MR. BRENDAN F WRINGE (Orcid ID : 0000-0002-9482-5534)
DR. ERIC C ANDERSON (Orcid ID : 0000-0003-1326-0840)

Article type : Resource Article

hybriddetective: a workflow and package to facilitate the detection of hybridization using genomic data in R

Brendan F. Wringe[1*], Ryan R. E. Stanley[1], Nicholas W. Jeffery[1], Eric C. Anderson[2], and Ian R. Bradbury[1]

[1] Science Branch, Department of Fisheries and Oceans Canada, 80 East White Hills Road, St. John's NL, A1C 5X1

[2] Fisheries Ecology Division, National Oceanic and Atmospheric Administration Southwest Fisheries Science Center, Santa Cruz, CA, 95060

[*] Corresponding author, bwringe@gmail.com

Running title: hybriddetective: hybrid detection workflow

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/1755-0998.12704

This article is protected by copyright. All rights reserved

Author Manuscript

Keywords: hybrid, introgression, population genetics, population structure, assignment tests, simulation

Author Manuscript

Abstract

The ability to detect and characterize hybridization in nature has long been of interest to many fields of biology and often has direct implications for wildlife management and conservation. The capacity to identify the presence of hybridization, and quantify the numbers of individuals belonging to different hybrid classes, permits inference on the magnitude of, and time scale over which, hybridization has been, or is occurring. Here we present an R package and associated workflow developed for the detection, with estimates of efficiency and accuracy, of multi-generational hybrid individuals using genetic or genomic data in conjunction with the program NEWHYBRIDS. This package includes functions for the identification and testing of diagnostic panels of markers, the simulation of multi-generational hybrids, and the quantification and visualization of the efficiency and accuracy with which hybrids can be detected. Overall, this package delivers a streamlined hybrid analysis platform, providing improvements in speed, ease of use and repeatability over current *ad hoc* approaches. The latest version of the package and associated documentation are available on GitHub (https://github.com/bwringe/hybriddetective).

Introduction

Detecting and elucidating patterns of hybridization between individuals from genetically distinct populations is of interest in many fields of biology (Abbott *et al.* 2013; Payseur & Rieseberg 2016; Todesco *et al.* 2016). Naturally occurring hybrid zones - areas where genetically distinct populations come into contact and create genetically (ad)mixed offspring - are important natural laboratories to study of the

This article is protected by copyright. All rights reserved

interplay between selection and recombination (Barton & Hewitt 1985; Burke & Arnold 2001; Hilbish *et al.* 2012). These areas have provided opportunities to glean information to further model, and test hypotheses related to speciation (Abbott *et al.* 2013; Barton 2013; Dowling & Secor 1997) and the maintenance of reproductive barriers (Albrechtová *et al.* 2012; Griebel *et al.* 2015; Landry *et al.* 2007), natural selection (Johnson *et al.* 2010; Pruvost *et al.* 2013), and genetic recombination. Hybridization can also have conservation, regulatory, and legal ramifications related to the genetic structure and integrity of populations (Allendorf *et al.* 2004; Benson *et al.* 2014; Boyer *et al.* 2008; Fitzpatrick *et al.* 2015; Rostgaard Nielsen *et al.* 2016), or the introgression of domesticated (Fraser *et al.* 2010; Kidd *et al.* 2009; Noren *et al.* 2005) or transgenic (Oke *et al.* 2013; Warwick *et al.* 2003) alleles into wild populations.

In some cases, hybrid individuals can be identified morphologically (de Oliveira *et al.* 2002; Ross & Cavender 1981; Solomon & Child 1978), however morphological classification is notoriously imperfect (Baumsteiger *et al.* 2005; Esquer-Garrigos *et al.* 2015; Hardig *et al.* 2000; Neff & Smith 1979) and does not allow for the classification of hybrid category (Lamb & Avise 1987) or the examination of the effect of genetic dosage (Kierzkowski *et al.* 2011; Rieseberg 1995). In contrast, the use of Mendelian genetic markers affords researchers the ability to not only identify individuals as hybrid or purebred, but also to characterize them to specific hybrid classes (e.g. pure, F_1, F_2 and backcrosses). This ability to quantify the types, and numbers of individuals of different hybrid classes present, allows inferences to be made on the magnitude of, and time scale over which, hybridization has been, or is occurring (Anderson & Thompson 2002; Brown *et al.* 2004; Godinho *et al.* 2015; Saarman & Pogson 2015).

Many statistical approaches have been put forward to use genetic markers to identify hybrids (Anderson 2009), and some of these have been incorporated into widely used, and cited software programs (e.g. NEWHYBRIDS [Anderson & Thompson 2002]; STRUCTURE [Hubisz *et al.* 2009]; GENODIVE [Meirmans & Van Tienderen 2004]). However, the analyses conducted by these programs is but one step in the path to go from individual genotypes, to the detection and assignment of

This article is protected by copyright. All rights reserved

Author Manuscript

those individuals to a hybrid class, with quantifiable levels of certainty. The process of performing hybrid analyses currently entails the use of multiple, standalone programs, many of which require data to be provided in a unique format (Lischer & Excoffier 2012; Stanley *et al.* 2017). Furthermore, the reliance on the user for file management, and for manually implementing individual analyses with separate programs in addition to affording opportunity for human error, leads to a disjunct analytical process with a steep learning curve that lacks the efficiency and repeatability of a true workflow.

Here we describe the R package *hybriddetective* and associated workflow for hybrid identification developed in the R computer language (R Development Core Team 2016). The package and workflow encompass every aspect of the hybrid identification procedure. Specifically, we include functions for (1) panel design, and the quantification of the efficiency, accuracy and power of panels of diagnostic markers; (2) error checking and diagnostics; and (3) quantification, and visualization of accuracy and assignment power of the selected panel(s). *hybriddetective's* simulation and panel selection functions have been designed to work in concert, as a workflow, to improve the accuracy, and reduce the overestimation of assignment certainty (Anderson & Thompson 2002), and concomitantly reduce high-grading bias (described in detail below; Anderson 2010). This package alleviates much of the complexity in the hybrid detection process, reduces the potential for human error, and at the same time offers significant speed improvements over previous *ad hoc* methodologies.

Description of the package

hybriddetective is compiled as an R package which facilitates a workflow within the R environment for the detection of hybrids based on genotypic/genomic information using the program NEWHYRIDS (Anderson & Thompson 2002), and provides a comprehensive and repeatable framework to move from genotypic data to the identification, with quantifiable certainty, of hybrid individuals.

This article is protected by copyright. All rights reserved

Author Manuscript

hybriddetective is comprised of 14 functions (Table 1), three example datasets, and a README. Function descriptions (Table 1), example data, and installation instructions are available online https://github.com/bwringe/hybriddetective. For an example of the *hybriddetective* workflow, see Jeffery *et al.* (2017), and Supplementary Figure 1 . We chose to implement hybrid detection using the program NEWHYRIDS (Anderson & Thompson 2002) because it permits the assignment of individuals to hybrid class (i.e. pure-bred, F_1, F_2, and back-crosses) and does not require *a priori* knowledge of the allele frequencies of the two populations being tested (Anderson & Thompson 2002). Moreover, NEWHYBRIDS is widely used, having been cited over 800 times as of the time of this writing.

Description of the workflow

The workflow can be broken down into three major elements: 1) data preparation, 2) error checking and diagnostics, and 3) quantification and analysis. **Data preparation** encompasses the process of selecting the *n* most informative loci from amongst the genotypic data available, and the simulation of multi-generational hybrids. After analyzing the simulated data with NEWHYBRIDS, **error checking and diagnostics** functions confirm that NEWHYBRIDS MCMC chains reached convergence and **quantification and analysis** functions test, quantify, and visualize the accuracy and assignment power of the selected panel(s). The workflow and the functions used in each step are illustrated and described in in Figure 1, and Table 1, respectively. We have also included a brief section on the implementation of (parallel) NEWHYBRIDS analyses using the related R package *parallelenewhybrid* (Wringe *et al.* 2017).

