WHODAD User Manual

Xiang Zhou

October 20, 2015

Contents

1 Introduction

1.1 What is WHODAD

WHODAD is a software package implementing the WHODAD method for paternity inference using low-coverage sequencing data [3]. WHODAD consists of two models – the WHODAD naive Bayes classifier and the WHODAD mixture model – that are used jointly for paternity assignment. These two models are implemented in the "whodad" C++ binary executable and the "whodad-mixture" R script, respectively. The WHODAD software package also contains a few tools for parsing files. These include the "parse-vcf" to parse vcf files into whodad format, "filter" to filter SNPs based on quality values (note that filtering can also be done using other programs, such as "vcftools", prior to parsing the vcf), and "make-pair-file" to extract mother child pair information. The WHODAD software package uses freely available open-source numerical libraries.

1.2 How to Cite WHODAD

Noah Snyder-Mackler, William H. Majoros, Michael L. Yuan, Amanda Shaver, Jacob B. Gordon, Gisela Kopp, Stephen Schlebusch, Jeffrey D. Wall, Susan C. Alberts, Sayan Mukherjee, Xiang Zhou and Jenny Tung (2015). Efficient genome-wide sequencing and low coverage pedigree analysis from non-invasively collected samples. bioRxiv.

1.3 Installing and Compiling WHODAD

If you have downloaded a binary executable, no installation is necessary. In some cases, you may need to use "chmod a+x whodad" before using the binary executable. In addition, notice that the end-of-line coding in Windows (DOS) is different from that in Linux, and so you may have to convert input files using the utility *dos2unix* or *unix2dos* in order to use them in a different platform.

If you want to compile WHODAD by yourself, you will need to download the source code, and you will need a standard C/C++ compiler such as GNU gcc. A sample Makefile is provided along with the source code, but you may need to change the library paths in the Makefile according to your system.

1.4 Acknowledgment

The current version of the software builds upon an early implementation of the naive Bayes classifier from William Majoros.

2 Input File Format

2.1 WHODAD Bayes Classifier

The WHODAD Bayes classifier requires three input files: a who file that contains genotype information for all individuals, a pair file that contains genotype information for mother-offspring dyads, and an eligibility list file. The first two files can be parsed from a vcf file using tools included in the WHODAD software package.

2.1.1 WHO File

This is a tab-delimited file containing genotype information for individuals (e.g. all father candidates together with mother and offsprings). Each row contains the genotype information for an individual. The first and second columns contain individual id. The remaining columns contain genotype for all loci. For each locus, the genotype information is represented by the probabilities (multiplied by 1,000) of the three genotypes together with the number of reads mapped to that locus. A value of Z is used to indicate missing data. For example, a value ”10,205,786,2” means that the genotype for the individual at the given locus is BB with probability 0.010, Bb with probability 0.205 and bb with probability 0.786 (notice that the probabilities are rounded and may not sum to one exactly). In addition, the individual has 2 reads mapped to this locus. An example who file with two individuals and three loci are as follows:

```
individual  GOS   Z   Z   10,205,786,2
individual  WEN   472,472,56,3   Z   Z
```

One can parse a vcf file into a who file using the program “parse-vcf”.

2.1.2 PAIR File

This is a tab-delimited file containing genotype information for mother-offspring dyads. Each mother-offspring dyad is contained in three rows: the first row contains the text ”mother-child:”; the second row contains the mother’s genotype; and the third row contains offspring’s genotype. All mother-offspring dyads are separately by an empty line. An example pair file with two mother-offspring dyads and three loci are as follows:

```
mother-child:
individual   WYN   Z   Z   651,331,18,1
individual   BIO   Z   Z   786,205,10,2

mother-child:
individual   WYN   495,495,10,3   379,379,243,1   Z
individual   BOT   Z   Z   Z
```

One can parse a who file into a pair file using the program “make-pair-file”.

2.1.3 List File

There are two list files used in WHODAD: the mother-offspring list file and the mating eligibility list. Both files are a two-column, tab-delimited list. The mother-offspring list file contains mother-offspring pairs that are needed for paternity assignment (the first columns lists mother ids and the second column lists children ids), while the mating eligibility list file contains all possible mother-father combinations (the first column lists mother ids and the second column lists all her possible mating candidates). An example file with two mothers and three father candidates is as follows:

```
WYN   JAG
WYN   VIP
WOB   VIP
WOB   PEC
```

You can set the mother id to NA if there is no genotype information for the mother.

