BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 20 2014, pages 2852–2859
doi:10.1093/bioinformatics/btu401

Genome analysis

Advance Access publication June 24, 2014

H^3M^2: detection of runs of homozygosity from whole-exome sequencing data

Alberto Magi[1,*], Lorenzo Tattini[1], Flavia Palombo[2], Matteo Benelli[3], Alessandro Gialluisi[4], Betti Giusti[1], Rosanna Abbate[1], Marco Seri[2], Gian Franco Gensini[1], Giovanni Romeo[2] and Tommaso Pippucci[2,*]

[1]Department of Experimental and Clinical Medicine, University of Florence, Florence 50019, [2]Medical Genetics Unit, Polyclinic Sant'Orsola-Malpighi, Department of Medical and Surgical Sciences, University of Bologna, Bologna 40138, [3]Diagnostic Genetic Unit, Careggi Hospital, Florence 50019, Italy and [4]Language and Genetics Department, Max Planck Institute for Psycholinguistics, Nijmegen 6525 EN, The Netherlands

Associate Editor: Michael Brudno

ABSTRACT

Motivation: Runs of homozygosity (ROH) are sizable chromosomal stretches of homozygous genotypes, ranging in length from tens of kilobases to megabases. ROHs can be relevant for population and medical genetics, playing a role in predisposition to both rare and common disorders. ROHs are commonly detected by single nucleotide polymorphism (SNP) microarrays, but attempts have been made to use whole-exome sequencing (WES) data. Currently available methods developed for the analysis of uniformly spaced SNP-array maps do not fit easily to the analysis of the sparse and non-uniform distribution of the WES target design.

Results: To meet the need of an approach specifically tailored to WES data, we developed H^3M^2, an original algorithm based on heterogeneous hidden Markov model that incorporates inter-marker distances to detect ROH from WES data. We evaluated the performance of H^3M^2 to correctly identify ROHs on synthetic chromosomes and examined its accuracy in detecting ROHs of different length (short, medium and long) from real 1000 genomes project data. H^3M^2 turned out to be more accurate than GERMLINE and PLINK, two state-of-the-art algorithms, especially in the detection of short and medium ROHs.

Availability and implementation: H^3M^2 is a collection of bash, R and Fortran scripts and codes and is freely available at https://sourceforge.net/projects/h3m2/.

Contact: albertomagi@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on February 25, 2014; revised on May 21, 2014; accepted on June 13, 2014

1 INTRODUCTION

Runs of homozygosity (ROH) are chromosomal stretches that in a diploid genome appear in the homozygous state, that is, display identical alleles at multiple contiguous loci.

The study of ROH can be relevant for both population and medical genetics. Genomic ROH patterns are governed by a number of factors, among which population genetic history (e.g. historical bottleneck, geographic isolation and population size), evolutionary forces (e.g. selective sweeps) (Sabeti *et al.*, 2007) and cultural habits or historical and geographical factors. Recent parental relatedness favors the formation of long ROH (several megabases) that occur because of IBD (identity by descent, when the two alleles at a locus match because they originate from the same common ancestor), as opposed to identity by state (when the two alleles at a locus match simply by coincidence). Homozygosity originating from the occurrence of individual IBD regions owing to parental relatedness (autozygosity) is known to possibly contain recessive highly penetrant deleterious disease-causing mutations surrounded by an unusually long homozygous haplotype. This is the principle that inspired homozygosity mapping in the study of rare recessive disorders affecting inbred individuals (Lander and Botstein, 1987). In outbred individuals, short (up to few hundreds of kilobases) or medium-sized ROH (from hundreds of kilobases to a few megabases) can surround disease-causing mutations as well (Hildebrandt *et al.*, 2009), playing a role in predisposition to disease through the effect of mildly deleterious recessive variants (Wang *et al.*, 2009).

To date, ROH detection has been achieved by microarray-based technologies. Currently available single nucleotide polymorphism (SNP)-array platforms contain millions of markers from the HapMap Project (International HapMap Consortium, 2003) and have a mean SNP-to-SNP distance of around 3 kb. The past few years have seen the emergence of several next-generation sequencing (NGS) platforms that are capable to sequence a full human genome per week at a cost 400-fold less than the previous methods. The advent of NGS platforms has revolutionized our ability of studying human genetic variation (Wang *et al.*, 2009) allowing the achievement of large-scale re-sequencing projects, such as the 1000 Genomes Project (1000GP) (1000 Genomes Project Consortium *et al.*, 2010) and the Cancer Genome Atlas (www.cancergenome.nih.gov). Recently, the 1000GP consortium, by combining low-coverage whole-genome sequencing (WGS) and high-coverage whole-exome sequencing (WES) of 1092 individuals from 14 populations, has genotyped ~38 million single nucleotide polymorphic positions (1000 Genomes Project Consortium *et al.*, 2012). This provided a genetic map characterized by a mean SNP-to-SNP distance of 73 bp and with >75% of the inter-marker distances under 200 bp. This catalog captures up to 98% of accessible SNPs with minor allele frequency of ≥1%.

*To whom correspondence should be addressed.

At present, the cost and the computational complexity still limit the routine use of WGS and make WES an effective alternative that has been successfully used for the discovery of single-nucleotide variants (Ng *et al.*, 2009), short indels and copy number variants (Magi *et al.*, 2013) and that can be applied to the detection of ROH (Pippucci *et al.*, 2011). However, the sparse nature of the WES target design makes this latter application challenging. The distance between adjacent 1000GP exomic SNPs ranges from 1 bp to 26 Mb (with an average value of around 500 bp), and consequently, large ROH may be covered by few and not uniformly spaced SNPs, whereas small and isolated ROH may display exceptionally high marker density. Currently available sliding-window methods, such as PLINK (Purcell *et al.*, 2007) or GERMLINE (Gusev *et al.*, 2009), were developed for the analysis of uniformly spaced SNP-array maps and do not fit easily to the analysis of the sparse and non-uniform distribution of WES SNP maps. The identification of long ROH, typically those detected in a context of consanguinity, may be faced by applying these methods to WES data (Pippucci *et al.*, 2011). However, medium or short ROH cannot be as easily captured from WES data by using traditional sliding-windows approaches.

