Computational and Structural Biotechnology Journal 18 (2020) 1956–1967

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

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

Alberto Magi [a,*], Tania Giangregorio [b], Roberto Semeraro [c], Giulia Carangelo [c], Flavia Palombo [b], Giovanni Romeo [b], Marco Seri [b,d], Tommaso Pippucci [d,*]

[a] Department of Information Engineering, University of Florence, Florence, Italy
[b] Department of Medical and Surgical Sciences, University of Bologna, Bologna, Italy
[c] Department of Experimental and Clinical Medicine, University of Florence, Florence, Italy
[d] Medical Genetics Unit, Sant'Orsola-Malpighi University Hospital, Bologna, Italy

ARTICLE INFO

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

ABSTRACT

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.

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 to 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],

* Corresponding authors.
E-mail addresses: albertomagi@gmail.com (A. Magi), tommaso.pippucci@gmail.com (T. Pippucci).

https://doi.org/10.1016/j.csbj.2020.07.003

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*^{3}*M*^{2}, a discrete-input, discrete-output Hidden Markov Model (HMM) obtained as a modification of our previous algorithm *H*^{3}*M*^{2} [19], and calculated for each of the so identified RoHs a logarithm of the odds (*RLOD*) score reflecting its probability to be autozygous (Fig. 1).

We packaged *DIDOH*^{3}*M*^{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^{3}M^2 algorithm

The HMM underlying *DIDOH*^{3}*M*^{2} (discrete-input, discrete-output homozygosity heterogeneous hidden Markov model) is a two-state HMM where the hidden states are the non-homozygous (S_1) and homozygous state (S_2) and the observations are the genotypes G_i assigned to each interrogated SNV i ($G_i \in \{G_{Homr}, G_{Het}, G_{Homa}\}$, where G_{Homr} is homozygous reference, G_{Het} is heterozygous, while G_{Homa} is homozygous alternative) along the length of the genome.

The emission matrix, B, has the following form:

$$B = \begin{pmatrix} 1-R_1 & R_1 \\ 1-R_2 & R_2 \end{pmatrix} \quad (1)$$

where R_1 and R_2 are the probabilities of finding a heterozygous SNV in non-homozygous and homozygous genomic regions, respectively. In practice, R_1 models the proportion of heterozygous SNVs in non-homozygous regions, while R_2 the presence of heterozygous SNVs in homozygous regions which results mainly from sequencing and alignment errors.

We incorporated the distance between adjacent SNVs (d_i) and the likelihood of each observed genotype (P_i) into the transition probabilities of the HMM by considering a modified transition matrix defined for $1 \leqslant i \leqslant n-1$ where n is the number of genomic markers:

$$A_i = \begin{pmatrix} 1-p_1(1-e^{-f_i}) & p_1(1-e^{-f_i}) \\ p_2(1-e^{-f_i}) & 1-p_2(1-e^{-f_i}) \end{pmatrix} \quad (2)$$