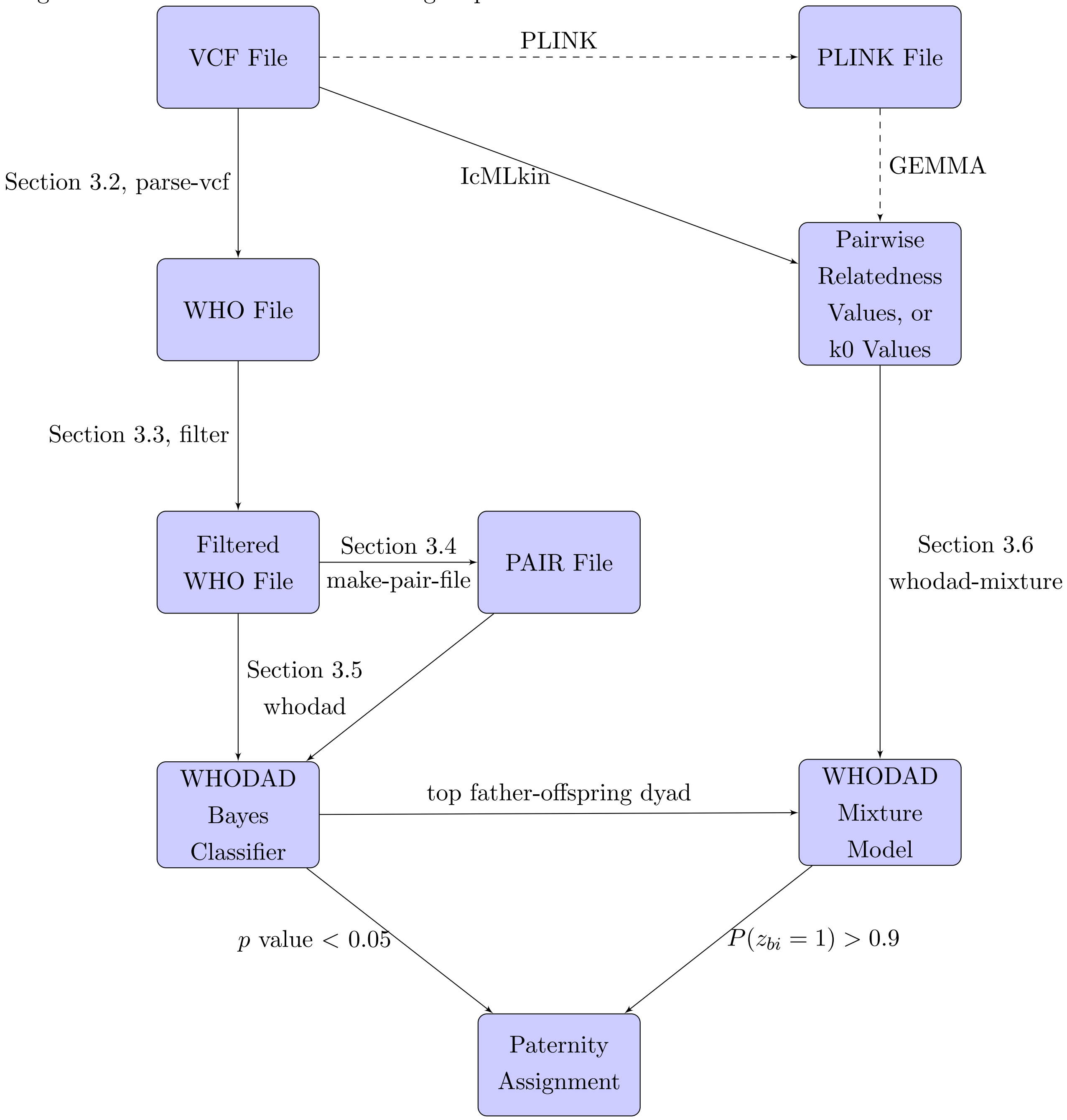
2.2 WHODAD Mixture Model

The WHODAD mixture model requires a list of pair-wise relatedness estimates for three categories of dyads: candidate-child dyads, top candidate-child dyads (determined by the WHODAD Bayes classifier), and mother-child dyads (optional). Since whodad-mixture is an R script, you can directly provide these values in three vectors and fit the model. You can obtain the relatedness estimates from GEMMA [5, 4, 6]. Please refer to the online manual of GEMMA (www.xzlab.org/software.html) for details. Alternatively, you can obtain the relatedness estimates from IcMLkin [1].

3 Running WHODAD

3.1 Overview

Running WHODAD consists of the following steps:



For converting vcf files to plink files in binary format, please refer to the PLINK [2] online manual. For computing pair-wise relatedness values from plink files using GEMMA, please refer to the GEMMA [5, 4, 6] online manual (www.xzlab.org/software.html). For computing k0 values from VCF files, please refer to IcMLkin [1] for details.

