Supplemental Materials to: "AUDACITY: a comprehensive approach for the detection and classification of Runs of Homozygosity in medical and population genomics."

Alberto Magi[1], Tania Giangregorio[2], Roberto Semeraro[3], Giulia Carangelo[3], Flavia Palombo[2], Giovanni Romeo[2], Marco Seri[2,4] and Tommaso Pippucci[4].

[1]Department of Information Engineering, University of Florence, Florence, Italy, [2]Department of Medical and Surgical Sciences, University of Bologna, Bologna, Italy, [3]Department of Experimental and Clinical Medicine, University of Florence, Florence, Italy, [4]Medical Genetics Unit, Sant'Orsola-Malpighi University Hospital, Bologna, Italy.

1000 Genomes Project

The 1000 Genomes Project (1000 Genomes Project Consortium, 2009) is the first project to sequence the genomes of a large number of people, to provide a comprehensive resource on human genetic variation.
In phase 1, the 1000 genomes project consortium, by combining low-coverage whole-genome sequencing (WGS) and high-coverage whole-exome sequencing (WES) of 1092 individuals from 14 populations drawn from Europe (TSI,Toscani in Italy; IBS,Iberian populations in Spain; GBR, British in England and Scotland; CEU, Utah residents (CEPH) with Northern and Western European ancestry; FIN, Finnish in Finland), East Asia (JPT, Japanese in Tokyo, Japan; CHB, Han Chinese in Beijing, China CHS,Han Chinese South), sub-Saharan Africa (YRI, Yoruba in Ibadan, Nigeria; LWK, Luhya in Webuye, Kenya; ASW,African Ancestry in Southwest) and the Americas (MXL, Mexican Ancestry in Los Angeles, California; CLM,Colombian in Medellin, Colombia; PUR,Puerto Rican in Puerto Rico), has identified around 38 million single nucleotide polymorphic positions, 1.4 million short insertions and deletions and more than 14,000 larger deletions (1000 Genomes Project Consortium, 2009).
In phase 3, the 1000 genomes project consortium, by combining low-coverage whole-genome sequencing (WGS) and high-coverage whole-exome sequencing (WES) of 2504 individuals from 26 populations from Europe (IBS, Iberian populations in Spain; TSI, Toscani in Italy; GBR, British in England and Scotland; CEU, Utah residents (CEPH) with Northern and Western European ancestry; FIN, Finnish in Finland), East Asia (CDX, Chinese Dai in Xishuangbanna, China; JPT, Japanese in Tokyo, Japan; CHB, Han Chinese in Beijing, China; CHS, Han Chinese South; KHV, Kinh in Ho Chi Minh City, Vietnam), South Asia (PJL, Punjabi in Lahore, Pakistan; STU, Sri Lankan Tamil in the UK; BEB, Bengali in Bangladesh; GIH, Gujarati Indian in Houston, TX; ITU, Indian Telugu in the UK), Africa (GWD, Gambian in Western Division, The Gambia - Mandinka; MSL, Mende in Sierra Leone; ESN, Esan in Nigeria; YRI, Yoruba in Ibadan, Nigeria; LWK, Luhya in Webuye, Kenya; ASW, African Ancestry in Southwest US; YRI, Yoruba in Ibadan, Nigeria; ACB, African Caribbean in Barbados) and the Americas (MXL, Mexican Ancestry in Los Angeles, California; CLM, Colombian in Medellin, Colombia; PEL, Peruvian in Lima, Peru; PUR, Puerto Rican in Puerto Rico).
All the variant calls of Phase 1 and Phase 3 are stored in VCF format and freely available at the 1000 genome project FTP site: `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/integrated_call_sets/` for Phase 1 and `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/` for Phase 3.
For both Phase 1 and Phase 3 datasets, we downloaded Variant calls from the 1000 genome project FTP site, we filtered out structural variants and small InDels and multiallelic SNVs using the VCFTools. For synthetic chromosomes generation and real data analysis we used VCF files from Phase 1 variant calls, while for "Characterization of RoH across worldwide populations" we used Phase 3 variant calls.

Synthetic Validations

To test the ability of our $DIDOH^3M^2$ to detect RoH of different sizes and constituted by different number of SNPs as a function of the distance between consecutive markers and the genotype probability of each SNP call, we performed an intensive simulation based on synthetic data. To this end, we generated synthetic chromosomes by using the genotype calls of bi-allelic SNPs of ten individuals of caucasian ancestry (CEU) sequenced by 1000GP consortium during Phase 1.
Each synthetic chromosome was generated as a stretch of 10,000 polymorphic positions in which:

- homozygous segments were simulated as N consecutive genotypes sampled from polymorphic positions predicted as homozygous by the 1000GP.
- non-homozygous segments were simulated by sampling $(10000 - N)$ genotypes from all the polymorphic positions of the Phase 1 individuals. Heterozygous segments were imposed to have SNPs in a heterozygous/homozygous ratio of 5 : 100. To this end, we sampled one heterozygous genotype every twenty homozygous SNPs.

The 5 : 100 ratio was imposed to simulate at best the actual heterozygous/homozygous proportion and to prevent the emergence of false positive homozygous segments.
In order to reproduce the complex architecture and distribution of homozygous and non-homozygous regions, we generated distances between adjacent SNPs as follows:

- The distances between consecutive SNPs in non-homozygous regions are sampled from the distribution of the distances between adjacent polymorphic positions in the human genome.
- The distances between adjacent polymorphic positions in homozygous regions are fixed to a predefined distance D.

