Bioinformatics Advance Access published April 22, 2016

Genome analysis

Estimating IBD tracts from low coverage NGS data

Filipe G. Vieira [1,*] Anders Albrechtsen [2] and Rasmus Nielsen [2,3]

[1] Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark
[2] Department of Biology, University of Copenhagen, Copenhagen, Denmark
[3] Department of Integrative Biology, University of California, Berkeley, USA

Associate Editor: Dr. Inanc Birol

Abstract

1 Motivation:

The amount of IBD in an individual depends on the relatedness of the individual's parents. However, it can also provide information regarding mating system, past history, and effective size of the population from which the individual has been sampled.

2 Results:

Here, we present a new method for estimating inbreeding IBD tracts from low coverage NGS data. Contrary to other methods that use genotype data, the one presented here uses genotype likelihoods to take the uncertainty of the data into account. We benchmark it under a wide range of biologically relevant conditions and show that the new method provides a marked increase in accuracy even at low coverage.

3 Availability:

The methods presented in this work were implemented in C/C++ and are freely available for non-commercial use from `https://github.com/fgvieira/ngsF-HMM`.

4 Contact:

fgvieira@snm.ku.dk

5 Introduction

The inference of inbreeding levels is of central importance in many studies of ecology, evolution and conservation biology. Inbred individuals often have lower fitness than offspring of unrelated parents (Ebert *et al.*, 2002) and a cumulative effect can reduce the population growth rate and probability of persistence (O'Grady *et al.*, 2006). Furthermore, the study of inbreeding levels on natural populations can shed light into the species' mating system and past history (Gibson *et al.*, 2006; Stevens *et al.*, 2012; Vieira *et al.*, 2013) and is important for understanding the distribution of genetic variation within and among populations and, consequently, the effect of natural selection (Charlesworth, 2003).

In addition to the genome-wide inbreeding coefficient of an individual, further information can be gained by examining the distribution of inbred (or Identity By Descent; IBD) regions throughout the genome. These regions are usually organized into tracts of homozygous genotypes that recombination breaks down over time. Their number and lengths are tightly coupled to population genetics processes, from fine-scale population structure, effective population size, selfing and recombination rate, to age and type of inbreeding and even relatedness of the parents (Gibson *et al.*, 2006). Briefly, short tracts reflect old inbreeding in the population (possibly due to large effective population sizes), while long tracts may reflect recent inbreeding either due to small effective population sizes or familial matings. Under certain conditions and carefully designed experiments, the identification of IBD tracts can even be used for other types of analyses, like mapping of recombination breakpoints from back-crossed individuals.

As common descent is always guaranteed for any pair of homologous loci, IBD is often defined relative to an expectation under a certain model. An example is IBD due to recent familial relationships relative to the expectation of genetic identity for two individuals sampled at random from the population. A common operational definition of inbreeding is the excess of homozygosity compared to the Hardy-Weinberg Equilibrium expectation. This corresponds to classical population genetic definitions dating back to Wright (1922) and is in effect the definition used here and in many previous papers (Leutenegger *et al.*, 2003; Vieira *et al.*, 2013), although inferences are done in a single individual by combining information from multiple sites.

Previous studies have addressed the problem of inferring IBD tracts using Hidden Markov Models (HMMs) (Leutenegger *et al.*, 2003) based on individual genotype data. Current Next-Generation Sequencing (NGS) technologies can sometimes produce data with high error rates due to random sampling of homologous base pairs, sequencing, or alignment errors (Nielsen *et al.*, 2011; Ross *et al.*, 2013). Furthermore, due to budget constraints, many NGS studies rely on low or medium depth of coverage sequence data ($< 5\times$ per individual), causing genotype calling to be made with a considerable amount of uncertainty. The uncertainty, especially from low depth data, can greatly bias inferences of IBD tracts, as sites

*to whom correspondence should be addressed

with sequencing errors can be mistaken for heterozygote sites thereby breaking up the tracts into smaller segments.

Many recent methods rely on probabilistic frameworks to account for these errors and accurately call SNPs and genotypes, even at low coverage (Martin *et al.*, 2010; Li, 2011; Nielsen *et al.*, 2012). These methods integrate the base quality score together with other error sources (e.g., mapping or sequencing errors) to calculate an overall *genotype likelihood* (Li *et al.*, 2009a,b; DePristo *et al.*, 2011). These likelihoods can be used directly or combined with priors to take into account the uncertainty associated with the data (for reviews see Nielsen *et al.*, 2011; O'Rawe *et al.*, 2015).

There are several methods available to infer inbreeding coefficients, but only **ngsF** (Vieira *et al.*, 2013) is suitable for low coverage NGS data; however, this program only estimates genome-wide inbreeding levels and is thus incapable of inferring IBD tracts. In this paper, we present a new method to estimate IBD tracts from low-coverage NGS data. We evaluate its accuracy using extensive biologically relevant simulations and apply it to two real datasets, one consisting of populations from the HAPMAP and 1000 Genomes projects, and another consisting of varieties of wild and domesticated rice. When comparing against genotype-based methods, the method presented here performs considerably better when estimating IBD tracts and per-individual inbreeding coefficients for coverages $< 3\times$.

