Norwegian University of Life Sciences
Faculty of Biosciences
Department of Animal and Aquacultural Sciences

Philosophiae Doctor (PhD)
Thesis 2021:7

Genomic methods in breeding programs: Parental assignment, triploid genomics and case-parental control modelling

Genomiske metoder i avlsprogrammer: Foreldretilordning, triploid genomikk og kasus-foreldrekontroll-modellering

Kim Erik Grashei

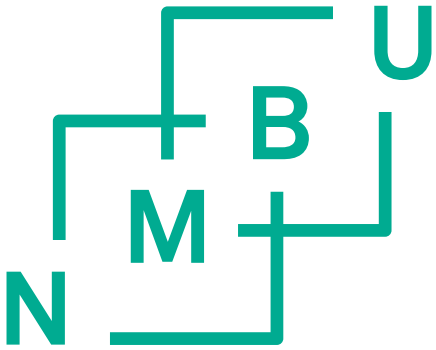
Genomic methods in breeding programs: Parental assignment, triploid genomics and case-parental control modelling

Genomiske metoder i avlsprogrammer: Foreldretilordning, triploid genomikk og kasus-foreldrekontroll-modellering

Philosophiae Doctor (PhD) Thesis
Kim Erik Grashei

Norwegian University of Life Sciences
Faculty of Biosciences
Department of Animal and Aquacultural Sciences

Ås (2021)



Thesis number 2021:7

ISSN 1894-6402

ISBN 978-82-575-1775-5

Supervisors and Evaluation Committee

PhD Supervisors

Prof. Theo Meuwissen

Faculty of Biosciences

Oluf Thesens vei 6, 1433 Ås

1433 Ås

Dr. Jørgen Ødegård

Faculty of Biosciences

Oluf Thesens vei 6, 1433 Ås

/

AquaGen AS

P.O. Box 1240, NO-7462 Trondheim

Prof. Sigbjørn Lien

Faculty of Biosciences / CIGENE

Oluf Thesens vei 6, 1433 Ås

1433 Ås

Dr. Thomas Moen

AquaGen AS

P.O. Box 1240, NO-7462 Trondheim

Prof. Thore Egeland

Faculty of Chemistry, Biotechnology and Food Science

Chr. M. Falsens vei 18, 1433 Ås

PhD Evaluation Committee

Prof. Luc L. Janss

Aarhus University

Denmark

Dr. Matthew Baranski

Mowi Genetics

Norway

Dr. Hanne Fjerdingby Olsen

Norwegian University of Life Sciences

Norway

Acknowledgements

The work presented here has been part of the project “Parentage assignment with high-density SNP genotypes: Tracing of escapees from fish farming and optimized breeding programs”. The project was coordinated by the Norwegian University of Life Sciences (NMBU) and was a collaboration between NMBU and AquaGen. The project was funded by The Research Council of Norway through the research programs NAERINGSPHD (project no. 251664) and HAVBRUK2 (project no. 245519). Additional funding was provided by AquaGen.

I want to thank all my supervisors: Dr. Jørgen Ødegård, Prof. Theo Meuwissen, Prof. Sigbjørn Lien, Prof. Thore Egeland and Dr. Thomas Moen. You have all had your part in influencing and enhancing both my PhD and my knowledge. A special thank you goes out to Jørgen for all the hours, and days, and weeks, spent discussing all sorts of things, PhD-related or not. You have truly gone above and beyond, and your passion for your field and willingness to discuss every minor detail to great length has been a great help and inspiration to me. I hope I have inspired you in some small way as well, and I hope we will continue to re-invent and improve the wheel for many years to come! Also, a special thanks to Theo, who seems to always have a solution ready when me and Jørgen have run into a roadblock chasing after new methodology. Your knowledge is truly astounding!

A big thank you to all the people at AquaGen for a great working environment with fun discussions and challenging tasks, you are part of what makes it fun to go to work every day. NMBU also deserves appreciation for all the help I have had with my PhD, and especially the ladies and men at the CIGENE laboratory who keeps the corridors “buzzing” with laughter and discussions that makes me smile so much, you are also a part of what makes work fun!

And last, but not least, I want to thank my family and friends. To Janne, who have stuck with this stubborn bastard for so many years, I love you and I would not be where I am without you. To Erik, you are my mischievous little energy-sucking, now one year old, son who I love above all else in this world. I can’t wait to see what wonders the future holds for you. To my mom, who have always been there for me

and have always supported my choices and my dreams, no matter how stupid they may have been, I love you and I am to this day and forever deeply appreciative of you. And to everyone else not mentioned already, friends, family, colleagues and those who have passed on, you have all had an impact on my life, and I am truly happy that I met you all.

Ås, 2020.

Kim E. Grashei

'Be what you would seem to be'

—or, if you'd like it put more simply—

'Never imagine yourself not to be otherwise than what it might appear to others that what you were or might have been was not otherwise than what you had been would have appeared to them to be otherwise.'

Lewis Carroll – Alice's Adventures in Wonderland

Table of Contents

1 Abbreviations and definitions

APT	Affymetrix Power Tools
BIC	Bayesian Information Criterion
BLUP	Best-Linear Unbiased Prediction
CIGENE	Centre for Integrative Genetics
DNA	Deoxyribonucleic acid
EBV	Estimated Breeding Value
EM	Expectation-maximization
ER	Exclusion Ratio
GBLUP	Genomic best-Linear Unbiased Prediction
GBS	Genotype by Sequence
GP	Genomic Prediction
GRL	Genomic Relationship Likelihood
GRM	Genomic Relationship Matrix
GWAS	Genome Wide Association Study
ICL	Integrated Complete Likelihood
IPN	Infectious Pancreatic Necrosis
LASSO	Least Absolute Shrinkage and Selection Operator
LD	Linkage Disequilibrium
LOD	Log-likelihood Ratio
LR	Likelihood Ratio
MAF	Minor Allele Frequency
MAS	Marker-Assisted Selection
MME	Mixed Model Equations
mother.ER	Mother Exclusion Ratio
NMBU	Norwegian University of life Sciences
QTL	Quantitative Trait Locus
RMSEP	Root Mean Squared Error of Prediction
SE	Standard Error
SNP	Single Nucleotide Polymorphism
TBV	True Breeding Value
TDGP	Transmission Disequilibrium Genomic Prediction
TDT	Transmission Disequilibrium Test

2 List of papers

Paper 1

Grashei, K.E., Ødegård, J. & Meuwissen, T.H.E. **Using genomic relationship likelihood for parentage assignment**. Genet Sel Evol 50, 26 (2018). https://doi.org/10.1186/s12711-018-0397-7

Paper 2

Grashei, K.E., Ødegård, J. & Meuwissen, T.H.E. **Genotype calling of triploid offspring from diploid parents**. Genet Sel Evol 52, 15 (2020). https://doi.org/10.1186/s12711-020-00534-w

Paper 3

Grashei, K.E., Ødegård, J. & Meuwissen, T.H.E. **Genomic prediction using a case-parental-control model**. Manuscript

3 Abstract

The use of high-density single nucleotide polymorphism (SNP) genotypes enables us to perform highly accurate parentage assignment. However, dependence between loci often results in using a subset of the data to obtain independent loci (likelihood-based parentage assignment), or just a fraction of the genotypes may be informative (exclusion-based parentage assignment). In this thesis, a novel method is suggested to perform parentage assignment using, at its core, genomic relationships which are estimated without the assumption of independence between the loci. Thus, all information from the SNP genotypes is used. In Paper 1, we show that the suggested method, called genomic relationship likelihood (GRL), obtains high accuracies when applied to high-density genotypes. The accuracy obtained by GRL is similar to the one obtained by the exclusion-based method we used for comparison, however with some differences as noted in Paper 1.

Genotyping triploid individuals who inherit two chromatids from the mother and a single chromatid from the father may be useful for species where escapees are an issue, such as in aquaculture. Genotyping triploids may also be useful for breeding programs where triploids are part of the product portfolio because genetic traits may differ between triploids and diploids. In Paper 2, we suggest a novel way of calling genotypes for triploids, and we use the called genotypes to assign the parents of triploid offspring. Due to the special inheritance pattern between mother and triploid offspring, direct assignment of mothers is shown to be both possible and useful. In addition, this inheritance pattern allowed us to map maternal crossovers which have occurred during meiosis.