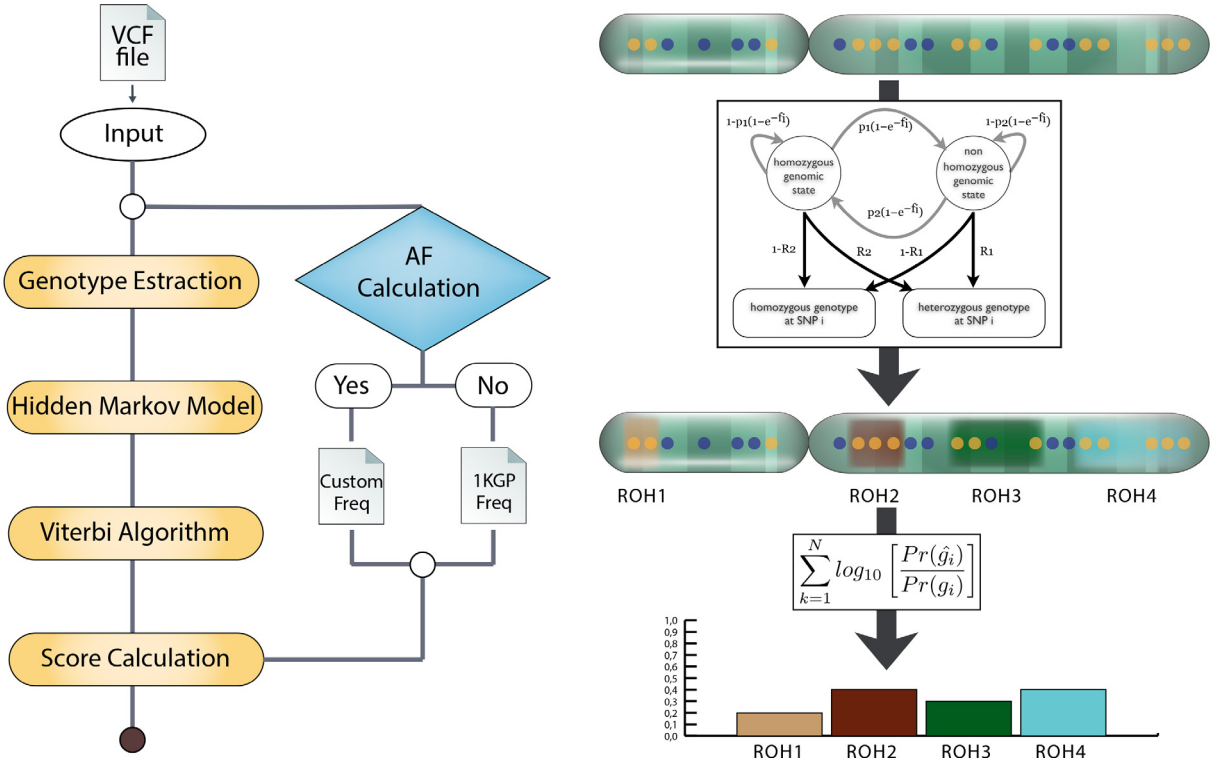

Fig. 1. AUDACITY workflow. Panel a: AUDACITY takes as input genotype data in VCF file format and follows 4 analysis steps for RoH identification by *DIDOH*^{3}*M*^{2} (Genotype Extraction, Hidden Markov Model, Viterbi algorithm) and classification by *RLOD* (Score calculation). For this latter step, 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*^{3}*M*^{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_1 (p_2) is the probability to shift from S_1 to S_2 (S_2 to S_1) in a homogeneous HMM, $f_i = (d_i/d_{Norm}) + (1 - P_i)/P_i^{P_{Norm}}), (d_{Norm}, d_{Norm})$ is the distance normalization parameter and P_{Norm} is the genotype likelihood normalization parameter.

As a result, we obtained a heterogeneous HMM, where the larger (smaller) *di* (*Pi*) and the greater the probability to shift from one state to another. d_{Norm} (the distance normalization parameter) and P_{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_1-S_2). $p1, p2, R_1, R_2$, together with d_{Norm} and P_{Norm}, are all set as parameters for $DIDOH^3M^2$ instead of being estimated by an expectation-maximization algorithm since they result to be useful to set the resolution of RoH detection in terms of region size and SNV number. Finally, we use the Viterbi algorithm to estimate the best sequence of S_1 and S_2 and to consequently associate each G_i to one of the two states, allowing to discriminate between homozygous and non-homozygous genomic regions and thus identifying RoH.

2.3. Evaluation dataset and performance comparison

In its Phase1, the 1000 Genomes Project (1KGP) consortium, by combining low-coverage whole-genome sequencing (WGS) and high-coverage whole-exome sequencing (WES) of 1092 individuals from 14 populations from Europe, East Asia, sub-Saharan Africa and the Americas, identified around 38 million single nucleotide polymorphic positions and 1.4 million short insertions and deletions [1].

In order to test the performance of our algorithm and the other three state of the art methods on real data analysis, we used them to analyze the WGS and WES genotype data of 200 individuals (50 of European ancestry, 50 of African ancestry, 50 of American ancestry and 50 of Asian ancestry) sequenced by 1000GP consortium (see Supplemental materials). For WGS data analyses we considered the complete set of biallelic SNVs (around 38 million), while for WES analyses we included all the SNVs falling within the range of the 1KGP exomic target regions.

To evaluate $DIDOH^3M^2$ ability to identify RoH from WES and WGS data and to compare its performance with respect to other three state of the art methods (PLINK, BCFtools, VCFtools, see Supplemental materials), we generated a gold standard RoH dataset by using the 1KGP SNV genotype calls of the aforementioned 200 individuals. To this end, we considered as gold standard RoH all the regions $\geq 100Kb$ and containing at least 200 consecutive homozygous SNVs.

To compare the performance of DH^3M^2 and existing tools we calculated precision and recall as follows:

- to calculate precision, we considered all the SNVs called within RoH by each of the 4 approaches and we then calculated the fraction of these SNVs called as homozygous also in the gold standard dataset;
- to calculate recall, we considered all the SNVs called in RoH in the gold standard dataset and we then calculated the fraction of these SNVs called as homozygous by each of the 4 approaches.

2.4. Generation of WGS/WES synthetic variant call sets in offspring to consanguineous unions

To simulate realistic WGS/WES data of offspring to consanguineous unions, we created synthetic variant call sets using a gene dropping strategy (Supplementary Fig. 4). To speed up the process of dropping dense genetic maps made of hundred thousands or million SNVs such as in WES or WGS, we followed the simulation framework of [11]. We generated a genetic map for each of the 22 autosomes by picking up from Rutgers Map v.3a (http://compgen.rutgers.edu/maps) one biallelic SNP having minor allele frequency (MAF) in the range 0.3–0.7 every about 0.05 cM. The so obtained SNP backbone (64 K autosomal SNPs) was used to simulate recombination patterns conditional on a disease-linked locus with the Markerdrop utility of the MORGAN v. 3.1.1 suitehttps://www.stat.washington.edu/thompson/Genepi/MORGAN/Morgan.shtml.

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 consanguineous 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 SNV 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 $\leqslant 0.01$.

We found that the minimum n was 50 and 60 for WGS and WES, respectively, and defined each of these stretches spanning $\geqslant 100kb$, 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 $\geqslant 0.05$.

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 r^2 threshold (0.5).

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

To calculate *RLOD*, $DIDOH^3M^2$ 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

RoHs 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:

$C^i_{AB} = \frac{A^i_{max}+B^i_{min}}{2}$

$C^i_{BC} = \frac{B^i_{max}+C^i_{min}}{2}$ where $A^i_{max}, B^i_{min}, B^i_{max}, C^i_{min}$ are the minimum and maximum RoH sizes for the three classes for population *i*, respectively.

3. Results

3.1. RoH identification

To evaluate the performance of $DIDOH^3M^2$ 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_{Norm} $(10^5, 10^6)$ 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_{Norm} $(10^3, 10^4)$ and larger p_1 values (0.2, 0.3) are recommended.