6 Methods

6.1 Optimization and decoding

To optimize the parameters from our model (Figure 1), we adopted a Maximum Likelihood iterative approach. Using F for the per-individual inbreeding coefficient, α for the transition rate, f for the allele frequencies, π for the most probable path (i.e. the inferred IBD tracts), and L_i for the likelihood at iteration i:

(0) Initialization:
  Set $i = 0$.
  Initialize F, α and f from a uniform distribution.
  Calculate initial likelihood L_0.

(1) Given f, jointly estimate F and α through the L-BFGS-B algorithm.

(2) Given F and α, estimate f using an EM algorithm.

(4) Check convergence:
  Set $i = i + 1$.
  Calculate Likelihood L_i.
  If $L_i - L_{i-1} > epsilon$ go to (1).

(5) Infer the most probable path π (decoding) with the Viterbi algorithm.

In more detail, on step (1), the L-BFGS-B algorithm (Zhu *et al.*, 1997) uses numerical derivatives and a new backward pass for each evaluation of the likelihood, that is in turn calculated from the forward algorithm. As for step (2), the EM algorithm was adapted from ANGSD (Korneliussen *et al.*, 2014) and estimates f directly from the genotype likelihoods using, as prior, the expected genotype frequencies under a certain inbreeding level; this is calculated from the probability of a given site being IBD (calculated from the forward and backward algorithms).

6.2 NGS data simulation

We performed extensive simulation to assess the performance of our method when estimating all parameters. Due to computational constraints, we simulated mapped sequencing data directly (rather than raw sequencing reads) similarly to previous studies (Kim *et al.*, 2011; Fumagalli *et al.*, 2013; Vieira *et al.*, 2013). When simulating genotypes, we have to account for the IBD state in each site. Therefore, we first sampled the true

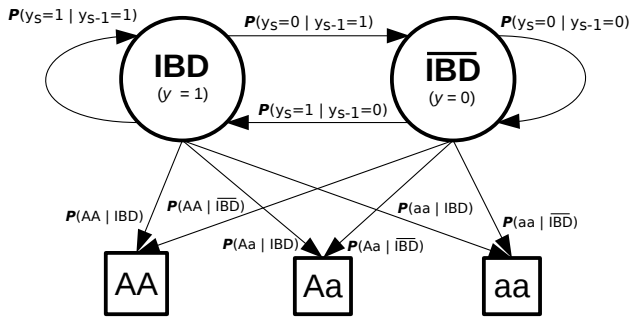

Fig. 1. HMM model. Illustration of the HMM model used. We used a two-state model to represent IBD and $\overline{IBD}$ (not IBD) genomic regions. The model switches between states with probability $P(y_{s-1}|y_s)$ and each state can emit the three different genotypes with probability $P(g|y_s)$ (where g represents a genotype and y_s the most probable state at site s).

IBD states directly by implementing a Markov model (Fig. 1), taking into account the distance between SNPs, per-individual inbreeding coefficient (F) and a transition rate (α). Allele frequencies were sampled uniformly between 0 and 0.5. We then sampled genotypes given the allele frequencies and the previously simulated IBD states. Third, and similarly to previous studies, we sampled the number of reads from a Poisson distribution with mean equal to the specified individual sequencing coverage and simulated sequencing errors by changing each read base to any of the other three nucleotides with probability $\epsilon/3$, where ϵ represents the error rate. We then calculated the genotype likelihoods assuming the GATK model (DePristo *et al.*, 2011). We note that, since we use the same model to simulate the data and make inferences, results from simulated data represent best case scenarios.

We simulated $10,000$ variable sites with distances sampled from a normal distribution with a mean of $100,000$ base pairs and standard deviation equal to $1/3$ of the mean. We assumed sample sizes of 10, 30 and 50 individuals, and average sequencing coverages of 0.5, 1, 3, 5 and $10\times$, with error rates of 0.5%, 1% and 2%, and varied the inbreeding coefficients from 0.0 to 1.0 in steps of 0.1, for a total of 495 combinations. With these parameter choices we focused on data sets with small sample sizes and low coverage, for which inference is harder, used realistic error rates (Glenn, 2011), and covered biologically relevant scenarios of inbreeding from < 0.07 in humans (Carothers *et al.*, 2006) and ~ 0.3 dogs (Kirkness *et al.*, 2003; Gray *et al.*, 2009) to $0.4 - 0.98$ in rice (Kovach *et al.*, 2007) and 0.757 in wasps (Chapman and Stewart, 1996).

6.3 Genotype calling

Genotypes were called following a Maximum-A-Posteriori (MAP) approach by internally setting the genotype with the highest posterior probability to 1 and the remaining genotypes to 0. Genotype's posterior probability was calculated from the genotype likelihood together with a prior (Li, 2011; Nielsen *et al.*, 2011, 2012), where the latter was either (1) a uniform distribution (GL_CG), the expected genotype frequencies under (2) HWE (HWE_CG) or (3) HWE assuming an inbreeding coefficient (HWE+F_CG). In this work, the inbreeding priors were calculated with the program **ngsF** Vieira *et al.* (2013), that estimated genome-wide inbreeding coefficients from low coverage NGS data. It is worth noting that under this genotype calling approach, we will only have undetermined (i.e. missing) genotypes when using a uniform prior since, in face of equally likely genotypes, the prior alone will determine the outcome.