In some situations, genotyping may be restricted to one category of a binary trait (cases), for example when there is a disease outbreak in a commercial population. In Paper 3, we show that it is possible to estimate heritability and to predict genomic breeding values even when using case-only genotypes in combination with their parental genotypes. The proposed method, called transmission disequilibrium genomic prediction (TDGP), is essentially a genome-wide generalization of the transmission disequilibrium test (TDT), with many of the same pros and cons.

4 Norsk sammendrag

Bruken av høytetthets enkelnukleotid-polymorfisme (SNP) genotyper muliggjør foreldretilordning med høy presisjon. Derimot så resulterer ofte avhengighet mellom loci i bruk av en delmengde av dataen for å oppnå uavhengige loci (sannsynlighetsbasert foreldretilordning), eller så kan bare en brøkdel av genotypene være informative (eksklusjonsbasert foreldretilordning). I denne avhandlingen foreslås en ny metode for bruk i foreldretilordning som, i dens kjerne, bruker genomiske slektskap estimert uten antagelse om uavhengige loci. Dermed blir all informasjon fra SNP genotypene brukt. I artikkel 1 viser vi at den foreslåtte metoden, kalt genomisk slektskapssanssynlighet (GRL), oppnår høye nøyaktigheter når den blir anvendt med høytetthets-genotyper. Nøyaktigheten oppnådd av GRL er lignende den som blir oppnådd av den eksklusjonsbaserte metoden vi brukte til sammenligning, men med noen forskjeller som er nevnt i artikkel 1.

Genotypering av triploide individer som arver to kromatider fra mor og én kromatid fra far kan være nyttig for arter hvor rømlinger er et problem, slik som i akvakultur. Genotypering av triploider kan også være nyttig for avlsprogrammer hvor triploider er del av produktporteføljen fordi genetiske egenskaper kan være forskjellige for triploider og diploider. I artikkel 2 foreslår vi en ny måte å avgjøre genotyper for triploider, og vi tilordner foreldrene til triploide avkom. På bakgrunn av det spesielle nedarvingsmønsteret mellom mor og triploid avkom blir det vist at direkte tilordning av mødre er mulig og nyttig. I tillegg tillot dette nedarvingsmønsteret oss å kartlegge rekombinasjoner som skjedde under meiosen til mødrene.

I noen situasjoner kan genotypering være begrenset til én kategori av en binær egenskap (tilfeller), for eksempel når det er et sykdomsutbrudd i en kommersiell populasjon. I artikkel 3 viser vi at det er mulig å estimere arvegrad og å predikere genomiske avlsverdier selv når det bare er brukt tilfelle-genotyper i kombinasjon med deres foreldregenotyper. Den foreslåtte metoden, kalt transmisjonslikevekt genomisk prediksjon (TDGP), er i hovedsak en genom-bred generalisering av transmisjonslikevekt testen (TDT), med mange av de samme fordelene og ulempene.

5 Synopsis

In all breeding programs the goal is to achieve genetic gain for selected traits over time. Some of the breeding candidates are selected as parents of a new generation, the offspring of these parents then become new breeding candidates for the next generation, and so on. New information is gathered each generation, which can be used to increase the accuracy of selection and thus increase the genetic gain.

Genotyping large numbers of individuals for use in a breeding program has become increasingly common. Selection candidates and individuals included in disease trials or slaughter tests are thus often genotyped, and genotyping individuals over multiple generations results in both parents and offspring having genotypes. Genomic parentage assignment can then be performed, and some of the reasons for doing so are discussed in this thesis. In addition to the rapid increase in number of genotyped individuals in recent years, the density of the genotypes is also increasing. In 2010 the cost of whole genome sequencing was $50,000[1], while the current costs have been significantly reduced [2], potentially resulting in whole genome sequencing replacing SNP genotypes in the relatively near future. However, when using medium/high-density genotypes or whole genome sequences, there will be substantial dependencies (linkage disequilibrium and co-segregation) among loci.

The methods for calling the genotypes are generally developed for diploid individuals. In salmon production, triploids are sometimes used since they are sterile and therefore cannot interbreed with wild salmonids, which protects the wild salmon populations from genetic introgression. However, triploids may still migrate into rivers and potentially disturb reproduction of wild salmon (see below). Furthermore, traits observed in triploids may differ from the same traits observed among their diploid conspecifics [3], potentially also involving genetic differences. Selective breeding must in any case be performed within the diploid population due to triploids being sterile, but may benefit from using triploid training data, which requires a proper method of genotyping triploids. Such benefits may come from

using both triploids and diploids for prediction of genomic breeding values, or for parentage assignment.

The last topic studied in this thesis is binary traits, for which only one of the binary categories (cases), and their assigned parents, are available for genotyping. In such situations, a method for estimating heritability and predicting genomic breeding values using genomic data of cases and their parents may be beneficial.

To solve the problems of parental assignment using high-density (and thus highly multicollinear) genotypes, calling of triploid genotypes and the use of case-only data in genomic prediction, three methods were invented: 1) parentage assignment using genomic relationship likelihoods (GRL), 2) triploid genotype calling and 3) a transmission disequilibrium genomic prediction (TDGP) model for estimation of genomic heritability and prediction of genomic breeding values using case-only genotypes, in combination with parental genotypes. The latter is a case-parental-control model. To the best of my knowledge, all three methods are novel.

5.1 Introduction

5.1.1 Why perform parentage assignment?

Parentage assignment is a useful tool to infer pedigree relationships for use in breeding programs (e.g. [4]) and to identify origin of escaped farmed fish (e.g. [5]).

The subject of escaped farmed fish is of special importance as AquaGen, a salmonid breeding company, in 2014 launched a new product called TRACK™. The TRACK™ product requires AquaGen to genotype all parent individuals of eyed Atlantic salmon (*Salmo salar*) egg deliveries using a medium density SNP chip (50-70k SNP markers). Any genotyped escapee that originates from a TRACK™ delivery can thus be assigned parents. However, applying classical assignment methods on dense SNP chip data has some pitfalls, such as assumption of independence between markers, genotyping errors and missing genotype calls (i.e. "no-calls"). With this in mind, AquaGen wanted to increase their knowledge about parentage assignment methods, especially when using dense genotypes, which is why it is a research subject in this thesis.

The risks of having fish escaping from a fish production facility include: 1) farmed escapees are in some cases known to migrate into rivers where their wild conspecifics are spawning and disturb their spawning rituals and, to some extent, cause nest destruction [6], 2) farmed escapees can successfully spawn with their wild conspecifics, resulting in genetic introgression of "farmed" alleles into the wild populations [7], 3) little is known about how farmed escapees affect the oceans, however, higher abundance of sea lice has been observed in captured escaped Atlantic salmon after one sea winter [8], and 4) the production facility may not be aware that fish are escaping, or the escape may be detected a long time after the incident (e.g. at slaughter), if at all (e.g. [9]). The TRACK™ product aims to address the above points by assigning unknown escapees to known parents and, consequently, tracking where they have, or should have, been during their production cycle.

In addition to identifying escaped farmed fish, parentage assignment is also useful for breeding- and documentation purposes. Examples include: 1) filtering genotypes

or markers based on deviances from Mendelian inheritance laws (e.g. [10]), which is only possible given known parentage, 2) performing transmission disequilibrium testing (TDT) comparing offspring and parental genotypes for genome-wide association studies (GWAS)[11], 3) mixing offspring groups before individual tagging is possible for use in breeding selection programs, 4) performing benchmark studies comparing different genetic stocks mixed before tagging is possible, and 5) predicting genomic breeding values using case-only genotypes (e.g. disease-affected individuals) and their parents as proposed in Paper 3.

5.1.2 Traditional parentage assignment methods

Traditionally, parentage assignment is performed using two main categories of methods: exclusion-based- and likelihood-based parentage assignment [12]. Parentage assignment can be done using various non-DNA methods such as biochemical markers or blood groups [13]. However, I will focus solely on the use of DNA markers in this thesis.