3.2 Parsing VCF Files

Programs in the WHODAD package take who file and pair file as input rather than the VCF file. Therefore, the VCF file must first be converted to a who file, from which a pair file can be extracted. Converting VCF file to a who file is accomplished by using the "parse-vcf" program. The "parse-vcf" program may need to be run on a machine with lots of memory if the VCF file is large.

The usage statement for "parse-vcf" is:

```
./parse-vcf [vcf file] [pseudocount] [who file] [min-distance] [max-distance] [max-#sites]
```

where min-distance (optional) and max-distance (optional) specify the minimal and maximal distance allowed between sites; if the probability for a particular genotype is zero, this probability will be replaced with a small pseudocount (e.g. 0.01) for numerical reasons.

3.3 Filtering WHO Files

The program "filter" can be used to eliminate unwanted loci in the who file based on some criterion, such as quality score or entropy. The usage statement is:

```
filter [options] [input who file] [output who file] [feature] [min] [max]
where [feature] is one of:
       ENTROPY : filter based on relative entropy (H/Hmax)
       ENTROPY_N : sort by RelEnt, then take top N loci (use dot for max)
       QUALITY : filter based on quality score (log(SNPqual))
       PRESENCE : filter based on percent nonmissing in population
       NUMBER : randomly choose <min> many loci (use dot for <max>)
   (use dot in place of min or max if none)
   -E = use evenness (H/Hmax) instead of relative entropy
```

3.4 Generating PAIR Files

The program "make-pair-file" can be used to extract a mother-offspring pair file from a who file. The usage statement is:

```
make-pair-file <input who file> <mother-child list file> <output pair file>
```

3.5 Fitting the WHODAD Bayes Classifier

Prediction with the WHODAD Bayes classifier is done using the “whodad” program:

```
whodad [options] [pair file] [who file]
  where: -e [eligibility list file] = use mating eligibility list
         -f X = start at index X ("first")
         -l X = end at index X-1 inclusive ("last")
         -m = use loci missing by mother
         -b = infer both the mother and the father
         -p X = emit posteriors of top X contenders
         -n X = the number of simulations used to compute the p-value (default: 100)
```

The output looks like this:

```
WYN   BIO   ELV   0.0303656   9495   0.00568502
WYN   BOT   ISR   0.0320056   20368   0.0496029
```

where the columns are: mother id, child id, top father candidate id, a delta value that measures the posterior difference between the top candidate and the second best candidate, number of loci, and *p*-value.

3.6 Fitting the WHODAD Mixture Model

Prediction with the WHODAD mixture model is done using the “WHODAD-mixture” function in the “whodad-mixture” R script. The function requires three numeric vectors:

- yb is a vector of pair-wise relatedness values (or k0 values) for all top candidate-offspring dyads
- y0 is a vector of pair-wise relatedness values (or k0 values) for all known mother-offspring dyads
- y1 is a vector of pair-wise relatedness values (or k0 values) for all other dyads.

Note that if there are no mother-offspring dyads then y0 is an empty vector, or y0=c().

For each each top candidate-offspring dyad, the function will output the posterior probability of it being a father-offspring dyad.

3.7 Paternity Assignment

With the output from the WHODAD Bayes classifer and the WHODAD mixture model, now we can assign paternity for all mother-offspring pairs. We recommend using both the p-value output from the WHODAD Bayes classifer and the posterior probability output from the WHODAD mixture model to assign paternity. In particular, if the top candidate-offspring dyad from the WHODAD Bayes classifer has a p-value below 0.05, and also has a posterior probability from the WHODAD mixture model above 0.90, then we can assign this candidate as the father for the offspring.

References

[1] Mikhail Lipatov, Komal Sanjeev, Rob Patro, and Krishna Veeramah. Maximum likelihood estimation of biological relatedness from low coverage sequencing data. *bioRxiv*, 2015:023374, 2015.

[2] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I. W. de Bakker, Mark J. Daly, and Pak C. Sham. PLINK: a toolset for whole-genome association and population-based linkage analysis. *The American Journal of Human Genetics*, 81:559–575, 2007.

[3] Noah Snyder-Mackler, William H. Majoros, Michael L. Yuan, Amanda Shaver, Jacob B. Gordon, Gisela Kopp, Stephen Schlebusch, Jeffrey D. Wall, Susan C. Alberts, Sayan Mukherjee, Xiang Zhou, and Jenny Tung. Efficient genome-wide sequencing and low coverage pedigree analysis from non-invasively collected samples. *bioRxiv*, 2015, 2015.

[4] Xiang Zhou, Peter Carbonetto, and Matthew Stephens. Polygenic modelling with Bayesian sparse linear mixed models. *PLoS Genetics*, 9:e1003264, 2013.

[5] Xiang Zhou and Matthew Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, 44:821–824, 2012.

[6] Xiang Zhou and Matthew Stephens. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature Methods*, 11:407–409, 2014.