Acta fytotechn zootechn, 23, 2020(4): 213–216
http://www.acta.fapz.uniag.sk

Original Paper

Comparison of identity by descent estimates with Plink and refinedIBD in dogs

Gábor Mészáros*, Mária Mészárosová, Anja Geretschläger
University of Natural Resources and Life Sciences, Vienna, Austria

Article Details: Received: 2020-05-26 | Accepted: 2020-07-10 | Available online: 2020-12-31

https://doi.org/10.15414/afz.2020.23.04.213-216

With the availability of dense SNP genotype data various types of estimation methods were developed to estimate relatedness of any two individuals, even in absence of traditional pedigrees. One of the most prominent method was the identity by descent (IBD), widely used in genetic diversity studies. IBD itself could be estimate using different approaches and software that might provide different results. The purpose of this study was to compare the estimates from two established software, probabilistic approach by Plink and a non-probabilistic approach based on haplotypes by refinedIBD. High density SNP genotypes from 98 Leonberger dogs were used to estimate IBD coefficients based on two data types: with one of the SNP markers in high linkage disequilibrium removed, as required by Plink, and SNP markers subjected only to standard quality control, as required by refinedIBD. The Pearson correlation coefficients from pairwise estimates were 0.97 when estimated with the same software and 0.84 between the two software and data types, as required by the respective user manuals. The numerical differences were clustered around zero (i.e. no to little difference) for half of the pairwise comparisons, and up to ±0.1 for the vast majority of cases. The most extreme differences were consistently estimated higher by Plink. Because of these differences a follow up investigation should be done, including pedigrees, as well as simulated data to provide a comprehensive analysis.

Keywords: SNP, Plink, refinedIBD, Leonberger, companion animals

1 Introduction

Individuals from the same family or the same population are related to each other due to shared ancestry (Weir et al., 2006). Traditionally the relatedness estimates were based on probabilities from conventional pedigrees, but with the availability of microsatellite, and later dense SNP genotypes allowed the use of more advanced methods. There are now many different ways to measure genetic similarity between individuals, as reviewed in Speed and Balding (2015).

In this work, we focus on the identity by descent (IBD), a relatedness measure that could be estimated with genetic markers, given the probabilities that two individuals share zero, one or two alleles at a locus (Weir et al., 2006). In contrast to identity by state (IBS), which simply identifies matching alleles between two individuals, regardless of their origin, the IBD refers to alleles that are the same due to their inheritance from a common ancestor. The numerical values of IBD range from zero for unrelated individuals to one for identical twins or clones, and in absence of inbreeding is broken down by recombination (Wright, 1922). The unique strength of IBD compared to other population genetics measures is in the efficiency to track distant relatives, when the IBD genome fragments are lost at an exponential rate per meiosis, while the decrease of their length is only linear to the reciprocal of the number of meiosis (Naseri et al., 2019).

Such probabilistic models are used among other software also by Plink (Chang et al., 2015), fitting a hidden Markov model for IBD status to determine posterior probabilities of IBD. A different approach is used by refinedIBD that uses genetic length and likelihood ratio for an IBD vs non-IBD model (Browning and Browning, 2013). A different handling of the linkage disequilibrium (LD) is also a notable difference between the two software. Plink does not account for LD, and requires an LD pruned data set for the computations. RefinedIBD incorporates the

***Corresponding Author:** Gábor Mészáros, University of Natural Resources and Life Sciences, Vienna, Austria

modelling of the LD within the run, so a non-pruned data set is required for the analysis.

The aim of this paper was to estimate IBD values in the Leonberger dog population, including the differences between Plink and refinedIBD programs.

2 Materials and methods

The Leonberger is a German dog breed established in the middle of the 19th century. The breed was named after a town near Stuttgart, where the founder of the breed lived. Leonberg has the lion in its city arms and breeding a lion-like dog resulted in the name Leonberger. The Leonberger was established by crossing a black and white Newfoundland (or Landseer) female with a long-haired male from the Hospice of the Great Saint Bernard (St. Bernard). After a few generations, a white Phyrenäen dog was added. It is assumed that the Leonberger breed has seven founders.

Actual number of Leonberger in Austria is around 500. The core of the breeding revolves around nine males and 11 females in ten active kennels. All Leonberger dogs used for breeding in Austria need to be genotyped because this is a prerequisite for the breeding permission. In case of interest from the Club full other animals have been also genotyped. Most of the dogs are from Austria, but the number of dogs from other countries increases. Reason for participation in genotyping from foreign dogs is to get an overview of the whole breed and find out if there are differences between Austrian dogs and those from other countries, such as UK, Sweden, USA, Germany, Switzerland and France.