To meet the need of an approach specifically tailored to WES data, which could overcome the inherent limitations of currently available tools, we developed a novel computational approach [homozygosity heterogeneous hidden Markov model (HMM), H^3M^2] for the identification of ROH. The algorithm is based on a heterogeneous HMM that, incorporating the distance between consecutive polymorphic positions into the transition probabilities matrix, is able to detect with high sensitivity and specificity ROH of every genomic size. The key feature of the algorithm is its heterogeneity, which makes it well-suited for WES data. To evaluate the capability of this novel method to detect ROH of different sizes, we applied it to the analysis of a synthetic dataset and compared its performance with that of PLINK and GERMLINE on real WES data. Another method for homozygosity IBD detection from sequences based on a HMM algorithm, IBDseq, has recently been reported (Browning and Browning, 2013). We did not compare H^3M^2 with IBDseq because this latter method is conceived explicitly for genome sequences and has never been applied to WES data.

As a whole, the results we obtained in these analyses demonstrated that H^3M^2 has the potentiality to capture most of the genomic ROH that overlap regions covered by exomic targets, and that it outperforms the existing algorithms especially in the detection of medium and short ROH.

2 METHODS

2.1 *B-allele* frequencies

As a measure of the homozygous/heterozygous genotype state of each polymorphic position *i*, we adopted the *B*-allele frequency (*BAF*). BAF_i is defined as the ratio between *B*-allele counts (N_B, the number of reads that match with the 1000GP alternate allele at position *i*) and the total number of reads mapped to that position (*N*, the depth of coverage):

$$BAF_i = \frac{N_B}{N} \quad (1)$$

BAF_i may thus assume values that belong to the interval [0,1]: when $BAF_i = 0$, all the reads aligned to position *i* match with the major allele; when $BAF_i = 1$, all the reads match with the minor allele; and when $BAF_i \neq 0, 1$, some reads match with the major allele, whereas some others match with the minor allele (see Supplementary Methods). It follows that BAF_i can predict the homozygous/heterozygous genotype state of each polymorphic position *i*:

- when $BAF_i \approx 0$, the polymorphic position is homozygous reference;
- when $BAF_i \approx 0.5$, the polymorphic position is heterozygous; and
- when $BAF_i \approx 1$, the polymorphic position is homozygous alternate.

To the scope of the present work, the use of the *BAF* measure has some advantages compared with that of NGS genotype calls. First, *BAF* calculation does not require computationally intensive steps like binary sequence alignment/map (Li *et al.*, 2009) file realignment and recalibration, which are widely adopted to improve reliability of variant calls. Second, genotype calling is usually performed for variant sites only, thereby an additional genotype calling step would be needed to cover also non-variant sites.

2.2 Map construction

To create a map of exomic SNPs, we included all the polymorphic positions discovered by 1000 Genomes Project Consortium (2012) falling into the range of the 1000GP exomic target regions (downloaded at http://www.1000genomes.org/) plus 1000 bp of sequences flanking both sides of each target region. As a result of this procedure, we selected 4 163 299 SNPs: the distance between adjacent SNPs ranged from 1 bp to 22 Mb with mean and median values of 47 and 686 bp, respectively, (and ≥75% of the distances under 110 bp).

2.3 H^3M^2 model and algorithm

To identify homozygous DNA segments, we decided to model *BAF* data by means of a discrete state HMM with continuous output. A discrete HMM with continuous output is characterized by the following elements:

- The number of hidden states, K, in the model. The states are denoted as $S = \{S_1, ..., S_K\}$ while q_i denotes the actual state at position *i* $(1 \leq i \leq n)$.
- The observed data $O = \{O_1, ..., O_N\}$.
- The initial state distribution, π, where $\pi_{1k} = P(q_1 = S_k)$.
- The emission probability distributions $b_k(i)$, that is, the probability of observing O_i at position *i* given the state S_k: $b_k(i) = P[O_i | q_i = S_k]$.
- The transition matrix, *A*, giving the probability of moving from one state to another, $A_{lm} = P(q_{i+1} = S_m | q_i = S_l)$ for $1 \leq i \leq n-1$ and $1 \leq l, m \leq K$.

To model our problem, we used a two-state HMM ($K = 2$) where the hidden states represent non-homozygous ($S_1 = non - Hom$) and homozygous ($S_2 = Hom$) states of the genome, and the observations are the *BAF* values at each polymorphic position *i* (BAF_i). The emission distributions are mixture of truncated Gaussian density with the following form:

- $P(BAF_i | q_i = S_1) = c_1 g_l^u(BAF_i; \theta_1) + c_2 g_l^u(BAF_i; \theta_2) + c_3 g_l^u(BAF_i; \theta_3)$
- $P(BAF_i | q_i = S_2) = c_1 g_l^u(BAF_i; \theta_1) + c_3 g_l^u(BAF_i; \theta_3)$

where c_k is the proportion of the *k*-th component in the mixture with $\sum_{k=1}^{3} c_k = 1$ and $\theta_1 = (\mu_1 = 0, F \cdot \sigma_1)$, $\theta_2 = (\mu_2 = 0.5, \sigma_2)$ and $\theta_3 = (\mu_3 = 1, F \cdot \sigma_3)$ are the means and the variances of the three truncated Gaussians. The lower and upper bounds are $l = 0$ and $u = 1$ for the three truncated Gaussians.

