ROHMM – A Flexible Hidden Markov Model Framework To Detect Runs of Homozygosity From Genotyping Data

Gökalp Çelik[1] and TIMUR TUNCALI[2]

[1]Ankara Yildirim Beyazit University
[2]Ankara University Faculty of Medicine

September 25, 2021

Abstract

Runs of long homozygous stretches (ROH) are considered to be the result of consanguinity and usually contain recessive deleterious disease causing mutations (Szpiech et al., 2013). Several algorithms have been developed to detect ROHs. Here, we developed a simple, alternative strategy by examining X chromosome non-pseudoautosomal region to detect the ROHs from next generation sequencing data utilizing the genotype probabilities and the Hidden Markov Model algorithm as a tool, namely ROHMM. It is implemented purely in java and contains both command-line and a graphical user interface. We tested ROHMM on simulated data as well as real population data from 1000G Project and a clinical sample. Our results have shown that ROHMM can perform robustly producing highly accurate homozygosity estimations under all conditions thereby meeting and even exceeding the performance of its natural competitors.

Introduction:

Runs of homozygosity (ROH) are long genomic stretches of homozygous genotypes particularly due to high consanguinity or inbreeding although they have also been observed in outbred populations (Cavalli-Sforza & Bodmer, 1999; Gibson, Morton, & Collins, 2006). It is known that ROH's contain much of the information related to recessive traits that help clinicians and researchers to correlate genotype – phenotype associations with respect to disease and population genetics (Bittles & Black, 2010; Ceballos, Joshi, Clark, Ramsay, & Wilson, 2018). Advancements in next generation sequencing (NGS) and availability to masses further accelerated the gene-disease associations and made homozygosity mapping using massively parallel sequencing data preferable to classical laborious STR mapping methods (Alsalem, Halees, Anazi, Alshamekh, & Alkuraya, 2013; Ceballos, Joshi, et al., 2018; Chahrour et al., 2012; Pippucci et al., 2013; Walsh et al., 2010).

Here we propose a strategy to estimate homozygous segments from error prone high density genotyping data especially from whole genome and whole exome sequencing. Our algorithm uses the HMM (Hidden Markov Model) approach with modifications. *ROHMM* 's dynamic HMM uses an observable pattern of hemizygosity in male X chromosomes as a model for homozygosity along the genome and female X chromosomes as a model for heterozygous segments. Allelic distances were also incorporated into the dynamic HMM algorithm as in *BioHMM* and *H3M2* , where the latter uses the former's exact algorithm (Magi et al., 2014; Marioni, Thorne, & Tavaré, 2006). We compared ROHMM to its natural competitors*H3M2* , *bcftools roh* and *PLINK* in terms of feature set in Table 1. The *ROHMM* software includes many enhancements to eliminate the need for different tools to filter and select the best data representing the sample set. *ROHMM* also lets users set their best estimator parameters freely compared to any other tools present.*ROHMM* is coded purely in Java and available from the github repository as a source code and compiled binaries.

Materials and Methods:

Deep coverage whole genome sequencing samples and collection of B-allele frequencies from X chromosome:

All high coverage whole genome sequencing samples (10 males and 15 females) from 1000G project phase3 were used to collect B-Allele frequencies (BAF) (Auton et al., 2015). GATK UnifiedGenotyper and HaplotypeCaller were used to collect BAF data from high quality SNPs from the X chromosome non-pseudoautosomal region (Van der Auwera et al., 2013). Initial collection of SNPs were filtered based on depth, strand balance and position bias using GATK FilterVariants. Final sets of variants were evaluated based on their BAF values and grouped into homozygous reference (BAF values 0.0-0.2), heterozygous (BAF values 0.2-0.8) and homozygous alternative (BAF values 0.8-1.0) categories.