Data preparation

Panel selection

This article is protected by copyright. All rights reserved

Author Manuscript

Panel selection is the process of selecting from amongst the available markers (i.e. thousands to several hundreds of thousands as produced by RAD-seq) a subset that together permit accurate identification of hybrids. In our workflow, the function *getTopLoc* is used to develop a panel of user defined size, of the most informative (based on global Weir and Cockerham (1984)'s F_{ST}) loci that are not in linkage disequilibrium (LD). Genotype data of individuals known (or suspected with high certainty (Oliveira *et al.* 2015)), to be of pure ancestry from the two populations potentially hybridizing are used as input for *getTopLoc*. *getTopLoc* first randomly creates two subsamples, each comprised of 50% of the individuals from each of the two populations, to create validation and training datasets. To prevent any "high-grading" bias (i.e. upward bias in the estimation of predictive capacity caused when the same data is used to both select and validate panels of markers), *getTopLoc* uses subsampling to ensure the same individuals are not used to create the panel and to validate it. The function uses the training dataset to calculate the global, locus-specific Weir and Cockerham's F_{ST} and ranks loci by this metric. Pairwise LD is then calculated using the training dataset for all loci within one or both populations at the users' discretion. During this process the r^2 threshold above which to consider a pair of loci to be in LD can be defined by the user. Any loci that are in LD are removed, because NEWHYBRIDS assumes no linkage, and each locus is treated as independent. *getTopLoc* returns a list of panel loci names , a list of individuals (IDs) in the validation dataset, and the genotypes of those individuals at the panel loci.

Importantly, random sampling selects the individuals in the training and validation datasets, so the individuals and corresponding panel can vary each time the function is run. The variance in global pairwise F_{ST}, and hence the loci returned between runs, will likely be greatest where sample sizes for one or both populations are small, and consequently subsampling is more apt to impart stochastic variances in allele/gene frequencies.

Construction of multi-generational simulated hybrids

Author Manuscript

This article is protected by copyright. All rights reserved

The next step in our workflow is to generate simulated multi-generational hybrid datasets using the genotypic data from the validation dataset exported by *getTopLoc*. The two simulation functions, *freqbasedsim_GTFreq* and *freqbasedsim_AlleleSample* differ in the way in which they create hybrids. *freqbasedsim_GTFreq* was designed to simulate individuals within the R environment analogously to the commonly used hybrid simulation program HYBRIDLAB (Nielsen *et al.* 2006). In *freqbasedsim_GTFreq*, like in HYBRIDLAB, individuals in generation *t+1* are created by sampling one allele per locus from the generation *t* parental populations, based on the allele frequencies in either population. Unlike HYBRIDLAB, *freqbasedsim_GTFreq* creates multi-generational hybrids, each time it is run, and requires only a single data file to do so. In controlled comparisons with HYBRIDLAB we find *freqbasedsim_GTFreq* to be more than 20X faster when creating multiple independent simulations (See Supplemental Table 1).

The other hybrid simulation function, *freqbasedsim_AlleleSample*, was designed with the intent of providing an additional simulation method. It first randomly subsamples a proportion of individuals from each of the two populations provided to it and only the alleles of these individuals will be available during the subsequent simulation. Secondly, to conduct the actual simulations, each locus in individuals in generation *t+1* is simulated by randomly sampling without replacement, one allele from among all the alleles present at that locus from one of the parental populations at time *t*, then combining it with an allele chosen in the same manner from the other parental population at time *t*. In this case, the number of individuals that can be simulated in a given hybrid generation is therefore dependent upon the number of individuals sampled in the first step.

(Parallel) NEWHYBRIDS analyses

For actual hybrid identification, we encourage users to take advantage of the R package **parallelnewhybrid**, which was developed to run NEWHYBRIDS in parallel thus providing significant speed improvements (Wringe *et al.* 2017). Furthermore, the error checking and analytical functions described below were designed to work

This article is protected by copyright. All rights reserved

Author Manuscript

with the file structure created by **parallelnewhybrid. parallelnewhybrid**, and documentation describing its installation and operation can be found at https://github.com/bwringe/parallelnewhybrid.

Error checking and diagnostics

Check Markov chain convergence

As with any MCMC process using Gibbs sampling, chain convergence in NEWHYBRIDS is dependent upon the 'topography' of the probability space relative to the starting point of the chain. Occasionally, the MCMC chains in NEWHYBRIDS analyses will fail to converge. In these cases NEWHYBRIDS will almost invariably report that (nearly) all individuals have the highest posterior probability of membership in the F_2 hybrid class, a result that is clearly erroneous. To this end, the function *nh_preCheckR* quickly checks the results of NEWHYBRIDS flagging those that may have failed to converge, and the function *nh_multiplotR* complements it by visualizing its results. *nh_preCheckR* inspects the NEWHYBRIDS output and identifies the individuals that are known to be pure-bred in origin, and checks that a user defined proportion of these individuals have not been assigned posterior probability of assignments (PofZ; Anderson 2003) to the F_2 hybrid class in excess of a user defined threshold. If these conditions are violated, the user is prompted to verify the(se) result(s). *nh_multiplotR* permits the user to visualize the cumulative posterior probability of assignment for all genotype frequency classes for each individual. *nh_multiplotR* can thus be used to confirm and compliment the results of *nh_preCheckR*, as well as quickly visualize the results of multiple NEWHYBRIDS analyses.

Quantification and analysis

Assess panel accuracy

This article is protected by copyright. All rights reserved

Author Manuscript

The next step in the workflow, after confirming convergence, is to assess the ability of NEWHYBRIDS to assign simulated individuals of known hybrid ancestry to the correct genotype frequency class given the genotypes of the individuals at the loci in the selected panel. Because it is impossible to statistically validate the assumed distribution of priors, and the efficacy of the loci in a panel *a priori* (Anderson 2003; Nielsen *et al.* 2006; Oliveira *et al.* 2008), simulations are often employed to evaluate power (Anderson 2003; Nielsen *et al.* 2006; Vähä & Primmer 2006). Also, Anderson and Thompson (2002), note that the power of NEWHYBRIDS to distinguish among genotype frequencies classes will vary across classes. Thus when evaluating a potential threshold value of posterior probability of assignment for assigning genotype frequency class membership, the effect of choice of posterior probability of assignment value on efficiency, accuracy and overall performance (Vähä & Primmer 2006), as well as on both Type I and Type II error should be considered simultaneously for each genotype frequency class, and for the differentiation of purebreds from any type of hybrid.

In order to allow researchers to better evaluate the effect of choice of critical posterior probability of assignment threshold (i.e. posterior probability value above which assignment to a given hybrid class is accepted) on assignment success, we have developed the function *hybridPowerComp*. *hybridPowerComp* calculates the number of individuals of known hybrid class correctly assigned over the total number of individuals known to belong to that class for posterior probability of assignment thresholds between 0.50 and 1.0 (i.e. number detected / number expected; "efficiency" sensu Vähä & Primmer 2006). This is done for each hybrid frequency class (Figure 2), as well as separately for the two parental classes, and all hybrids classes considered together (i.e. posterior probability of assignment for hybrid is the sum of all of F_1, F_2, BC1, BC2). In addition, *hybridPowerComp* calculates and plots the number of individuals correctly assigned to a class over the total number of individuals assigned to that class (i.e. "accuracy" sensu Vähä & Primmer 2006)(Figure 3), and the “power” (i.e. the product of "efficiency" and "accuracy" sensu Vähä & Primmer 2006) of the panel . Similarly, the number of individuals wrongly deemed to belong to hybrid genotype frequency classes divided

This article is protected by copyright. All rights reserved

Author Manuscript

by the total number of known pure individuals (i.e. type I error; Burgarella *et al.* 2009), and the proportion of individuals misclassified (i.e. type II error) are assessed and plotted. *hybridPowerComp* allows visualization of the distribution of posterior probability of assignment values by plotting them for each genotype frequency class, as well as for all hybrid classes considered together (refer to Supplementary Table 2 for a list of the plots produced by *hybridPowerComp*).

The function *nh_panel_delta_plotR* complements *hybridPowerComp* by visualizing the efficacy of different panel sizes for each genotype frequency class and can be used during the assessment of panel accuracy phase of the workflow.