F is a parameter used to modulate the spread of the two truncated Gaussian distributions with mean $\mu_1 = 0$ and $\mu_3 = 1$. *F* can take values in the range $[1, \infty]$, and the larger is its value the larger is the probability to include in homozygous regions *BAF* values that strongly deviate from 0 and 1. In practice, the parameter *F* allows our model to recognize ROH taking into account sequencing and alignment errors that, in complex regions of the genome, could generate BAF values that belong to the g_2 distribution (heterozygous state). The expression of a truncated Gaussian density g with lower bound l and upper bound u can be easily derived from the density of a non-truncated Gaussian:

$$g_l^u(BAF_i;\theta) = \frac{f(BAF_i;\theta)}{F(u;\theta) - F(l;\theta)} I_l^u(BAF_i), \tag{2}$$

where $f(\bullet;\theta)$ and $F(\bullet;\theta)$ represent the density and cumulative distribution functions of a non-truncated Gaussian of parameter $\theta = (\mu, \sigma^2)$ and $I_l^u(BAF_i) = 1$ if BAF_i belongs to the interval $[l, u]$ and $I_l^u(BAF_i) = 0$ otherwise.

Finally, to take into account the distance between consecutive polymorphic positions $d = (d_1, d_2, ..., .d_{n-1})$, we decided to incorporate them into the transition probabilities matrix A_i defined for $1 \leq i \leq n-1$:

$$A_i = \begin{pmatrix} 1 - p_1(1 - e^{-f_i}) & p_1(1 - e^{-f_i}) \\ p_2(1 - e^{-f_i}) & 1 - p_2(1 - e^{-f_i}) \end{pmatrix} \tag{3}$$

where p_1 (p_2) represents the probability of moving from State 1 to State 2 (from State 2 to State 1) in the homogeneous HMM, $f_i = d_i / d_{Norm}$ and d_{Norm} is the distance normalization parameter. The parameter d_{Norm} modulates the effect of genomic distance d_i on the transition probabilities: the larger is d_{Norm} the smaller is the probability to jump from one state to another.

To estimate the parameters of the heterogeneous HMM, we developed a two-step computational recipe based on expectation-maximization (EM) and Viterbi algorithms. In the first step, we estimate the variances σ_k and the proportions c_k of the mixture of three truncated Gaussians, whereas in the second step we estimate the best state sequence (we associate each BAF value to particular states) by using the Viterbi algorithm. The inputs to the algorithm are the sequence of BAF values $BAF = (BAF_1, BAF_2, .., BAF_N)$, the distance between consecutive polymorphic positions $d = (d_1, d_2, ..., .d_{n-1})$, the values of the parameters F, p_1, p_2 and d_{Norm}, and the output are genomic regions with consecutive SNPs in homozygous and heterozygous states. We decided to use the two transition probabilities p_1 and p_2 as algorithm parameters instead of estimating them with an EM algorithm, as we do believe that they can be useful in setting the resolution of our algorithm (the capability of our computational method to detect ROH of different size and with different number of SNPs).

To estimate the parameters σ_k and c_k of the Gaussian mixture model, we make use of the classical EM algorithm (Dempster *et al.*, 1977). In brief, denoting with Z_{ki} the hidden state for BAF_i (Z_{ki} is a random variable equal to 1 if BAF_i belongs to state k and 0 otherwise), we can define the conditional probabilities $\tau_{ki} = Pr(Z_{ki} = 1 | BAF_i)$. The basic ingredient of the EM family of algorithms is the iterative application of an expectation step followed by a likelihood maximization step. EM starts with initial values ($c_k^{(0)}$, $\sigma_k^{(0)}$) for the parameters and iteratively performs the two steps until convergence. In the E-step, the conditional probabilities τ_{ki} are computed. Given the parameters estimated at h-th iteration, $c_k^{(h)}$ and $\theta_k^{(h)} = (\mu_k^{(h)}, \sigma_k^{(h)})$, the conditional probabilities τ_{ki}^{h+1} are obtained with the following formula:

$$\tau_{ki}^{h+1} = \frac{c_k^{(h)} g_l^u(m_i;\theta_k^h)}{\sum_{i=1}^{N} c_k g_l^u(m_i;\theta_l)}. \tag{4}$$

In the M-step, the proportions of the components in the mixture and the empirical estimators of the mean and the variance are computed. In particular at the iteration $(h+1)$ of the M-step, we compute $\sigma_k^{(h+1)}$ and $c_k^{(h+1)}$ with the following formulas:

$$\sigma_k^{(h+1)} = \frac{\sum_{i=1}^{N} \tau_{ki}^{h+1} (m_i - \mu_k^{(h+1)})^2}{\sum_{i=1}^{N} \tau_{ki}^{h+1}}, \tag{5}$$

$$c_k^{(h+1)} = \frac{\sum_{i=1}^{N} \tau_{ki}^{h+1}}{N}. \tag{6}$$

Finally, once all the parameters of the mixture of Gaussians have been estimated, we apply the Viterbi algorithm to find the best state sequence and consequently to associate each BAF value to one of the two (non-homozygous/homozygous) states, thus identifying ROH.

3 RESULTS

3.1 BAF properties and distributions

To evaluate the capability of BAF to predict the homozygous/heterozygous SNP state, and of BAF profiles to discriminate between homozygous and heterozygous DNA segments, we studied the distribution of BAF values by analyzing WES data of three individuals (NA12878, NA12891 and NA12892) sequenced by the 1000GP consortium and previously genotyped with SNP-array technologies by the HapMap consortium (see Supplementary Methods). For each position i of the SNP-map, we compared the BAF value with HapMap genotypes and with genotypes independently generated by SAMtools (Li *et al.*, 2009) and the Genome Analysis ToolKit (GATK) (McKenna *et al.*, 2010) on WES data (see Supplementary Methods). All the analyses were performed by progressively filtering out SNPs covered less than a defined threshold (5×, 10×, 15× and 20×). The results of these comparisons are reported in panels a, b and c of Figure 1 and Supplementary Figures 1–3, and clearly show strong correlation between BAF and SNP genotypes calls. The 'Violin Plots' reported in panels a, b and c of Figure 1 illustrate the capability of BAF values to predict the genotype calls made by SAMtools (R = 0.988), GATK (R = 0.993) and SNP-array data (R = 0.99). Moreover, as expected, higher the coverage over the polymorphic position higher the correlation between BAF and genotype calls (Supplementary Figs 1–3). Based on these results, we performed all the downstream analyses by using only SNPs covered $\geq$ 10.