As a further step, to test $DIDOH^3M^2$ in the analysis of real genotype 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 represent the proportion of heterozygous markers that our HMM tolerates in a homozygous region. Larger values of R_2 allows 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_2) in small segments (with a fraction of heterozygous markers smaller than R_2) 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 $R_2 = 1/100000$), $DIDOH^3M^2$ 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 ($DIDOH^3M^2$, 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 $DIDOH^3M^2$ 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 $DIDOH^3M^2$ 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) $DIDOH^3M^2$ 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 $DIDOH^3M^2$, 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:

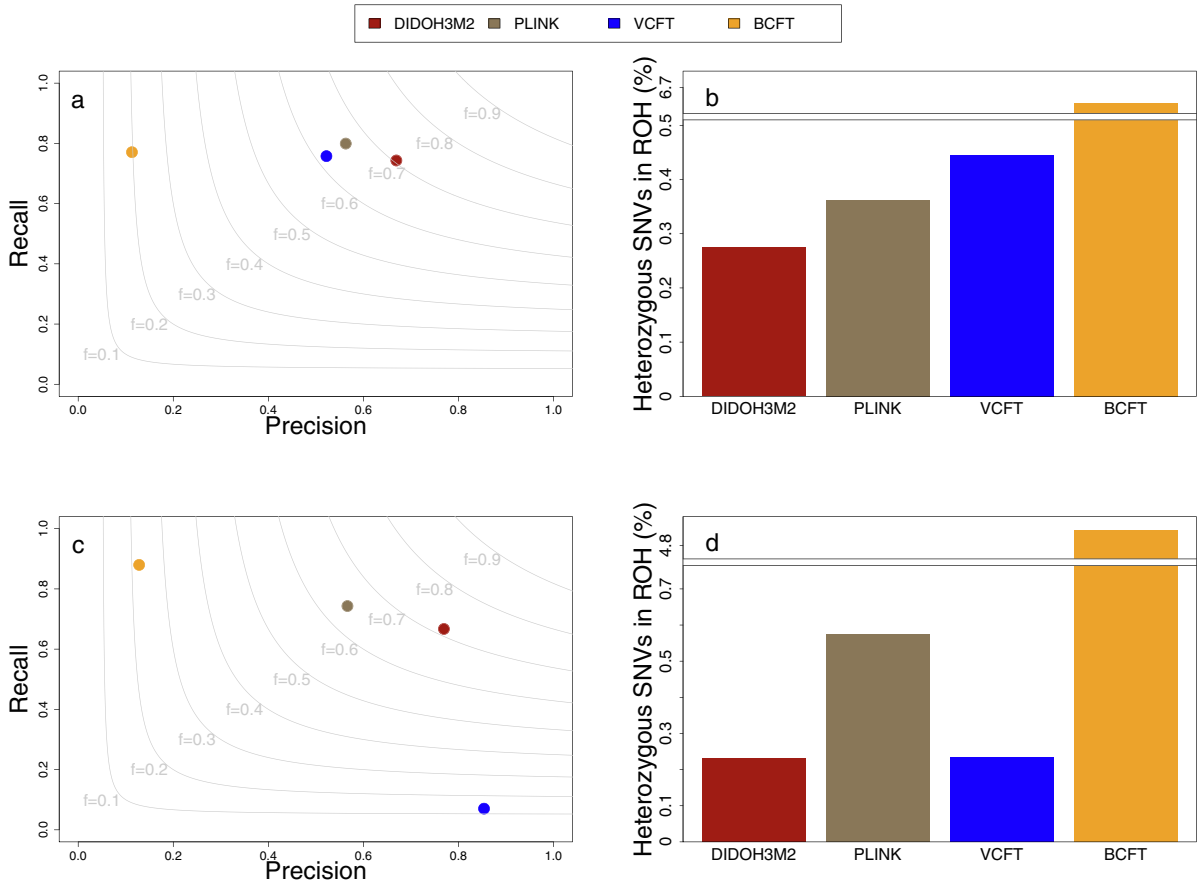

Fig. 2. Performance comparison between DIDOH3M2, PLINK, BCFtools and VCFtools on the WGS and WES data of the 200 individuals sequenced by the 1000 Genomes Project Consortium. Panels a and c report the results of the precision-recall analysis for WGS and WES data respectively. The bar plots of panels b and d report the fraction of heterozygous single nucleotide variants that belong to all ROHs detected by the four algorithms. The performance of the *DIDOH*^{3}*M*^{2} and PLINK algorithms have been reported for the parameter settings the gave the best results in terms of trade-off between precision and recall. For WGS data (panels a and b) DIDOH3M2 obtained the best results with $R_2 = 1/1000$, $R_1 = 2/100$, $p_1 = 0.1$, $p_2 = 0.1$, $P_{Norm} = 1$, $d_{Norm} = 100000$, while PLINK with heterozygote allowance 1, kb threshold 200 and SNP threshold 1/1000. For WES data (panels c and d) DIDOH3M2 obtained the best results with $R_2 = 1/10000$, $R_1 = 4/100$, $p1 = 0.1$, $p2 = 0.1$, $P_{Norm} = 1$, $d_{Norm} = 100000$, while PLINK with heterozygote allowance 1, kb threshold 100 and SNP threshold 1/1000.