6.4 Metric of accuracy

We calculated error rates associated with all estimated parameters: F, α, IBD tract (π), allele frequencies (f), and genotype calls (g). For genotype calls, the associated error was calculated as the proportion of miscalled genotypes, while for the other parameters we used the RMSD defined as:

$$RMSD = \sqrt{\frac{1}{S}\sum_{S}(X_{true} - X_{est})^2} \quad (1)$$

where, X_{true} and X_{est} are the true and estimated values of the parameters, and S the total number of estimates (number of individuals for F and α, and all sites across all individuals for π; missing values were ignored. All plots were made using R package *ggplot2* (Wickham, 2009).

6.5 Analysis of real data

6.5.1 Human dataset

We selected individuals from the LWK (Luhya in Webuye, Kenya) and GIH (Gujarati Indian from Houston, Texas) populations present in both the HAPMAP (The International HapMap 3 Consortium, 2010) and 1000 Genomes (The 1000 Genomes Project Consortium, 2012) datasets, on a total of 86 and 94 samples from the LWK and GIH, respectively, with sequencing coverages ranging from $3.13\times$ up to $13.98\times$, with an average of $6.2\times$ (Sup. Table 1 and 2). For computational reasons, we restricted our analyses to chromosomes 8 and 11 on the LWK and GIH populations, respectively (Sup. Figure 6). We selected all sites with a Minor Allele Frequencies (MAF) of at least 0.05 and, since our method does not explicitly account for linkage disequilibrium (LD) among sites, devised two datasets to assess the impact of LD on the IBD inferences: one composed of all sites in the HAPMAP dataset (HAPMAP) and another (HAPMAP-LD) where sites in LD were pruned with PLINK (Purcell *et al.*, 2007) based on pairwise LD (*–indep-pairwise 50 5 0.2*). To assess the impact of sample size and sequencing coverage on the estimates, apart from the original 1000 Genomes dataset, we also analyzed downsampled datasets. For the coverage we downsampled the original dataset to 50% and 25% of the original total number of reads; for the sample size, we downsampled it to 15, 30 and 50 individuals. We extracted the sites for each of these datasets and, using the originally mapped reads, used 'ANGSD' (Korneliussen *et al.*, 2014) to calculate genotype likelihoods (*-baq 1 -C 50 -minMapQ 15 -minQ 10 -minInd (N_IND/2) -GL 1 -doGlf 2 -doMajorMinor 1 -doMaf 1 -SNP_pval 1e-6*). Briefly, we used the SAMtools formula (Li *et al.*, 2009a) to calculate the genotype likelihoods using only reads with a root mean square (RMS) mapping quality > 15, and sites with a base quality > 10, where data was present in at least half the individuals and with a high probability of being SNPs. To assess the accuracy of our method, we inferred IBD tracts on all of the above mentioned 36 datasets (2 populations, 3 sample sizes, 3 coverages, and 2 sets of sites) and calculated the proportion of sites with correctly assigned IBD states. Since we don't know the true IBD state of these samples, benchmarked the performance of our method dealing with low-coverage data (using genotype likelihoods) relative to the performance of full genotyping data, using the our HMM algorithm in both cases.

6.5.2 Rice dataset

For the rice example, we used the 113 wild, 72 *indica* and 79 *japonica* (both tropical and temperate) accessions from the MiniCore collection (Wang *et al.*, 2016). We used the original data comprising 52,838 SNPs evenly distributed across the rice genome. Briefly, genotype likelihoods were calculated with ANGSD (options "-baq 1 -C 50 -minMapQ 20 -minQ 20 -GL 1 -doGlf 2") and called SNPs with a significance level of approx. 0.0001 for rejecting the hypothesis of the site being non-polymorphic. We restricted our analyses to sites with a MAF of at least 0.05 and, to account for LD, we randomly selected a representative SNP for every 5kb region. In this case we simply inferred the IBD tracts, since we don't have a high quality reference to compare to. Since the level of population structure of these species is unclear, we analyzed the wild, *indica* and *japonica* accessions separately (Sup. Figures 7, 8 and 9).

7 Results

7.1 Model for IBD tracts

To estimate Identity By Descent (IBD) tracts along a genome, we extended the approach of Leutenegger *et al.* (2003) which uses an HMM model with two hidden states (not inbred, $y = 0$ or $\overline{IBD}$; inbred, $y = 1$ or IBD) for each polymorphic site (s) of the genome (Figure 1). However, instead of the observed data being genotypes, we will use genotype likelihoods directly to infer IBD regions (or the most probable hidden state path $\hat{\pi}$) across the genome.

The transition probabilities between site s and $s-1$ represent the probability of switching between IBD states ($y \in \{0, 1\}$) and can be obtained from an instantaneous rate matrix Q. This rate matrix is naturally parametrized in terms of an inbreeding coefficient (F) and a transition rate per Mb (α) which depends on the recombination rate and the time to common ancestor(s) in the underlying pedigree:

$$Q = \begin{array}{c|c|c} State(y) & \overline{IBD} & IBD \\ \hline \overline{IBD} & 1-\alpha F & \alpha F \\ \hline IBD & \alpha(1-F) & 1-\alpha(1-F) \end{array} \quad (2)$$