As a further step, to evaluate the capability of BAF profiles to discriminate between homozygous and non-homozygous regions of the genome, we studied the distributions of BAF in different sets of genomic regions. To this end, we defined as homozygous and non-homozygous gold standard regions, those classified as such by PLINK on HapMap calls (see Supplementary Material). The four sets in which we studied BAF distribution were (i) all the regions of the genome, (ii) the homozygous regions only, (iii) the non-homozygous regions only and (iv) the non-pseudoautosomal X chromosome of male individuals. The results reported in panels d, e, f and g of Figure 1 reveal that the BAF distribution across all the polymorphic positions can be well approximated by a mixture of three truncated normal distributions:

$$g_l^u(BAF_i; c, \theta) = \sum_{k=1}^{3} c_k g_l^u(BAF_i;\theta_k), \tag{7}$$

where c_k is the proportion of the k-th component in the mixture with $\sum_{k=1}^{3} c_k = 1$ and $\theta_k = (\mu_k, \sigma_k)$ are the means and the variances of the three Gaussians with $\mu_1 = 0$, $\mu_2 = 0.5$ and $\mu_3 = 1$. The lower and upper bounds are l = 0 and u = 1 for the three truncated Gaussians.

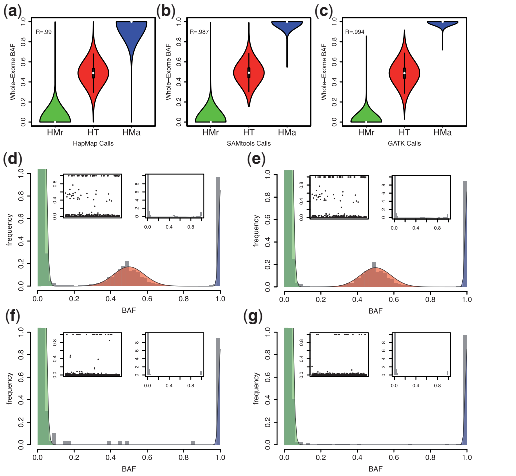

Fig. 1. BAF data distribution. Panels a, b and c show the distributions of BAF values against the genotype calls generated by the HapMap consortium on SNP-array data (**a**), the genotype calls made by SAMtools (**b**) and the genotype calls made by GATK (**c**). For each genotype caller, the distribution of *BAF* values is reported for homozygous reference calls (HMr), heterozygous calls (HT) and homozygous alternative calls (HMa). *R* is the Pearson correlation coefficient. Panels d–g show the distribution of *BAF* values in all the regions of the genome (**d**), in heterozygous regions (**e**), in homozygous regions (**f**) and in the X chromosome of male individuals (**g**). For each panel, the main plot reports the zoomed histogram, the left subplot shows the *BAF* values against genomic positions, whereas the right subplot shows the entire histogram of *BAF* values

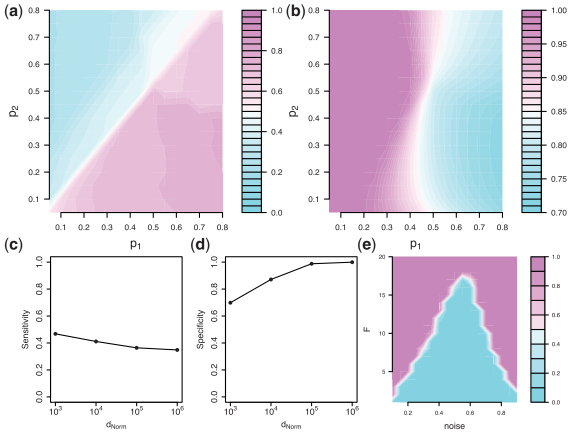

Fig. 2. H^3M^2 algorithm and parameter settings on synthetic chromosomes. The contour plots of panels a and b show the sensitivity and specificity of H^3M^2 for different combinations of values of p_1 and p_2 parameters. Panel c and d show the sensitivity and specificity of H^3M^2 against d_{Norm}. Panel e shows the performance of H^3M^2 as a function of the parameter F

Figure 1 also shows that in non-homozygous regions (Fig. 1e) $c_k \neq 0$ for all the three distributions, whereas in homozygous regions (Fig. 2f and g) $c_2 = 0$. Thus, homozygous and non-homozygous regions can be discriminated based on *BAF* distribution as signature. Genomic regions with *BAF* values generated by Equation 7 with $c_2 \neq 0$ can be classified as non-homozygous, whereas genomic regions with *BAF* values generated by Equation 7 with $c_2 = 0$ can be classified as homozygous.

3.2 Synthetic validations

To test the ability of H^3M^2 to detect ROH of different sizes and constituted by different number of SNPs as a function of the distance between consecutive markers, we performed an intensive simulation based on synthetic data. To this end, we generated synthetic chromosomes starting from the *BAF* data of the three samples used in previous section. *BAF* data from the non-pseudoautosomal X chromosome of the male individual NA12891 were used to simulate homozygous DNA segments, whereas *BAF* data from the autosomal chromosomes of all the three samples were used to simulate heterozygous segments.

Each synthetic chromosome was generated as a stretch of 2000 polymorphic positions in which:

- Homozygous segments were simulated as N consecutive data sampled from the *BAF* values of the non-pseudoautosomal X chromosomes regions of male individuals.
- Non-homozygous segments were simulated by sampling $(2000 - N)$ data from the *BAF* values of the autosomal chromosomes of the three aforementioned individuals.

  Heterozygous segments were imposed to have SNPs in a heterozygous/homozygous ratio of 1:9. To this end, we sampled one *BAF* value from SNPs called as heterozygous by GATK every nine sampled SNPs called as homozygous by GATK.

The 1:9 ratio was imposed to simulate at best the actual heterozygous/homozygous proportion and to prevent the emergence of false-positive (FP) homozygous segments. To reproduce the complex architecture and distribution of homozygous and heterozygous WES 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 WES polymorphic positions.
- The distances between adjacent polymorphic positions in homozygous regions are fixed to a predefined distance D.