Combine simulated and experimental data for analysis

Once the panel and critical posterior probability of assignment threshold(s) have been finalized, the experimental/unknown data can be analyzed. Combining simulated data with the unknown/experimental data (1) assists with the interpretation of results in the absence of known individuals, and (2) allows the user the option to designate the genotype frequency class membership of known individuals, to improve assignment power (Anderson 2003; Anderson & Thompson 2002).

The function *nh_analysis_generateR* allows researchers to specify both the unknown and experimental genotype data to analyze and the simulated data to combine with it, thus facilitating reproducibility of analyses as well as the ability to use the same simulated dataset(s) from which the critical posterior probability of assignment values were determined. The function *nh_analysis_simulateR_generateR* permits users to quickly create analysis-ready datasets when panel development and/or more conservative simulation methodology are not required. This function uses the frequency based simulation algorithm and simulation options of *freqbasedsim_GTFreq* to create simulated hybrids based on supplied genotype data, and then merge them with experimental or unknown genotypes.

Author Manuscript

This article is protected by copyright. All rights reserved

Conclusions

Here we have shown that the use of *hybriddetective* as part of a workflow in the detection of hybrids has clear and quantifiable benefits over the generally *ad hoc*, methods normally used. *hybriddetective* provides researchers an efficient platform for reproducible analyses of hybridization within the R computational language. Furthermore, the interoperability of *hybriddetective* for the simulation of multi-generation hybrid datasets and the separate R package *parallelnewhybrid* (Wringe *et al.* 2017) to efficiently and automatically execute runs of NEWHYBRIDS in parallel, makes it tractable to quantify the expected variability in hybrid assignment success.

In conclusion, we have created an R package and associated workflow for the detection, with quantifiable accuracy, efficiency and power, of multi-generational hybrid individuals using genetic or genomic data with the program NEWHYBRIDS. This package includes functions for the development and testing of diagnostic panels of markers, the simulation of multi-generational hybrids, and the quantification and visualization of the accuracy with which (simulated) hybrids can be detected. Use of this package offers improvements in the repeatability, speed, and ease of use over conventional approaches.

Acknowledgements

The authors wish to thank Marion Sinclair-Waters, Justine Létourneau, and Anne-Laure Ferchaud for their help bug-checking the code. We also thank Thierry Gosselin for encouraging us to publish this package. The manuscript was greatly improved by comments from Sarah Lehnert and three anonymous reviewers. This work was supported by a Natural Sciences and Engineering Research Council Strategic Project Grant, a Natural Sciences and Engineering Research Discovery Grant, and Canadian Healthy Oceans Network, and Fisheries and Oceans Canada

This article is protected by copyright. All rights reserved

Author Manuscript

funding (International Governance Strategy; Programme for Aquaculture Regulatory Research; Genomics Research and Development Initiative) to I.R.B.

Author contributions

B.F.W. wrote the manuscript and the package code, and developed the supporting documentation and example data files hosted on GitHub. R.R.E.S, N.F.W., E.C.A., and I.R.B. all contributed to the initial concept, development of the code, and associated documentation, as well as assisting in writing of the manuscript.

Data Accessibility

The package, user manual, README, example workflow, and example data sets are all available online from https://github.com/bwringe/hybriddetective.

Author Manuscript

This article is protected by copyright. All rights reserved

References

Abbott R, Albach D, Ansell S*, et al.* (2013) Hybridization and speciation. *Journal of Evolutionary Biology* **26**, 229-246.

Albrechtová J, Albrecht T, Baird SJE*, et al.* (2012) Sperm-related phenotypes implicated in both maintenance and breakdown of a natural species barrier in the house mouse. *Proceedings of the Royal Society B-Biological Sciences* **279**, 4803-4810.

Allendorf FW, Leary RF, Hitt NP*, et al.* (2004) Intercrosses and the US Endangered Species Act: Should hybridized populations be included as westslope cutthroat trout? *Conservation Biology* **18**, 1203-1213.

Anderson EC (2003) User's guide to the program NewHybrids Version 1.1 beta. Department of Integrative Biology, University of California, Berkeley, Berkeley, California.

Anderson EC (2009) Statistical methods for identifying hybrids and groups. In: *Population genetics for animal conservation* (eds. Bertorelle G, Bruford MW, Hauff HC, Rizzoli A, Vernesi C), pp. 25-41. Cambridge University Press, New York.

Anderson EC (2010) Assessing the power of informative subsets of loci for population assignment: standard methods are upwardly biased. *Molecular Ecology Resources* **10**, 701-710.

Anderson EC, Thompson EA (2002) A model-based method for identifying species hybrids using multilocus genetic data. *Genetics* **160**, 1217-1229.

Barton NH (2013) Does hybridization influence speciation? *Journal of Evolutionary Biology* **26**, 267-269.

Barton NH, Hewitt GM (1985) Analysis of hybrid zones. *Annual Review of Ecology and Systematics* **16**, 113-148.

Baumsteiger J, Hankin D, Loudenslager EJ (2005) Genetic analyses of juvenile steelhead, coastal cutthroat trout, and their hybrids differ substantially from

This article is protected by copyright. All rights reserved

Author Manuscript

field identifications. *Transactions of the American Fisheries Society* **134**, 829-840.

Benson JF, Patterson BR, Mahoney PJ (2014) A protected area influences genotype-specific survival and the structure of a *Canis* hybrid zone. *Ecology* **95**, 254-264.

Boyer MC, Muhlfeld CC, Allendorf FW (2008) Rainbow trout (*Oncorhynchus mykiss*) invasion and the spread of hybridization with native westslope cutthroat trout (*Oncorhynchus clarkii lewisi*). *Canadian Journal of Fisheries and Aquatic Sciences* **65**, 658-669.

Brown KH, Patton SJ, Martin KE*, et al.* (2004) Genetic analysis of interior Pacific Northwest *Oncorhynchus mykiss* reveals apparent ancient hybridization with westslope cutthroat trout. *Transactions of the American Fisheries Society* **133**, 1078-1088.

Burgarella C, Lorenzo Z, Jabbour-Zahab R*, et al.* (2009) Detection of hybrids in nature: application to oaks (*Quercus suber* and *Q. ilex*). *Heredity* **102**, 442-452.

Burke JM, Arnold ML (2001) Genetics and the fitness of hybrids. *Annual Review of Genetics* **35**, 31-52.

de Oliveira AC, Garcia AN, Cristofani M, Machado MA (2002) Identification of citrus hybrids through the combination of leaf apex morphology and SSR markers. *Euphytica* **128**, 397-403.

Dowling TE, Secor CL (1997) The role of hybridization and introgression in the diversification of animals. *Annual Review of Ecology and Systematics* **28**, 593-619.

Esquer-Garrigos Y, Hugueny B, Ibañez C*, et al.* (2015) Detecting natural hybridization between two vulnerable Andean pupfishes (*Orestias agassizii* and *O. luteus*) representative of the Altiplano endemic fisheries. *Conservation Genetics* **16**, 717-727.

Fitzpatrick BM, Ryan ME, Johnson JR, Corush J, Carter ET (2015) Hybridization and the species problem in conservation. *Current Zoology* **61**, 206-216.

This article is protected by copyright. All rights reserved

Author Manuscript

Fraser DJ, Minto C, Calvert AM, Eddington JD, Hutchings JA (2010) Potential for domesticated-wild interbreeding to induce maladaptive phenology across multiple populations of wild Atlantic salmon (*Salmo salar*). *Canadian Journal of Fisheries and Aquatic Sciences* **67**, 1768-1775.

Godinho R, López-Bao JV, Castro D*, et al.* (2015) Real-time assessment of hybridization between wolves and dogs: combining noninvasive samples with ancestry informative markers. *Molecular Ecology Resources* **15**, 317-328.

Griebel J, Giessler S, Poxleitner M*, et al.* (2015) Extreme environments facilitate hybrid superiority - the story of a successful *Daphnia galeata x longispina* hybrid clone. *PLoS One* **10**, e0140275.