$$RLOD = \sum_{k=1}^{N} log_{10}\left[\frac{Pr(\hat{g}_i)}{Pr(g_i)}\right] \quad (3)$$

where g_i is the observed homozygous genotype and $\hat{g}_i$ is the most likely genotype at SNV *i*. $\Pr(g_i)$ and $\Pr(\hat{g}_i)$, the probabilities of the observed and most likely genotypes, respectively, are calculated following Hardy-Weinberg's law. If *A* and *a* are the two possible alleles of any SNV *i* with frequency $f(A) = p$ and $f(a) = q$, and $p \geqslant q$, then $Pr(g_i) = p^2$ when g_i is *AA* and $Pr(g_i) = q^2$ when g_i is *aa*. On the other hand, $Pr(\hat{g}_i) = p^2$ when $p > 0.66$ and $Pr(\hat{g}_i) = 2pq$ when $p < 0.66$. When $Pr(g_i) = Pr(\hat{g}_i)$ at all the *N* homozygous SNVs within the RoH, $RLOD = 0$. Conversely when $Pr(g_i) < Pr(\hat{g}_i)$ at any of the *N* SNVs, $RLOD > 0$ and both the g_i likelihoods at any SNV and the number of SNVs with $Pr(g_i) < Pr(\hat{g}_i)$ affect *RLOD*. *RLOD* therefore inversely reflects the cumulative frequency of the SNV alleles found homozygous within the RoH.

To evaluate how well *RLOD* predicts autozygosity we created synthetic call sets of autosomal WES biallelic SNVs in the simulated offspring to consanguineous unions. We constructed 100 pedigrees with genealogy loops formed by one among the unions between 1st, 2nd and 3rd degree cousins (1C, 2C and 3C) with a single offspring ("index offspring") to the consanguineous parents (Supplemental Fig. 5).

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), 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) and created four *gF* 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 *gF* 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 chance. We took this into account when we performed *DIDOH*^{3}*M*^{2} analysis. First, in addition to the initial 100 Kb threshold, we introduced two more stringent thresholds representing the rough size limit below which RoH are likely under LD (500 Kb) and above which they are likely autozygous (1500 Kb) [25]. Second, we performed another *DIDOH*^{3}*M*^{2} analysis on a subset of the original marker map after removing SNVs with $LD \geqslant 0.5$ (Supplementary materials). 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 SNV 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 to SNVs outside the coding or the near-coding sequences. Finally, to measure the accuracy with which *RLOD* predicts autozygosity, we calculated precision and recall for each RoH-call set as follows:

- As precision, we calculated the fraction of RoH calls by *DIDOH*^{3}*M*^{2} that have any overlap with true autozygous RoH;
- As recall, we calculated the fraction of true autozygous RoH called by *DIDOH*^{3}*M*^{2}.

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 gF range decreases (Fig. 3).

For any *gF* 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 $LD \geqslant 0.5$. In WES data, progressive regression of the trade-off point is observed from higher to lower *gF* ranges (Fig. 3a–d), while in WGS *RLOD* improved performance even more markedly than in WES without apparent loss in accuracy in lower *gF* ranges (Fig. 3e–h).

DIDOH^3*M*^{2} 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 [37,7,30,2]. 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 *gF* 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*^{3}*M*^{2}. 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*^{3}*M*^{2} 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* outdistances 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 *gF* 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*^{3}*M*^{2}

To show the potentiality of our computational approach to explore genomic patterns of homozygosity in human populations