Known SNP data and simulation of synthetic truth sets:

1000G project phase3 integrated SNP data were used to generate the set of data for simulation as well as real data analysis (Auton et al., 2015). INDELs and complex multiallelics were removed due to higher error rates and sequence complexity as in some other work published elsewhere (Magi et al., 2014; Narasimhan et al., 2016). To test the performance of our algorithm, we generated true homozygous stretches using allele frequency data from 1000G CEU individuals (99 individuals) inside a viterbi scheme. To generate a variety of homozygous stretches, the percentage of homozygous sites were limited to a discrete value indicating total autozygosity for the sample [0.02 – 0.12] as well as the transition probabilities were adjusted by 10 fold at each simulation step between 1/100000 and 1/2500000. Generated synthetic calls were merged into individual VCF files for synthetic benchmarks. To make the most out of synthetic benchmarks we also introduced noise in the form of extraneous heterozygous sites or homozygous sites inside random positions of all VCF files. 5 to 10 percent of homozygous reference allele sites were converted to heterozygous sites and vice versa. Resulting VCF files included up to 10 percent more heterozygous or more homozygous sites compared to their original state.

ROHMM's HMM algorithm:

ROHMM uses a 2-state HMM to infer homozygosity from genotyping data in VCF format. *ROHMM* 's algorithm uses the following notions;

1. 2 states representing homozygous (ROH) and non-homozygous regions (NonROH).
2. Genotype at any given position i, G_i
3. Genotype likelihood at any given position is calculated by the variant caller represented in PL or GL format, "GL_i".
4. Allele distribution probability of the given genotype derived from X chromosome non-pseudoautosomal regions at given state, "P(Genotype|State)".

Using allele distribution probabilities and genotype likelihoods from GL or PL fields populated by variant callers within VCF FORMAT tags or assigning user-defined PL value for the missing entries, we generated emission probabilities per site as follows.

Emission probabilities can also be calculated using population allele frequencies as in *bcftools roh* yet as an optional method of operation for *ROHMM* . The Allele Frequency Model is also included for the sake of comparison.

Transition probabilities of *ROHMM* are similar to logarithmic decay function introduced by Marioni and colleagues (Marioni et al., 2006). This function calculates dynamic transition probabilities between 2 adjacent loci as an exponential function therefore the longer the distance the larger the probability to disconnect from a previous state. This logarithmic distance decay function is summarized below.

Standard transition probability *stdtrans* is set to a default value of 0.1. Alternatively *ROHMM* also has the ability to use fixed transition probabilities given by the user but the default is the distance decay function.

The initial state probabilities of *ROHMM* is set to 0.5 to avoid any bias towards any state unlike other methods described (Magi et al., 2014; Narasimhan et al., 2016). *ROHMM* uses a viterbi decoding function to infer homozygous and non-homozygous states based on expectation maximization and calculates the

average posterior forward-backward scores for any inferred interval for quality scoring. Results are presented as a 6-column BED file indicating state, average posterior score and the number of sites used to infer the state.

Performance Testing on Simulated Data:

ROHMM , *bcftools roh* (v1.9) and *PLINK* (v1.9) were used to infer ROHs from synthetic chromosomes generated by simulation. Parameters used to infer ROHs from simulated datasets were as below

bcftools roh -O r –AF-tag AF1KG_CEU -M 100 -m ˜/genetic_map_{CHROM}_combined_-b37.txt -o bcftools_inference.bed input.vcf

plink –homozyg –vcf input.vcf

java -jar -Xmx16G rohmmcli.jar -hmm ADM -V input.vcf -O output

java -jar -Xmx16G rohmmcli.jar -hmm AFM -V input.vcf -O output

ADM and AFM parameters used in *ROHMM* 's command line correspond to the allele distribution model and allele frequency model respectively. False positive and false negative error rates were calculated as below

Performance Testing on Real Population Data:

1000G phase 3 from data was used to test *ROHMM* 's inference capabilities on real population data. *ROHMM* was run with parameters used to test the simulated data performance. Inferred homozygosity within subpopulations were compared against the inbreeding coefficient calculated via the method of moments estimator using*PLINK het* and in-house script. Meta analysis of inferences also included comparison of different ROH categories among different sub and super populations as indicated by others (Keller, Visscher, & Goddard, 2011; Magi et al., 2014; Narasimhan et al., 2016; Pemberton et al., 2012). Inferred ROHs were also compared against the heterozygosity measure as mentioned by Samuels and colleagues (Samuels et al., 2016).

