersed throughout the genome. As an anticipation of this, RLOD was helpful in identifying a disease-causing synonymous variant in NARS2 inside the mRoH. As synonymous changes are usually assigned low priority in that they are not predicted to alter the protein product, this variant was initially discarded by our workflow for candidate variant filtering. However it later emerged as of interest when we incorporated RLOD into our variant classification algorithm, since the NARS2-surrounding RoH was ranked 7th by RLOD instead of 12th by size, eventually leading to diagnosis when the patient’s phenotype matched to literature reports of other patients with mutations in this gene [31,33,35].

As shown in our Precision-Recall plots of Fig. 3, RLOD is the major factor to improve autozygosity prediction from both WES and WGS data. Size and LD cut-offs play a role especially in WES and gain relevance for lower g\textsuperscript{F} ranges. Using these cut-offs is safe in population studies where the end-point is to obtain an estimate in terms of RoH number or length of genomic autozygosity. We would however recommend caution especially for gene-mapping and mutation-detection purposes, because their use may lead to loose the mRoH for the sake of autozygosity prediction accuracy.

Since users may perform DIDOH\textsuperscript{H}M\textsuperscript{2} analysis on samples of different sizes, we wanted to evaluate the extent to which the specification of allele frequencies calculated from smaller to larger sample sizes could affect RLOD. The use of allele frequencies derived from increasing sample sizes indicates that DIDOH\textsuperscript{H}M\textsuperscript{2} analysis is reliable also when carried out in small cohorts or populations. This is particularly important for analyses involving samples from populations that are not referenced in large variant databases, so that retrieving allele frequencies from such resources may lead to dangerously alter RLOD calculation.

To evaluate the capability of DIDOH\textsuperscript{H}M\textsuperscript{2} to prioritize the mRoH in patients affected with autosomal recessive disorders, we simulated patients’ genomes as offspring to 1C, 2C and 3C consanguineous parents. Since the well-known dispersion around the mean of g\textsuperscript{F} values across pedigrees with identical genealogy loops, we introduced here the median g\textsuperscript{F} values for each genealogy as thresholds to separate groups of pedigrees based on the actual genomic inbreeding rather than relying on pedigrees. Such an expedient was instrumental to extract deeper information on how RLOD and size perform when analyzing data of patients characterized by different inbreeding levels. As shown in our simulations, the RLOD outperforms size in prioritizing the mRoH for any genealogy loop, whether it is of the 1st or of the higher ranking position. Looking at g\textsuperscript{F} ranges, conversely, the performance of the 2 approaches are substantially indistinguishable for the highest interval (0.066-1), suggesting that RLOD does not provide valuable advantage over size in patients with very high inbreeding levels, irrespective of their reported consanguinity. Indeed, these individuals bear multiple RoH extending up to tens of Mb which usually receive the highest RLOD scores of the genome. In such a situation, which is often the rule in highly consanguineous communities, RLOD and size result to be therefore equivalent. Otherwise, our findings clearly show that RLOD improves mRoH prioritization for any other g\textsuperscript{F} range and overall, therefore it could be successfully used as substitute for size in the gene-mapping process.

Medium to small RoHs are thought to contribute to complex diseases and quantitative traits, and their role is increasingly investigated. As RLOD was capable of prioritizing mRoHs that belong to these classes, it may prove useful in prioritizing also the multiple loci that in a patient genome accumulate detrimental homozygous alleles contributing to disease additively.

As for ROH analysis in 6 1KGP populations, we obtained results consistent with previous studies based on SNP arrays [25], demonstrating suitability of AUDACITY to enable reliable analysis of ROH distribution across human populations. Our previous tool H\textsuperscript{2}M\textsuperscript{2} [19] has been used to profile ROHs in large collections of samples from populations known for their high inbreeding degrees [29]. We believe that AUDACITY, with its improved workflow for straightforward processing of multiple sample VCF data, will be of greater help to carry out such large-scale projects.

5. Conclusion

In conclusion, AUDACITY is a comprehensive approach for the analysis of RoHs from NGS data, either WES or WGS, tailored for applications in medical as well as population genomics. It proved to outperform existing tools in the accuracy to detect RoHs and RLOD, the autozygosity prediction score it incorporates, is suitable to prioritize regions relevant for traits and diseases. Its ability to handle data in VCF format responds to the emerging need of reli-
able and rapid RoH characterization in ever larger WGS data sets, that are becoming increasingly available to researchers that aim to enlighten the effect of RoHs in conferring risk for complex diseases and in shaping the genome of human populations.