In likelihood-based parentage assignment, the likelihood ratio (LR) conditional on two different hypotheses is often used (the following equation is a special case of the model used by Marshall *et al*. in [14]):

$$LR = \frac{P(g_c|g_s, g_d, H_1)}{P(g_c|H_0)}$$

where g_c is the child genotype, g_s and g_d are the genotypes of the candidate sire and dam, respectively, H_0 is the null hypothesis (often that the candidate parents are random individuals from the population) and H_1 is the alternative hypothesis that the candidate parents are the true parents of the child. Thus, for LR > 1 the alternative hypothesis is more likely, and for LR < 1 the null hypothesis is more likely. As in [14], a LOD score may be used across independent loci: $LOD = \log(\prod_i LR_i) = \sum_i log(LR_i)$, where *i* is locus. A threshold is used by Marshall *et al*. in [14] for the statistic $\Delta = LOD_1 - LOD_2$ to assign parents, where LOD_1 and LOD_2 are the scores for the most likely and second-most likely set of parent candidates, respectively. A threshold for the LR, or for Δ, at which the null hypothesis is rejected may be found empirically by analyzing multiple datasets or by simulation as in [14]. However, in human forensics the weight of the LR may be set intuitively given all of

the evidence [15]. Parents can be assigned categorically (i.e. yes/no) or fractionally (i.e. assigns partially to multiple parent candidates based on relative- or posterior likelihoods) using likelihood-based methods [12]. If the true parents are not in the dataset, their genotypes can be imputed using parental reconstruction [e.g. 12, 16]. As noted by Jones *et al.* in [12], the genetic markers used in likelihood-based parentage assignment are most often assumed to be independent due to the complexity that arises when non-independence is assumed. Thus, markers used in likelihood-based methods should be filtered in such a way that there is no dependence between them. Consequently, the number of markers is effectively reduced from potentially tens- or hundreds of thousands to tens or hundreds. To compensate for some of the loss in information due to the requirement of marker independence, highly polymorphic markers such as microsatellites may be used in likelihood-based parentage analyses [e.g. 17, 18]. Due to the relatively low number of microsatellite loci used, genotyping errors can decrease the accuracy of parentage assignment [19]. Small SNP panels of (relatively) independent SNPs has been shown to be effective for parentage assignment [20]. SNPs are generally less prone to genotyping errors and mutations compared to microsatellites [21], although more SNPs are generally needed to achieve the same power due to the SNPs being less polymorphic. However, using multiple linked SNP markers as 'super' markers to increase the polymorphism is also an option [22]. Bayesian posterior probability models can be used to incorporate the possibility of the candidate parent(s) not being present in the dataset, e.g. if the fraction of sampled parents to the total number of parents is known or can be estimated [23]. The Bayesian framework also incorporates the possibility of including other population-level variables such as age, sex or location. Such models are called full probability parentage assignment models [12, 24]. A Bayesian likelihood-based method with assumption of independent loci where several hypothetical relationships are compared with the hypothesis of "offspring-parent" was developed by Whalen *et al.* in [25] (based on the work by Huisman in [26]). The parental assignments done by Whalen *et al.* in [25] had overall high accuracy, with increasing accuracy with increasing number of SNPs and genotype by sequence (GBS) coverage. However, the rate of false positives was high when true parents were excluded from the dataset, and increasingly so

when the number of loci increased towards 50 000, well beyond what can be considered as a set of independent loci.

The other method of parentage assignment is by exclusions. Exclusions are observed deviations from Mendelian laws of inheritance between offspring and candidate parent(s) [12]. For duos of a single parent and its offspring, opposite homozygotes are not possible by laws of Mendelian inheritance. A single exclusion is in theory enough to reject a parent candidate as the true parent. However, genotyping errors or mutations may introduce false exclusions. Because of this, a threshold on the number of allowed exclusions must be chosen, where any candidate parent-offspring pair exceeding this threshold is rejected as a true parent-offspring duo. The threshold on the number of exclusions may be estimated empirically by using sets of known offspring-parent duos, or by simulation [e.g. 27]. Another way of estimating the exclusion threshold is by plotting the number of exclusions for the three most likely parent candidates of each offspring when at least some offspring are expected to have one or both parents in the dataset. In this situation, at least two distinct distributions should be observed, i.e. one where the top parent candidate(s) is/are the true parent(s) (few exclusions) and one or more where the parent candidates are false (more exclusions). Thus, the exclusion threshold is estimated to be somewhere between the two (hopefully separated) distributions, as was done in Paper 2. The expected number of exclusions between true offspring-parent pairs is dependent not only on genotyping errors, but on call-rate as well. For any reduction in call-rate, the number of expected exclusions decreases for all duos of offspring and parent. Consequently, it is useful to divide the number of exclusions by the number of called genotypes, i.e. using exclusion ratios (ERs) instead of number of exclusions to reject false duos of offspring and parent candidates:

$$ER = \frac{\#exclusions}{\#calls}$$

ER was used in place of exclusions in papers 1 and 2.

Only duos of parents and offspring have been mentioned so far, however trios of mother-father-offspring are also possible to use in both exclusion- and likelihood-

based parentage assignments. For such trios, as with duos, offspring should not be oppositely homozygous to any of its parents. In addition, offspring where both parents are the same homozygote should be homozygous for the same allele, while offspring of two opposingly homozygous parents should be heterozygous. Any deviation from these rules results in an exclusion. Thus, trios contain more information than duos for both exclusion- and likelihood-based parentage assignments. Because exclusion-based methods do not use information from heterozygous parents, likelihood-based parentage assignments are more powerful than exclusion-based assignments. However, exclusion-based assignments are considered more easily interpretable compared with likelihood-based assignments [e.g. 28]. A key difference between exclusion- and likelihood-based parentage assignments is that exclusions are summed while likelihoods are multiplied across loci. Factors such as genotype errors, call rates and relatedness between offspring and candidate parent(s) affect both the exclusion sums and the likelihood products. However, the interpretation of the likelihood-products depends on the model used, and especially on the common assumption of independence between markers, while no such assumption is used in the exclusion-based assignment. Thus, when using medium- or high-density marker panels for parentage assignments, exclusions (or ERs) are commonly used due to the complications that arise for dependent markers [e.g. 29].

Using traditional methods, inbreeding will tend to increase the likelihoods and decrease the number of exclusions when considering close non-parental relatives as candidate parents. For example, if a candidate parent is an uncle or aunt of an inbred offspring.

While likelihood- and exclusion-based methods are predominant in parentage assignments, there is another method that uses linear regression of the offspring genotypes on the parental candidate genotypes [30]. Because the offspring genotype is expected to be equal to the average of the two parental genotypes, a threshold can be used for the linear regression slope between the parental- and offspring genotypes to perform parentage assignment. Such a threshold has to be chosen much the same way as the threshold for exclusion-based assignment. The method of

regressing offspring on parental genotypes is interesting, but as it is relatively new it still remains unclear if it can challenge likelihood- or especially exclusion-based parentage assignment methods.

5.1.3 Background on triploid salmon production

Triploid salmon are sterile, and therefore they are used by the aquaculture industry to limit the impact of escaped farmed salmon. Triploid farmed salmon are produced by pressurizing newly fertilized salmon eggs to prevent the second polar body from leaving the secondary oocyte during meiosis [31, 32]. Thus, a pair-set of sister chromatids are passed down for each chromosome from the mother to the triploid offspring, while a single set of chromosomes is passed down from the father. The sister chromatids passed down from the mother are identical except for any recombinations that might have happened during the prophase of the meiosis in the mother.

Even though genetic introgression of farmed triploid salmon alleles into wild populations is impossible due to the sterility of triploids, some may still escape from sea net pens and migrate into rivers [33]. Here, the escapees may cause damage to the wild salmon spawning grounds, they may decrease the success rate of spawning wild salmon by taking up space in the river, or triploid males may even initiate spawning in wild diploid females [34]. Therefore, it is important to also being able to track triploid escapees, e.g. by parentage assignment. However, assigning parents to a triploid offspring is non-trivial since the methods of genotype calling and parentage assignment of diploids do not apply to triploids without major modifications.