Finally, to simulate calling errors and biases of real sequencing data, for each marker of the synthetic chromosomes we randomly sampled genotype likelihood from the calls generated by the 1000GP Phase 1 data.
We performed simulations with $N = (100, 200, 500, 1000, 2000)$ and $D =$ (10 bp, 100 bp, 1 Kb, 10 Kb, 100 Kb) and for each combination of N and D we generated 100 synthetic chromosomes: all the synthetic datasets were analyzed by using different values of the parameters d_{Norm} ($D_{Norm} = 10^3$, 10^4, 10^5, 10^6), P_{Norm} (from 0.5 to 5 by 0.5), p_1 and p_2 (from 0.05 to 0.8 by 0.05), R_1 (2/100, 3/100, 4/100, 5/100) and R_2 (1/100, 1/1000, 1/10000, 1/100000 and 1/1000000).
To evaluate the performance of $DIDOH^3M^2$ for different parameter settings, we calculated sensitivity (true positive rate, TPR) and specificity (1-FPR, false positive rate). TPR was defined as the number of markers inside the synthetic RoH called by our approach as homozygous divided by the total number of markers inside the synthetic RoH. FPR was defined as the number of markers outside the synthetic RoH called by $DIDOH^3M^2$ as homozygous divided by the total number of markers outside the synthetic ROH.
Supplemental Figure 1.a-b show that R_2 has little effect on sensitivity and specificity (for values of $R_2 \leq 1/10000$ sensitivity and specificity only depend on R_1). On the other hand, R_1 has strong effect on the global performance of our algorithm: the larger (smaller) R_1 the larger (smaller) is sensitivity (specificity).
Concerning parameters p_1 and p_2, the results of Supplemental Figure 1.c-d are in accordance with those obtained for the classical H^3M^2 algorithm (Magi *et al.*, 2014), where the larger p_2 the smaller the range of p_1 values that ensure high sensitivity (Supplemental Figure 1.c). In particular, when $p_2 = 0.1$, almost any value of p_1 guarantees the best performance in term of sensitivity. On the other hand, for values of p_1 larger than 0.4 the specificity of our method drastically decreases (Supplemental Figure 1.d). Finally, Supplemental Figure 1.e-h show that also D_{Norm} and P_{Norm} have strong effect on global performance of $DIDOH^3M^2$. The larger D_{Norm} the smaller (larger) is sensitivity (specificity), on the other hand while increasing the value of P_{Norm} reduces the number of false positives, it also reduces the sensitivity of our algorithm, and the value of P_{Norm} that gives the best trade-off between sensitivity and specificity is 1.
As a further test, to evaluate the capability of $DIDOH^3M^2$ to detect RoH of different size and comprising different number of SNPs, we calculated True Positive Rate (TPR) and False Positive Rate (FPR) as follows: a detected ROH is considered a true positive (TP) if has any overlap with a synthetic RoH, while it is

considered a false positive (FP) if it has no overlap with a synthetic RoH. These analyses were performed by setting $p_2 = 0.1$, $p_1 \in [0.05, 0.3]$, $d_{Norm} = (10^3, 10^4, 10^5, 10^6)$, $R_1 = (2/100, 3/100, 4/100, 5/100)$ and $R_2 = (1/100, 1/1000, 1/10000, 1/100000, 1/1000000)$. As expected, the larger the number of SNPs falling within a given ROH, the higher the probability to correctly identify the homozygous segment (Supplemental Figure 2.a.d.). Similarly, the larger the distance between adjacent positions, the higher the probability to call the region as homozygous (Supplemental Figure 2.b.e.).
A detailed analysis of Supplemental Figure 2.a-c reveals that R_1 parameter has strong effect on the the capability of of $DIDOH^3M^2$ to detect ROH made of small number of markers. High values of the parameter R_1 increase the resolution of the algorithm but also the total number of false positive events. On the other hands, parameter R_2 has little effect on the performance of our method, however setting R_2 $ge1/10000$ reduces the total number of false positive events and increase the capability of our method in detecting small RoHs (Supplemental Figure 2.d-f).
The results of these simulations (SupplementalFiguresDIDOH3M2.pdf) also show that parameter p_1 regulates the resolution of $DIDOH^3M^2$, and increasing p_1 allows to detect smaller ROHs at the expenses of an higher number of FP events.
The Figures of SupplementalFiguresDIDOH3M2.pdf file also show that the parameter D_{Norm} rules the capability of $DIDOH^3M^2$ to detect homozygous segments characterized by variable SNP densities. When D_{Norm} is set to large values $(10^5, 10^6)$, $DIDOH^3M^2$ is not able to detect homozygous segments made of even hundreds of densely distributed SNPs and increasing p_1 has poor effect on the resolution of the algorithm. On the contrary, when D_{Norm} is set to small values $(10^3, 10^4)$, $DIDOH^3M^2$ becomes able to detect homozygous segments made of densely distributed SNPS, and increasing p_1 has a relevant effect on resolution.
Taken as a whole, these results suggest 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). On the other hand, to increase the resolution of the algorithm and detect small ROHs, small D_{Norm} (10^3, 10^4) and large p_1 values (0.2, 0.3) are recommended.

PLINK, VCFtools and BCFtools settings on WES and WGS data

PLINK

PLINK implements an algorithm (`--homozyg` option) that scan each chromosome by moving a fixed size window along the whole length of the genome in search of stretches of consecutive homozygous SNPs. A given SNP is considered to potentially be in an ROH by calculating the proportion of completely homozygous windows that encompass that SNP. If this proportion is higher than a defined threshold, the SNP is designated as being in a ROH. The `--homozyg` option of PLINK allows to set several parameters that include: Sliding window size in SNPs, Sliding window size in kb, Heterozygote allowance and Window threshold to call a RoH.
To run PLINK on the 200 individuals, we converted the information stored in VCF format into MAP and PED formats usable by PLINK using VCFtools. For the detection of homozygous segments we used the `--homozyg` option specifying the following parameter settings for both WES and WGS datasets:

- Heterozygote allowance (`--homozyg-window-het`) 0/1
- Sliding window size in SNPs (`--homozyg-window-snp`) 100
- Window threshold to call a RoH (`?homozyg-window-threshold`) 0.05
- Sliding window size in kb (`--homozyg-window-kb`) 50
- Minimum SNP density to call a RoH (`--homozyg-density`) 50
- Maximum gap before splitting RoH (`--homozyg-gap`) 1000
- Kb threshold to call a RoH (`--homozyg-kb`) 50/100/200