Data sharing statement
All WES and WGS data used for algorithm validation, simulation and real data analyses, and population study are publicly available as part of the 1000 Genomes Project (ftp://ftp.1000genomes.ebi.ac.uk/vol1/phase3/data/). WES of the 13 patients were not provided with consent for data sharing.

Declaration of Competing Interest
The authors declare that they have no known conflicting financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement
Alberto Magi: Conceptualization, Methodology, Software, Writing - original draft. Tania Giangregorio: Software, Writing - original draft. Roberto Sereneto: Formal analysis. Giulia Carangelo: Formal analysis. Flavia Palombo: Formal analysis. Giovanni Romeo: Supervision. Marco Seri: Supervision. Tommaso Pipucci: Conceptualization, Methodology, Software, Writing - original draft.

Appendix A. Supplementary data
Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.csbj.2020.07.003.

References
[1] 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, McVean GA. An integrated map of genetic variation from 1,092 human genomes. Nature 491 (7422): 2012: 56–65.
[2] Alkuraya FS. The application of next-generation sequencing in the Harel T, Yoon WH, Garone C, Gu S, et al. Recurrent de novo and biallelic Joshi PK, Esko T, Mattsson H, Eklund N, Gandin I, et al. Directional dominance Pemberton TJ, Absher D, Feldman MW, Myers RM, Rosenberg NA, Li JZ. McQuillan R, Leutenegger AL, Abdel-Rahman R, Franklin CS, Pericic M, Barac-Narasimhan V, Danecek P, Scally A, Xue Y, Tyler-Smith C, Durbin R. BCFtools/Christofidou P, Nelson CP, Nikpay M, Qu L, Li M, Loley C, Debiec R, Braund PS, Broman KW, Weber JL. Long homozygous chromosomal segments in reference Carr IM, Flintoff KJ, Taylor GR, Markham AF, Bonthron DT. Interactive visual Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Kirin M, McQuillan R, Franklin CS, Campbell H, McKeigue PM, Wilson JF. Genomic runs of homozygosity record population history and consanguinity. Lenz T, Lambert C, DeRosse P, Burdick KE, Morgan TV, Kane JM, Kuchlerapiti R, Malhotra AK. Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. PNAS 2007; 104:19942–7.
[3] Magi A, Tattini L, Palombo F, Benelli M, Galliuzzi A, Guisti B, Abbate R, Seri M, Genisi GF, Romeo G, Pippucci T. H3M2: detection of runs of homozygosity from whole-exome sequencing data. Bioinformatics 2014;30(20): 2852–9.
[4] McQuillan R, Leuteneger AL, Abdel-Rahman R, Franklin CS, Pericic M, Barac-Lauc L, Smolej-Naranic N, Janicijevic B, Polasek O, Tenesa A, Macleod AK, Farrington SM, Rudan P, Hayward C, Vitart V, Rudan I, Wild SH, Dunlop MG, Water AF, Campbell H, Wilson JF. Runs of homozygosity in European populations. Am J Hum Genet 2008;3(3): 359–72.
[5] Nalis MA, Simon-Sanchez J, Gibbs JR, Puisant-Ruz C, Bras JT, Tanaka T, Martarin M, Schouls W, Weitz C, Harris TB, Ferrutz L, Hardy J, Singleton AB. Measures of autozygosity in disease: globalization, urbanization, and its implications for medical genetics. PLoS Genet 2009;5(3): e1000415.
[6] Nalis MA, Guerreiro RJ, Simon-Sanchez J, Bras JT, Traynor BJ, Gibbs JR, Laurent H, Le Hyaric Y, Singleton AB. Extended tracts of homozygosity identify novel candidate genes associated with late-onset Alzheimer’s disease. Neurogenetics 2009;10(3): 183–90.
[7] Narasimhan V, Danecek P, Scally A, Xue Y, Tyler-Smith C, Durbin R, BCfTools/ RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. Bioinformatics 2016;32(11): 1749–51.