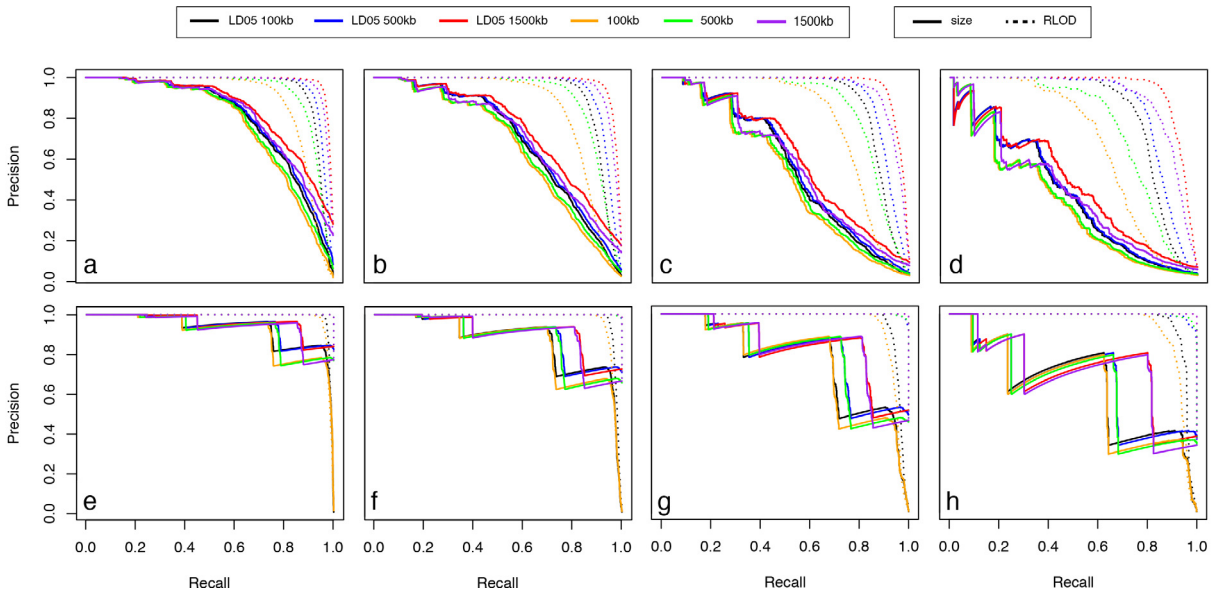

Fig. 3. Performance comparison between RLOD and RoH size to identify true autozygosity. Results of the analysis carried out in the simulated WES/WGS of offspring to consanguineous parents. Precision-recall plots of WES (panels a–d) and WGS (panels e–h) data are shown for the 4 different *gF* ranges from high (left) to low (right) inbreeding levels: F1: 0.066–1 (a, e); F2: 0.023–0.066 (b, f); F3: 0.00105–0.023 (c, g); F4: 0–0.0105 (d, h). RLOD and ROH size performances are depicted as dotted and continuous lines, respectively, while colors indicate different combinations of size and LD thresholds applied to the analysis.

Table 1
Statistical tests for assessing significance of disease-linked RoH ranking position in simulations for different consanguinity loops. McNemar test is used to assess significance for how many times the disease-linked RoH ranked as 1st by *RLOD* rather than size. Wilcoxon test is used to assess significance for how many times the disease-linked RoH ranks by RLOD higher or equal to that by size.

Statistical Test	allC	1C	2C	3C
McNemar test	0.01172	0.8026	0.06137	0.08086
wilcoxon test	0.000005885	0.2652	0.0003025	0.00001339

Table 2
Statistical tests for assessing significance of disease-linked RoH ranking position in simulations for different *gF* ranges. McNemar test is used to assess significance for how many times the disease-linked RoH ranked as 1st by *RLOD* rather than size. Wilcoxon test is used to assess significance for how many times the disease-linked RoH ranks by RLOD higher or equal to that by size.

Statistical Test	allF	F1	F2	F3	F4
McNemar test	0.01172	1	0.4795	0.505	0.00596
wilcoxon test	0.000005885	0.4226	0.05076	0.002154	9.08e-06

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 $p_2 = 0.1, p_1 = 0.1$, $d_{Norm} = 10^5, R_1$=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 $R = 0.34$,