Data from 98 dogs Leonberger dogs from Austria were genotyped with the Illumina CanineHD BeadChip, with total of 211,830 SNPs. The data set itself was relatively small, but realistic in size for endangered breeds or companion animals which do not have large data sets due to the lack of routine genotyping.

The initial data was subjected to quality control using Plink 1.9 (Chang et al., 2015), where animals and SNPs with more than 10% missingness were removed, as well as SNPs with minor allele frequency below 1% and those not adhering to Hardy-Weinberg distribution with p-value of 10-7. After the quality control 137,632 SNPs and 98 dogs remained.

For calculation of IBD values the Plink and the refinedIDB version 17 Jan20.102 (Browning and Browning, 2013) software were used with the data set after the quality control. An additional set of analyses was done with a pruned date set, where SNPs in a linkage disequilibrium of r^2 = 0.7 or higher were removed in 50 kb windows, via-indep-pairwise option, as recommended on the Plink manual. This pruning procedure removes 172,741 SNPs, which left only 36,226 SNPs in 98 dogs after the quality control. In the paper the term "data type" was used to collectively refer to quality controlled non-pruned and quality controlled pruned SNP data sets. For the refinedIBD software a constant recombination rate of 1 cM per Mb was assumed for the computations with the default parameter settings. Post processing and data visualization was done in R (R Core Team, 2019).

3 Results and discussion

The Pearson correlations of IBD values between the two software (Plink and refinedIBD) and the two data handling procedures (pruned and non-pruned SNP data) are shown in table 1. The highest agreement with correlations 0.96 and 0.97 were between the runs of the same software using pruned and non-pruned data sets. The correlation between IBD values from Plink with pruned and refinedIBD with non-pruned data, as suggested by the softwares' authors was 0.84. Although this correlation was high in general, it was far from unity, suggesting that there could be markedly different estimates in individual cases.

Also, the high correlations between the IBD values using the same software with different data types would suggest large agreement, the individual values could be markedly different, as shown in Figure 1, with IBD values in the range of ±0.1 for Plink (ibdPP – ibdPnP). An arbitrary threshold of 0.1 was also defined by Taylor et al. (2019) as an acceptable error rate, albeit with a much lower SNP count.

Interestingly, using pruned data in refinedIBD resulted into IBD differences up to negative 0.15, or more in an

Table 1 Correlations between IBD estimates from Plink and refinedIBD using pruned and non-pruned data sets. All correlations were highly significant

	Plink (pruned)	Plink (non-pruned)	RefinedIBD (pruned)
Plink (non-pruned)	0.97		
RefinedIBD (pruned)	0.85	0.84	
RefinedIBD (non-pruned)	0.84	0.84	0.96

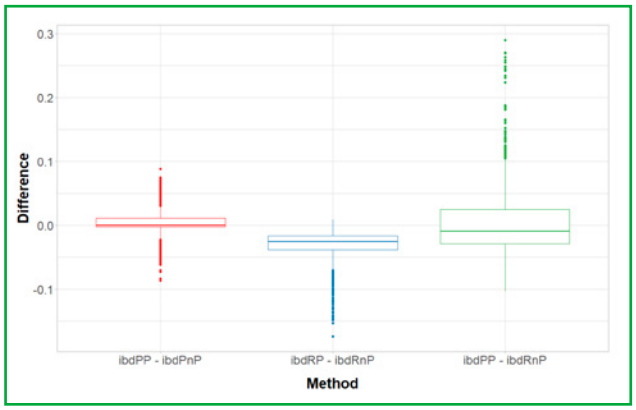

Figure 1 Differences between IBD estimates from the four estimation methods
ibdPP – IBD computed in Plink with pruned data; ibdPnP – IBD computed in Plink with non-pruned data; ibdRP – IBD computed in refinedIBD with pruned data; ibdRnP – IBD computed in refinedIBD with non-pruned data

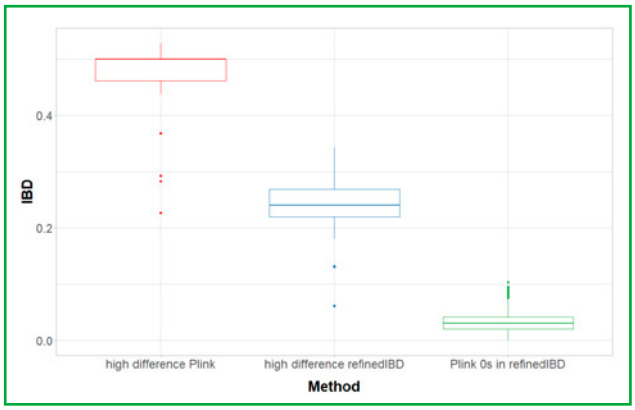

Figure 2 Distribution of IBD estimates with high differences between Plink and refinedIBD software. The "Plink 0s" were the dog pairs among which the IBD was estimated as 0 by Plink

extreme case. The reason for the different distribution of refinedIBD compared to Plink was most likely due to the IBD estimation methodology, relying on haplotype segments. With a much lower SNP count many of the relevant haplotypes were not detected. It should be noted here, however, that according to the user manuals of the two software non-pruned data should not be used with Plink and pruned data should not be used with refinedIBD. This comparison is merely to demonstrate the effect of the incorrect data type use.