However, we note that the α parameter can only be estimated when $0 < F < 1$. Based on the instantaneous rate matrix, the transition probabilities of the continuous time Markov chain are given by Leutenegger *et al.* (2003):

$$\begin{aligned} P(y_s = 0|y_{s-1} = 0) &= (1 - e^{-\alpha t})(1 - F) + e^{-\alpha t} \\ P(y_s = 0|y_{s-1} = 1) &= (1 - e^{-\alpha t})(1 - F) \\ P(y_s = 1|y_{s-1} = 0) &= (1 - e^{-\alpha t})F \\ P(y_s = 1|y_{s-1} = 1) &= (1 - e^{-\alpha t})F + e^{-\alpha t} \end{aligned} \quad (3)$$

The emission probabilities at site s, $P(X_s|y_s)$, represent the probability of the observed data (i.e. overlapping reads and corresponding base qualities) at that site (X_s) given the current state (y_s). If we were dealing with called genotypes, this would be straightforward since the observed data would be the genotype but here, since we are dealing with genotype likelihoods, we have to integrate over all possible genotypes:

$$P(X_s|y_s) = \sum_g P(X_s|g)P(g|y_s) \quad (4)$$

where, for site s, $P(X_s|g)$ is the genotype likelihood for genotype g, and $P(g_s|y_s)$ the probability of observing genotype g on state y_s. Assuming that all sites are bi-allelic with two alleles represented by A and a and, there are only three possible genotypes, so:

$$\begin{aligned} P(AA|y_s) &= (1 - f_s)^2 + f_s(1 - f_s)y_s \\ P(Aa|y_s) &= 2f_s(1 - f_s) - 2f_s(1 - f_s)y_s \\ P(aa|y_s) &= f_s^2 + f_s(1 - f_s)y_s \end{aligned} \quad (5)$$

where y_s is the IBD state and f_s the minor allele (a) frequency at site s, which is also estimated from the data (see below). From a set of individuals, we jointly estimate the allele frequencies f, and the per-individual inbreeding coefficients F and transitions rates α.

7.2 Estimating inbreeding from simulated data

To assess the accuracy of our method, we applied it to a simulated dataset covering a wide range of biologically relevant scenarios. For each simulated scenario, we used our HMM method and assessed the estimates' accuracy when based both on genotype likelihoods (GL) and called genotypes (GL_CG, HWE_CG and HWE+F_CG). On every case, we estimated the per-individual inbreeding coefficient (F), transition rate (α), most probable path ($\hat{\pi}$) and allele frequencies (f), together with their associated RMSD.

7.2.1 Estimating IBD tracts and transition probabilities

We benchmarked the accuracy of the new method on the above mentioned simulated dataset for inferring IBD tracts ($\hat{\pi}$). In all cases we used the Viterbi algorithm to infer IBD tracts, i.e. the most probable path between the model's two states. Overall, the analyses based on genotype likelihoods directly vastly outperform analyses based on called genotypes (Figures 2 and Sup. Figures 1). In fact, we get broadly the same accuracy level at $0.5\times$ from genotype likelihoods than at $3\times$ when calling genotypes accounting for inbreeding. Calling genotypes with priors that do not take inbreeding into account (uniform or HWE priors) reduces the accuracy of the estimates further, specially at low depth sequencing.

When inferring the transition rate (α) the overall trend is the same as for IBD tracts, except that all methods based on called genotypes perform quite poorly for coverages $< 3\times$ and, with higher error rates, even $< 5\times$ (Figure 3 and Sup. Figure 2). When calling genotypes assuming HWE in particular, it is impossible to make any inference at coverages $< 3\times$ (and why data points are missing for these coverages). Due to the extremely low coverage, the HWE prior will lead to high genotype calling errors (30%; Sup. Figure 5) that disrupt IBD tracts and preclude their identification. For any given depth, inferences directly from genotype likelihoods always outperform all other methods, even looking across different depths.

7.2.2 Estimating individual inbreeding coefficients

The individual inbreeding coefficients is given from the model and, as previously shown, can be used as a prior in Bayesian analyses (Vieira *et al.*, 2013) and can be particularly important when dealing with low depth-of-coverage NGS data sets. Overall we see that all approaches perform quite well at a sequencing coverage depth of $10\times$, except when assuming HWE on highly inbred samples. This is true even at $> 5\times$ (Figure 4 and Sup. Figure 3). At $< 3\times$ the two methods based on called genotypes perform quite poorly, illustrating the biases that this type of approach entails when dealing with low coverage data. At this coverage range, only genotype likelihood-based methods have acceptable accuracies, with the new method presented here slightly outperforming **ngsF** (Vieira *et al.*, 2013) at low sequencing depth ($< 3\times$) and small sample sizes ($n <= 30$).

7.2.3 Estimating allele frequencies

Allele frequencies form the backbone of most population genetics methods and, as such, their accurate estimation is of high importance. The method presented here was adapted from ANGSD and, as such, performs similarly. Overall, genotype likelihood-based methods seem to perform better than those based on called genotypes and be somewhat robust to various levels of inbreeding and priors (Sup Figure 4).