Trait differences between diploids and triploids have been shown to exist in salmonids [35-37]. Parentage analysis has been done by Taylor *et al.* by use of microsatellite marker genotypes [38]. Prior to the work done in Paper 2, there has, to my knowledge, been no work done which directly compares the quantitative genomics of different ploidies by use of SNP genotypes. However, the differences in trait genetics of diploid and triploid salmonids has been analyzed by use of known families [38-42] or by use of transcriptomics [43, 44]. To be able to study the

differences in trait genetics between diploids and triploids using SNPs, a method for calling SNP genotypes of triploids is needed.

5.1.4 Brief summary of some traditional prediction methods used in breeding

Phenotypic selection

The most basic way of selection for breeding purposes is phenotypic selection, also called mass selection, where parents for the next generation are selected using own phenotype only. Phenotypic selection is often normally based on traits that are easily recorded on individual breeding candidates, such as growth [e.g. 45, 46, 47]. Phenotypic selection has been done for thousands of years. The phenotype is a relatively accurate predictor of the true breeding value for traits of high heritability, while the accuracy is poor with low heritabilities. Furthermore, phenotypic selection is also restricted to traits that can be measured on individuals that can subsequently be used as breeders. For example, resistance against specific infectious diseases is typically not suited for phenotypic selection.

Pedigree-based selection

The covariance between phenotypes and family relatedness can be used to increase the accuracy of selection compared to simply using the phenotype as the predicted breeding value. Training data that include known half- and full-sib groups can be phenotyped to accurately predict family-based breeding values by estimating family- (or dam- and sire-) effects. Family-based breeding values enables selection for invasive traits, such as carcass traits. Henderson formulated the mixed model equations (MME) through his groundbreaking papers from 1953 [48] and 1959 [49] (see [50] by Searle for an overview of Henderson's work). Henderson's work built on the work done by Crump [51, 52] and Eisenhart [53]. This work allowed joint estimation and prediction of fixed and random effects, respectively, to obtain best linear unbiased predictions (BLUP) of breeding values, using the pedigree to form a random covariance structure. Assuming that variance components are known, it is possible to predict breeding values despite the $n < p$ problem where there are more unknow variables than observations. Consequently, breeding experiments designed

to produce equally sized half- and/or full sib groups are not needed to predict family breeding values as long as there is a pedigree describing the numerator relationship between all included individuals. It can be noted that any pedigree-based method will produce identical predicted breeding values for all full-sibs that do not have phenotypes or phenotyped offspring. When this is the case, within-family genetic variation is not captured, limiting the maximum obtainable accuracy to ~0.71 [54].

Marker-assisted selection (MAS)

SNPs in LD with Quantitative trait loci (QTLs), or indeed the QTLs themselves, can be used to select individuals in a breeding- or production program. Genome wide association studies (GWAS) can be used to identify such genetic markers. For example, a QTL explaining most of the genetic variance in infectious pancreatic necrosis (IPN) in Atlantic salmon was found by Houston *et al.* [55], and, in parallel, found, confirmed and fine-mapped by Moen *et al.[56]*. The IPN QTL was subsequently included in the AquaGen breeding program by use of MAS. Different methods for GWAS can be used depending on the available information. For example, for high-density SNP data, each candidate SNP can be tested for significance by including it as a fixed- [e.g. 57] or random- [e.g. 58, 59, 60] effect in the model and adjusting for polygenic effects. Another example is analysis of case-only data, i.e. where genotyping is restricted to individuals having a certain phenotypic condition (e.g. mortalities during a disease outbreak) and their parents. In such situations, the transmission disequilibrium test (TDT) can be used to test for deviations from parental expectation in Mendelian inheritance of alleles in the case-individuals [11].

Integrating MAS into a breeding program in combination with conventional (e.g. genomic) breeding values is an option when the goal is to obtain changes in specific QTL allele frequencies [61].

Genomic prediction using GBLUP

The genomic best linear unbiased prediction (GBLUP)[62], using dense genome-wide SNP markers, has in recent years been increasingly used due to its improved accuracy compared with traditional pedigree-based BLUP (described above) [e.g.

63, 64, 65]. GBLUP also provides the possibility of predicting individual genomic breeding values even for individuals without own phenotypes or recorded offspring. In many cases, this allows individual prediction early in life, e.g. even before phenotypes are available for offspring or even siblings. In its simplest form, genomic prediction (as GBLUP) simply replaces the pedigree relationship matrix in the mixed model equation (MME) system with a genomic relationship matrix (GRM). The GRM is a matrix of genomic relationships between duos of all genotyped individuals estimated using e.g. one of VanRaden's [66] methods. Thus, GBLUP uses individual information (within- and between family information), as opposed to regular BLUP which uses pedigree information. The general model used in GBLUP is:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u} + \boldsymbol{e}$$

where $\boldsymbol{y}$ is a vector of phenotypes, $\boldsymbol{X}$ is an incidence matrix linking the fixed effects with the phenotypes, $\boldsymbol{\beta}$ is a vector of fixed effects, $\boldsymbol{u} \sim N(\boldsymbol{0}, \sigma_g^2 \boldsymbol{G})$ is a vector of random individual effects having covariance structure $\boldsymbol{G}$ (=GRM) scaled by the genetic variance σ_g^2 and $\boldsymbol{e} \sim N(\boldsymbol{0}, \sigma_e^2 \boldsymbol{I})$ is a vector of residuals with common residual error variance σ_e^2.

SNP-BLUP and its equivalence with GBLUP

While GBLUP predicts individual effects (i.e. genomic breeding values), a non-weighted SNP-BLUP[67] predicts marker effects using the model:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Mg} + \boldsymbol{e}$$

where $\boldsymbol{y}$, $\boldsymbol{X}$, $\boldsymbol{\beta}$ and $\boldsymbol{e}$ are as above, $\boldsymbol{M}$ is a matrix of genotypes adjusted for two times the allele frequency and $\boldsymbol{g} \sim N(\boldsymbol{0}, \sigma_m^2 \boldsymbol{I})$ is a vector of marker effects where each marker effect is assumed to be identically and independently distributed, and $\sigma_m^2 = \frac{\sigma_g^2}{2\boldsymbol{p}'(\boldsymbol{1}-\boldsymbol{p})}$, where $\boldsymbol{p}$ is a vector of allele frequencies and σ_g^2 is as above. In 2009, Goddard [68] and Strandén & Garrick [69] showed that the product $\boldsymbol{Mg}$ is in reality $\boldsymbol{u}$ from the section above. Thus:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Mg} + \boldsymbol{e} = \boldsymbol{X\beta} + \boldsymbol{u} + \boldsymbol{e}$$

Consequently, SNP-BLUP and GBLUP are equivalent statistical models.

There are numerous methods that expand on SNP-BLUP which are not explored in this thesis, such as Bayes A, Bayes B, Bayes C, Bayes D, Bayes R, LASSO and Elastic net [67, 70-73].

Aims

The main aim of this thesis is to use parental genomic information to optimize breeding programs, with an emphasis on aquaculture breeding programs. Modern breeding programs have many facets. Consequently, the aims were focused into three directions: (1) creating a novel method for parentage assignment using, at its core, genomic relationships which are widely used in modern breeding programs to see if improvements from current parentage assignment methodology could be found, (2) developing a method for calling genotypes in triploid Atlantic salmon and assigning diploids parents to triploid offspring and (3) developing a method for estimating heritability and predicting genomic breeding values using genotypes from only cases and their parents. Although the emphasis was on aquaculture breeding programs, an overall aim was to generalize the methods such that they may be used in non-aquaculture breeding programs, and even by geneticists not involved in breeding at all.

In the following I describe (1) a method for parentage assignment using genomic relationship likelihoods (GRL), (2) genotype calling for triploids using a modified version of the mixture models implemented by the 'mclust' R package, parentage assignment using triploid offspring genotypes and other uses of triploid genotypes and (3) estimation of heritability and prediction of genomic breeding values using a case-parental-control model coined transmission disequilibrium genomic prediction (TDGP). The latter three numbered points address the three numbered aims in the previous paragraph, respectively. Perhaps deviating from most other PhD introductions is the extra emphasis put on the future perspectives, where I try to explore possible topics to expand on the work done in the thesis. Thus, the future perspectives should not necessarily be interpreted as testable hypotheses, but rather directions of interest which may or may not be explored in the future.