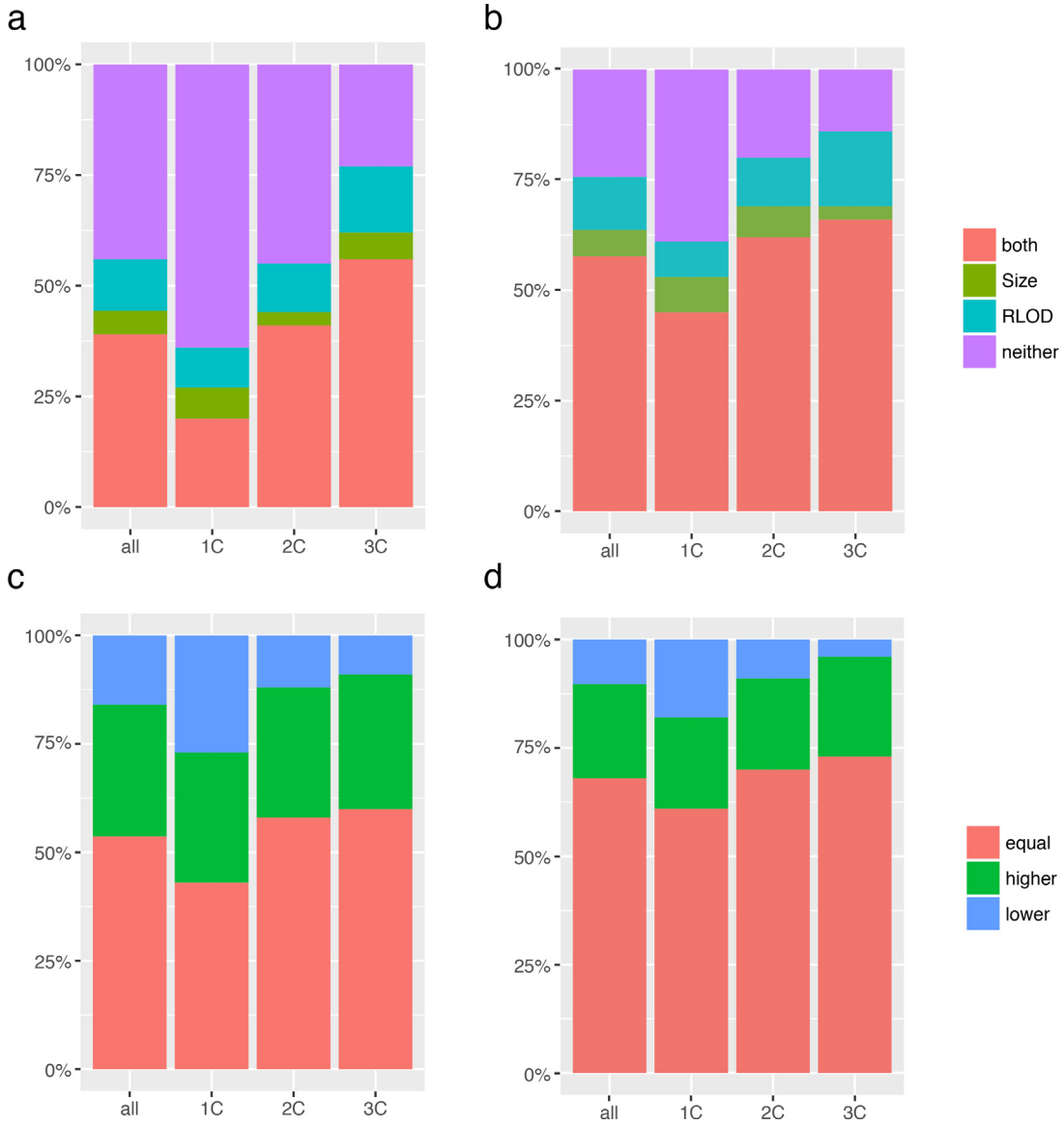

Fig. 4. Mutation-surrounding (ms) RoH prioritization by *RLOD* and RoH size in simulations. Results of the analysis carried out in the simulated WES (a and c) and WGS (b and d) of offspring to consanguineous parents are shown, as a whole (all) or split into the 3 genealogical loops: 1C (first cousins), 2C (second cousins), 3C (third cousins). Panels (a and b) reports the percentage of times the msRoH ranked as 1st among all the identified ROH by both or neither of the two measures, while panels (c and d) the percentage of times the msRoH ranked higher, equal or lower by *RLOD* than size.

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 athttps://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-

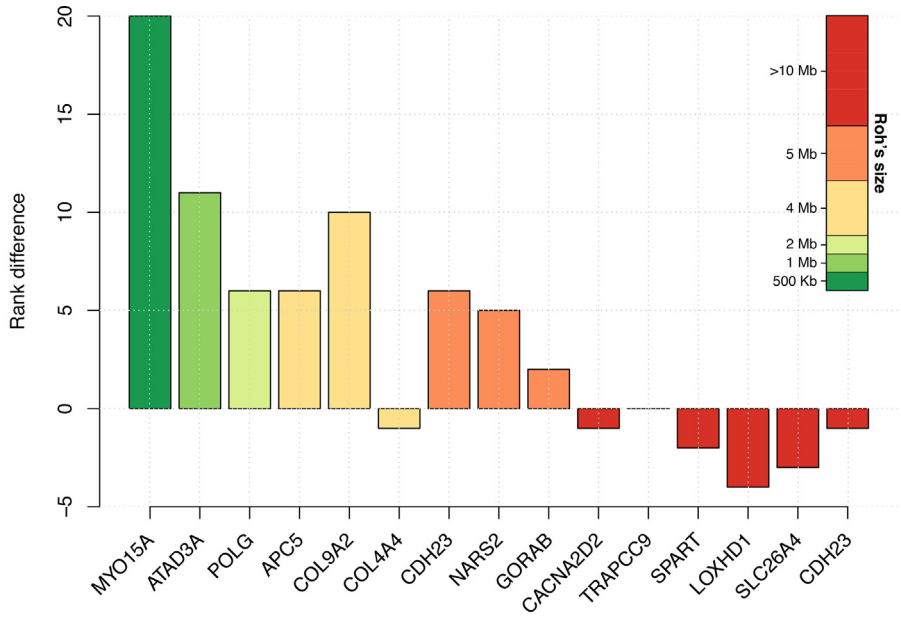

Fig. 5. Mutation-surrounding (ms) ROH prioritization by *RLOD* and RoH size in real WES. The difference in msRoH ranking position by *RLOD* and RoH size is shown for WES data of 15 patients in whom the homozygous disease-causing variant had been identified as part of a research project or of a diagnostic process. Colors represent different size ranges of the msRoH.