7.2.4 Effect of inbreeding on genotype calling

A common downstream analysis of NGS data is the identification of genotypes at each position for all individuals (genotype calling). The methods developed here can provide improved Bayesian genotype-calling in low depth-of-coverage data, by providing a more appropriate prior that takes inbreeding into account. However, several factors in addition to inbreeding, can affect genotype calling, including high error rates, sequencing coverage, and small sample sizes. To assess their impact, we calculated genotype posterior probabilities with our HMM method, called genotypes, and compared these with the previously mentioned called genotypes datasets: GL, HWE_CG and HWE+F_CG (see section 6). All methods show overall comparable error rates for calling genotypes (Sup. Figure 5), but assuming a uniform prior gives a high proportion of undetermined genotypes (e.g. 60% and 37% at $0.5\times$ and $1\times$, respectively). If we use an informative prior, the undetermined genotypes' levels are drastically reduced. Assuming HWE yields a relatively constant error rate across inbreeding levels but, in highly inbred samples, being able to incorporate inbreeding into the prior (either local or global) can drastically reduce genotype calling errors by as much as 79% (from 0.215 to 0.046) when analysing 10 fully inbred individuals at $5\times$ (Sup. Figure 5).

Dividing genotypes into homozygous and heterozygous calls, it is clear they are differently affected. Homozygous genotypes tend to have constant error rates independently of inbreeding levels, except when assuming HWE, where higher inbred samples have higher error since the prior makes it difficult to call low frequency homozygotes. Using a uniform prior gives very low error rates but at the expenses of high levels of undetermined genotypes, while both informative inbreeding priors have similar performances.

Heterozygous genotypes are typically the most difficult genotypes to call and, as such, have considerably higher error rates. In this case, a uniform prior is the worst method with both high error rates and undetermined genotypes' levels. Assuming an inbreeding prior performs similarly to HWE for low inbred samples but performs worse for highly inbred ones (i.e. $0.5 < F < 0.9$), since the prior penalizes heterozygote genotypes. In this case, the method presented here presents another advantage as it actually allows the use of different priors for regions that have been inferred to be IBD or $\overline{IBD}$. To sum up, the method largely has the advantages of an inbreeding prior for homozygote genotypes and a HWE prior for heterozygotes genotypes in non-IBD regions.

7.3 Application to real data

In addition to simulated data, we also analysed two publicly available datasets of low coverage NGS data: a human dataset composed of populations from HAPMAP and the 1000 Genomes projects, as well as a recently published rice dataset.

7.3.1 HAPMAP and 1000 genomes human datasets

In the analyses of the human dataset, we took advantage of the fact that some individuals genotyped for the HAPMAP project (The International HapMap 3 Consortium, 2010) have also been sequenced at low coverage on the 1000 Genomes (The 1000 Genomes Project Consortium, 2012). This gives us a perfect test dataset, as we have both low-coverage sequencing data and high quality genotypes to use as reference. We selected the LWK and GIH populations, since they are included in both datasets, both were expected to be composed of unrelated individuals, but where recent studies have shown some history of inbreeding (e.g. Stevens *et al.*, 2012; Gazal *et al.*, 2015).

We inferred IBD tracts on the 1000 Genomes populations comparing results across several combinations of sample size and coverage, with and without LD pruning (see Methods). We assume HAPMAP genotypes are called correctly and, as such, that IBD tracts inferred from them represent the gold standard. Comparing to it, we show that the HMM method can accurately infer IBD tracts even under cases of extremely low coverage and small sample sizes ($1/4$ of 1000 Genomes coverage and 30 individuals) (Sup. Figure 6). In fact, both sample size and coverage seem to have a relatively small effect on IBD inference (the latter when sites in LD have been pruned), in comparison to the presence of sites in LD. That said, pruning of sites in LD can greatly reduce error rates by as much as an

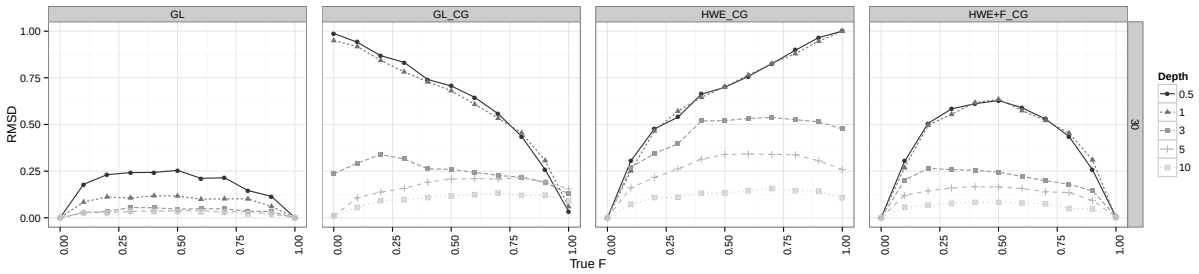

Fig. 2. Estimation of IBD tracts. Performance of the HMM method to infer IBD tracts for a sample size of 30 individuals, a transition rate of 0.01, and $10,000$ variable sites simulated with a 0.5% error rate. Columns represent the different analytical approaches, from genotype likelihoods (1st column), called genotypes assuming no prior (2nd column), and called genotypes assuming a prior based on genotype frequencies under HWE (3rd column) or a prior assuming inbreeding (last column). The different lines represent RMSD under different simulated sequencing depths.