5.2 Methods

5.2.1 Parentage assignment using Genomic Relationship Likelihood (GRL)

In Paper 1, we present what we believe is a novel method of parentage assignment using genomic relationship likelihoods (GRL). The GRL method does not assume independence between loci as it uses genomic relationships at its core (see Paper 1 for a reflection regarding normality of genomic relationships). Thus, the use of GRL with highly dependent markers such as with high-density SNP panels or even whole genome sequences is possible. In theory, the GRL method can work with all types of genetic data which can be used to estimate relationships between individuals. However, in Paper 1 we have used VanRaden's [66] first method of calculating realized genomic relationships based on SNP data, i.e.

$$r_{ij} = \frac{\sum_k^c (m_{ik} - 2p_k)(m_{jk} - 2p_k)}{2\sum_k^c p_k(1 - p_k)}$$

where r_{ij} is the realized genomic relationship between individuals i and j, m_{ik} and m_{jk} are the genotypes (0, 1 or 2) for individuals i and j at locus k, respectively and p_k is the allele frequency at locus k. The VanRaden realized genomic relationship has been embraced by the scientific community through the use of genomic prediction (GP) or genome-wide association studies (GWAS), [e.g. 74, 75, 76]. However, studies of parentage assignment using VanRaden's genomic relationships seems scarce. In Paper 1, we formed residuals by subtracting realized genomic relationships by their expectations given a trio structure of an offspring with two true parents:

$$e_{O,O} = r_{O,O} - E\left(r_{O,O}|TP\right) = r_{O,O} - \left(1 + 0.5r_{P_1,P_2}\right)$$

$$e_{O,P_1} = r_{O,P_1} - E\left(r_{O,P_1}|TP\right) = r_{O,P_1} - 0.5\left(r_{P_1,P_1} + r_{P_1,P_2}\right)$$

$$e_{O,P_2} = r_{O,P_2} - E\left(r_{O,P_2}|TP\right) = r_{O,P_2} - 0.5\left(r_{P_2,P_2} + r_{P_1,P_2}\right)$$

where $e_{O,O}$, e_{O,P_1} and e_{O,P_2} are the residuals for offspring to offspring, offspring to first parent candidate and offspring to second parent candidate, respectively, while $r_{O,O}$, r_{O,P_1}, r_{O,P_2}, r_{P_1,P_1}, r_{P_2,P_2} and r_{P_1,P_2} are the realized genomic relationships between

offspring and itself, offspring and first candidate parent, offspring and second candidate parent, first parent candidate with itself, second parent candidate with itself and between first and second parent candidates, respectively. Note that $E(r_{O,O}|TP)$, $E(r_{O,P_1}|TP)$ and $E(r_{O,P_2}|TP)$ are the expected genomic relationships between offspring and itself, offspring and first parent candidate and offspring and second parent candidate, respectively, conditional that P_1 and P_2 are the true parents of O (TP stands for true parents). A property of using such residuals is that inbreeding is accounted for, e.g. if two full siblings are mated, the relationship their offspring has with itself is expected to increase by half the realized relationship between the parents.

The genomic relationship likelihood (GRL) is calculated for each trio of offspring and two parent candidates:

$$GRL = -\frac{1}{2}(\boldsymbol{e} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{e} - \boldsymbol{\mu})$$

where $\boldsymbol{e}$ is a 3x1 vector of residuals $e_{O,O}$, e_{O,P_1} and e_{O,P_2}, and $\boldsymbol{\mu} = E(\boldsymbol{e})$ is the expected values of the residuals. Note that the form of the GRL statistic above is the same as the exponent in the multivariate normal distribution function. $\Delta GRL = GRL_1 - GRL_2$ is a second statistic used by the assignment procedure where GRL_1 and GRL_2 are the highest and second highest GRL values achieved for an offspring across all candidate parent-offspring trios, respectively. A threshold for ΔGRL is set at 6.9, which implies that the most likely parent-pair should be at least $1000(\approx e^{6.9})$ times more likely than the second-most likely parent pair, see Paper 1. Then, a two-step approach where step 1 (the allele dropping step) provides the initial estimates for $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and the final estimate of the GRL_1 threshold, while step 2 (the iteration step) updates $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ until the number of assignments starts to decrease. Thus, the test dataset can be used directly to estimate the $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and GRL_1 parameters while performing the parentage assignments, removing the need to have a separate training dataset for estimation purposes. However, using the estimated parameters on future datasets is only possible under the assumption that the future datasets have a similar genotype error rate as the dataset used to estimate the parameters

(see Paper 1). When estimation of the parameters is done using the test dataset, the GRL method is robust for genotype error rates of at least 3%, as shown in Paper 1.

5.2.2 Genotyping triploids

For any kind of genotype-based genetic analysis we first need to call the genotypes of the individuals. At the moment, Thermo Fisher's software for calling genotypes does not support triploid DNA. However, triploid allele signal intensities can still be produced the same way as for diploid individuals. At least one method of calling genotypes using the allele signal intensities of polyploid organisms exists [77, 78]. However, we chose to build on the functionality in the R package 'mclust' [79], to create, to the best of my knowledge, a novel way of calling genotypes. The choice of using mclust was due to our familiarity with its functionality and its substantial documentation, frequent updates, and extensive use. The Bayesian information criterion (BIC) [80] and the integrated complete likelihood (ICL) [81] statistics may be used for model selection. The BIC statistic penalizes complex models to adjust for overfitting, while ICL, in addition, penalizes clusters which are not well separated. High quality SNP markers are expected to have clear cluster separation, harmonizing well with the ICL penalization. When looking at multiple models for each marker, we found that ICL was a suitable statistic for identifying the optimum model for genotype calling (see Paper 2). In addition, marker quality was assessed by:

$$\Delta ICL = ICL_1 - ICL_2$$

where ICL_1 and ICL_2 is, respectively, the highest and second highest ICL values achieved using different number of genotype clusters for a specific marker. Intuitively, having a low ΔICL means that the top two models having different number of clusters are of similar quality, and thus genotype calls will be uncertain. Oppositely, having a high ΔICL indicates that the top model is much more likely to model the correct number of genotype clusters.

5.2.3 Parentage assignment in triploids

Parentage in triploids can be assigned in much the same way as for diploids using exclusions (see Background). However, the triploid genotypes do not only allow us

to assign parents to the triploids, but also allow us to distinguish between mothers and fathers in the assignment. This is due to the fact the triploid offspring, for each locus, inherits two alleles from the mother, but only one allele from the father. Hence, the sex of each parent can be identified directly from the genotypes without using sex markers. This is done by looking at the heterozygous genotypes in the triploid offspring: for an offspring having genotype “AAB” the true mother has necessarily contributed with at least one “A”-allele, implying that the maternal genotype cannot be “BB”. Likewise, for an offspring having genotype “ABB”, the maternal genotype cannot be “AA”. However, for the same triploid genotypes, the father contributes a single allele to its offspring and, considering a duo exclusion, can have any genotype without violating the laws of Mendelian inheritance. Such “mother-specific exclusions” were used to construct exclusion ratios by dividing by the number of calls in mother and offspring, and coined “mother.ER”.

5.2.4 Maternal recombination linkage map using triploid inheritance pattern

The pressure-induced triploid salmon offspring carry two sets of alleles inherited from their mothers, where the sets of alleles are sister chromatids. This means that, prior to crossover during meiosis, the alleles are identical, i.e. “AA” or “BB”. However, if recombination with the homologous chromosome occurs, part of the sister chromatid is swapped for a different homologous chromatid, possibly ending up with two different alleles, i.e. “AB”. Thus, for triploid offspring-mother-father trios at loci where the father is homozygous (“AA” or “BB”) and the mother is heterozygous (“AB”) the maternal inheritance in the triploid offspring can be deduced to be either “AA”, “AB” or “BB”. Inheritance of an “AB” maternal genotype implies that a maternal recombination event has occurred somewhere between the centromere and the locus in question because both the maternal alleles are represented. If yet another maternal crossover occurs further away from the centromere, the offspring again inherits maternal alleles of either “AA” or “BB”.

5.2.5 Transmission disequilibrium genomic prediction (TDGP)

Using a categorical trait when predicting breeding values with SNP-BLUP/GBLUP or family-based methods normally involves the sampling of individuals from at least two levels (e.g. deceased and survivors). These levels must be assigned a value on a discrete scale (e.g. 1 and 0, respectively)[63, 82, 83]. If there are exactly two levels, the trait is called binary.