For the following parameter, we specified different parameter settings for WES and WGS datasets:

- SNP threshold to call a RoH (`--homozyg-snp`) 50/250/500 for WES dataset, 500/1000/2000 for WGS dataset

BCFtools

To detect RoH, BCFtools/RoH uses a 2 state hidden Markov model (HMM). The 2 hidden states represent extended homozygosity (H) and non-homozygosity (N) within the sample. The emission probability is determined by the Hardy-Weinberg model and depends on the minor allele frequency of any marker and on the genotype likelihoods provided by the variant calling algorithm.
The HMM takes as input the genotype data, where genotypes are represented by RR for a homozygous site matching the reference, RA for a heterozygous site and AA for a homozygous alternate (non-reference) site. H segments can only include RR and AA sites, while N tracts can include sites of any genotype.
BCFtools/RoH option allows to set the 2 transition probabilities from autozygous to Hardy-Weinberg (–az-to-hw) state and from Hardy-Weinberg to autozygous state (–hw-to-az) as starting parameters. For comparison analyses, after testing several combinations of transition probabilities we selected the configuration achieving the best results in terms of precision and recall: –hw-to-az 0.01, –az-to-hw 0.01. BCFtools was run on the 200 individuals separately by using the allele frequency estimated by the 1000GP consortium and stored in the AF tag of the INFO field and ignoring genotype likelihoods (–GTs-only 0).

VCFtools

The –LRoH option of VCFtools (`http://vcftools.sourceforge.net/man_latest.html`) implements the HMM algorithm described in (Auton *et al.*, 2009) that allows to detect long Runs of Homozygosity (LRoH).
The HMM consists of two states for each SNP, which represent LROH or heterozygous region respectively. For each state, the emission probabilities at each SNP are dependent on the probability of observing a heterozygote (based on the heterozygosity of the SNP within the population) and the estimated rate of genotyping error. Transition probabilities between the two states are a function of the per-generation recombination rate between SNPs and the (assumed) number of generations since a common ancestor of the two chromosomes. A LROH is called when the HMM reports the homozygous state as being the most likely state in a region of at least 1cM and containing at least 50 SNPs with a minimum minor allele frequency of 5%.
The –LRoH option of VCFtools does not allow for any parameter setting.

Real data analysis

In order to test the performance of our method for the identification of homozygous segments on real data, we applied $DIDOH^3M^2$ to the WGS and WES genotype data of 200 individuals (50 CEU of European ancestry, 50 YRI of African ancestry, 50 PUR of American ancestry and 50 CHS of Asian ancestry) sequenced by 1000GP consortium during Phase 1 by using the following parameter settings: $p_2 = 0.1$, $p_1 = 0.1$, $d_{Norm} = 10^5$, $\mathrm{R}_1 = (1/100, 2/100, 3/100, 4/100, 5/100)$ and $R_2 = (1/1000, 1/10000, 1/100000)$.
As a first step, we studied the RoHs identified by our method in terms of their cumulative global size and number for both WGS and WES data. 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 Figure 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 hands, 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.
Subsequently, we compared the results of $DIDOH^3M^2$ with those obtained by the other three tools that use genotype calls for RoHs identification: PLINK, the RoH option of BCFTools and the LROH option of VCFTools. To allow for a comprehensive evaluation of the performance of PLINK, we defined six different parameter configurations for this tool (see "PLINK, VCFtools and BCFtools settings on WES and WGS data" section).
VCFTools identified an average of 500 Mb of homozygous segments (25,000 RoHs) for WGS and less than 50 Mb (500 RoHs) for WES data. Although we tested several parameters configurations, the results obtained by BCFTools are completely unreliable, since it detected more than 2 Gb of homozygous segments with both WGS and WES genotype data.
By using the most conservative configuration (`--homozyg-window-het` 0 and `--homozyg-window-threshold` 200 kb), PLINK (`--homozyg-snp` 2000 for WGS and `--homozyg-snp` 500 for WES) detected an average of around 10 Mb (tens of ROHs) of homozygous segments for WGS and 50 Mb for WES (few RoHs). On the other hand, using the less stringent configuration (`--homozyg-window-het` 1 `--homozyg-window-threshold` 50 kb), PLINK (`--homozyg-snp` 500 for WGS and `--homozyg-snp` 50 for WES) detected an average of 600 Mb for WGS and more than 1 Gb for WES data (thousands of RoHs).
As a further step, to evaluate $DIDOH^3M^2$ ability to identify ROH from WES and WGS data and to compare its performance with respect to the other three state of the art methods, we generated a gold standard dataset of RoHs by using the genotype calls generated by the 1000GP consortium for the aforementioned 200 individuals. For WGS we considered the entire map of biallelic single nucleotide variants discovered by the 1000GP (around 38 millions of markers), while for WES we included only the around 1.5 millions of SNVs that belong to coding sequence of the genome (see methods of main manuscript for more details).
For both WGS and WES experimental design, we considered as genuine RoHs all the regions larger than 100 kb and containing at least 200 consecutive markers in homozygous state. To test the performance of the four methods we calculated precision and recall in the following manner:

- To calculate precision, we considered all the polymorphic positions called in ROHs by each of the four methods and we then calculated the fraction of these positions that were called as homozygous also in the gold standard datasets.
- To calculate recall, we considered all the polymorphic position called in ROH in the gold standard dataset and we then calculated the fraction of these positions called as homozygous by each of the four state of the art methods.