Hardig TM, Brunsfeld SJ, Fritz RS, Morgan M, Orians CM (2000) Morphological and molecular evidence for hybridization and introgression in a willow (*Salix*) hybrid zone. *Molecular Ecology* **9**, 9-24.

Hilbish TJ, Lima FP, Brannock PM*, et al.* (2012) Change and stasis in marine hybrid zones in response to climate warming. *Journal of Biogeography* **39**, 676-687.

Hubisz MJ, Falush D, Stephens M, Pritchard JK (2009) Inferring weak population structure with the assistance of sample group information. *Molecular Ecology Resources* **9**, 1322-1332.

Jeffery NW, DiBacco C, Wringe BF*, et al.* (2017) Genomic evidence of hybridization between two independent invasions of European green crab (Carcinus maenas) in the Northwest Atlantic. *Heredity (Edinb)*.

Johnson JR, Fitzpatrick BM, Shaffer HB (2010) Retention of low-fitness genotypes over six decades of admixture between native and introduced tiger salamanders. *BMC Evolutionary Biology* **10**, 147.

Kidd AG, Bowman J, Lesbarreres D, Schulte-Hostedde AI (2009) Hybridization between escaped domestic and wild American mink (*Neovison vison*). *Molecular Ecology* **18**, 1175-1186.

Kierzkowski P, Pasko L, Rybacki M, Socha M, Ogielska M (2011) Genome dosage effect and hybrid morphology - the case of the hybridogenetic water frogs of the Pelophylax esculentus complex. *Annales Zoologici Fennici* **48**, 56-66.

This article is protected by copyright. All rights reserved

Author Manuscript

Lamb T, Avise JC (1987) Morphological variability in genetically defined categories of anuran hybrids. *Evolution* **41**, 157-165.

Landry CR, Hartl DL, Ranz JM (2007) Genome clashes in hybrids: insights from gene expression. *Heredity* **99**, 483-493.

Lischer HEL, Excoffier L (2012) PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics* **28**, 298-299.

Meirmans PG, Van Tienderen PH (2004) GENOTYPE and GENODIVE: two programs for the analysis of genetic diversity of asexual organisms. *Molecular Ecology Notes* **4**, 792-794.

Neff NA, Smith GR (1979) Multivariate-analysis of hybrid fishes. *Systematic Zoology* **28**, 176-196.

Nielsen EE, Bach LA, Kotlicki P (2006) HYBRIDLAB (version 1.0): a program for generating simulated hybrids from population samples. *Molecular Ecology Notes* **6**, 971-973.

Noren K, Dalen L, Kvaloy K, Angerbjorn A (2005) Detection of farm fox and hybrid genotypes among wild arctic foxes in Scandinavia. *Conservation Genetics* **6**, 885-894.

Oke KB, Westley PAH, Moreau DTR, Fleming IA (2013) Hybridization between genetically modified Atlantic salmon and wild brown trout reveals novel ecological interactions. *Proceedings of the Royal Society B-Biological Sciences* **280**.

Oliveira R, Godinho R, Randi E, Alves PC (2008) Hybridization versus conservation: are domestic cats threatening the genetic integrity of wildcats (*Felis silvestris silvestris*) in Iberian Peninsula? *Philosophical Transactions of the Royal Society B-Biological Sciences* **363**, 2953-2961.

Oliveira R, Randi E, Mattucci F*, et al.* (2015) Toward a genome-wide approach for detecting hybrids: informative SNPs to detect introgression between domestic cats and European wildcats (*Felis silvestris*). *Heredity* **115**, 195-205.

Author Manuscript

This article is protected by copyright. All rights reserved

444 Payseur BA, Rieseberg LH (2016) A genomic perspective on hybridization and
445 speciation. *Molecular Ecology* **25**, 2337-2360.
446 Pruvost NBM, Hollinger D, Reyer H-U (2013) Genotype-temperature interactions on
447 larval performance shape population structure in hybridogenetic water frogs
448 (P*elophylax esculentus* complex). *Functional Ecology* **27**, 459-471.
449 R Development Core Team (2016) *R: A language and environment for statistical*
450 *computing* R Foundation for Statistical Computing, Vienna, Austria.
451 Rieseberg LH (1995) The role of hybridization in evolution: old wine in new skins.
452 *American Journal of Botany* **82**, 944-953.
453 Ross MR, Cavender TM (1981) Morphological Analyses of Four Experimental
454 Intergeneric Cyprinid Hybrid Crosses. *Copeia* **1981**, 377-387.
455 Rostgaard Nielsen L, Brandes U, Dahl Kjaer E, Fjellheim S (2016) Introduced Scotch
456 broom (*Cytisus scoparius*) invades the genome of native populations in
457 vulnerable heathland habitats. *Molecular Ecology* **25**, 2790-2804.
458 Saarman NP, Pogson GH (2015) Introgression between invasive and native blue
459 mussels (genus *Mytilus*) in the central California hybrid zone. *Molecular*
460 *Ecology* **24**, 4723-4738.
461 Solomon DJ, Child AR (1978) Identification of juvenile natural hybrids between
462 Atlantic salmon (*Salmo salar* L) and trout (*Salmo trutta* L). *Journal of Fish*
463 *Biology* **12**, 499-&.
464 Stanley RR, Jeffery NW, Wringe BF, DiBacco C, Bradbury IR (2017) genepopedit: a
465 simple and flexible tool for manipulating multilocus molecular data in R.
466 *Molecular Ecology Resources* **17**, 12-18.
467 Todesco M, Pascual MA, Owens GL*, et al.* (2016) Hybridization and extinction.
468 *Evolutionary Applications*.
469 Vähä JP, Primmer CR (2006) Efficiency of model-based Bayesian methods for
470 detecting hybrid individuals under different hybridization scenarios and with
471 different numbers of loci. *Molecular Ecology* **15**, 63-72.
472 Warwick SI, Simard MJ, Légère A*, et al.* (2003) Hybridization between transgenic
473 *Brassica napus* L. and its wild relatives: *Brassica rapa* L., *Raphanus*

This article is protected by copyright. All rights reserved

Author Manuscript

raphanistrum L., *Sinapis arvensis* L., and *Erucastrum gallicum* (Willd.) OE Schulz. *Theoretical and Applied Genetics* **107**, 528-539.

Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population-structure. *Evolution* **38**, 1358-1370.

Wringe BF, Stanley RR, Jeffery NW, Anderson EC, Bradbury IR (2017) parallelnewhybrid: an R package for the parallelization of hybrid detection using newhybrids. *Molecular Ecology Resources* **17**, 91-95.

Author Manuscript

This article is protected by copyright. All rights reserved

Table 1 – Functions included in the *hybriddetective* R package, a synopsis of their purpose, and which of the three major elements they are used in.

Function Name	Synopsis	Main Use
getTopLoc	Creates a panel comprised of the *n* (user-specified) most informative (based on highest loci-specific F_{ST}s), markers not in linkage disequilibrium. The function randomly assigns half the individuals in each of the two populations to be used to calculate loci-specific Weir and Cockerham's F_{ST}s[1], and returns the genotypes at the *n* loci of the other half to be used to test the efficacy of the panel to avoid high-grading bias[2].	Data Preparation
freqbasedsim_GTFreq	Creates simulated multi-generational (i.e. Pure 1, Pure 2, F_1, F_2, BC1, BC2) hybrids based on the allele frequencies in the two populations provided. The user can specify the number of individuals in each of the hybrid classes to be created.	Data Preparation
freqbasedsim_AlleleSample	Creates simulated multi-generational hybrids by randomly sampling, without replacement, two alleles per loci from a proportion of the individual genotypes provided. The user is able to specify the proportion of genotypes to sample, as well as the number of individuals of each hybrid class to create.	Data Preparation
nh_analysis_generateR	Merges a file composed of simulated hybrid genotypes with a file	Data