In Paper 3 the afflicted (case) individuals are given phenotype 1, while the non-afflicted (non-cases) are given phenotype 0. Consequently, for SNP-BLUP, the right-hand side of the MME reduces to be a function of the difference between the expected allele frequencies (i.e. for entire offspring group) and the allele frequencies observed in the afflicted group. However, there are situations when only individuals from a single level of a categorical trait are practical to sample, e.g. sampling deceased individuals from a disease outbreak, as described further in Paper 3. A novel case-parental-control model was therefore developed which allows analysis of traits where genotyping is restricted to case-individuals and their parents. This method was coined Transmission Disequilibrium Genomic Prediction (TDGP). TDGP can be said to be the genomic prediction-equivalent of the transmission disequilibrium test (TDT)[11]. As with TDT, TDGP uses the genotyped case- (e.g. afflicted) individuals and their (genotyped) parents to identify deviations in observed inherited alleles in the cases from what is expected given the parental genotypes. Consequently, TDGP requires known and genotyped parents. With this condition satisfied, genotyping controls (non-afflicted individuals) is not needed, thus saving the cost of genotyping these.

5.3 Discussion

5.3.1 Assigning parents using GRL

The results from Paper 1 shows that the GRL method has significantly higher true-positive- and true-negative rate when using medium/high-density SNP data compared with Colony2 [84], which assumes independence of the loci. GRL has very similar results compared with the exclusion-based approach used in Paper 1. However, the GRL parameter estimation procedure was used to estimate the proper ER threshold for rejecting a candidate trio, which capitalizes on the GRL method's ability to estimate parameters using the test dataset. When using an ER threshold estimated from a dataset with 3% genotyping errors on a dataset with 1% genotyping errors, the exclusion-based method had a small increase in false assignments. However, under the same circumstances, the GRL method chose to not assign any trios. Thus, when doing a parentage assignment and not performing training on the test dataset, a decision has to be made whether a small (but unknown) increase in assignment error is acceptable, or if all trios should rather be rejected. In situations where training can be performed on the test dataset, i.e. the test dataset contains a sufficient number of true (but unknown) offspring-mother-father trios, the GRL method has an accuracy near 100%.

Existence of clones is a particular problem with respect to parentage assignment. Ordinary exclusion-based methods will not be able to distinguish between a true parent and a clone of the offspring because none of their true genotypes will inflict exclusions when compared with the "offspring". In contrast, GRL is able to distinguish between clones and true parents, as the relationship between the offspring and the clonal parent candidate will deviate from expectation (i.e. increasing the genomic residual). For example, if a non-inbred clone is inserted as the first parent, the genomic relationship between offspring and the "parent" will be $r_{O,P_1} \approx 1$, while the relationship between the "parent" and itself and offspring and itself are both $r_{P_1,P_1} = r_{O,O} \approx 1$. Consequently, the residual between offspring and parent (assuming no relatedness between offspring/p1 and p2, i.e. $r_{P_1,P_2} = r_{O,P_2} \approx 0$) is $e_{O,P_1} = r_{O,P_1} - 0.5\left(r_{P_1,P_1} + r_{P_1,P_2}\right) \approx 1 - 0.5(1 + 0) = 0.5 \gg 0$. Thus, the residual is inflated, resulting in a (very) poor GRL value.

Future perspective – duo parentage assignment using GRL:

In this thesis, parentage assignment using GRL has been restricted to situations where genotypes of both mother and father are available. However, it is theoretically possible to use a GRL-based approach to perform parentage assignment of duos of an offspring and a single parent using the relationship between the parent candidate and the offspring. We know that (using the notation from above):

$$E\left(r_{O,O}|TP\right) = 1 + 0.5r_{P_1,P_2}$$

The higher the relationship between the parents, the more inbred their offspring is expected to be. By replacing $E\left(r_{O,O}|TP\right)$ by the *realized* $r_{O,O}$ we can estimate r_{P_1,P_2} (the relationship between the two parents) as:

$$0.5r_{P_1,P_2} = E\left(r_{O,O}|TP\right) - 1$$

$$\Rightarrow \hat{r}_{P_1,P_2} = 2\left(r_{O,O} - 1\right)$$

where $\left(r_{O,O} - 1\right)$ is the realized inbreeding level of the offspring. Note that $\hat{r}_{P_1,P_2}$ is an estimate for the relationship between the true parents of the offspring based on the offspring's own genotype and does not depend on any specific parent candidate(s).

We also know from above that $E\left(r_{O,P_1}|TP\right) = 0.5(r_{P_1,P_1} + r_{P_1,P_2})$. By substituting $\hat{r}_{P_1,P_2}$ into the latter expression we get:

$$\begin{aligned} e_{O,P_1} = r_{O,P_1} - E\left(r_{O,P_1}|TP, r_{P_1,P_2} = \hat{r}_{P_1,P_2}\right) &= r_{O,P_1} - 0.5\left(r_{P_1,P_1} + \hat{r}_{P_1,P_2}\right) \\ &= r_{O,P_1} - 0.5\left(r_{P_1,P_1} + 2\left(r_{O,O} - 1\right)\right) = r_{O,P_1} - \left(0.5r_{P_1,P_1} + r_{O,O} - 1\right) \end{aligned}$$

Note that, using duo GRL, there is no way to separate full-sibs from true parents except for rare cases where either the offspring or the true parent are severely inbred. Still, full-sibs inserted as candidate parents will be excluded with the duo exclusion method. A clone of the offspring will not be assigned as a parent using duo GRL, while it will not be excluded as a parent using duo exclusions. Consequently, using both duo exclusion and duo GRL for parentage assignment may increase the overall assignment accuracy. Neither duo exclusions nor GRL (except in rare inbreeding cases) can identify who is the offspring and who is the parent when

generation for the individuals is unknown. Duo parentage assignment based on GRL deserves further investigation.

5.3.2 Calling genotypes in triploids

Loci having a large ΔICL tend to have fewer uncalled genotypes (NoCalls) and lower ER, i.e. fewer Mendelian exclusions, between offspring and assigned parents (see Paper 2). Genotypes were called for diploids as well as triploids to compare the method in Paper 2 with Affymetrix Power Tools (APT), the (Thermo Fisher) proprietary software used for calling genotypes in diploids. The results show that the method in Paper 2 needs to filter out ~7,500 SNPs by using the ΔICL statistics to obtain similar ER as when using APT for genotype calling in diploids. However, the method in Paper 2 is useable both for diploids and triploids, and it could easily be modified to call genotypes for polyploids as well.

Future perspective - Can ICL be used by the EM algorithm to improve genotype calling accuracy?

The "mclust" R package [79] uses the expectation-maximization (EM) algorithm [85] to estimate parameters for Gaussian finite mixture models. For each model that is tested, the EM algorithm iterates towards a local maximum likelihood estimate of the model parameters, which is hopefully the global maximum. When all defined models have achieved such parameter estimates, the models are penalized by the number of parameters using BIC and by (lack of) cluster separation using ICL [81]. Then, the model having the highest (log) likelihood after penalization may be assumed to be the correct model. However, the chosen model might be misdirected by outliers or noise in the data that contributes to the EM algorithm getting stuck with poor local maximum likelihood estimates. It would be interesting to see if incorporating the ICL penalization directly into the EM algorithm would increase the genotyping accuracy. Thus, each iteration likelihood can perhaps be replaced with an iteration score which includes the ICL penalization, or the use the ICL penalization directly.

Future perspective - Quantitative trait differences of diploids and triploids:

Triploid and diploid individuals have previously been compared using family information (see "Background on triploid salmon production" above). However,

comparing the same trait, e.g. growth, in diploids and triploids using medium- or high-density SNP genotypes has to my knowledge never been done before publication of Paper 2. Kjøglum *et al.* [86] used genotyped diploid and triploid individuals, where the triploid genotypes were called using the results from Paper 2, to show that growth in Atlantic salmon was likely the same genetic trait for diploids and triploids. Kjøglum *et al.* [86] points out that adjusting for the difference in allele inheritance from sire and dam, and for the covariance of the alleles inherited maternally, makes triploid quantitative genetics less trivial than indicated in other studies [e.g. 38, 40]. Even though growth genetics may be largely the same for diploid and triploid Atlantic salmon, it is not necessarily so for other traits, or in other species.