The plots of panels (a) and (c) of Figure 2 of the main manuscript show that the performance of $DIDOH^3M^2$ is mainly governed by changes in parameter R_1. Setting R_1=5/100 and 4/100 gives high recall rate at the expenses of precision, while using R_1=1/100 improves precision and drastically decrease recall. On the other hands, R_2 has little effect on both precision and recall. The combination of parameters that ensure the best trade-off between precision and recall is different for WGS and WES experimental design: for WES we found the best setting as R_1=2/100 and R_2=1/1000, for WGS R_1=4/100 and R_2=1/1000.
As previously reported in (Magi *et al.*, 2014), also the performance of PLINK is profoundly altered by changes in parameter configurations. The high recall rate reached by PLINK with less stringent parameter settings (`--homozyg-window-het` 1 `--homozyg-window-threshold` 50 kb) is obtained paying a tremendous cost in terms of precision. On the other hand, attempts to improve precision adopting conservative parameter configurations (`--homozyg-window-het` 0 and `--homozyg-window-threshold` 200 kb), lead to a drastic deterioration of recall rates.
Regarding the two other tools studied in this paper, although VCFTools obtained good performance for WES data, all the other simulations clearly show that the homozygosity algorithms at the base of these two software packages are not well suited for this kind of analysis and need computational improvements.
Finally, in order to study the accuracy of the four algorithms, we examined the proportion of heterozygous

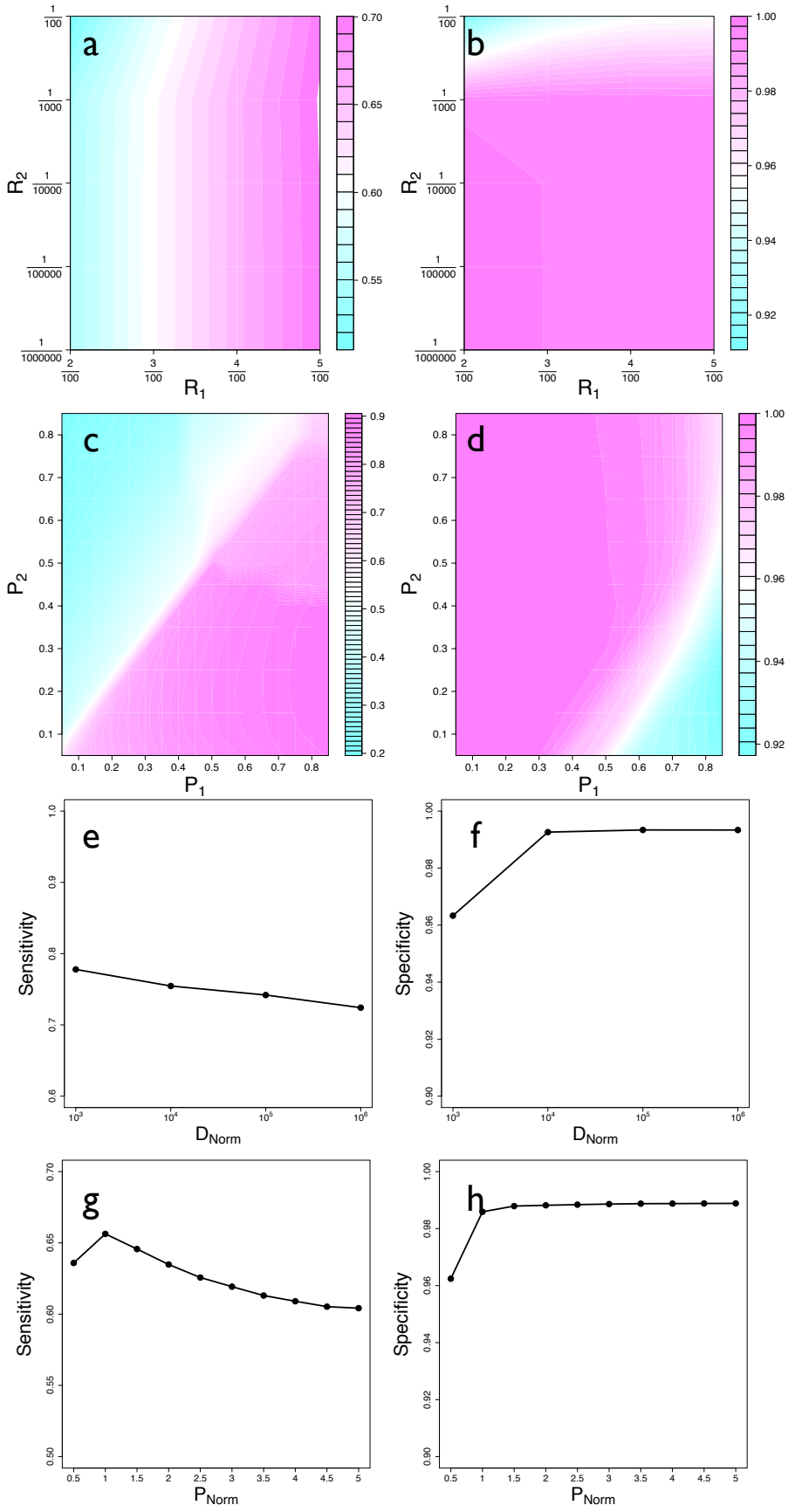
variants within the ROHs detected by each of the four methods and we found that the ROHs detected by $DIDOH^3M^2$ are characterized by the smallest fraction of heterozygous variants.