The visualization of differences between the runs with appropriate data types with Plink and refinedIBD is shown in the third boxplot ("ibdPP – ibdRnP") in figure 1. The box itself denotes that 50% of the differences (2,376 of the 4,753 unique comparisons) were close to zero, as both software produced similar estimates. The other 50% of the cases, however, were outside of this range, with the whiskers of the boxplot denoting the largest value within 1.5 times interquartile range above the 75% and below the 25% percentile. This threshold was in the 0.1 range in both directions. A small fraction of the comparisons between software showed even more extreme values. These large differences were all higher than zero that indicates the numerically higher estimate from the Plink.

The pairs of dogs with high IBD difference above 0.15 were identified and visualized again in figure 2. The total number of such cases was 21 out of the 4,753 unique comparisons, so a very small proportion. Still, it is important to point out these dogs, so the best mating advice could be given to the owners. From the distribution of these high difference IBD values, it was apparent that PLINK consistently gives higher estimates compared to refinedIBD. These estimates were as high as 0.5, which would indicate parent – offspring, or full sib relationships, which was not confirmed by refinedIBD. Unfortunately, the conventional pedigree records were not available, so it could not be checked which of the software is closer to the recorded reality.

Another interesting phenomenon was the large number of IBD relationships that Plink has put to zero. Given the generally high relatedness in dog populations, and after feedback loops from the dog breeders these IBD values were cross-checked with refinedIBD (Figure 1, Plink0s in refinedIBD). Indeed, the vast majority of these relationships had a non-zero estimate, albeit most of them below 0.05. In some cases, however the IBD relationships could go as high as 0.1.

4 Conclusions

In this paper we compared IBD estimates from two well established software, Plink and refinedIBD. The correlations coefficients between the estimates were high, although far from unity. The calculated differences were scattered around zero for half of the comparison, with the overwhelming majority within ±0.1. A small fraction of the comparisons resulted into high differences, with the plots indicating higher estimates from Plink.

Based on these observations we conclude a large degree of agreement between the two software, although the animals with the large differences should be followed up in order to provide precise advice for the breeders. Moreover, a follow up investigation should be done preferably using genotypes from multiple breeds and species, including pedigrees, as well as simulated data to provide a comprehensive analysis.

Acknowledgements

The authors would like to thank the Leonberger association of Austria, and the individual breeders to share the genotypes data. The DNA extractions and technical processing of the genotypes data by the FERAGEN GmbH company is greatly appreciated.

References

BROWNING, B.L. and BROWNING, S.R. (2013). Improving the Accuracy and Efficiency of Identity-by-Descent Detection in Population Data. *Genetics*, 194(2), 459–471. https://doi.org/10.1534/genetics.113.150029

CHANG, C.C., CHOW, C.C., TELLIER, L.C., VATTIKUTI, S., PURCELL, S.M. and LEE, J.J. (2015). Second-Generation PLINK: Rising to the Challenge of Larger and Richer Datasets. *GigaScience*, (4) 7. https://doi.org/10.1186/s13742-015-0047-8

NASERI, A., LIU, X., TANG, K., ZHANG, S. and ZHI, D. (2019). RaPID: Ultra-Fast, Powerful, and Accurate Detection of Segments Identical by Descent (IBD) in Biobank-Scale Cohorts. *Genome Biology*, 20(1), 143. https://doi.org/10.1186/s13059-019-1754-8

R CORE TEAM (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

SPEED, D. and BALDING, D.J. (2015). Relatedness in the Post-Genomic Era: Is It Still Useful? *Nature Reviews Genetics*, 16(1), 33–44. https://doi.org/10.1038/nrg3821

TAYLOR, A.R., JACOB, P.E., NEAFSEY, D.E. and BUCKEE, C.O. (2019). Estimating Relatedness Between Malaria Parasites. *Genetics*, 212(4), 1337–1351. https://doi.org/10.1534/genetics.119.302120

WEIR, B.S., ANDERSON, A.D. and HEPLER, A.B. (2006). Genetic Relatedness Analysis: Modern Data and New Challenges. *Nature Reviews Genetics*, 7(10), 771–780. https://doi.org/10.1038/nrg1960

WRIGHT, S. (1922). Coefficients of Inbreeding and Relationship. *The American Naturalist*, 56(645), 330–338.