Author Manuscript

This article is protected by copyright. All rights reserved

Author Manuscript

	containing the genotypes of unknown/experimental individuals to produce a file suitable to ascertain the hybrid class of the unknowns. The user is able to specify which hybrid classes from the simulated dataset to include in the output.	Preparation
nh_analysis_simulateR_generateR	Creates a simulated multi-generational hybrid reference dataset from user provided data, and then merges it with the genotypes of unknown/experimental individuals. This function will create a new simulated dataset each time it is run using the same simulation methodology as *freqbasedsim_GTFreq*.	Data Preparation
nh_subsetR	Removes subsets of desired loci from NEWHYBRIDS formatted files so that the efficacy of panels of various sizes can be assessed.	Data Preparation
nh_Zcore	Allows the user to assign known hybrid category designations to individuals in NEWHYBRIDS formatted files	Data Preparation
nh_preCheckR	Checks all NEWHYBRIDS results within a directory and flags those that show evidence that the Markov chain may have failed to converge. This is done by evaluating the proportion of known Pure Population 1 or 2 individuals in which the posterior probability of assignment to F_2 exceeds a threshold. The user may specify both the proportion of individuals and the PofZ threshold.	Error Checking and Diagnostics
nh_mulitplotR	Creates a cumulative probability of assignment plot for each	Error Checking

This article is protected by copyright. All rights reserved

Author Manuscript

	NEWHYBRIDS result within a user-specified directory. Compliments *preCheckR* by allowing visually verification of Markov chain (non-) convergence.	and Diagnostics
nh_plotR	Plots the cumulative probability of assignment of a single NEWHYBRIDS result. Also allows the user to match plotting colours between analyses when NEWHYBRIDS reverses which population it designates Population 1 and 2.	Quantification and Analysis
hybridPowerComp	Evaluates the accuracy[3] and efficiency[3] with which NEWHYBRIDS assigns individuals of known hybrid class to the correct class across a range of minimum posterior probability thresholds from 0.50 to 0.99. Calculates the number of individuals wrongly assigned to hybrid genotype frequency classes over the total number of known pure individuals (type I error)[4] , and the proportion of individuals misclassified (type II error). The distribution of PofZ values for each genotype frequency class, as well as for all hybrid classes considered together is plotted. The effect of varying panel sizes on each of these variables is also evaluated. Plots are returned as .pdf and .jpg files, and all data frames constructed for plotting are exported.	Quantification and Analysis
nh_accuracy_checkR	Evaluates the accuracy with which NEWHYBRIDS assigns individuals of known hybrid class to the correct class for a single analysis at three	Quantification and Analysis

This article is protected by copyright. All rights reserved

	minimum posterior probability thresholds (PofZ >= 0.05, 0.75 and 0.90). This function is meant to compliment *hybridPowerComp*.	
nh_panel_delta_plotR	Plots the genotype class assignment (class with max. PofZ) of individuals among panels of different size. Allows visualization of the stability of individual assignments to compliment the proportion of correct assignments returned by *hybridPowerComp*.	Quantification and Analysis
nh_build_Example_Data	Writes example NEWHYBRIDS results to be evaluated with the function *hybridPowerComp*	

[1] Weir and Cockerham (1984)

[2] Anderson (2010)

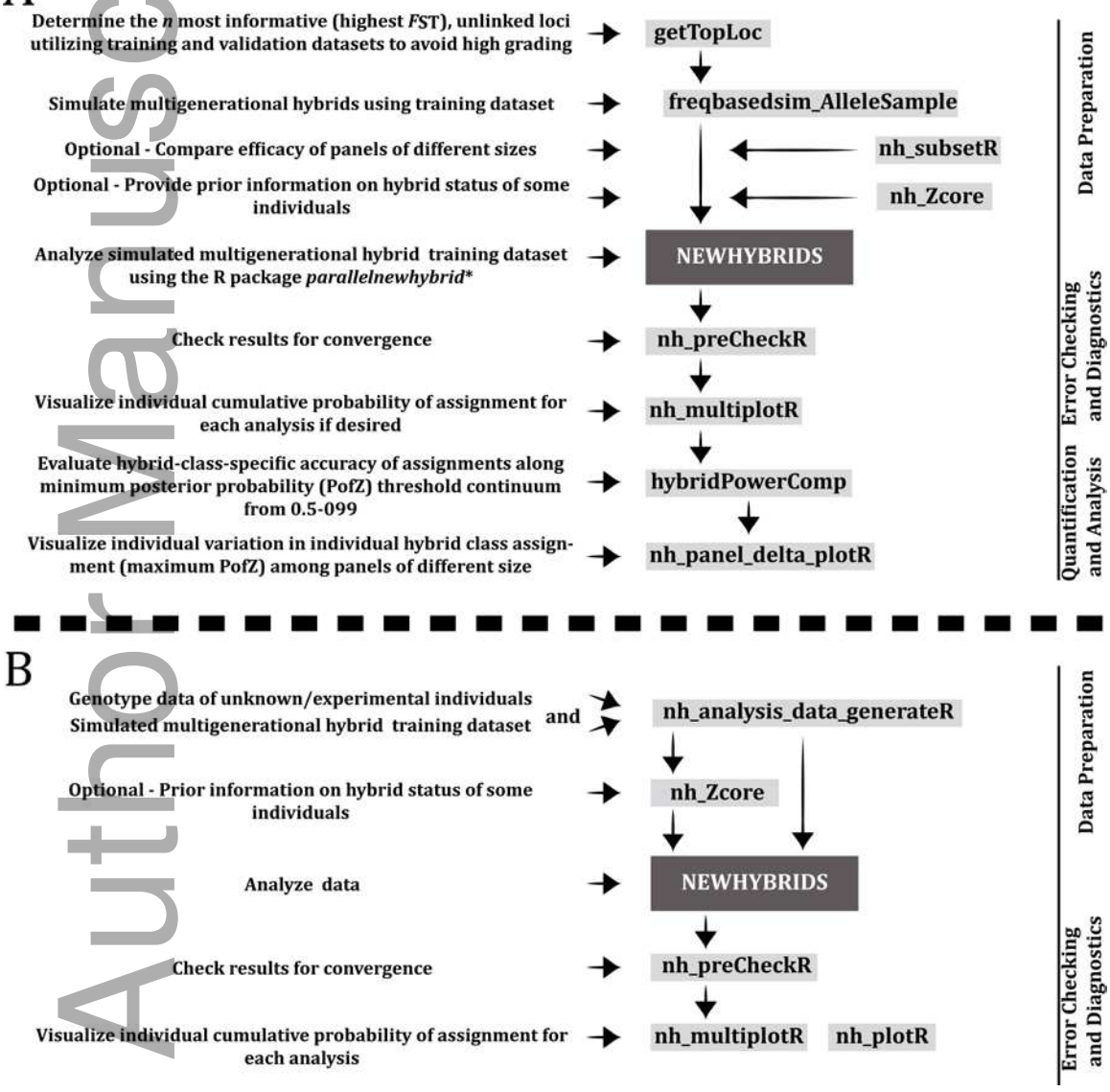
[3] Vaha and Primmer (2006)

[4] Burgarella et al. (2009)