We performed simulations with $N =$ (50, 100, 200, 300, 400, 500, 600, 700, 800, 900 and 1000) and $D =$ (10 bp, 100 bp, 1 kb, 10 kb and 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 parameter d_{Norm} ($d_{Norm} = 10^3$, 10^4, 10^5 and 10^6), p_1 and p_2 (from 0.05 to 0.8 by 0.05).

To evaluate the performance of H^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 H^3M^2 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 H^3M^2 as homozygous divided by the total number of markers outside the synthetic ROH. The results of these analyses are summarized in Figure 2.

Figure 2a and b report the sensitivity and specificity for all the combinations of the p_1 and p_2 parameters. Figure 2a shows that the larger p_2 the smaller the range of p_1 values that ensure high sensitivity. 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, Figure 2b demonstrates that for values of $p_1 > 0.4$ the specificity of our method drastically decreases. In summary, because global performance of H^3M^2 is a result of the trade-off between sensitivity and specificity, we argue that the best performance can be obtained by setting $p_1 = 0.1$ and $p_2 \in [0.1, 0.3]$. Figure 2c and d report the results of this analysis as a function of the d_{Norm} parameter and show that d_{Norm} has strong effect on global performance of H^3M^2: the larger is d_{Norm} the smaller (larger) is sensitivity (specificity).

To study the effect of the parameter F in modulating the capability of H^3M^2 to tolerate sequencing and alignment errors in the detection of ROH, we built another synthetic dataset. The synthetic chromosomes of this dataset were generated with the same procedure described above, apart from the central SNP of the homozygous stretch. With the purpose of reproducing increasing error rates, BAF values ranging from 0.1 to 0.9 were assigned to the central SNP. We applied H^3M^2 to this synthetic dataset using different values of the parameter F ($F \in [1, 20]$), and the results are reported in the contour plot of Figure 2e. Each point in the plot represents the fraction of times the algorithm detects a unique ROH instead of splitting it in two ROH. These results show that larger F values make H^3M^2 more tolerant of positions characterized by higher sequencing and alignment error rates that appear to be heterozygous sites: as an example, $F = 10$ induces the algorithm to include in a homozygous region BAF values as large as 0.35.

As a further test, to evaluate the capability of H^3M^2 to detect ROH of different size and comprising different number of SNPs, we calculated TPR and FPR as follows: a detected ROH is considered a true positive if it has any overlap with a synthetic ROH, whereas it is considered a FP if it has no overlap with a synthetic ROH. These analyses were performed by setting $p_2 = 0.1$, $p_1 \in [0.05, 0.3]$ and $d_{Norm} = (10^3, 10^4, 10^5, 10^6)$. As expected, the results (Fig. 3) show that the larger the number of SNPs falling within a given ROH the higher the probability to correctly identify the homozygous region. The same holds, considering the distance between adjacent positions (D): the larger D the higher the probability to call the region as homozygous.

A detailed analysis of Figures 3a and b reveals that the sensitivity of H^3M^2 increases with increasing values of the parameter p_1. However, setting $p_1 \geq 0.1$ has little effect on the resolution of H^3M^2 (i.e. the capability of the algorithm to detect ROH constituted by a small number of SNPs), while it favors the detection of FP events (Fig. 3c). On the other hand, the results of Figure 3d–f show that the parameter d_{Norm} strongly affects the performance of the H^3M^2 algorithm in terms of both sensitivity and specificity: the smaller d_{Norm} the higher the probability to detect small ROH (ROH characterized by high SNP density) and FP events. The results of these analyses (Supplementary Figs 4–8) also show that the parameter d_{Norm} rules the capability of H^3M^2 to detect homozygous segments characterized by variable SNP densities. When d_{Norm} is set to large values (10^5, 10^6), H^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), H^3M^2 is able to detect homozygous segments made of densely distributed SNPs, and increasing p_1 has a relevant effect on resolution.

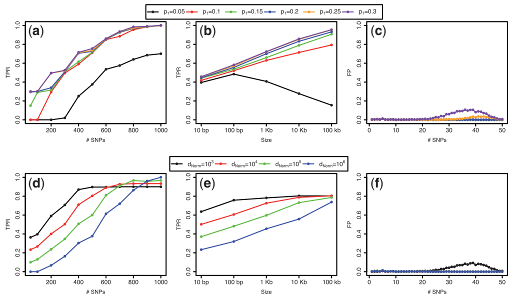

Fig. 3. Performance evaluation of the H^3M^2 algorithm in the detection of ROHs on synthetic chromosomes. Panels a–c report the performance of H^3M^2 as a function of parameter p_1, whereas panels d–f as a function of parameter D_{Norm}. Panels a and d show TPR versus the number of SNPs within the detected 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 FP ROH detected by the H^3M^2

Taken as a whole, these results suggest that to detect large homozygous segments and limit the false discovery rate, large values of d_{Norm} (10^5, 10^6) and small values of p_1 (0.1) should be used. On the other hand, to detect homozygosity with a high level of resolution, small d_{Norm} (10^3, 10^4) and large p_1 values (0.2, 0.3) are recommended.

3.3 Real data analysis

To test the proposed computational pipeline for the identification of homozygous segments on real data, we analyzed the WES data of 15 individuals (five CEU, Utah residents with ancestry from Northern and Western Europe, five JPT, Japanese in Tokyo, and five YRI, Yoruba in Ibadan) sequenced by 1000GP consortium (see Supplementary Methods) by using the following parameter settings: $p_2 = 0.1$, $p_1 = [0.1, 0.2, 0.3]$ and $d_{Norm} = [10^3, 10^4, 10^5]$.