Performance testing on a clinical case data:

ROHMM 's performance on a clinical case was tested on the sample published by Pippucci and colleagues (Pippucci et al., 2013). Raw sequencing data was downloaded from the repository indicated in the publication (Pippucci et al., 2013). Raw reads were quality checked and mapped onto GRCh37 according to the GATK best practices (Van der Auwera et al., 2013). Variant calling and filtering were done using GATK 4.0.*ROHMM* 's and *H3M2's* results were compared side by side for the concordance of ROH calls.

Results:

Generation of allele distribution models:

In order to generate the allele distribution model for *ROHMM* we utilized the hemizygosity of male X chromosome non-pseudoautosomal region. Hemizygosity of haploid genomes have been used to detect sequencing errors by others (Li, 2014) whereas BAF distribution of male X have been shown to mimic that of homozygous regions within autosomes by Magi et. al. (Magi et al., 2014). We hypothesized allele distribution of male X chromosome non-pseudoautosomal region should be able to infer any long runs of homozygosity using a 2-state Hidden Markov Model. BAF distributions from X chromosome non-pseudoautosomal regions were collected from 25 whole genome samples (˜30x coverage) from 1000G project phase3 using all valid biallelic SNP locations included in the latest version of gnomAD 2.1.1 dataset (Karczewski et al., 2020). Through this collection we compared male and female non-pseudoautosomal regions of X against the BAF distribution of the whole genome and homozygous regions determined using 1000G omni2.5 array data and PLINK software. Comparison indicated a high level of correlation between homozygosity and male X chromosome where as female X chromosome was in high concordance with whole genome distribution of BAF (Figure 1).

Benchmarks with Synthetic Data Sets:

ROHMM 's performance was compared against *PLINK* and*bcftools roh* with our synthetic datasets. Under various homozygosity and erroneous site levels *ROHMM* 's allele distribution model showed comparable performance against its competitors under both genome and exome simulated data scenarios. Additionally allele frequency model implemented in *ROHMM* performed similarly if not better under all conditions compared to *bcftools roh*(Figure 2).

To test the stability and performance of *ROHMM* with various levels of data density we used randomly down-sampled synthetic exome samples from our simulated datasets. *ROHMM* 's false positive rate did not increase more than 0.06 % and false negative rate did not increase more than 3.3 %. Additionally *ROHMM* 's alternative allele frequency model showed even lesser changes in both error types which was comparable to what was published for *bcftools roh*(Narasimhan et al., 2016) (Figure 3).

Performance on Real Data Sets:

To test the performance of *ROHMM* on real datasets we used 1000G integrated phase3 data. Homozygous regions of different classes were inferred from this data and we compared distributions of each class and overall homozygosity among different continents and sub-populations listed here. Initial comparison of exome scale data and genome scale data performance of *ROHMM* indicated that *ROHMM* is able to detect homozygosity at a comparable level from both types of data as it was the case when we tested down-sampled data from our synthetic benchmarks (R ˜ 0.96, p<2.2e-16) (Figure 4A).

In order to make sure that inferred homozygous regions were real or close to real homozygosity among individuals, we decided to compare the inbreeding coefficient calculated by method of moments estimator (F_{mom}) against F_{ROH} which was defined as the ratio of sites within homozygous stretches over all sites present at each individual. This comparison was performed by many others before studying the effects of inbreeding on populations as well as small pedigrees (Keller et al., 2011; Narasimhan et al., 2016; Rosenberg, Pemberton, Li, & Belmont, 2013). We noticed that when F_{mom} was compared against the F_{ROH}calculated from the total homozygosity detected by *ROHMM* we obtained a low level of correlation even when we used allele frequency model and *bcftools roh* itself. When F_{ROH} was calculated using homozygous regions longer than 0.5 kilobases the correlation between F_{mom} and F_{ROH} was more pronounced especially for superpopulations with higher consanguinity (R ˜ 0.9, p<2.2e-16) (Figure 4B-4C). This result was also parallel to what others have published before (Keller et al., 2011) but in contradiction to what was reported by Narasimhan et. al. (Narasimhan et al., 2016). This discriminant behavior between reports may need further investigation.