Author Manuscript

This article is protected by copyright. All rights reserved

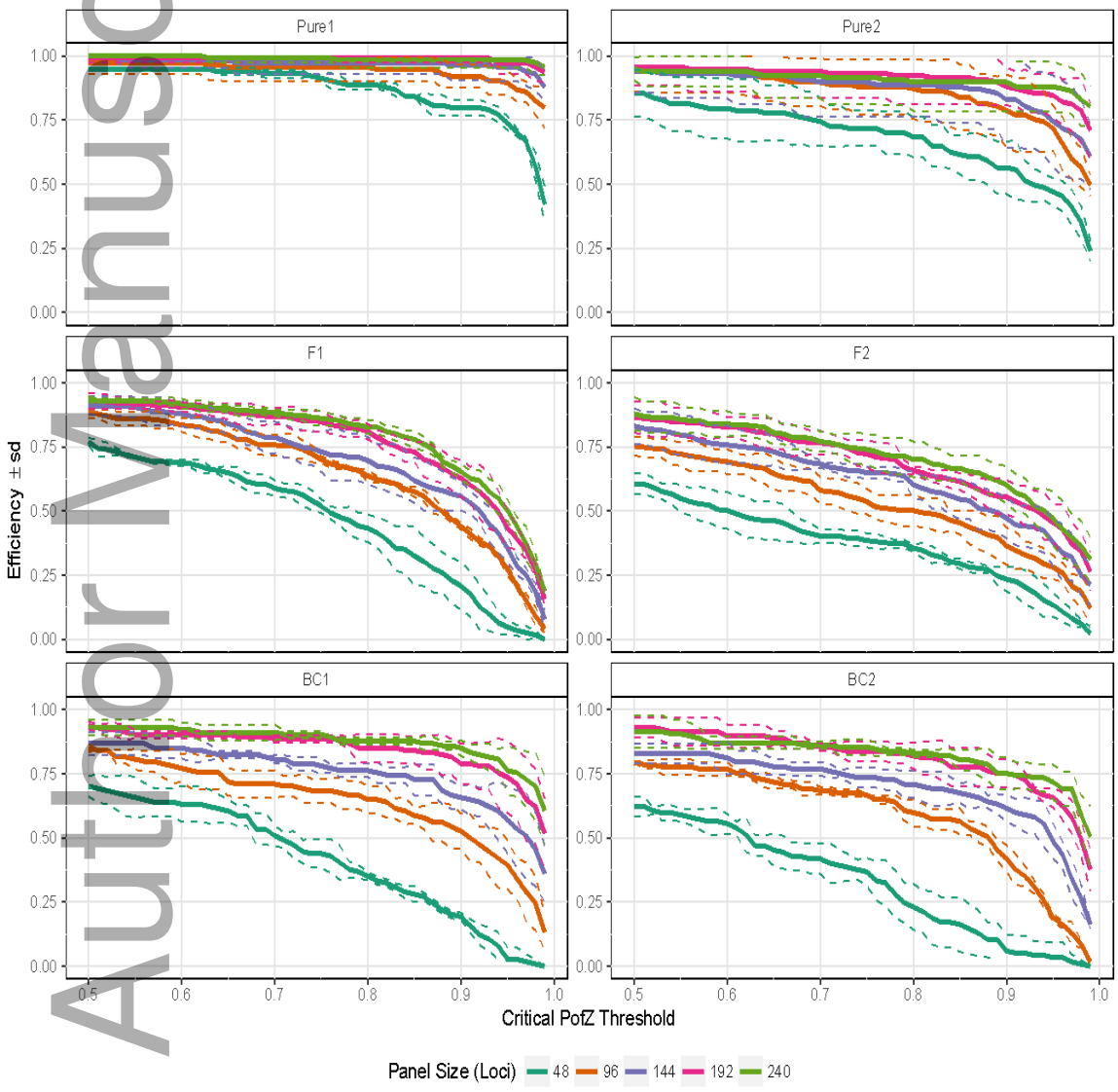
Figure 1. Schematic of the hybrid detection workflow and the associated functions (grey boxes) for: A, the development and quantification of the efficiency and accuracy of diagnostic panels of loci, and B the analysis of unknown/experimental data to detect hybrid individuals.



* (Wringe *et al.* 2017)

This article is protected by copyright. All rights reserved

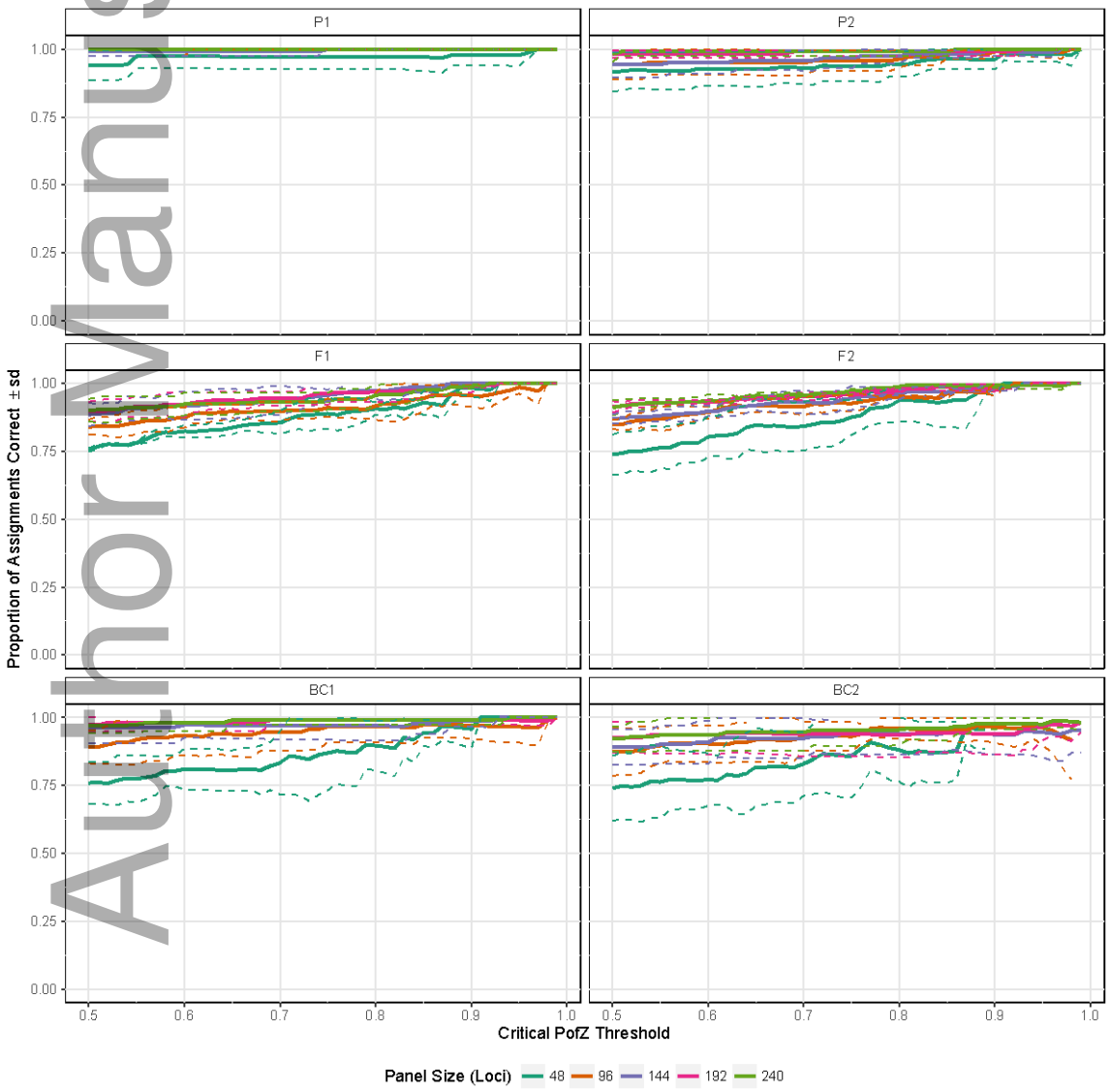
Figure 2. Plot of the efficiency of assignment for each of the six genotype frequency classes at critical posterior probability of assignment thresholds from 0.5 to 1.0 for diagnostic panels of various size. Each genotype frequency class is show in an individual facet, with abbreviations at its top. The solid coloured lines are the mean efficiency, and the dotted line the standard deviation of three independently simulated datasets, each analyzed in triplicate. Panel sizes and their corresponding colours are shown in the legend. The x-axis is the posterior probability of assignment threshold.



Author Manuscript

This article is protected by copyright. All rights reserved

Figure 3. Plot of accuracy of assignment for each of the six genotype frequency classes for various panel sizes at critical posterior probability of assignment threshold values ranging from 0.5 to 1.0. Genotype frequency class abbreviations are as in Supplementary figure 1, and each class is displayed in a single facet. The solid coloured lines are the mean accuracy, and the dotted lines the standard deviation of three independently simulated datasets, each analyzed in triplicate. The panel sizes, and their representative colours are shown in the legend. The x-axis is the critical posterior probability of assignment threshold.



Author Manuscript

This article is protected by copyright. All rights reserved

Table 1 – Functions included in the *hybriddetective* R package, a synopsis of their purpose, and which of the three major elements they are used in.