First, we studied the detected ROH in terms of both cumulative global size and fraction of SNPs within them. As expected (Fig. 4a–f), the larger the value of the parameter p_1 the larger the total size of the detected ROH, and accordingly the larger the total fraction of SNPs within the detected ROH. Conversely, the smaller is d_{Norm} the larger is the total size and the total fraction of SNPs detected by H^3M^2. By setting the most conservative set of parameters ($d_{Norm} = 10^5$ and $p_1 = 0.1$), H^3M^2 detected an average of around 160 Mb (10% of SNPs) in the YRI individuals, 330 Mb in the CEU (18% of SNPs) and 380 in the JPT (21% of SNPs), whereas using more inclusive parameters ($d_{Norm} = 10^3$, $p_1 = 0.3$), we detected 860 Mb (38% of SNPs) for YRI, and around 1.25 Gb for both CEU and JPT individuals (50% of SNPs).

Subsequently, we compared the results of H^3M^2 with those obtained by PLINK and GERMLINE on the GATK calls. To allow for a comprehensive evaluation of the performance of the two tools, we defined six different parameter configurations for each of the tools (see Supplementary Methods). By using the most conservative configuration, PLINK (`--homozyg-snp` 500 and `--homozyg-window-het` 0) detected around 11 Mb

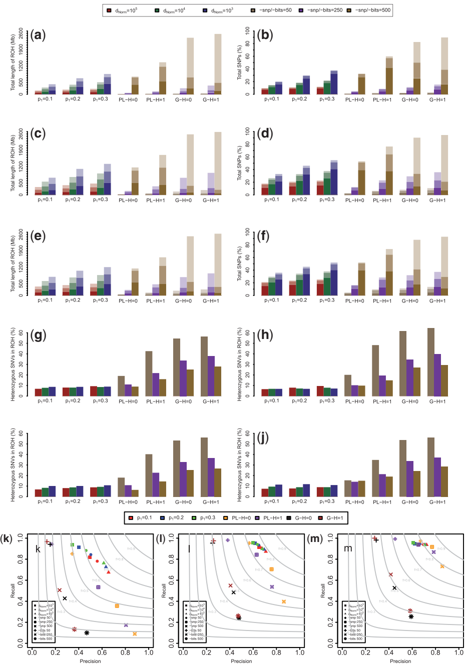

Fig. 4. Performance comparison between H^3M^2, GERMLINE and PLINK on the WES data of the 15 individuals sequenced by the 1000 Genomes Project Consortium. The bar plots of panels a, c and e show the total length of ROH detected by the three approaches in the YRI (**a**), CEU (**c**) and JPT (**e**) individuals. The bar plots of panels b, d and f show the total percentage of SNPs that belong to the ROH detected by the three approaches in the YRI (**b**), CEU (**d**) and JPT (**f**) individuals. Each bar of the bar plot is colored with three different graduated shading: dark shading represents small ROHs ($ROHs \leq 500kb$), lighter shading represents medium ROHs ($500kb < ROHs \leq 1500Kb$) and light shading represents large ROHs ($ROHs > 1500kb$). The bar plots of panels g–j report the fraction of heterozygous single nucleotide variants that belong to all (**g**), small (**h**), medium (**i**) and large (**j**) ROHs detected by the three algorithms. The performance of the H^3M^2 algorithm have been reported for different settings of the D_{Norm} (10^3, 10^4, 10^5) and p_1 (0.1, 0.2, 0.3) parameters. The performance of PLINK have been reported for different values of heterozygote allowance (PL-H = 0and PL-H = 1) and different values of SNP threshold to call a ROH/sliding window size in SNPs (-snp = 50, 250 and 500). The performance of GERMLINE has been reported for different values of mismatching heterozygote allowance (G-H = 0 and G-H = 1) and different window size in SNPs (-bits = 50, 250 and 500). Panels k–m report the results of the precision–recall analysis for small, medium and large ROHs, respectively

(1% of the analyzed SNPs) of ROH for the YRI individuals, 26 Mb (3% of SNPs) for the CEU and 53 Mb (4% of SNPs) for JPT, whereas GERMLINE (`-bits` 500 and `err_het` 0) identified 32 Mb (2% of analyzed SNPs) for the YRI, 141 Mb (8% of SNPs) for CEU and 202 Mb (11% of SNPs) for JPT. On the other hand, using the less stringent configuration (`--homozyg-snp` 50, `--homozyg-window-het` 1 for PLINK and `-bits` 50, `-err_het` 1 for GERMLINE), PLINK detected around 1.35 Gb (60% of analyzed SNPs) for each YRI, 1.7 Gb (76% of SNPs) for CEU and 1.62 Gb (73%) for JPT, whereas GERMLINE identified 2.6 Gb (90% of analyzed SNPs) for YRI, 2.7 Gb (94.5% of SNPs) for CEU and 2.65 Gb (92% of SNPs) for JPT.

The total ROH length per individual detected with the most conservative parameter setting of H^3M^2 is in general agreement with previously published work that studied homozygosity at a population level (Auton *et al.*, 2009; Kirin *et al.*, 2010; Pemberton *et al.*, 2012). The YRI population has the shortest ROH per individual, reflecting the longer time over which recombination has been breaking haplotypes in this sub-Saharan African population, whereas the individuals in East Asia populations (JPT) have a slightly greater cumulative length of ROH than the other populations, which most likely reflects smaller founder population sizes in these populations.