References

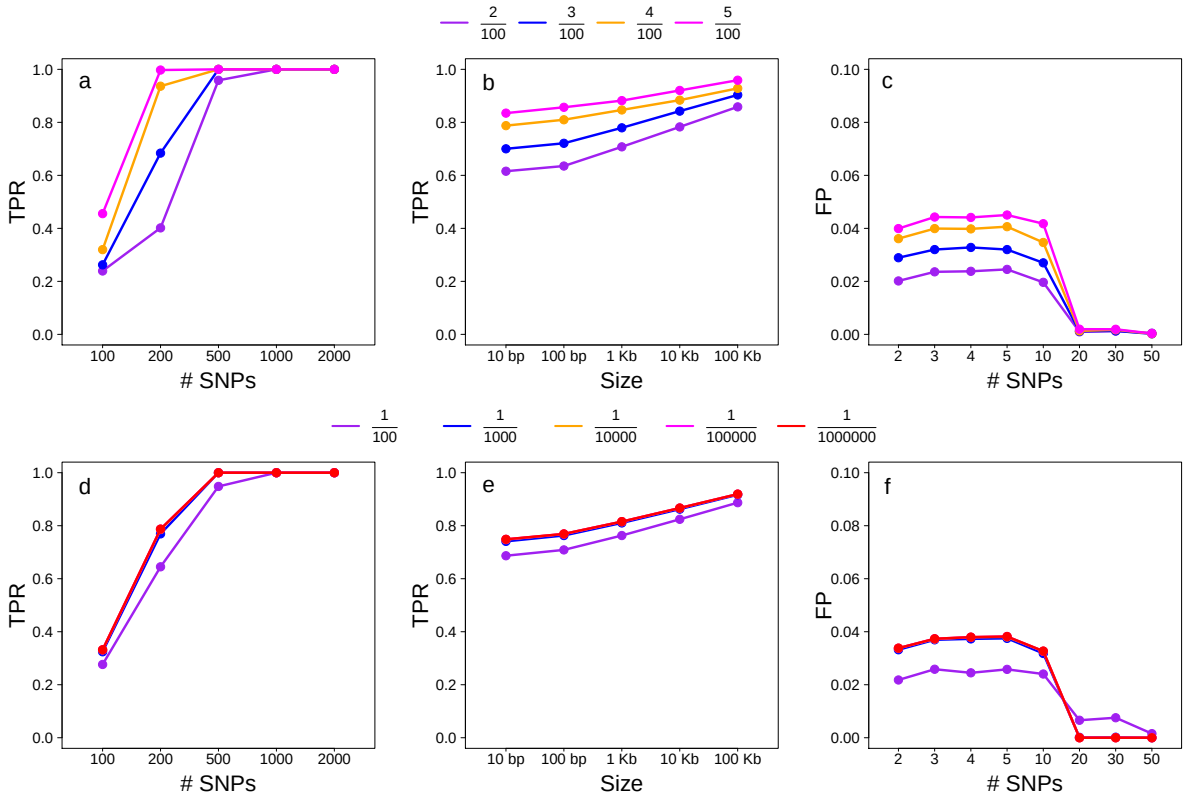
1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature.* **491(7422)**:56-65.

Auton A, Bryc K, Boyko AR, Lohmueller KE, Novembre J, Reynolds A, Indap A, Wright MH, Degenhardt JD, Gutenkunst RN, King KS, Nelson MR, Bustamante CD. Global distribution of genomic diversity underscores rich complex history of continental human populations. Genome Res. 2009 May;19(5):795-803.

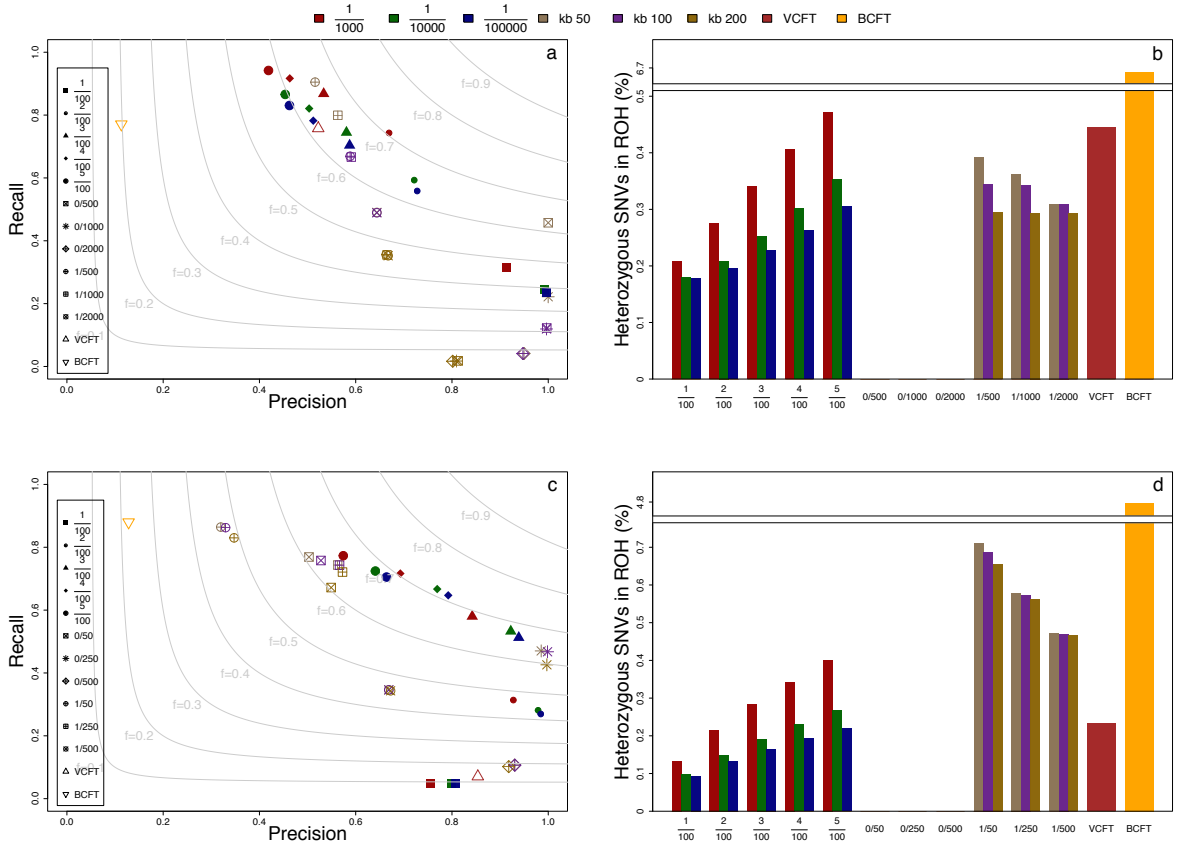
Magi A, Tattini L, Palombo F, Benelli M, Gialluisi A, Giusti B, Abbate R, Seri M, Gensini GF, Romeo G, Pippucci T. H3M2: detection of runs of homozygosity from whole-exome sequencing data. Bioinformatics. 2014 Oct 15;30(20):2852-9.