Future perspective - Can triploids be used to investigate dominance?

Since triploids have three alleles, the assumption of additivity of allele effects, often used by genomic selection methods, may perhaps be more precisely investigated by using triploids compared with using diploids, since there are more possible genotypes and thus more interaction terms between different genotypes. Estimation of allele dominance may be dependent on the environment. For example, for IPN disease in salmon, the favorable allele for reduced mortality in the challenge test was deemed largely dominant, however the authors stated that the dominance may be subject to the environmental pathogen dose [87]. Thus, having an additional allele could be beneficial when making inferences as to the degree of dominance displayed in one or multiple different environments.

Future perspective - Single QTL effect differences in triploids and diploids:

When performing a genome wide association study (GWAS), the intention is to estimate an allele substitution effect, and identify its significance, for a genetic marker, e.g. a SNP. The method may adjust for effects not related to the marker such as e.g. sex, location or relatedness among individuals. It would be interesting to investigate if the triploid genotype “AAA” for a known QTL would have a different effect compared with the diploid genotype “AA” due to the additional “A” allele. Similarly, comparing triploid- (“AAB” and “ABB”) with diploid (“AB”) heterozygotes to investigate the possible ploidy-effect of heterozygosity. If an allele substitution

effect is truly and observably additive relative to diploids, and the trait is the same for diploids and triploids, the addition of a third allele in the triploids should increase the genetic variance. If this is the case, then perhaps GWAS-es using triploids can be more powerful when contrasting opposing genotypes “AAA” versus “BBB” than when using diploids and contrasting “AA” versus “BB”.

5.3.3 Assigning parents to triploids

After a strict marker quality filtering, 13 906 SNPs were retained from a total of 56 177 for use with parentage assignment in Paper 2. Out of 379 triploid offspring, all were assigned at least one parent and 304 were assigned both parents. Lacking assignments were likely due to some of the parents being absent from the candidate parent dataset. The assignment of mothers and the maternal recombination linkage map both indicate a highly accurate parentage assignment, see below.

5.3.4 Assigning mothers to triploids

When the maternal assignments were compared with the general parentage assignment, no mothers were found to be assigned as fathers, or vice versa. Out of the 304 triploid offspring that were assigned both parents, all were assigned a mother and a father (i.e. none were assigned two apparent mothers or two apparent fathers). Out of the 75 triploid offspring that were assigned a single parent, 14 were assigned a mother and 61 were assigned a father. All assigned parents were consistent with regards to sex. These results indicate that triploid parentage assignment was accurate, and the sex of each parent was accurately predicted.

5.3.5 Maternal recombination linkage map using triploid inheritance pattern

The rate of maternal crossovers observed in triploid offspring were estimated by the fraction of maternally inherited heterozygous alleles, see Methods and Paper 2. The linkage map shown in Paper 2 coincides well with the findings of Lien *et al*. in [88], and serves as a validation of both the parental- and maternal assignments which were both necessary for this analysis.

Future perspective - Differences in triploid genetic variance due to distance between QTLs and centromeres:

The two maternally inherited alleles in triploids are identical sister chromatids except for any crossovers that might have occurred. In Paper 2 we found that the chance of observing a maternal crossover near the centromere is low. The chance of observing a maternal crossover seems to increase from ~0% to ~100% when moving away from the centromere along the chromosome. Generally, a second crossover seems not to occur except for in the larger chromosomes, indicating crossover interference. This means that the triploid offspring should, generally, be more homozygous around the centromere and more heterozygous further away from the centromere. Consequently, traits controlled mainly by genes close to the centromere in triploids may have increased genetic variation compared with traits mainly controlled by genes further away from the centromere. However, this is only true if the causal variants of the traits adhere to allele additivity relative to the diploids. For example, a QTL genotype of "AAA" in triploids would then be expected to give a more extreme effect than "AA" in diploids. There are two hypotheses to investigate: 1. if the genetic variance in triploids deviates from what is expected if the trait adheres to allele additivity relative to the diploids, adjusted for crossover rate, and 2. whether the allele substitution effects are the same in triploids and diploids.

Future perspective - What can we learn by studying the allele inheritance pattern between induced triploids and their mothers?

Imputing SNP genotypes can be done by building a haplotype library and identifying which haplotypes are inherited by the offspring from each parent, including any possible crossovers that might have occurred during parental meiosis [89]. However, the haplotype library and parental- and offspring haplotype prediction might contain errors. Using induced triploids, the maternal crossovers can be identified directly from the genotypes for markers where the father is homozygous, and the mother is heterozygous. This information may be used to predict both maternal and triploid offspring haplotypes, and thus increasing phasing accuracy. In addition, if triploids are already induced and genotyped, as they were in Paper 2, this facilitates investigations into individual female (maternal) variation in recombination rate.

5.3.6 Transmission disequilibrium genomic prediction (TDGP)

From Paper 3, TDGP seems to work best when the population is large and the case incidence is low, i.e. a sizeable number of cases despite low incidence. This is likely due to the cases being phenotypically, and thus likely also genetically, more extreme with regards to the trait in question. Such extreme cases are more likely to be enriched for alleles that increase the probability of becoming a case.

TDGP performs genomic predictions using within-family information only. The predicted marker effect from TDGP depends on whether a systematic deviation from neutral inheritance is observed within all families for a marker due to LD with QTLs affecting risk. In comparison, ordinary SNP-BLUP (or equivalently GBLUP) uses both within- and between-family information (see Paper 3 and [e.g. 90]), i.e. SNP effects may not only capture markers in LD with QTLs, but also polygenic family differences. Consequently, accuracy of SNP-BLUP is expected to decline if applied to individuals being distantly related to the training population. Because the TDGP marker effects depend on LD between their associated markers and causative QTLs, the predictive ability is not expected to decline with decreasing relationship in the same fashion as for SNP-BLUP.

In practical use of TDGP, genomic assignment of parents to offspring is highly recommended to ensure that offspring genotypes are matched with correct parental genotypes.

Future perspective – Using TDGP with high-density SNP chips or whole genome sequencing:

Ordinary SNP-BLUP uses SNP markers to fit QTLs affecting the trait (using markers in LD with the QTLs) as well as polygenic family effects, using markers to capture the relationship structure in the population. In contrast, TDGP does not attempt to capture the relationship structure of the population but rather uses markers solely to fit QTL effects based on markers in LD with the QTLs. Thus, one may hypothesize that increasing the density of SNPs may increase accuracy of TDGP, as the QTLs will be more likely to have markers in close LD. However, this is not supported by the results of this thesis. Figure 1 shows that the accuracy of TDGP asymptotes at an

accuracy of 0.70-0.75 at around 15 000 SNPs across a genome of 30 chromosomes (1 Morgan/chromosome, see Paper 3 for more detailed information). This asymptote is regardless of whether QTLs are included among the SNPs or not.