As we sought for different measures for testing performance under real data, a direct comparison against heterozygosity measure turned out to be a better performer. Heterozygosity measure is defined as the ratio of heterozygous sites against all homozygous non-reference sites present per individual as defined by Wang and others (Samuels et al., 2016; Wang, Raskin, Samuels, Shyr, & Guo, 2015). This measure has been tested for its usefulness when comparing populations and individuals for disease resistance and recessive phenotype associations. According to those reports heterozygosity ratio is more robust when compared to homozygosity ratio which was reported to be density dependent. Our measurements and others have also confirmed that when the number of sites is reduced, the power to detect true homozygosity is diminished (Figure 3A-3B). We decided to compare our results against heterozygosity measure and surprisingly *ROHMM* 's Allele Distribution Model showed significant correlation between heterozygosity measure of individual populations and F_{ROH} inferred from total sites within inferred homozygous segments. Previous reports from Samuels and colleagues indicated an inverse correlation albeit with a lower R^2 value. *ROHMM* 's inferences showed much higher correlation between F_{ROH} and heterozygosity measure (R < -0.9, p<2.2e-16). For South Asian populations where consanguinity is much higher this correlation coefficient is almost the same even when F_{ROH} is calculated from much less dense exome data (R < -0.9, p<2.2e-16) (Figure 5A-5B). When overall homozygosity ratios are compared between populations and sub-populations', *ROHMM's* inferences also show the differeces between subpopulations as indicated by other reports (Figure 5C-5D). Additionally, heterozygosity measure graphs by Wang et. al and homozygosity ratio calculated from *ROHMM* shows almost a perfect mirror image of each other (Figure 5C-5D compared to Figures 2A-2E from Wang et. al. (Wang et al., 2015)).

As a final comparison, we investigated the distribution of homozygous segments captured by *ROHMM* allele distribution model and allele frequency model. When we compared homozygous stretches longer than 0.5Kb and 1.5Mb we noticed that distribution of sites resemble each other regardless of the model used by *ROHMM* (Figure 6). This result further supports the idea that the allele distribution model is as useful as the allele frequency model when used with population scale data.

Performance on single sample clinical data:

To test the performance of *ROHMM* on single sample clinical data we used the publicly available case data from Pippucci and colleagues (Pippucci et al., 2013). This test was also a means to show the unique abilities of *ROHMM* in a single sample case where homozygous reference sites are almost always not available from the variant call format. We tested *ROHMM* using 3 different settings to infer homozygosity from this data. *ROHMM* was able to detect the long homozygous stretch containing the *CACNA2D2* NM_006030.4:c.1295del mutation in the proband as it was detected by the original study (Pippucci et al., 2013). Authors of the original study used a predefined set of SNPs to infer homozygosity and we tried simulating a similar input using a BED file containing the same set of SNPs with*ROHMM* 's spike-in function (Setting3). We observed that the spike-in functionality further enabled the detection of potentially cryptic short ROH's that are otherwise not visible from the variant sites only data (Figure 7).

Discussions:

Efforts to detect homozygosity from genotyping data have resulted in many different tools and algorithms. Sliding window and Hidden Markov model approaches have been proposed as means to estimate homozygous segments from various different data types (Ceballos, Hazelhurst, & Ramsay, 2018; Howrigan, Simonson, & Keller, 2011). Sliding window approaches have been useful especially when working with dense genotyping arrays where allele densities are usually uniform and error rates are low compared to sequencing based methods. *GERMLINE* and*s* are two representatives of the early sliding window algorithms where the latter is still widely used by many studies utilizing homozygosity mapping (Gusev et al., 2009; Purcell et al., 2007). However both tools have been particularly targeted for dense genotyping arrays and their performance under sparse and error prone data types generated by next generation sequencing is questionable. Earlier algorithms using HMM approaches were also present, yet their primary target is high quality dense genotyping array data and their applicability to next generation sequencing data is limited (Leutenegger et al., 2003; Marioni et al., 2006). Newer HMM approaches like *H3M2* , *Filtus* and*bcftools roh* mostly target sparse and error prone next generation sequencing data. *H3M2* uses a predefined set of SNPs along with a heterogenous HMM to incorporate allelic distances as in BioHMM (Marioni et al., 2006) and gaussian mixture probabilities of B-allele frequencies to calculate genotypic probability under different states. *Filtus*uses a modified version of Leutenegger's algorithm to detect autozygosity in next generation sequencing data (Vigeland, Gjøtterud, & Selmer, 2016). *bcftools roh* on the other hand uses allele frequencies as genotypic probabilities and utilizes genome wide recombination maps to calculate state transitions between consecutive allele positions. Both approaches have advantages over using sliding window algorithms when used with next generation sequencing data (Magi et al., 2014; Narasimhan et al., 2016).

Here we present *ROHMM* as a flexible HMM implementation for homozygosity mapping using high throughput sequencing data.*ROHMM* 's unique approach relies on observed allele distributions in X chromosome non-pseudoautosomal regions in male and female samples. Utilization of different approaches were present in other tools namely*H3M2* , *bcftools roh* , *Filtus* . *ROHMM* 's design approach resembles the strategy in between *H3M2* and*bcftools roh* with the additional user friendliness from the graphical user interface. *H3M2's* design is not suitable for population scale data, whereas lack of proper allele frequencies and recombination maps limits *bcftools roh's* functionality under the condition of limited number samples. *ROHMM* on the other hand is free from these limitations and can be utilized freely and flexibly on all types of data.

ROHMM 's performance under simulated data showed that *ROHMM*is vastly superior to sliding window algorithms. False negative rate of sliding window algorithms especially under sparse genotyping data is limiting their usability. *ROHMM* on the other hand can perform stably even when data density is further

lowered. During our simulated data tests we observed a direct correlation between $F_{ROH}$and F_{mom} however we noted that this correlation may not be used as a direct measure of performance under real data unlike what was reported by others. Obviously our simulation data does not contain the linkage disequilibrium present in the actual population data and assumes that all sites were present individually but there may be other reasons. One possibility is that others might have pruned the 1000G data to an extend that was not reported in detail. Secondly imputation within 1000G data might have introduced excessive heterozygosity to regions that were not properly genotyped by high throughput sequencing. Both possibilities may need further investigation. Nevertheless, the ratio of sites within homozygous stretches above 500Kb show high correlation with the F_{mom} calculated from 1000G data. When we compared exome and genome inferences of homozygous stretches above 500Kb we noticed the similar levels of correlation reported by others, further supporting the stable and robust performance of *ROHMM* . Surprisingly heterozygosity ratio showed a much pronounced correlation with *ROHMM* 's inferences. Previous studies showing correlation albeit with a lesser "R^2" values suggest that homozygosity inference methods used by those studies are sub-optimal hence supporting the *ROHMM* 's precision and accuracy under real data. Additional support comes from the distribution of long homozygous stretches inferred by *ROHMM* 's allele distribution model and allele frequency model. When compared against each other homozygous stretches above 500Kb, especially above 1.5Mb shows high concordance between two models suggesting that allele distribution model can be used for population scale data. Narasimhan and colleagues reported that the power to detect true autozygosity diminishes with the reduced number of samples as the emission states are dependent on calculated allele frequencies (Narasimhan et al., 2016). Since allele distribution model is not affected from population size and allele frequencies, this may further indicate that *ROHMM* 's default model may be even more suitable to any size of population or cohort data. On the clinical data,*ROHMM* was able to detect homozygosity signals within a single sample and enhancements implemented within *ROHMM* enabled to fine tune the inference even further especially for shorter segments that are not evident from the VCF data only.