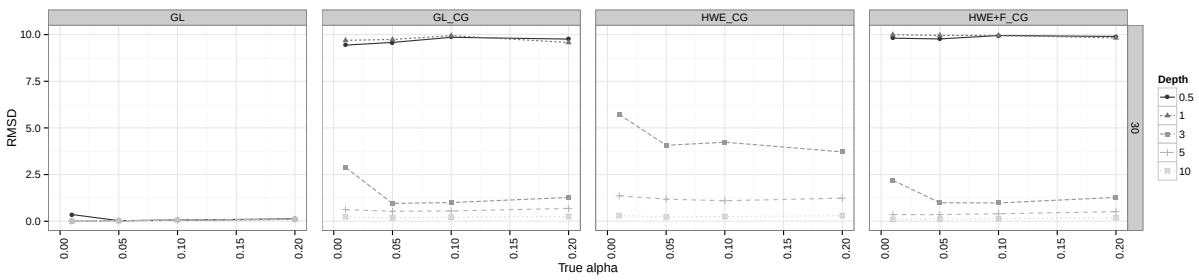

Fig. 3. Estimation of transition rate. Performance of the HMM method when estimating α for a sample size of 30 individuals, an inbreeding coefficient of 0.5, and $10,000$ variable sites simulated with a 0.5% error rate. Columns represent the different analytical approaches: genotype likelihoods (1st column), called genotypes assuming no prior (2nd column), and called genotypes assuming a prior based on genotype frequencies under HWE (3rd column) or a prior assuming inbreeding (last column). The different lines represent RMSD under different simulated sequencing depths.

order of magnitude, from (e.g.) an average proportion of miss-identified sites of 0.182 to 0.020, when analysing 30 individuals from the GIH population at $1/4$ of their original coverage.

7.3.2 MiniCore rice dataset

As a second example, we used the recently published MiniCore dataset, composed of both domesticated and wild accessions sequenced at low coverage (Wang *et al.*, 2016). There are several species of wild rice but the *O. rufipogon* species complex is thought to be the closest to domesticated rice (*Oryza sativa*) (e.g. Grillo *et al.*, 2009; Wei *et al.*, 2012), which is further divided into two subspecies (*O. s. japonica* and *O. s. indica*). Wild and domesticated accessions have markedly different selfing rates, ranging in the wild between $50-95\%$ (Morishima *et al.*, 1984; Oka, 1988; Gao *et al.*, 2002; Phan *et al.*, 2012), and in cultivated between $95-100\%$.

Our estimates show wild rice with a wide range of inbreeding values, from totally outbred to almost fully inbred (Sup. Figure 7), while cultivated rice accessions were all almost totally inbred. These estimates are also reflected in the IBD tract inferences, in which cultivated accessions have IBD tracts spanning whole chromosomes(Sup. Figure 9 and 8).

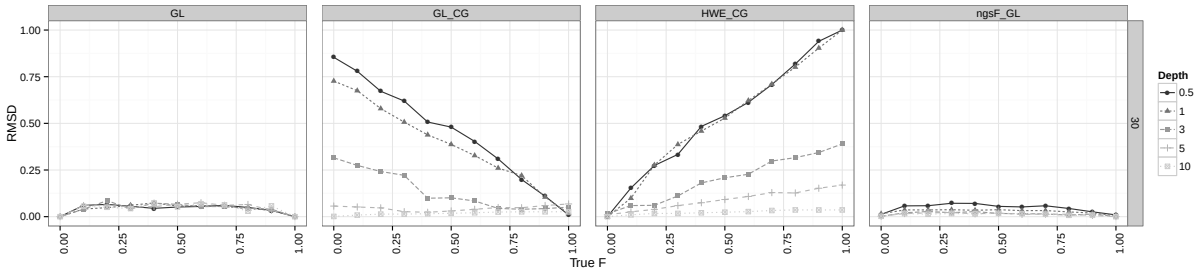

Fig. 4. Estimation of per-individual inbreeding coefficients. Performance of the HMM method to infer genome-wide per-individual inbreeding coefficients for a sample size of 30 individuals, a transition rate of 0.01, and $10,000$ variable sites simulated with a 0.5% error rate. Columns represent the different analytical approaches, from genotype likelihoods (1st column), called genotypes assuming no prior (2nd column), and called genotypes assuming a prior based on genotype frequencies under HWE (3rd column). The last column represents the inference of per individual F from genotype likelihoods (as in 1st column) but using the previously developed method ngsF. The different lines represent RMSD under different simulated sequencing depths.

8 Discussion

The levels of inbreeding in an individual is an important parameter in population genomic studies, since it can reflect mating system, selfing rates, population size and past population history. Theoretically, the best way to infer it is through the pedigree, but pedigrees are not available in much studies. However, even in these cases, inferences based on pedigrees can be biased due to incomplete knowledge of the pedigree. In addition, pedigrees provide expected levels of inbreeding, but these may differ from true genetic levels of inbreeding due to the stochasticity of allelic segregation and recombination. This has been demonstrated by recent analyses identifying higher than expected ROH prevalence in unrelated individuals from outbred populations (Gibson *et al.*, 2006; The International HapMap Consortium, 2007), supporting a recent claim that estimates based on markers are more accurate than expected values inferred from pedigrees (Kardos *et al.*, 2015).