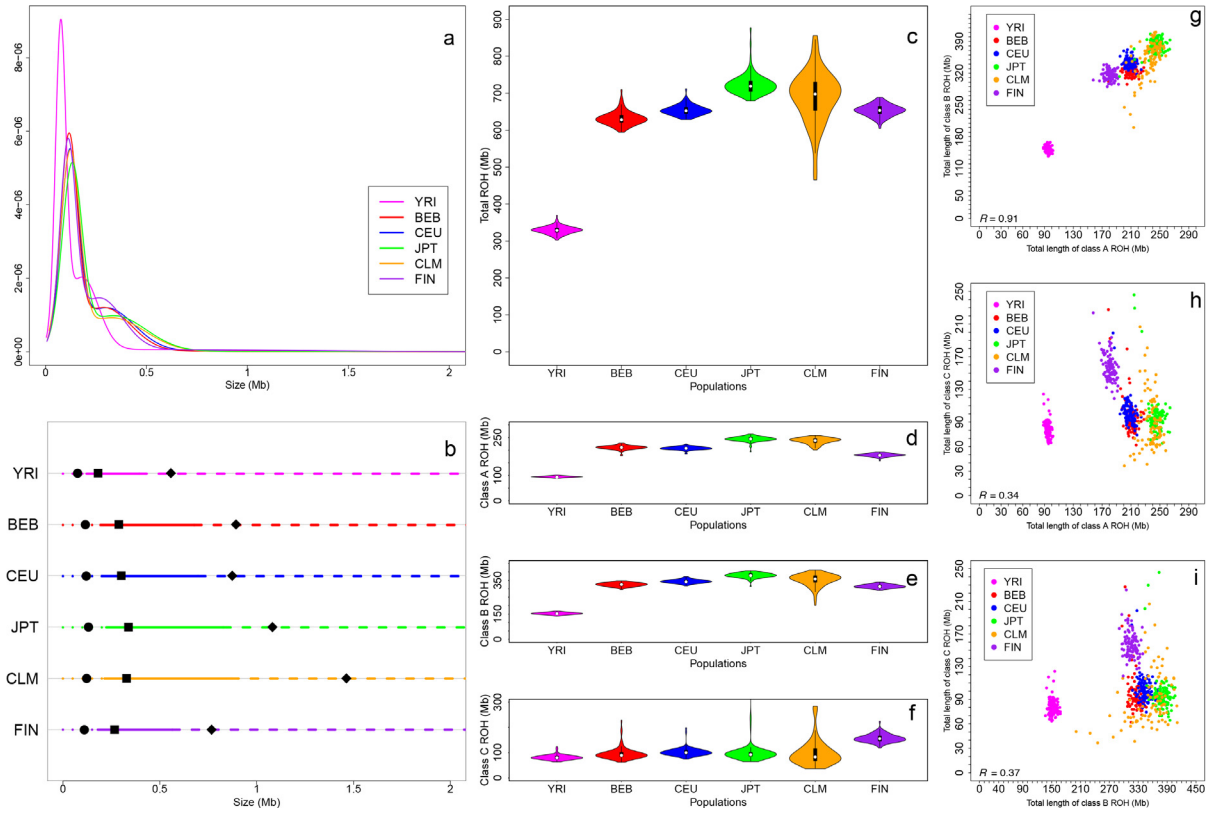

Fig. 6. Genomic patterns of homozygosity in 6 human populations. a Gaussian kernel density estimates of the ROH size distribution in each of the 6 populations (YRI, BEB, CEU, JPT, CLM and FIN). b Mean of each ROH class (A, B and C) and the boundaries between classes A and B and classes B and C for each 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*^{3}*M*^{2} overcomes the major limitation of our former tool *H*^{3}*M*^{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*^{3}*M*^{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*^{3}*M*^{2} outperforms popular existing tools in the accuracy to detect RoH from both WES and WGS. In particular, as the previous *H*^{3}*M*^{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 msRoH 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 msRoH. 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 resulted to match 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 *gF* 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 msRoH for the sake of autozygosity prediction accuracy.

Since users may perform *DIDOH*^{3}*M*^{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*^{3}*M*^{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*^{3}*M*^{2} to prioritize the msRoH 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 *gF* values across pedigrees with identical genealogy loops, we introduced here the median *gF* 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 msRoH for any genealogy loop, whether it is of the 1st or of the higher ranking position. Looking at *gF* 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 msRoH prioritization for any other *gF* 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 msRoHs 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*^{3}*M*^{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 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/ftp/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 competing 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 Semeraro:** Formal analysis. **Giulia Carangelo:** Formal analysis. **Flavia Palombo:** Formal analysis. **Giovanni Romeo:** Supervision. **Marco Seri:** Supervision. **Tommaso Pippucci:** Conceptualization, Methodology, Software, Writing - original draft.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, athttps://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, Marth GT, 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 autozygosity mapping of human recessive diseases. Hum Genet 2013;132 (11):1197–211.

[3] Alsalem AB, Halees AS, Anazi S, Alshamekh S, Alkuraya FS. Autozygome sequencing expands the horizon of human knockout research and provides novel insights into human phenotypic variation. PLoS Genet 2013;9(12): e1004030.

[4] Bittles AH, Black ML. Evolution in health and medicine Sackler colloquium: consanguinity, human evolution, and complex diseases. Proc Natl Acad Sci USA 2010;26(107 Suppl 1):1779–86.

[5] Belkadi A, Pedergnana V, Cobat A, Itan Y, Vincent QB, Abhyankar A, Shang L, El Baghdadi J, Bousfiha A; Exome/Array Consortium, Alcais A, Boisson B, Casanova JL, Abel L. Whole-exome sequencing to analyze population structure, parental inbreeding, and familial linkage. Proc Natl Acad Sci USA 14;113 (24);2016:6713–8..

[6] Broman KW, Weber JL. Long homozygous chromosomal segments in reference families from the centre d'Etude du polymorphisme humain. Am J Hum Genet 1999;65(6):1493–500.

[7] Carr IM, Flintoff KJ, Taylor GR, Markham AF, Bonthron DT. Interactive visual analysis of SNP data for rapid autozygosity mapping in consanguineous families. Hum Mutat 2006;27(10):1041–6.

[8] Ceballos FC, Joshi PK, Clark DW, Ramsay M, Wilson JF. Runs of homozygosity: windows into population history and trait architecture. Nat Rev Genet 2018;19 (4):220–34.

[9] Christofidou P, Nelson CP, Nikpay M, Qu L, Li M, Loley C, Debiec R, Braund PS, Denniff M, Charchar FJ, Arjo AR, Trgout DA, Goodall AH, Cambien F, Ouwehand WH, Roberts R, Schunkert H, Hengstenberg C, Reilly MP, Erdmann J, McPherson R, Knig IR, Thompson JR, Samani NJ, Tomaszewski M. Runs of homozygosity: association with coronary artery disease and gene expression in monocytes and macrophages. Am J Hum Genet 2015;97(2):228–37.

[10] Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R. 1000 Genomes Project Analysis Group. The variant call format and VCFtools. Bioinformatics 27 (15);2011:2156–8..

[11] Gazal S, Sahbatou M, Babron MC, Gnin E, Leutenegger AL. FSuite: exploiting inbreeding in dense SNP chip and exome data. Bioinformatics 2014;30 (13):1940–1.

[12] Jalkh N, Sahbatou M, Chouery E, Megarbane A, Leutenegger AL, Serre JL. Genome-wide inbreeding estimation within Lebanese communities using SNP arrays. Eur J Hum Genet 2015;23(10):1434.