Subsequently, we studied the properties of the detected ROH by classifying them into three classes: ROH smaller than 500 kb (A), ROH in the interval (500 kb, 1.5 Mb) (B) and ROH >1.5 Mb (C), grossly following the classification adopted by (Pemberton *et al.*, 2012). The first class (A) gathers short ROH mainly governed by local LD patterns; the second class (B) is that supposedly resulting from background relatedness in the population; and finally the third class (C) includes ROH originated from recent parental relatedness. The results obtained in this analysis completely reflect those of the synthetic data. When d_{Norm} is large (10^5), increasing p_1 poorly affects the total amount of ROH detected in the three classes: Class A ranges between 101 and 141 Mb, Class B between 176 and 233 Mb and Class C between 232 to 310 Mb. Conversely, when d_{Norm} is small (10^3) the parameter p_1 has strong effect on the total size of ROH detected in each of the three classes: Class A ranges from 200 to 460 Mb, Class B ranges from 311 to 520 Mb and Class C from 392 and 567 Mb. With regard to SNP proportions (Fig. 4b, d and f), when d_{Norm} is large, increasing p_1 has little effect on the total fraction of SNPs involved in ROH regions (independently of the class), whereas for small values of d_{Norm}, increasing p_1 drastically inflates the total fraction of SNPs within Class A ROH but has limited effect on the other two classes. All these results are explained by the fact that d_{Norm} dictates the resolution of the algorithm. Large values of d_{Norm} enable the algorithm to detect only large ROH or small ROH featuring a large number of SNPs. In this situation, increasing p_1 only improves the capability to detect Class B or C ROH containing a small number of markers. On the other hand, small d_{Norm} values enable H^3M^2 to efficiently detect also Class A ROH. In this case, increasing p_1 enables H^3M^2 to identify the huge amount of small ROH made of small number of markers. On the three ROH class analysis, PLINK provided results similar to those of the present approach, while GERMLINE efficiently detected only large ROH.

As a further step, to study the accuracy of the three algorithms, we examined the proportion of heterozygous variants independently called by GATK that in each of the 15 HapMap individuals overlapped the ROH detected by each of the three methods. Globally, as well as for any ROH class separately, ROH detected by H^3M^2 are characterized by the smallest fraction of heterozygous GATK calls, with the sole exception of Class B ROH where one of the PLINK parameter configurations (`--homozyg-snp` 500 and `--homozyg-window-het` 0) performs slightly better (Fig. 4g–j). These results demonstrate the capability of our algorithm in detecting genuine homozygous segments with respect to the other methods.

Finally, to evaluate H^3M^2 ability to identify ROH from WES data and to compare its performance with respect to the other two state-of-the-art methods, we used the ROH detected by PLINK on the HapMap data as gold standard regions. The 1.6 million HapMap SNPs are rather uniformly distributed across the entire genome with a mean SNP-to-SNP distance of around 2 kb. Conversely, the 4.2 million SNPs that we used are densely clustered in the exome target regions with a mean distance of 686 bp. It follows that these two different experimental designs may show limited overlap, making it difficult to compare results obtained by the respective analysis. To overcome this drawback, we evaluated only those polymorphic positions interrogated by both platforms and calculated precision and recall in the following manner:

- To calculate precision, we considered all the polymorphic positions called in ROH by each of the three methods that belong to the HapMap dataset, and we then calculated the fraction of these positions that were called as homozygous also in the gold standard dataset.
- To calculate recall, we considered all the polymorphic position called in ROH in the gold standard dataset that belong to WES experimental design, and we then calculated the fraction of these positions called as homozygous by each of the three state-of-the-art methods.

These analyses were performed for the three A, B, C ROH classes separately and the results are reported in Figure 4k–m. The precision–recall plots of Figure 4 show that the combination of large values of p_1 (0.2, 0.3) and small values of d_{Norm} (10^3, 10^4) increases the recall rate at the expense of the precision of H^3M^2. On the other hand, the combination of small p_1 values (0.1) and large d_{Norm} values (10^5) increases the precision of H^3M^2, while it decreases the recall rate. Moreover, changes in parameter configuration significantly perturb the performance of H^3M^2 for Class A, but not for Class B and C, ROH. The combination of parameters that ensure the best trade-off between precision and recall is $p_1 = 0.1$ and $d_{Norm} = 10^5$. Unlike H^3M^2, the performance of the two state-of-the-art algorithms is profoundly altered by changes in parameter configurations, whatever the ROH class. The high recall rate reached by GERMLINE and PLINK with less stringent parameter settings (`-bits` 50, `--homozyg-snp` 50) is obtained paying a tremendous cost in terms of precision at least for A and B ROH classes. On the other hand, attempts to improve precision adopting conservative parameter configurations (`-bits` 500 and `--homozyg-snp` 500) lead to a drastic deterioration of recall rates. With its best parameter configuration (`--homozyg-snp` 250, `--homozyg-window-het` 1), PLINK achieves a performance comparable with H^3M^2 in the detection of Class C ROH (Fig. 4m), but not of Class A and B (Fig. 4k–l). Taken as a whole, these results disclose how H^3M^2 outperforms the existing state-of-the-art methods in terms of both precision and recall (sensitivity and specificity), and how H^3M^2 performances are more robust with respect to changes in parameter configurations.

4 DISCUSSION AND CONCLUSION

In this article, we present a novel approach for the detection of ROH from individual WES data. The major computational issue we had to deal with was the non-uniform distribution of DNA markers in the exomic space. Current state-of-the-art methods for ROH detection have been conceived to be used with SNP-array data, which in principle rely on genomic maps featuring equally spaced SNPs. This represents an intrinsic limitation to the application of such approaches to SNP maps retrieved from exome-targeted designs. If this issue can be overcome when focusing on ROH as large as megabases (Class C in the present article) (Pippucci *et al.*, 2011), it can be less easily resolved when handling regions of smaller size (Class A and B in the present article). To meet the need of an approach tailored to WES data and develop a method that could efficiently capture exomic ROH of any size, we designed the H^3M^2 algorithm that incorporates the distances between consecutive SNPs into the transition matrix of the heterogeneous HMM.