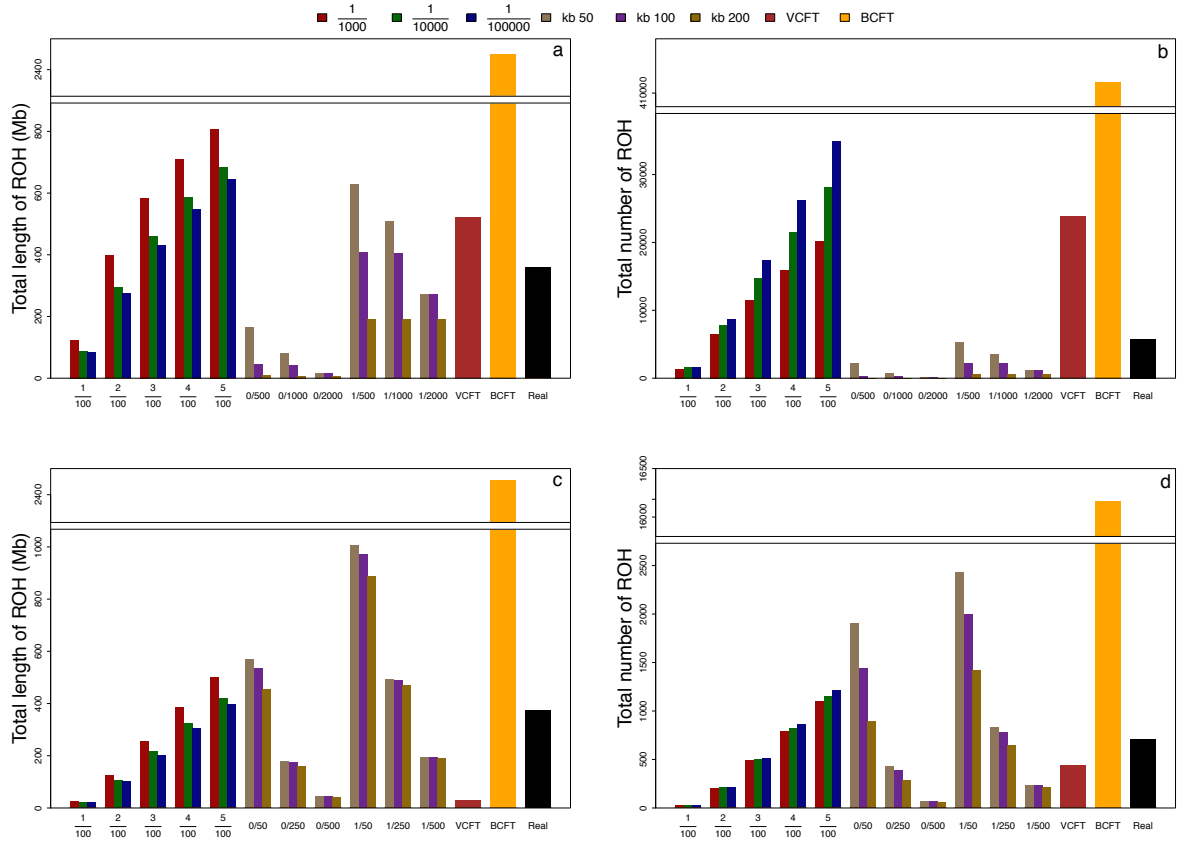
Supplemental Figure 1: $DIDOH^3M^2$ algorithm and parameter settings on synthetic chromosomes. The contourplots of panels a and b show the sensitivity and specificity of $DIDOH^3M^2$ for different combinations of values of R_1 and R_2 parameters. Panel c and d show the sensitivity and specificity of $DIDOH^3M^2$ for different combinations of values of p_1 and p_2 . Panel e-h shows the sensitivity and specificity of $DIDOH^3M^2$ as a function of the parameter D_{Norm} and P_{Norm}.



