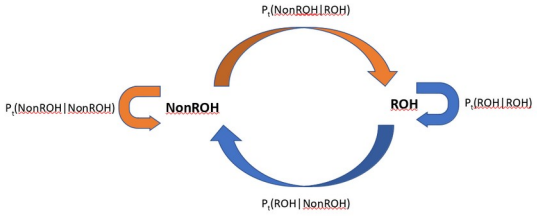
P17.096.A - ROHMM, a flexible HMM approach for detecting homozygosity using next generation sequencing data – Supplementary Data

G. Çelik[1], T. Tuncalı[2]
[1]Health Sciences Institute, Ankara Yildirim Beyazit University, Ankara, Turkey, [2]Department of Medical Genetics, Ankara University School of Medicine, Ankara, Turkey.

ROHMM's Markov Model:

A 2-state Markov chain was used as a basis to generate a model for ROHMM. ROH and NonROH states define possible genomic states. Genotype probability calculated from Allele Distribution Probability (derived from 25 whole genome samples) and genotype likelihood (PL or GL - if provided from vcf file or from user defined uncertainty level) defines the emission probabilities at each state. Transition probabilities used in ROHMM can be dynamic or static. Dynamic parameters use a natural logarithm function that calculates transition probabilities between 2 states within 2 loci as a function of distance. Natural logarithm function used in dynamic parameters is adopted from H3M2's dynamic transition parameters (Magi et al. 2014). Alternatively users can define fixed probabilities for state transitions which may work more efficiently (in terms of algorithm speed and false positive and negative rates) under dense genotyping data such as whole genome sequencing (data not shown). Figure 1 shows the structure of the markov chain and the formula for the transition probabilities.



$$P_t(NonROH|ROH) = P(stdtrans) * (1 - e^{\frac{-dist}{NormFactor}})$$

$$P_t(NonROH|NonROH) = 1 - P_t(NonROH|ROH)$$

$$P_t(ROH|NonROH) = P(stdtrans) * (1 - e^{\frac{-dist}{NormFactor}})$$

$$P_t(ROH|ROH) = 1 - P_t(ROH|NonROH)$$

Figure 1: Transition states and the structure of the markov chain. *Stdtrans* and *NormFactor* parameters are set by the user. As opposed to H3M2 ROHMM uses a single default transition parameter for stdtrans in dynamic parameters.

Emission probabilities are calculated using the formulas below. Genotype likelihoods at each position i (GLi) were used from VCF input or from user defined uncertainty value if not present. Formula for emission probability calculation is derived from bcftools roh (Narasimhan et al. 2016). ROHMM allows users to define their own allele distribution probability that they can calculate from their own data or empirically define to their own preference. ROHMM also provides Allele Distribution Model parameters derived by authors of the tool as default.

$$P_e(Gi|Si = ROH) = P(HOMREF|ROH) * P(GLi|HOMREF) + P(HOMVAR|ROH) * P(GLi|HOMVAR)$$

$$P_e(Gi|Si = NonROH) = P(HOMREF|NonROH) * P(GLi|HOMREF) + P(HET|NonROH) * P(GLi|HET) + P(HOMVAR|NonROH) * P(GLi|HOMVAR)$$

Alternatively ROHMM can use allele frequencies (derived from sample population or obtained from appropriate INFO tags provided by the VCF file) as emission probabilities of genotypes under different states. This functionality is similar to bcftools roh with changes that removes bias at starting probabilities (ROHMM uses 0.5 vs bcftools roh uses F_{mom} calculated as a function of allele frequencies) and ability to consume both PL or GL values provided by VCF file. ROHMM also adopts on-the-fly filtering of genotype loci using a BED or a VCF file provided as known alleles. As an extension to variant filtering function, ROHMM also has a spike-in function which allows insertion of homozygous reference sites into the call set for those known genotype loci when there is no call from the sample. Spike-in function helps ROHMM to detect boundaries of ROH sites more precisely as well as helps inferring cryptic ROH sites where only a small number of homozygous loci present as evidence.

References:

Magi, Alberto, Lorenzo Tattini, Flavia Palombo, Matteo Benelli, Alessandro Gialluisi, Betti Giusti, Rosanna Abbate, et al. 2014. "H3M2: Detection of Runs of Homozygosity from Whole-Exome Sequencing Data." *Bioinformatics (Oxford, England)* 30 (20): 2852–59. https://doi.org/10.1093/bioinformatics/btu401.

Narasimhan, Vagheesh, Petr Danecek, Aylwyn Scally, Yali Xue, Chris Tyler-Smith, and Richard Durbin. 2016. "BCFtools/RoH: A Hidden Markov Model Approach for Detecting Autozygosity from next-Generation Sequencing Data." *Bioinformatics* 32 (11): 1749–51. https://doi.org/10.1093/bioinformatics/btw044.