We compared H^3M^2 performances with those of two methods based on sliding–window algorithms, GERMLINE and PLINK. Previous application of GERMLINE to WES data offered poor specificity/sensitivity trade-offs to isolate even long IBD segments beyond 10 cM (Zhuang *et al.*, 2012). We confirmed this observation: the highest recall rates reached by GERMLINE were obtained at a tremendous expense in precision. Conversely, demanding higher precision caused an indiscriminate fall of recall rates (Fig. 4k–m). As expected, based on what was already reported for SNP-array data (Howrigan *et al.*, 2011), PLINK algorithm holds well even on WES data, and behaved better than GERMLINE independently of the ROH class (Fig. 4k–m). H^3M^2, using the parameter configuration that guarantee the most appropriate F-measure according to the synthetic data analysis ($p_1 = 0.1$ and $d_{Norm} = 10^5$), outperforms both GERMLINE and PLINK for A and B classes. The previously reported (Pippucci *et al.*, 2011) ability of PLINK to detect long exomic ROH emerges also from the present analysis, where its performance in the detection of Class C is excellent with two different parameter configurations (`--homozyg-snp` 50, `--homozyg-window-het` 0 and `--homozyg-snp` 250, `--homozyg-window-het` 1) and highly comparable with that of H^3M^2. Survey of the heterozygous genotypes independently called by GATK within detected regions emphasized how H^3M^2 ensures the most accurate global ROH identification. The gain in performance of H^3M^2 in the detection of Class B and Class A ROHs should not be undervalued. As it has been recently highlighted, despite their small size, such regions may be biologically relevant (Pemberton *et al.*, 2012) and surround a causative mutation (Hildebrandt *et al.*, 2009).

In ROH studies, it might be useful to take into account for LD, especially for Class A ROHs. This has been done in different ways (McQuillan *et al.*, 2008; Nothnagel *et al.*, 2010). Similarly, it might be desirable to identify autozygous ROHs according to the probability of ROHs to be IBD (Pemberton *et al.*, 2012). H^3M^2 detects all ROHs, without making such functional distinctions. All these approaches can be applied to ROHs identified by H^3M^2 as part of downstream analysis.

A notable advantage of H^3M^2 is that its performance appears to be little altered by changes in parameter configuration. This is a particularly important property in the context of WES-based analyses, where the number and distribution of markers can vary extensively according to target design and experimental yield. Performance of sliding-window methods appears to be severely affected by this variability. Parameter changes affect H^3M^2

performance less drastically than they do for PLINK and GERMLINE, indicating that H^3M^2 results are robust across a wide range of parameter configurations. This is more evident for Class B and C, where all the F-measures fall in the range [0.7, 0.8], than in Class A, where F-measures span the range [0.5, 0.7].

Another important feature of H^3M^2 is its fast computational performance and its basic hardware requirements. Relying on *BAF* profiles instead of NGS genotype calls for the detection of ROH, the analysis of an aligned WES experiment with a mean coverage of 100× requires around 20 min on a single 2.4 GHz processor with 2 Gb of RAM. Conversely, preparation of input genotype calls for tools like GERMLINE or PLINK requires at least few hours in the same machine.

In conclusion, H^3M^2 is a WES-based ROH-detection algorithm well-suited for direct identification and classification of exon-rich homozygous regions of every size. It outperforms GERMLINE and PLINK applied to the same training dataset, and most importantly it is less sensitive to parameter specification, ensuring that analysis results are not severely affected by the chosen parameter configuration as in GERMLINE and PLINK. The present work supports the use of H^3M^2 for homozygosity mapping where it can efficiently replace SNP-arrays, and in studies of ROH variation and variation content across individuals in human populations.

Funding: This work was supported by the Italian Ministry of Health, Young Investigators Programme, Project GR-2009-1574072 to T.P.

Conflict of interest: none declared.

REFERENCES

1000 Genomes Project Consortium. *et al*. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.

1000 Genomes Project Consortium. *et al*. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.

Auton,A. *et al*. (2009) Global distribution of genomic diversity underscores rich complex history of continental human populations. *Genome Res*., **19**, 795–803.

Browning,B.L. and Browning,S.R. (2013) Detecting identity by descent and estimating genotype error rates in sequence data. *Am. J. Hum. Genet*., **93**, 840851.

Dempster,A.P. *et al*. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc*., **39**, 1–38.

Gusev,A. *et al*. (2009) Whole population, genome-wide mapping of hidden relatedness. *Genome Res*., **19**, 318–326.

Hildebrandt,F. *et al*. (2009) A systematic approach to mapping recessive disease genes in individuals from outbred populations. *PLoS Genet*., **5**, e1000353.

Howrigan,D.P. *et al*. (2011) Detecting autozygosity through runs of homozygosity: a comparison of three autozygosity detection algorithms. *BMC Genomics*, **2**, 460.

International HapMap Consortium. (2003) The International HapMap Project. *Nature*, **426**, 789–796.

Kirin,M. *et al*. (2010) Genomic runs of homozygosity record population history and consanguinity. *PLoS One*, **5**, e13996.

Lander,E.S. and Botstein,D. (1987) Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science*, **236**, 1567–1570.

Li,H. *et al*. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Magi,A. *et al*. (2013) EXCAVATOR: detecting copy number variants from whole-exome sequencing data. *Genome Biol*., **14**, R120.

McKenna,A. *et al*. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*., **20**, 1297–1303.

McQuillan,R. *et al*. (2008) Runs of homozygosity in European populations. *Am. J. Hum. Genet*., **83**, 359–372.

Ng,S.B. *et al*. (2009) Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, **10**, 272–276.

Nothnagel,M. *et al*. (2010) Genomic and geographic distribution of SNP-defined runs of homozygosity in Europeans. *Hum. Mol. Genet*., **19**, 2927–2935.

Pemberton,T.J. *et al*. (2012) Genomic patterns of homozygosity in worldwide human populations. *Am. J. Hum. Genet*., **10**, 275–292.

Pippucci,T. *et al*. (2011) EX-HOM (EXome HOMozygosity): a proof of principle. *Hum. Hered*., **72**, 45–53.

Purcell,S. *et al*. (2007) PLINK: a toolset for whole genome association and population based linkage analyses. *Am. J. Hum. Genet*., **81**, 559–575.

Sabeti,P.C. *et al*. (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature*, **449**, 913–918.

Wang,S. *et al*. (2009) Genome-wide autozygosity mapping in human populations. *Genet. Epidemiol*., **33**, 172–180.

Zhuang,Z. *et al*. (2012) Detecting identity by descent and homozygosity mapping in whole-exome sequencing data. *PLoS One*., **7**, e47618.