Function Name	Synopsis	Main Use
getTopLoc	Creates a panel comprised of the *n* (user-specified) most informative (based on highest loci-specific F_{ST}s), markers not in linkage disequilibrium. The function randomly assigns half the individuals in each of the two populations to be used to calculate loci-specific Weir and Cockerham's F_{ST}s[1], and returns the genotypes at the *n* loci of the other half to be used to test the efficacy of the panel to avoid high-grading bias[2].	Data Preparation
freqbasedsim_GTFreq	Creates simulated multi-generational (i.e. Pure 1, Pure 2, F_1, F_2, BC1, BC2) hybrids based on the allele frequencies in the two populations provided. The user can specify the number of individuals in each of the hybrid classes to be created.	Data Preparation
freqbasedsim_AlleleSample	Creates simulated multi-generational hybrids by randomly sampling, without replacement, two alleles per loci from a proportion of the individual genotypes provided. The user is able to specify the proportion of genotypes to sample, as well as the number of individuals of each hybrid class to create.	Data Preparation
nh_analysis_generateR	Merges a file composed of simulated hybrid genotypes with a file	Data

Author Manuscript

This article is protected by copyright. All rights reserved

Author Manuscript

	containing the genotypes of unknown/experimental individuals to produce a file suitable to ascertain the hybrid class of the unknowns. The user is able to specify which hybrid classes from the simulated dataset to include in the output.	Preparation
nh_analysis_simulateR_generateR	Creates a simulated multi-generational hybrid reference dataset from user provided data, and then merges it with the genotypes of unknown/experimental individuals. This function will create a new simulated dataset each time it is run using the same simulation methodology as *freqbasedsim_GTFreq*.	Data Preparation
nh_subsetR	Removes subsets of desired loci from NEWHYBRIDS formatted files so that the efficacy of panels of various sizes can be assessed.	Data Preparation
nh_Zcore	Allows the user to assign known hybrid category designations to individuals in NEWHYBRIDS formatted files	Data Preparation
nh_preCheckR	Checks all NEWHYBRIDS results within a directory and flags those that show evidence that the Markov chain may have failed to converge. This is done by evaluating the proportion of known Pure Population 1 or 2 individuals in which the posterior probability of assignment to F_2 exceeds a threshold. The user may specify both the proportion of individuals and the PofZ threshold.	Error Checking and Diagnostics
nh_mulitplotR	Creates a cumulative probability of assignment plot for each	Error Checking

This article is protected by copyright. All rights reserved

Author Manuscript

	NEWHYBRIDS result within a user-specified directory. Compliments *preCheckR* by allowing visually verification of Markov chain (non-) convergence.	and Diagnostics
nh_plotR	Plots the cumulative probability of assignment of a single NEWHYBRIDS result. Also allows the user to match plotting colours between analyses when NEWHYBRIDS reverses which population it designates Population 1 and 2.	Quantification and Analysis
hybridPowerComp	Evaluates the accuracy[3] and efficiency[3] with which NEWHYBRIDS assigns individuals of known hybrid class to the correct class across a range of minimum posterior probability thresholds from 0.50 to 0.99. Calculates the number of individuals wrongly assigned to hybrid genotype frequency classes over the total number of known pure individuals (type I error)[4] , and the proportion of individuals misclassified (type II error). The distribution of PofZ values for each genotype frequency class, as well as for all hybrid classes considered together is plotted. The effect of varying panel sizes on each of these variables is also evaluated. Plots are returned as .pdf and .jpg files, and all data frames constructed for plotting are exported.	Quantification and Analysis
nh_accuracy_checkR	Evaluates the accuracy with which NEWHYBRIDS assigns individuals of known hybrid class to the correct class for a single analysis at three	Quantification and Analysis

This article is protected by copyright. All rights reserved

	minimum posterior probability thresholds (PofZ >= 0.05, 0.75 and 0.90). This function is meant to compliment *hybridPowerComp*.	
nh_panel_delta_plotR	Plots the genotype class assignment (class with max. PofZ) of individuals among panels of different size. Allows visualization of the stability of individual assignments to compliment the proportion of correct assignments returned by *hybridPowerComp*.	Quantification and Analysis
nh_build_Example_Data	Writes example NEWHYBRIDS results to be evaluated with the function *hybridPowerComp*	

[1] Weir and Cockerham (1984)

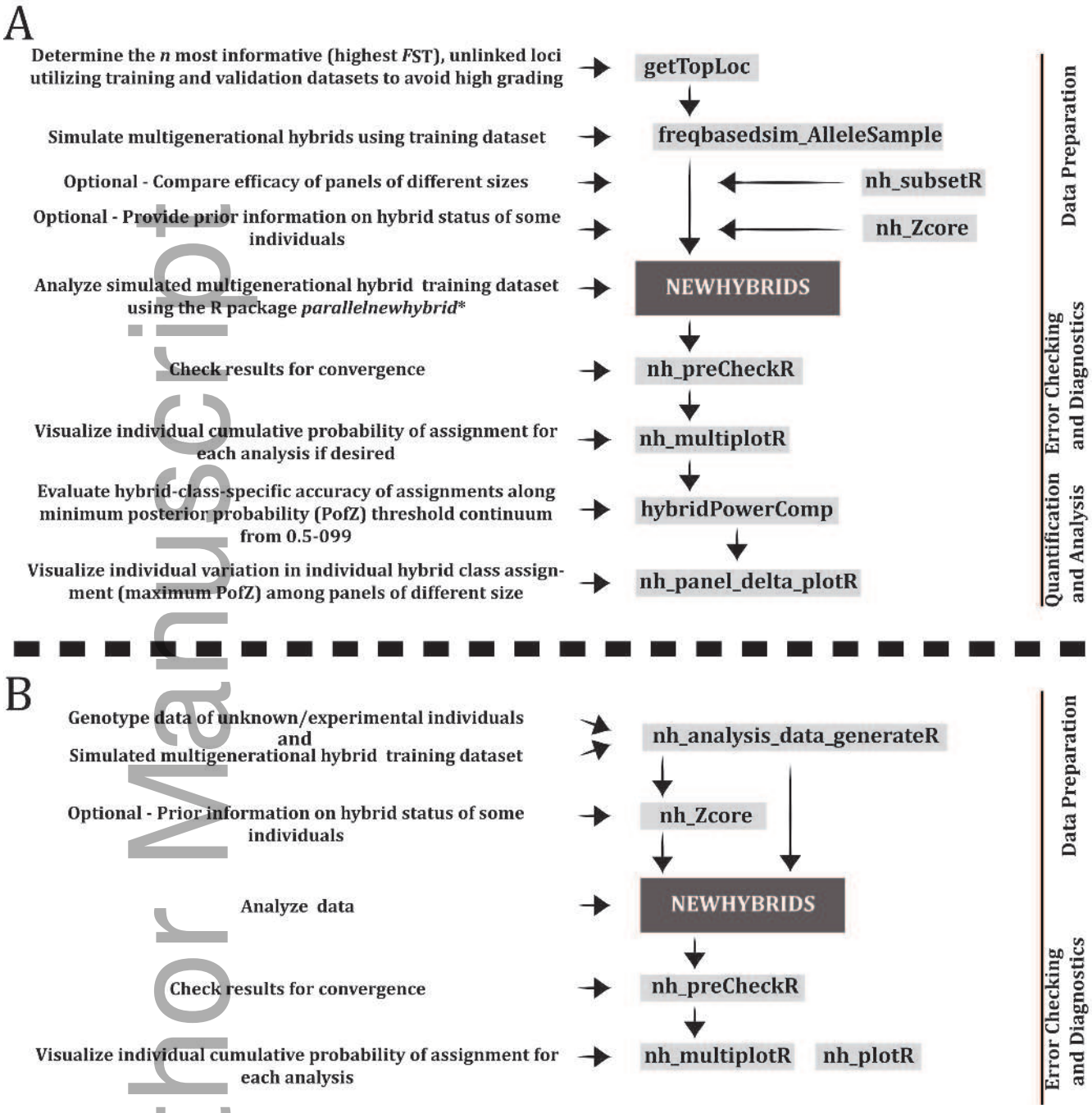
[2] Anderson (2010)

