# FISHR COMMAND LINE INSTRUCTIONS

## Example usage:

ErrorFinder23.5 -bmatch FISHR.errhigh.bit96.ErrHom2.ErrHet0.Option1.MinCM0.5.gp1.N1k.bmatch -bsid FISHR.errhigh.bit96.ErrHom2.ErrHet0.Option1.MinCM0.5.gp1.N1k.bsid -bmid FISHR.errhigh.bit96.ErrHom2.ErrHet0.Option1.MinCM0.5.gp1.N1k.bmid -reduced 100 2.0 -ped-file /work/KellerLab/mmkeller/Find.Optimal.Pihats/Simulated.Data/Final.Phased.Data2/Beagle.Phased.Group1.1k.ped -window 50 -ma-threshold 0.2 -gap 2 -pct-err-threshold 0.8 -trueCM 6 -trueSNP 500 -count.gap.errors TRUE -PIE.dist.length 3 -output.type finalErrorsOutput -log.file s8v2t1 | gzip > s8v2t1.gz

## Commands representations:

- **-bmatch:** This is the path to the bmatch file to be used by ErrorFinder.
- **-bsid:** This is the path to the bsid file.
- **-bmid:** This is the path to the bmid file.
- **-ped-file:** This is the path to the ped file.
- **-reduced (snp)(cM):** The reduced option takes two values, the first is the minimum acceptable SNP length, and the second is the minimum acceptable cM length. Every SH will be checked against these values. If a SH has a SNP length OR a cM length below the reduced value, it will be dropped and will not undergo any further processing by ErrorFinder.
- **-window:** This specifies the length in SNPs of the sliding window that is used in the moving averages calculations. We will describe the moving averages

algorithm, and why we would want to use it, in more detail in a later section. Generally speaking, a value of 50 snps is used by default.

- **-ma-threshold:** This is also used in moving average calculations. Basically, after processing all TrulyIBD data in the data set, a moving averages value is found for each of those TrulyIBD segments, and is stored in a sorted array. This value (between 0.0 and 1.0) is an index into that array. It is saying, “take the Xth percentile of the trulyIBD moving averages array”.For example, if we have 10 truly ibd ma values sorted as [0.1,0.1,0.2,0.2,0.2,0.5,0.6,0.7,0.8,0.9] then a -ma-threshoold value of 70 will give the 70th percentile value, in this case, 0.6.
- **-empirical-ma-threshold:** Supplying this parameter bypasses the -ma-threshold calculations entirely, and allows you to specify the threshold you wish to use for moving averages without calculating it via truly ibd segments. It should be noted that the use of -ma-threshold and -empirical-ma-threshold are **mutually exclusive**.
- **-gap:** SHs from GERMLINE can be broken up by opposite homozygote SNPs when the underlying segment is IBD due to SNP call errors or to deletions. FISHR looks for same-person SHs that have a gap of K SNPs between them and puts them back together under the assumption that two long SHs that are separated by K or fewer SNPs are likely to be IBD.
- **-trueCM M (6):** M is the length in cM for a SH to be deemed truly IBD (for finding pct-err-threshold; see above). If few SHs are expected to exist in the dataset of this cM length (due to a small sample size or high error SNP data), then M may need to be lt 6. Check the line “No of matches assumed to be IBD for deriving -pct-err-threshold and (if applicable) holdout-threshold” in the log file; if this is fewer than 100, then consider lowering M. However, if M gets too small (e.g., lt 5), then non-IBD SHs begin to be counted, increasing the threshold and resulting in more false positive SH calls

- **-trueSNP N (600):**N is the length in SNPs for a SH to be deemed truly IBD (for finding pct-err-threshold; see above). If few SHs are expected to exist in the dataset of this SNP length (due to a small sample size or high error SNP data), then N may need to be lt 600. Check the line “No of matches assumed to be IBD for deriving -pct-err-threshold and (if applicable) holdout-threshold” in the log file; if this is fewer than 100, then consider lowering N. However, if N gets too small (e.g., lt 400), then non-IBD SHs begin to be counted, increasing the threshold and resulting in more false positive SH calls.
- **-log.file R:**Name of the log file, which outputs parameters, timing, and other relevant information.
- **-output.type S:**Controls what type of output is produced. For most situations, this should be “full”; other options are useful for understanding which SH were dropped and why. [In brackets are the number of SHs assuming 10K SHs following GERMLINE, 8K following consolidation, and 2K that were not dropped. Further, say 500 rows were dropped following holdout step]. Further stil, this option can be used with newer versions of ErrorFinder to give additiona input files to specify weights for shared haplotypes based on base pair ranges or cM ranges. S can be any of the following:
    - **finalOutput [default] (aka “Full”)** - Outputs the final called SHs. The columns of this file are pers1, pers2, bp.distance.start, bp.distance.end, no.of.snps.in.match, and cm.distance [2k rows from above; or 1.5k if holdout was used]
    - **FullPlusDropped** - Works exactly like full but also includes the SH's that were dropped and the reason for the drop. [8k rows]
    - **weightedOutput** - Weights SH based on a weighting algorithm
    - **weightedOutputBP** - User must also provide an input file that provides a range of basepairs and corresponding weights

- **weightedOutputCM** - Similar to weightedOutputBP, except that the input file contains ranges of cM and their corresponding weights.