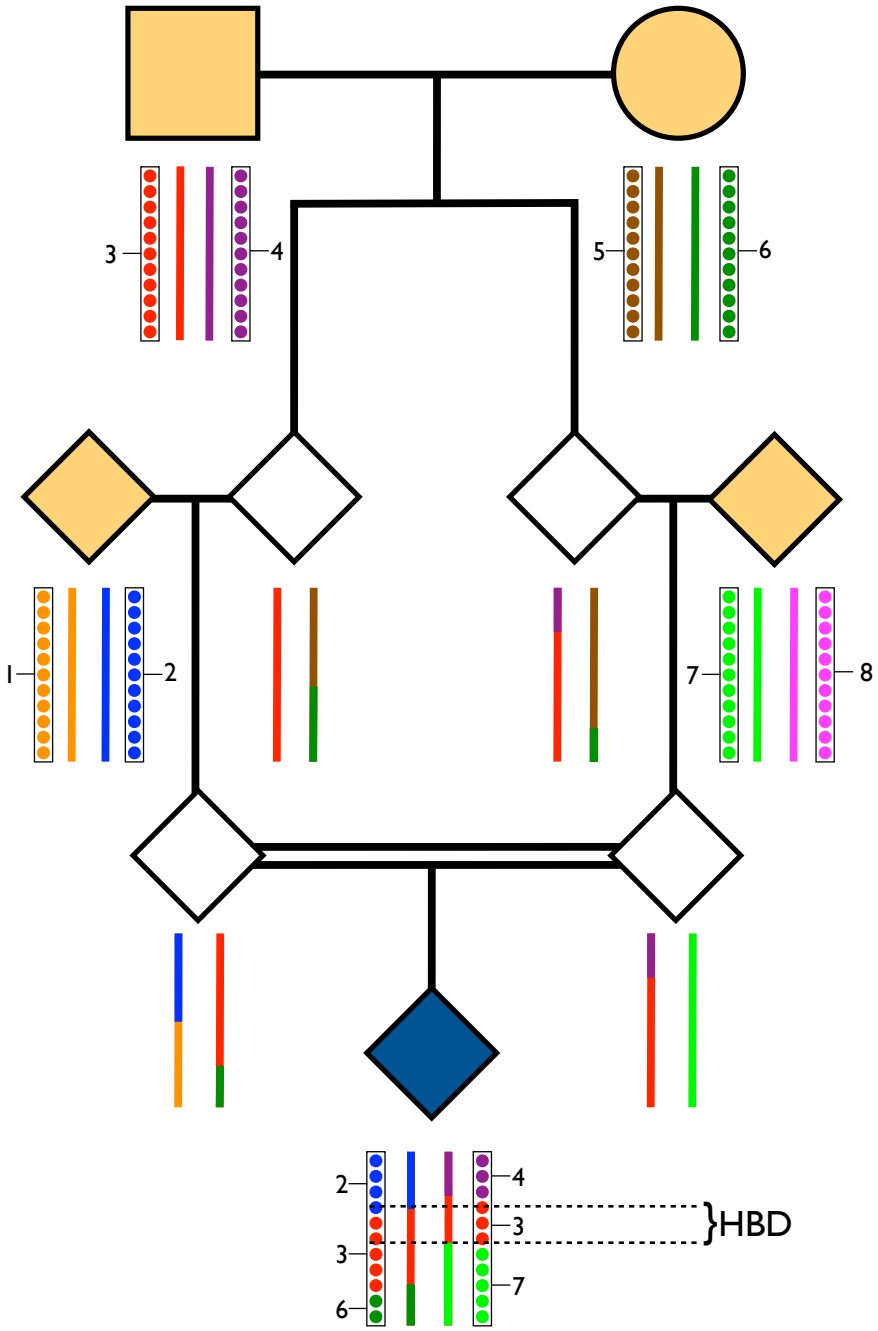
Supplemental Figure 2: Performance evaluation of the $DIDOH^3M^2$ algorithm in the detection of ROHs on synthetic chromosomes. Panels a, b and c report the performance of $DIDOH^3M^2$ as a function of parameter R_1, while panels d, e and f as a function of parameter R_2. Panels a and d show TPR vs the number of SNPs within the detected ROH. Panels b and e show the TPR as a function of the distance between consecutive polymorphic positions in the detected ROH. Panels c and f show the number of False Positive (FP) ROH detected by the $DIDOH^3M^2$.



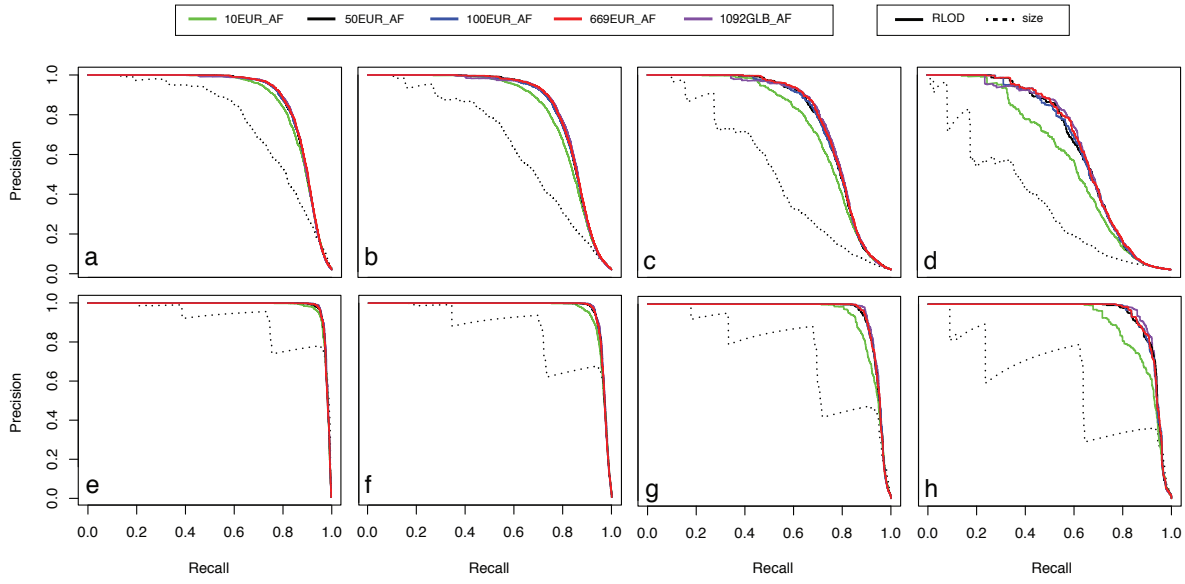
Supplemental Figure 3: Performance comparison between $DIDOH^3M^2$, PLINK, BCFtools and VCFtools on the WGS and WES data of the 200 individuals sequenced by the 1000 Genomes Project Consortium. Panels a and c report the results of the precision-recall analysis for WGS and WES data respectively. The bar 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 $DIDOH^3M^2$ algorithm have been reported for different settings of the R_2 ($R_2 = 1/1000$,$R_2 = 1/10000$,$R_2 = 1/100000$) and R_1 (1/100, 2/100, 3/100, 4/100, 5/100) parameters and $p_1 = 0.1$, $p_2 = 0.1$, $P_Norm = 1$, $d_{Norm} = 100000$. The performance of PLINK have been reported for different values of heterozygote allowance (PL-H = 0 and PL-H = 1), different values of kb threshold (-kb 50/100/200) and different values of SNP threshold to call a ROH (-snp= 50, 250, 500 for WES, -snp=500, 1000, 2000 for WGS).