[8] Palombo F, Al-Wardy N, Ruscone GA, Oppo M, Kindi MN, Angius A, Al Lamki K, Cirotto G, Giangregorio T, Benelli M, Magi A, Seri M, Gasparini P, Cucca F, Fattobian S, Al Khabori M, Al Khabori M, Pippucci T, Romeo G, A novel founder MY015A frameshift duplication is the major cause of genetic hearing loss in Oman. J Hum Genet 62(2): 2017:259-264..
[9] Pemberton TJ, Absher D, Feldman MW, Myers RM, Rosenberg NA, Li JZ, Conservation patterns of homozygosity in worldwide human populations. Am J Hum Genet 2012;102(7): 275–92.
[10] Pippucci T, Magi A, Giuliacci A, Romeo G. Detection of runs of homozygosity from whole-exome sequencing data: state of the art perspectives for coarse clustering and epidemiological studies. Hum Hered 2014;77(1– 4): 63–72.
[11] Pippucci T, Parmeggiani A, Palombo F, Maresca A, Angius A, Crisponi L, Cucca F, Ligouri R, Valentino ML, Seri M, Carelli V. A novel null homozygous mutation confirms CACNA2D2 as a gene mutated in epileptic encephalopathy. PLoS One 8(12): 2013:e82154. .
[12] 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 association and population based linkage analyses. Am J Hum Genet 2007;81:559–75.
[13] Scott EM, Hallez A, Itan Y, Spencer EG, He Y, Azab MA, Gabriel SB, Belkadi A, Boisson B, Abel L, Clark AG, Greaves ME. G:East Variome Collaboration. Alkuraya FS, Casanova JL, Gleeson G. Characterization of Greater Middle Eastern genetic variation for enhanced disease gene discovery. Nat Genet. 2016;48(9): 1071–6.
[14] Seelew D, Schuette M, Hildebrandt F, Nrnberg P. HomozygosityMapper—an interactive approach to homozygosity mapping. Nucleic Acids Res 37(Web Server issue): 2009:W593–9.
[15] Simon M, Richard EM, Wang X, Shahzad M, Huang VH, Quispe TA, Potiriu M, Li J, Castelli MV, Davila A, Nabil MA, Al-Shehri S, Yu M, Garajz C, Chang R, Al-Shehri N, Newman WG, Abdenur J, Starr A, Hegde R, Dorn T, Busch A, Park E, Wu J, Schwenzer H, Flierl A, Florint C, Sissler M, Khan SN, Li R, Guran MX, Friedman TB, Wu D, Procaccio V, Rizzanudin S, Wallace DC, Ahmed ZM, Huang T, Rizzanudin S. Mutations of human NARS2, encoding the mitochondrial asparaginyl-tRNA synthetase, cause nonsyndromic deafness and Leigh syndrome. PLoS Genet 2015;25:11(3):e1005997.
[16] Simon-Sanchez J, Kilarasi LL, Nalis MA, Martinez M, Schulte C, et al. Cooperative genome-wide analysis shows increased homozygosity in early onset Parkinsons disease. PLoS One 2012;7(3):e28787.
[33] Sofou K, Kollberg G, Holmström M, Dvila M, Darin N, Gustafsson CM, Holme E, Oldfors A, Tulinius M, Asin-Cayuela J. Whole exome sequencing reveals mutations in NARS2 and PARS2, encoding the mitochondrial asparaginyl-tRNA synthetase and prolyl-tRNA synthetase, in patients with Alpers syndrome. Mol Genet Genomic Med 2015;3(1):59–68.

[34] Szpiech ZA, Xu J, Pemberton TJ, Peng W, Zilker S, Rosenberg NA, Li JZ. Long runs of homozygosity are enriched for deleterious variation. Am J Hum Genet 2013;93(1):90–102.

[35] Vanlander AV, Menten B, Smet J, De Meirleir L, Sante T, De Paepe B, Seneca S, Pearce SF, Powell CA, Vergult S, Michotte A, De Lette E, Vantomme L, Minczuk M, Van Coster R. Two siblings with homozygous pathogenic splice-site variant in mitochondrial asparaginyl-tRNA synthetase (NARS2). Hum Mutat 2015;36(2):222–31.

[36] Wall JD, Tang LF, Zerbe B, Kvale MN, Kwok PV, Schaefer C, Risch N. Estimating genotype error rates from high-coverage next-generation sequence data. Genome Res 2014;24(11):1734–9.

[37] 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(5):889–96.

[38] Zhang L, Yang W, Ying D, Cherny SS, Hildebrandt F, Sham PC, Lau YL. Homozygosity mapping on a single patient: identification of homozygous regions of recent common ancestry by using population data. Hum Mutat 2011;32(3):345–53.