Here, we have developed a method that can reliably estimate individual Identical By Descent (IBD) tracts and inbreeding coefficient, directly from genomic data, without requiring any knowledge of the underlying genealogy. Other methods exist for this (Leutenegger *et al.*, 2003; Hall *et al.*, 2012) but all were developed for SNP chip data, which has much lower error rates than low coverage NGS data. NGS technologies have revolutionized genetics by providing fast, cheap and reliable large-scale DNA sequencing data. However, the per base pair error rate in NGS data is still considerably higher than in Sanger sequencing or chip-based genotyping technologies (Glenn, 2011). As a consequence, researchers usually sequence at high sequencing depths but this comes at an increased financial, computational, and storage cost. Furthermore, due to the ever-growing demand for larger sample sizes, many NGS studies rely on low coverage NGS sequence data ($< 5\times$). As such, the availability of methods that can properly handle this data will help researchers make more cost-effective choices in the trade-off between sample size and sequencing depth.

The method presented here facilitates the estimation of IBD tracts from low coverage NGS data. We evaluate its performance through both simulated and real data analyses. When compared to genotype-calling-based methods, the improvement in accuracy when estimating IBD tracts and individual inbreeding coefficients is considerable for sequencing depths $< 3\times$. Apart from the previously mentioned use of this methods in population genomic studies, we note that there is another possible application of the method: mapping of recombination breakpoints in backcrosses between inbred lines. In backcross data, each individual in the backcross generation effectively has an inbreeding coefficient of $F = 0.5$. The method presented here allows the estimation of recombination rates for such data by estimation of the parameter α. More importantly, the posterior decoding algorithm provides estimates of the genomic location of recombination breakpoints. This provides, in combination with low-coverage NGS sequencing of a backcross generation, an efficient design for mapping recombination breakpoints from model species such as yeast.

Acknowledgements

We would like to thank Thorfinn Korneliussen for helpful discussions and assistance on the use of ANGSD, and Shyam Gopalakrishnan for helpful discussions.

Funding: Funding for this work was supported by a DFF-MOBILEX grant to FGV (DFF-1325-00136), and a Villum Foundation fellowship to AA.

Disclosure Declaration

The authors declare that they have no conflicts of interest.

References

Carothers, A. D., Rudan, I., Kolcic, I., Polasek, O., Hayward, C., Wright, A. F., Campbell, H., Teague, P., Hastie, N. D., and Weber, J. L. (2006). Estimating human inbreeding coefficients: comparison of genealogical and marker heterozygosity approaches. *Ann. Hum. Genet.*, **70**(Pt 5), 666–76.

Chapman, T. W. and Stewart, S. C. (1996). Extremely high levels of inbreeding in a natural population of the free-living wasp Ancistrocerus antilope (Hymenoptera: Vespidae: Eumeninae). *Heredity (Edinb).*, **76**(1), 65–69.

Charlesworth, D. (2003). Effects of inbreeding on the genetic diversity of populations. *Philos. Trans. R. Soc. London B Biol. Sci.*, **358**(1434), 1051–1070.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., and Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, **43**(5), 491–8.

Ebert, D., Haag, C., Kirkpatrick, M., Riek, M., Hottinger, J. W., and Pajunen, V. I. (2002). A selective advantage to immigrant genes in a Daphnia metapopulation. *Science*, **295**(5554), 485–488.

Fumagalli, M., Vieira, F. G., Korneliussen, T. S., Linderoth, T., Huerta-Sánchez, E., Albrechtsen, A., and Nielsen, R. (2013). Quantifying population genetic differentiation from next-generation sequencing data. *Genetics*, **195**(3), 979–92.

Gao, L.-z., Schaal, B. A., Zhang, C.-h., Jia, J.-z., and Dong, Y.-s. (2002). Assessment of population genetic structure in common wild rice Oryza rufipogon Griff. using microsatellite and allozyme markers. *Theor. Appl. Genet.*, **106**(1), 173–80.

Gazal, S., Sahbatou, M., Babron, M.-C., Génin, E., and Leutenegger, A.-L. (2015). High level of inbreeding in final phase of 1000 Genomes Project. *Sci. Rep.*, **5**, 17453.

Gibson, J., Morton, N. E., and Collins, A. (2006). Extended tracts of homozygosity in outbred human populations. *Hum. Mol. Genet.*, **15**(5), 789–795.

Glenn, T. C. (2011). Field guide to next-generation DNA sequencers. *Mol. Ecol. Resour.*, **11**(5), 759–69.

Gray, M. M., Granka, J. M., Bustamante, C. D., Sutter, N. B., Boyko, A. R., Zhu, L., Ostrander, E. A., and Wayne, R. K. (2009). Linkage disequilibrium and demographic history of wild and domestic canids. *Genetics*, **181**(4), 1493–505.

Grillo, M. A., Li, C., Fowlkes, A. M., Briggeman, T. M., Zhou, A., Schemske, D. W., and Sang, T. (2009). Genetic architecture for the adaptive origin of annual wild rice, oryza nivara. *Evolution (N. Y).*, **63**(4), 870–83.

Hall, N., Mercer, L., Phillips, D., Shaw, J., and Anderson, A. D. (2012). Maximum likelihood estimation of individual inbreeding coefficients and null allele frequencies. *Genet. Res. (Camb).*, **94**(3), 151–61.

Kardos, M., Luikart, G., and Allendorf, F. W. (2015). Measuring individual inbreeding in the age of genomics: marker-based measures are better than pedigrees. *Heredity (Edinb).*, **115**(1), 63–72.