[13] Joshi PK, Esko T, Mattsson H, Eklund N, Gandin I, et al. Directional dominance on stature and cognition in diverse human populations. Nature 2015;523 (7561):459–62.

[14] Harel T, Yoon WH, Garone C, Gu S, et al. Recurrent de novo and biallelic variation of ATAD3A, encoding a mitochondrial membrane protein, results in distinct neurological syndromes. Am J Hum Genet 2016;99(4):831–45.

[15] Hildebrandt F, Heeringa SF, Rschendorf F, Attanasio M, Nrnberg G, Becker C, Seelow D, Huebner N, Chernin G, Vlangos CN, Zhou W, O'Toole JF, Hoskins BE, Wolf MT, Hinkes BG, Chaib H, Ashraf S, Schoeb DS, Ovunc B, Allen SJ, Vega-Warner V, Wise E, Harville HM, Lyons RH, Washburn J, Macdonald J, Nrnberg P, Otto EA. A systematic approach to mapping recessive disease genes in individuals from outbred populations. PLoS Genet 2009;5(1):e1000353.

[16] Keller MC, Simonson MA, Ripke S, Neale BM, Gejman PV, Howrigan DP, Lee SH, Lencz T, Levinson DF, Sullivan PF. Schizophrenia Psychiatric Genome-Wide Association Study Consortium. Runs of homozygosity implicate autozygosity as a schizophrenia risk factor. PLoS Genet 2012;8(4):e1002656.

[17] Kirin M, McQuillan R, Franklin CS, Campbell H, McKeigue PM, Wilson JF. Genomic runs of homozygosity record population history and consanguinity. PLoS One 2010;5(11):e13996.

[18] Lencz T, Lambert C, DeRosse P, Burdick KE, Morgan TV, Kane JM, Kucherlapati R, Malhotra AK. Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. PNAS 2007;104:19942–7.

[19] Magi A, Tattini L, Palombo F, Benelli M, Gialluisi A, Giusti B, Abbate R, Seri M, Gensini GF, Romeo G, Pippucci T. H3M2: detection of runs of homozygosity from whole-exome sequencing data. Bioinformatics 2014;30(20):2852–9.

[20] 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(3):359–72.

[21] Nalls MA, Simon-Sanchez J, Gibbs JR, Paisan-Ruiz C, Bras JT, Tanaka T, Matarin M, Scholz S, Weitz C, Harris TB, Ferrucci L, Hardy J, Singleton AB. Measures of autozygosity in decline: globalization, urbanization, and its implications for medical genetics. PLoS Genet 2009;5(3):e1000415.

[22] Nalls MA, Guerreiro RJ, Simon-Sanchez J, Bras JT, Traynor BJ, Gibbs JR, Launer L, Hardy J, Singleton AB. Extended tracts of homozygosity identify novel candidate genes associated with late-onset Alzheimer's disease. Neurogenetics 2009;10(3):183–90.

[23] 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.

[24] Palombo F, Al-Wardy N, Ruscone GA, Oppo M, Kindi MN, Angius A, Al Lamki K, Girotto G, Giangregorio T, Benelli M, Magi A, Seri M, Gasparini P, Cucca F, Sazzini M, Al Khabori M, Pippucci T, Romeo G. A novel founder MYO15A frameshift duplication is the major cause of genetic hearing loss in Oman. J Hum Genet 62(2);2017:259-264..

[25] 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;10:275–92.

[26] Pippucci T, Magi A, Gialluisi A, Romeo G. Detection of runs of homozygosity from whole exome sequencing data: state of the art and perspectives for clinical, population and epidemiological studies. Hum Hered 2014;77(1–4):63–72.

[27] Pippucci T, Parmeggiani A, Palombo F, Maresca A, Angius A, Crisponi L, Cucca F, Liguori 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..

[28] 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;81:559–75.

[29] Scott EM, Halees A, Itan Y, Spencer EG, He Y, Azab MA, Gabriel SB, Belkadi A, Boisson B, Abel L, Clark AG. Greater Middle East Variome Consortium, Alkuraya FS, Casanova JL, Gleeson JG. Characterization of Greater Middle Eastern genetic variation for enhanced disease gene discovery. Nat Genet 2016;48(9):1071–6.

[30] Seelow D, Schuelke M, Hildebrandt F, Nrnberg P. HomozygosityMapper–an interactive approach to homozygosity mapping. Nucleic Acids Res 37(Web Server issue):2009;W593–9..

[31] Simon M, Richard EM, Wang X, Shahzad M, Huang VH, Qaiser TA, Potluri P, Mahl SE, Davila A, Nazli S, Hancock S, Yu M, Gargus J, Chang R, Al-Sheqaih N, Newman WG, Abdenur J, Starr A, Hegde R, Dorn T, Busch A, Park E, Wu J, Schwenzer H, Flierl A, Florentz C, Sissler M, Khan SN, Li R, Guan MX, Friedman TB, Wu DK, Procaccio V, Riazuddin S, Wallace DC, Ahmed ZM, Huang T, Riazuddin S. Mutations of human NARS2, encoding the mitochondrial asparaginyl-tRNA synthetase, cause nonsyndromic deafness and Leigh syndrome. PLoS Genet 2015;25:11(3):e1005097.

[32] Simon-Sanchez J, Kilarski LL, Nalls 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, Holmstrm 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, Zllner 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 Latter 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 PY, 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.