[3] Vaha and Primmer (2006)

[4] Burgarella et al. (2009)

Author Manuscript

This article is protected by copyright. All rights reserved

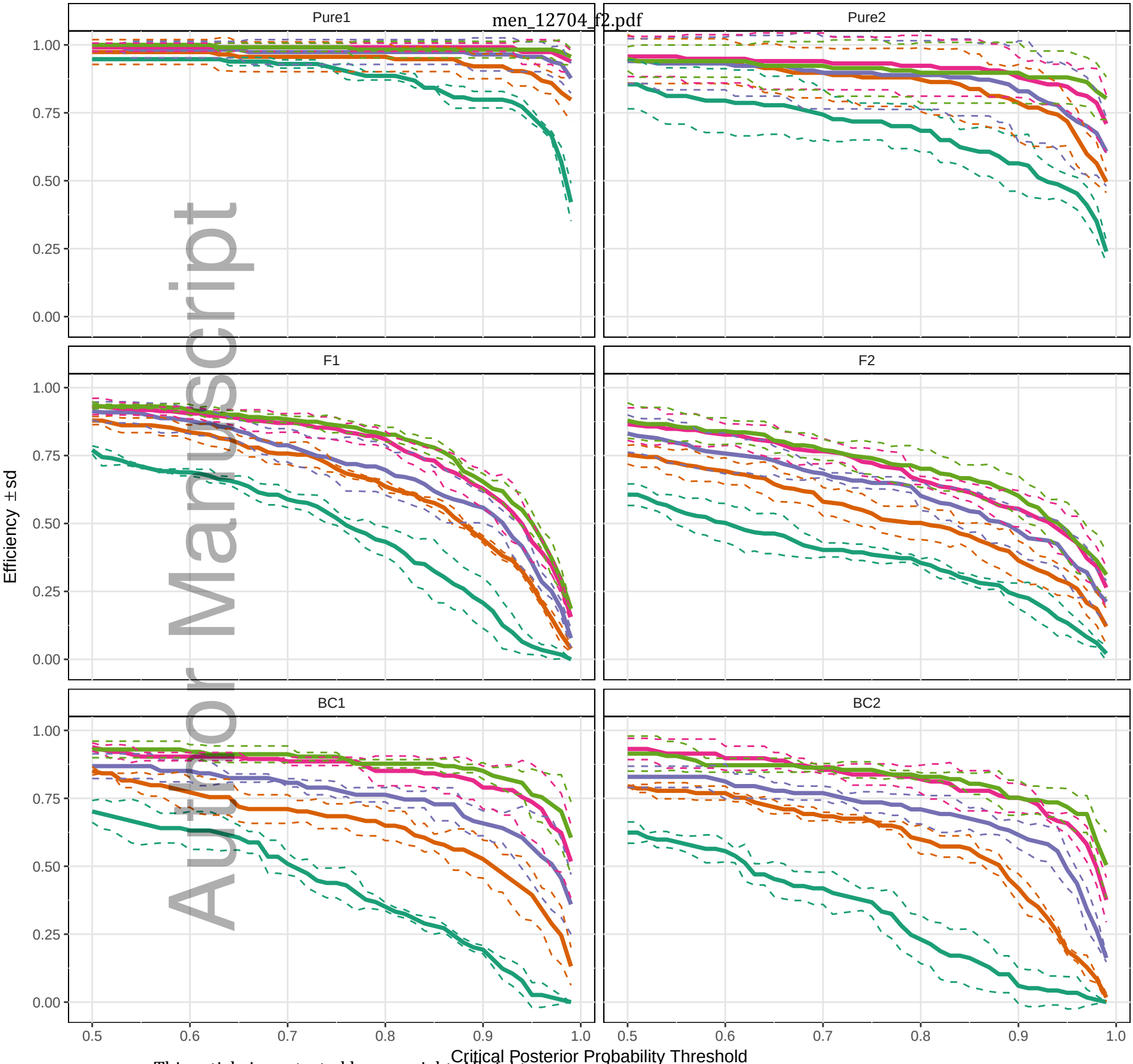


men_12704_f1.tif

Author Manuscript

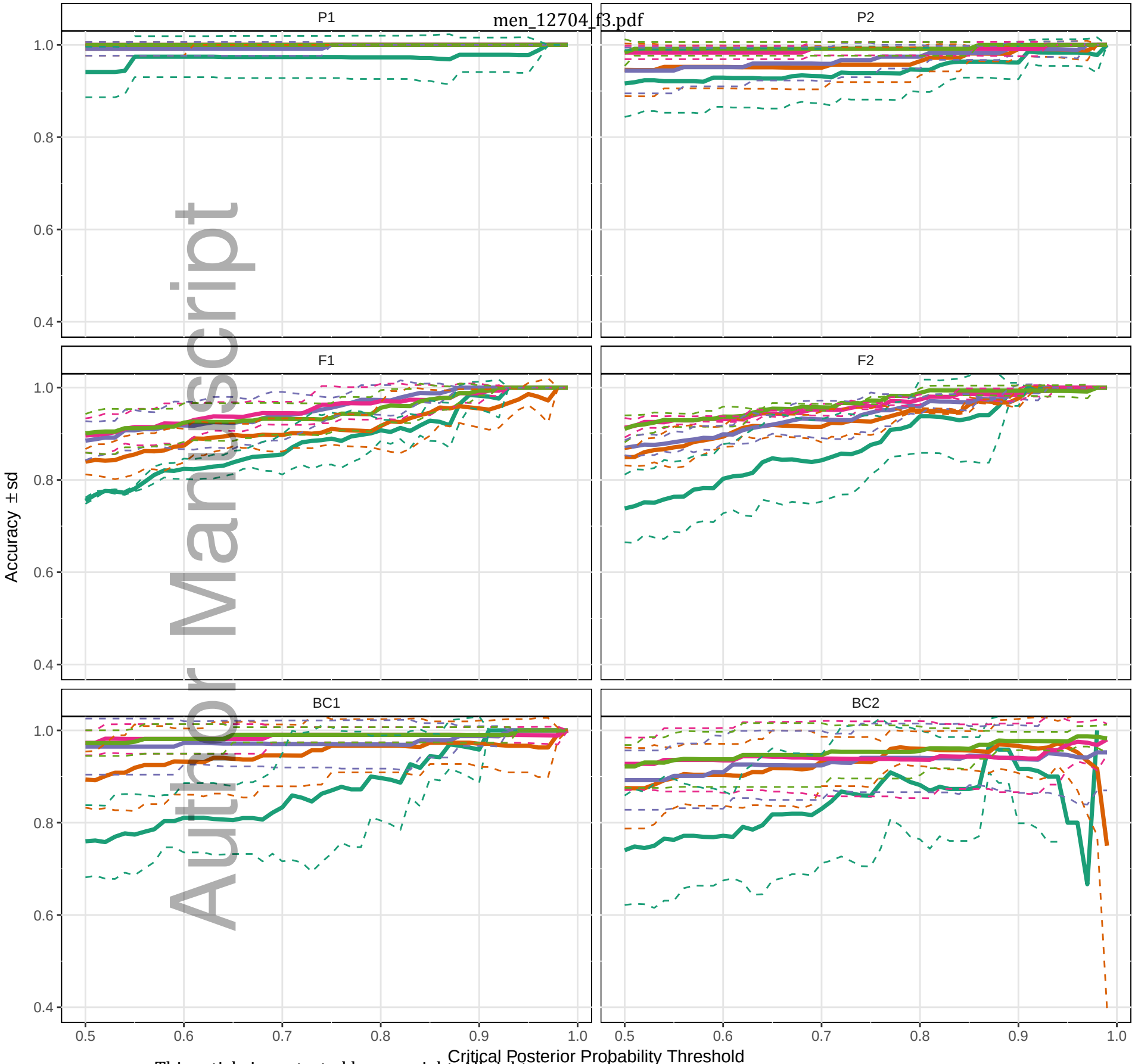
This article is protected by copyright. All rights reserved

men_12704_f2.pdf



Author Manuscript

This article is protected by copyright. All rights reserved



Author Manuscript

This article is protected by copyright. All rights reserved