We recommend *ROHMM* to any user for detecting homozygosity with confidence. We believe that the unique qualities presented here will be make *ROHMM* a go to tool for all kinds of homozygosity analyses.

Source code and binary availability:

ROHMM 's source code and precompiled binary are freely available from*https://github.com/gokalpcelik/ROHMMCLI* . Compilation under various operating systems only requires Java1.8 SDK and above.

Conflict of Interest:

None declared.

Funding Sources and Acknowledgements:

No funding sources were used for this study.

References:

Alsalem, A. B., Halees, A. S., Anazi, S., Alshamekh, S., & Alkuraya, F. S. (2013). Autozygome sequencing expands the horizon of human knockout research and provides novel insights into human phenotypic variation.*PLoS Genetics* , *9* (12), e1004030. doi: 10.1371/journal.pgen.1004030

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., ... Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature* , *526* (7571), 68–74. doi: 10.1038/nature15393

Bittles, A. H., & Black, M. L. (2010). Consanguinity, human evolution, and complex diseases. *Proceedings of the National Academy of Sciences* , *107* (suppl 1), 1779–1786. doi: 10.1073/pnas.0906079106

Cavalli-Sforza, L. L., & Bodmer, W. F. (1999). *The Genetics of Human Populations* . Courier Corporation.

Ceballos, F. C., Hazelhurst, S., & Ramsay, M. (2018). Assessing runs of Homozygosity: A comparison of SNP Array and whole genome sequence low coverage data. *BMC Genomics* , *19* (1), 106. doi: 10.1186/s12864-018-4489-0

Ceballos, F. C., Joshi, P. K., Clark, D. W., Ramsay, M., & Wilson, J. F. (2018). Runs of homozygosity: Windows into population history and trait architecture. In *Nature Reviews Genetics* . doi: 10.1038/nrg.2017.109

Chahrour, M. H., Yu, T. W., Lim, E. T., Ataman, B., Coulter, M. E., Hill, R. S., ... Walsh, C. A. (2012). Whole-exome sequencing and homozygosity analysis implicate depolarization-regulated neuronal genes in autism. *PLoS Genetics* , *8* (4), e1002635. doi: 10.1371/journal.pgen.1002635

Gibson, J., Morton, N. E., & Collins, A. (2006). Extended tracts of homozygosity in outbred human populations. *Human Molecular Genetics* , *15* (5), 789–795. doi: 10.1093/hmg/ddi493

Gusev, A., Lowe, J. K., Stoffel, M., Daly, M. J., Altshuler, D., Breslow, J. L., ... Pe'er, I. (2009). Whole population, genome-wide mapping of hidden relatedness. *Genome Research* , *19* (2), 318–326. doi: 10.1101/gr.081398.108

Howrigan, D. P., Simonson, M. A., & Keller, M. C. (2011). Detecting autozygosity through runs of homozygosity: A comparison of three autozygosity detection algorithms. *BMC Genomics* , *12* . doi: 10.1186/1471-2164-12-460

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., ... MacArthur, D. G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans.*Nature* , *581* (7809), 434–443. doi: 10.1038/s41586-020-2308-7

Keller, M. C., Visscher, P. M., & Goddard, M. E. (2011). Quantification of Inbreeding Due to Distant Ancestors and Its Detection Using Dense Single Nucleotide Polymorphism Data. *Genetics* , *189* (1), 237–249. doi: 10.1534/genetics.111.130922

Leutenegger, A.-L., Prum, B., Génin, E., Verny, C., Lemainque, A., Clerget-Darpoux, F., & Thompson, E. A. (2003). Estimation of the Inbreeding Coefficient through Use of Genomic Data. *American Journal of Human Genetics* , *73* (3), 516–523.