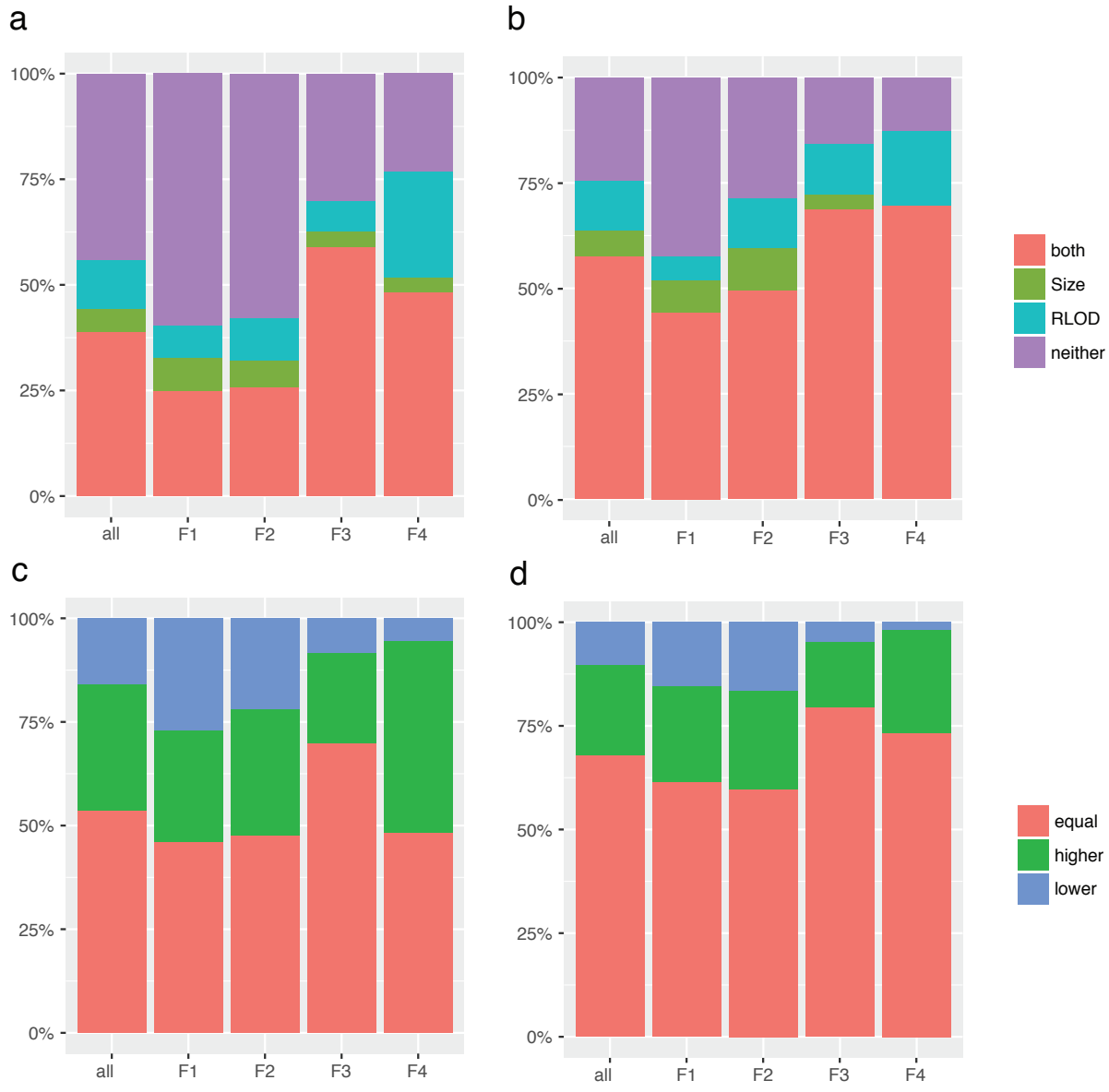
Supplemental Figure 4: Performance comparison between $DIDOH^3M^2$, PLINK, the RoH option of BCFTools and the LROH option of VCFTools on the WES and WGS data of 200 individuals sequenced by the 1000 Genomes Project Consortium. The bar-plots of panels a and c show the average total length of ROH detected by the four approaches in WES (a) and WGS (c) data respectively. The bar-plots of panels b and d shows the average total number of ROHs detected by the four approaches in WES (b) and WGS (d) data respectively.



Supplemental Figure 5: Generation and identification of ROHs in simulted genome maps of offspring to consanguineous parents. SNP backbones were built on a genetic map generated by picking up SNPs from Rutgers map markers and assigned to family founders (orange-colored individuals) to simulate pairs of chromosomes. Markerdrop simulates recombination patterns along the pedigree (conditional on a disease-linked locus), associating founder-tracking labels (numbers) with each dropping chromosome haplotype. 1KG haplotypes were assigned to the SNP backbones of the founders, so that each 1KGP SNV becomes associated with a founder-tracking label. 1KGP haplotypes were superimposed on the SNP backbones of the index offspring (blue-colored individual) according to the recombination patterns traced by Markerdrop. Simulated 1KGP genotypes in the index offspring were used to detect ROH by DIDOH3M2 and to calculate RLOD for each ROH. Autozygous ROH were identified as those ROH with alleles of both haplotypes associated with the same founder-tracking label in the index offspring.



Supplemental Figure 6: Performance comparison between *RLOD* and RoH size to identify true autozygosity with allele frequencies calculated on different sample sizes. Results of the analysis carried out in the simulated 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; F2: 0.023-0.066; F3: 0.00105-0.023; F4: 0-0.0105. *RLOD* and RoH size performances are depicted as dotted and continuous lines, respectively, while colors indicate that allele frequencies applied to *RLOD* calculation were obtained on different sample compositions and sizes (EUR: Europeans; GLB: Global; AF: Allele Frequencies).



Supplemental Figure 7: 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 4 different gF ranges. Panels a and b report the percentage of times the msRoH ranked as 1st among all the identified ROH by both or neither of the two measures, while panels d and f report the percentage of times the msRoH ranked higher, equal or lower by *RLOD* than size.