Kim, S. Y., Lohmueller, K. E., Albrechtsen, A., Li, Y., Korneliussen, T., Tian, G., Grarup, N., Jiang, T., Andersen, G., Witte, D., Jorgensen, T., Hansen, T., Pedersen, O., Wang, J., and Nielsen, R. (2011). Estimation of allele frequency and association mapping using next-generation sequencing data. *BMC Bioinformatics*, **12**(1), 231.

Kirkness, E. F., Bafna, V., Halpern, A. L., Levy, S., Remington, K., Rusch, D. B., Delcher, A. L., Pop, M., Wang, W., Fraser, C. M., and Venter, J. C. (2003). The dog genome: survey sequencing and comparative analysis. *Science*, **301**(5641), 1898–903.

Korneliussen, T., Albrechtsen, A., and Nielsen, R. (2014). ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics*, **15**, 356.

Kovach, M. J., Sweeney, M. T., and McCouch, S. R. (2007). New insights into the history of rice domestication. *Trends Genet.*, **23**(11), 578–87.

Leutenegger, A.-L., Prum, B., Génin, E., Verny, C., Lemainque, A., Clerget-Darpoux, F., and Thompson, E. A. (2003). Estimation of the inbreeding coefficient through use of genomic data. *Am. J. Hum. Genet.*, **73**(3), 516–23.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, **27**(21), 2987–93.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009a). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**(16), 2078–9.

Li, R., Li, Y., Fang, X., Yang, H., Wang, J., Kristiansen, K., and Wang, J. (2009b). SNP detection for massively parallel whole-genome resequencing. *Genome Res.*, **19**(6), 1124–32.

Martin, E. R., Kinnamon, D. D., Schmidt, M. A., Powell, E. H., Zuchner, S., and Morris, R. W. (2010). SeqEM: an adaptive genotype-calling approach for next-generation sequencing studies. *Bioinformatics*, **26**(22), 2803–10.

Morishima, H., Sano, Y., and Oka, H. I. (1984). Differentiation of perennial and annual types due to habitat conditions in the wild riceOryza perennis. *Plant Syst. Evol.*, **144**(2), 119–135.

Nielsen, R., Paul, J. S., Albrechtsen, A., and Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.*, **12**(6), 443–51.

Nielsen, R., Korneliussen, T., Albrechtsen, A., Li, Y., and Wang, J. (2012). SNP calling, genotype calling, and sample allele frequency estimation from New-Generation Sequencing data. *PLoS One*, **7**(7), e37558.

O'Grady, J. J., Brook, B. W., Reed, D. H., Ballou, J. D., Tonkyn, D. W., and Frankham, R. (2006). Realistic levels of inbreeding depression strongly affect extinction risk in wild populations. *Biol. Conserv.*, **133**(1), 42–51.

Oka, H. I. (1988). *Origin of cultivated rice*. Elsevier Science/Japan Scientific Societies Press, Tokyo.

O'Rawe, J. A., Ferson, S., and Lyon, G. J. (2015). Accounting for uncertainty in DNA sequencing data. *Trends Genet.*, **31**(2), 61–66.

Phan, P. D. T., Kageyama, H., Ishikawa, R., and Ishii, T. (2012). Estimation of the outcrossing rate for annual Asian wild rice under field conditions. *Breed. Sci.*, **62**(3), 256–62.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., and Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**(3), 559–575.

Ross, M. G., Russ, C., Costello, M., Hollinger, A., Lennon, N. J., Hegarty, R., Nusbaum, C., and Jaffe, D. B. (2013). Characterizing and measuring bias in sequence data. *Genome Biol.*, **14**(5), R51.

Stevens, E. L., Baugher, J. D., Shirley, M. D., Frelin, L. P., and Pevsner, J. (2012). Unexpected Relationships and Inbreeding in HapMap Phase III Populations. *PLoS One*, **7**.

The 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.

The International HapMap 3 Consortium (2010). Integrating common and rare genetic variation in diverse human populations. *Nature*, **467**(7311), 52–8.

The International HapMap Consortium (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**(7164), 851–61.

Vieira, F. G., Fumagalli, M., Albrechtsen, A., and Nielsen, R. (2013). Estimating inbreeding coefficients from NGS data: Impact on genotype calling and allele frequency estimation. *Genome Res.*, **23**(11), 1852–61.

Wang, H., Xu, X., Vieira, F. G., Xiao, Y., Li, Z., Wang, J., Nielsen, R., and Chu, C. (2016). GWAS using low coverage sequencing in the rice minicore collection reveals alleles underlying convergent evolution during rice domestication. *Genome Biol.*, **submitted**.

Wei, X., Qiao, W.-H., Chen, Y.-T., Wang, R.-S., Cao, L.-R., Zhang, W.-X., Yuan, N.-N., Li, Z.-C., Zeng, H.-L., and Yang, Q.-W. (2012). Domestication and geographic origin of Oryza sativa in China: insights from multilocus analysis of nucleotide variation of O. sativa and O. rufipogon. *Mol. Ecol.*, **21**(20), 5073–87.

Wickham, H. (2009). *Ggplot2: Elegant Graphics for Data Analysis*. Springer.

Wright, S. (1922). Coefficients of Inbreeding and Relationship.

Zhu, C., Byrd, R. H., Lu, P., and Nocedal, J. (1997). Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans. Math. Softw.*, **23**(4), 550–560.