Li, H. (2014). Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* ,*30* (20), 2843–2851. doi: 10.1093/bioinformatics/btu356

Magi, A., Tattini, L., Palombo, F., Benelli, M., Gialluisi, A., Giusti, B., ... Pippucci, T. (2014). H3M2: Detection of runs of homozygosity from whole-exome sequencing data. *Bioinformatics (Oxford, England)* , *30* (20), 2852–2859. doi: 10.1093/bioinformatics/btu401

Marioni, J. C., Thorne, N. P., & Tavaré, S. (2006). BioHMM: A heterogeneous hidden Markov model for segmenting array CGH data.*Bioinformatics* , *22* (9), 1144–1146. doi: 10.1093/bioinformatics/btl089

Narasimhan, V., Danecek, P., Scally, A., Xue, Y., Tyler-Smith, C., & Durbin, R. (2016). BCFtools/RoH: A hidden Markov model approach for detecting autozygosity from next-generation sequencing data.*Bioinformatics* , *32* (11), 1749–1751. doi: 10.1093/bioinformatics/btw044

Pemberton, T. J., Absher, D., Feldman, M. W., Myers, R. M., Rosenberg, N. A., & Li, J. Z. (2012). Genomic patterns of homozygosity in worldwide human populations. *American Journal of Human Genetics* ,*91* (2), 275–292. doi: 10.1016/j.ajhg.2012.06.014

Pippucci, T., Parmeggiani, A., Palombo, F., Maresca, A., Angius, A., Crisponi, L., ... Carelli, V. (2013). A novel null homozygous mutation confirms CACNA2D2 as a gene mutated in epileptic encephalopathy. *PLoS ONE* , *8* (12). doi: 10.1371/journal.pone.0082154

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R. R., Bender, D., ... Sham, P. C. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses.*American Journal of Human Genetics* , *81* (3), 559–575. doi: 10.1086/519795

Rosenberg, N. A., Pemberton, T. J., Li, J. Z., & Belmont, J. W. (2013). Runs of homozygosity and parental relatedness. *Genetics in Medicine* , *15* (9), 753–754. doi: 10.1038/gim.2013.108

Samuels, D. C., Wang, J., Ye, F., He, J., Levinson, R. T., Sheng, Q., ... Guo, Y. (2016). Heterozygosity ratio, a robust global genomic measure of autozygosity and its association with height and disease risk. *Genetics* , *204* (3), 893–904. doi: 10.1534/genetics.116.189936

Szpiech, Z. A., Xu, J., Pemberton, T. J., Peng, W., Zöllner, S., Rosenberg, N. A., & Li, J. Z. (2013). Long Runs of Homozygosity Are Enriched for Deleterious Variation. *American Journal of Human Genetics* , *93* (1), 90–102. doi: 10.1016/j.ajhg.2013.05.003

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., ... DePristo, M. A. (2013). From FastQ data to high confidence variant calls: The Genome Analysis Toolkit best practices pipeline. *Current Protocols in Bioinformatics ,43* , 11.10.1-33. doi: 10.1002/0471250953.bi1110s43

Vigeland, M. D., Gjøtterud, K. S., & Selmer, K. K. (2016). FILTUS: A desktop GUI for fast and efficient detection of disease-causing variants, including a novel autozygosity detector.*Bioinformatics* , *32* (10), 1592–1594. doi: 10.1093/bioinformatics/btw046

Walsh, T., Shahin, H., Elkan-Miller, T., Lee, M. K., Thornton, A. M., Roeb, W., ... Kanaan, M. (2010). Whole exome sequencing and homozygosity mapping identify mutation in the cell polarity protein GPSM2 as the cause of nonsyndromic hearing loss DFNB82. *American Journal of Human Genetics* , *87* (1), 90–94. doi: 10.1016/j.ajhg.2010.05.010

Wang, J., Raskin, L., Samuels, D. C., Shyr, Y., & Guo, Y. (2015). Genome measures used for quality control are dependent on gene function and ancestry. *Bioinformatics (Oxford, England)* , *31* (3), 318–323. doi: 10.1093/bioinformatics/btu668

Hosted file

`Table1_20210831.docx` available at https://authorea.com/users/436671/articles/538832-rohmm-a-flexible-hidden-markov-model-framework-to-detect-runs-of-homozygosity-from-genotyping-data