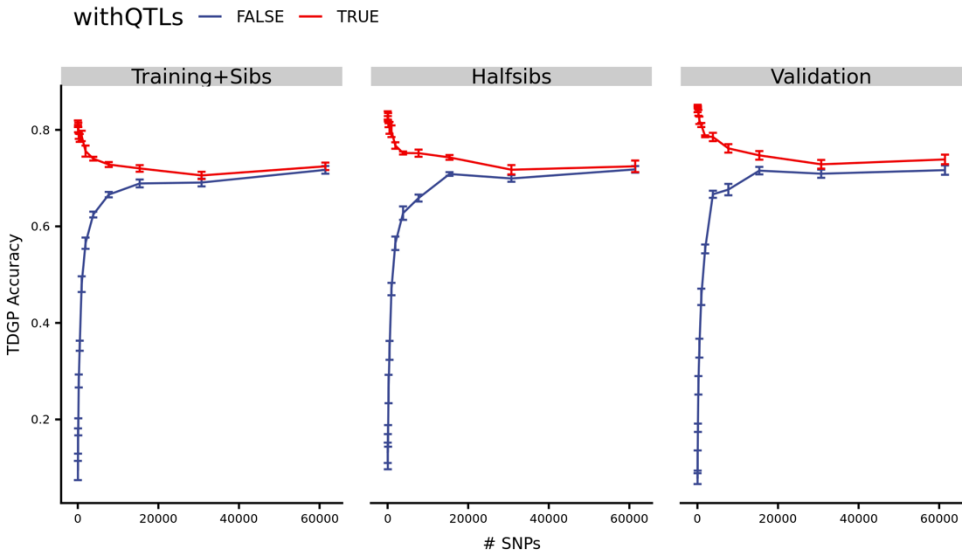


Figure 1: Mean accuracy of TDGP for SNP densities between 30 and 61,440. The red and blue lines indicate mean accuracy when QTLs are included and not included, respectively, among the SNPs used to perform the analysis. The three column-wise facets are from the same populations used in Paper 3. Briefly, mean accuracies for the individuals included in the training set and their sibs (left), mean accuracies for half-sibs of the training population (middle) and mean accuracies for a validation population distantly related with the training population (right). The trait is binary with 5% case incidence and 100 000 individuals in total. Five replications were done to calculate the mean $\pm$ SE of the accuracies for each set of SNPs. Accuracies are calculated as the correlation between predicted- and true genomic breeding value.

As shown in Figure 1, increasing the number of SNPs when the QTLs are already included on the SNP chip reduces the accuracy. Although TDGP uses LD-based information, all SNPs are *a priori* assumed to explain a fraction of the genetic variance, as with ordinary SNP-BLUP. Thus, adding more SNPs to the analysis in addition to the causative QTLs implicitly means that the actual QTLs *a priori* are assumed to explain a smaller fraction of the genetic variance. The results from Figure 1 indicate that, in the given simulation, the highest accuracy can be achieved by having just the QTLs in the dataset. Consequently, for dense markers, combining TDGP with SNP selection algorithms may increase the theoretically highest obtainable accuracy above what is achieved by using only TDGP. In addition,

increasing the number of training individuals should further increase the obtainable accuracy.

Using within- and between family information in the same model to obtain increased accuracy:

The SNP-BLUP method, through equivalence with GBLUP, deems both within- and between family information equally important, and thus both sources of information are implicitly given equal weights. As discussed in Paper 3, TDGP uses only within-family information, which is based on LD between SNPs on the SNP chip and QTLs which might or might not be on the SNP chip.

Assume, for a moment, a scenario where we create a SNP chip where none of the SNPs are in LD with any QTLs of a complex trait. In this unlikely scenario, the TDGP method would not be expected to achieve any accuracy. However, SNP-BLUP could achieve higher accuracy than the pedigree-based accuracy by (indirectly) taking the genomic relationships between individuals into account, which can be estimated using the SNPs having no LD with the QTLs. The latter point is supported by Ødegård *et al.* [91], Tsai *et al.* [92] and Correa *et al.* [93] who all achieved approximately the same or significantly better than the pedigree-accuracy when using GBLUP with 0.5K-1K markers. When so few markers are used (< 1K), much or all of the LD with QTLs may be lost.

Imagine instead a second scenario where we make a SNP chip consisting of all the true QTLs for the same trait as above, in addition to having SNPs which are not in LD with any QTLs. In this scenario, the TDGP method could, in theory, predict all QTL effects correctly. However, TDGP would still not achieve 100% accuracy as genetic variance is, *a priori*, assumed to be distributed over all SNPs, potentially giving the non-QTL SNPs some apparent effect (see Figure 1). The SNP-BLUP method uses both LD-based- and family-based information. Thus, in a scenario where all QTLs are among the SNPs on the chip, SNPs that are not in LD with any QTLs may still separate good families from bad ones. Consequently, these “family” SNP markers can achieve a marker effect estimate which accurately serves to predict good and bad families in the current data material. Because of this, SNP-BLUP can work even without having any LD with true QTLs. However, the “family” SNP effects will fail to

predict more distant families who were not part of the marker effect estimate. The "family" marker effects may be the very reason that stops SNP-BLUP from achieving 100% accuracy when the true QTLs are included on the SNP chip.

Let us assume a third, more likely, scenario where some of the SNPs are in LD with some of the QTLs, and perhaps some QTLs are even found on the SNP chip, i.e. a combination of the two preceding scenarios. In such a scenario, not all QTLs are in LD with SNPs on the chip, and some QTLs may only be partly in LD with the SNPs, thus the TDGP method will not be able to predict effects for all QTLs. Due to relatedness between individuals, the SNP-BLUP method can achieve higher accuracy than TDGP through collinearities between SNPs which explains relatedness between families. However, such false SNP effects results in poor predictions of genomic breeding value for individuals with low relatedness with the training population, as shown in Paper 3. Thus, perhaps a weighing factor for within- vs. between-family information can be used to increase the overall accuracy for both highly- and lowly related individuals. This idea comes as a result of discussions with Jørgen Ødegård. Such a weighing factor could be estimated using cross validation, where the estimate used is the one that achieves the lowest root mean squared error of prediction (RMSEP). From Paper 3:

$$\mathbf{T} = \mathbf{Q} - \frac{1}{2}(\mathbf{Z}_\mathrm{s} + \mathbf{Z}_\mathrm{d})\mathbf{Q}_\mathrm{p}$$

where $\mathbf{T}$ is the offspring genotype matrix centered by mean parental genotypes, $\mathbf{Q}$ and $\mathbf{Q}_p$ are the genotype matrices (with values 0, 1 and 2) for offspring and parents, respectively and $\mathbf{Z}_\mathrm{s}$ and $\mathbf{Z}_\mathrm{d}$ are appropriate incidence matrices connecting offspring with their sires and dams, respectively (see Paper 3). Using ordinary SNP-BLUP, $\boldsymbol{M}$ rather than $\boldsymbol{T}$ is used, where $\boldsymbol{M}$ is a genotype matrix centered by allele frequencies rather than by mean parental genotypes. The adjusted genotypes in $\boldsymbol{M}$ contains both within- and between-family information, while $\boldsymbol{T}$ only contains within-family information. Consequently:

$$\boldsymbol{B} = \boldsymbol{M} - \boldsymbol{T} = \frac{1}{2}(\mathbf{Z}_\mathrm{s} + \mathbf{Z}_\mathrm{d})\mathbf{M}_\mathrm{p}$$

where $\boldsymbol{B}$ is a matrix of genotypes adjusted so they contain only between-family information, $\mathbf{Z}_s$ and $\mathbf{Z}_d$ are the incidence matrixes linking offspring genotypes with sire and dam genotypes, respectively, and $\mathbf{M}_p$ are the parent genotypes adjusted for two times the parental allele frequencies. An ordinary SNP-BLUP model (see Introduction) can thus be written as:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Mg} + \boldsymbol{e} = \boldsymbol{X\beta} + (\boldsymbol{B} + \boldsymbol{T})\boldsymbol{g} + \boldsymbol{e}$$

From the above equation we see that from using $\boldsymbol{Mg} = (\boldsymbol{B} + \boldsymbol{T})\boldsymbol{g}$, the within- and between-family information is equally weighted. However, as questioned above, this may perhaps not produce the most accurate breeding values. To investigate this further, a model which weights the within- and between-family information can be used:

$$\boldsymbol{y} = \boldsymbol{X\beta} + (\alpha\boldsymbol{B} + (2 - \alpha)\boldsymbol{T})\boldsymbol{g} + \boldsymbol{e}$$

where α is a weighing factor between 0 and 2 for within- and between-family information. In the above model, $\alpha = 0$ means only within-family information is used (i.e. as with TDGP), $\alpha = 2$ means only between-family information is used and $\alpha = 1$ means equal weighing of within- and between-family information (i.e. as with SNP-BLUP).

Another model, suggested by Jørgen Ødegård, is:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Bg}_B + \boldsymbol{Tg}_T + \boldsymbol{e}$$

where the change from above is that separate SNP effects are predicted using between- ($\boldsymbol{g}_B$ SNP effects) and within- ($\boldsymbol{g}_T$ SNP effects) family information. The latter model allows SNPs in LD with QTLs to achieve higher relative SNP effect predictions in $\boldsymbol{g}_T$ compared with those in $\boldsymbol{g}_B$, while the opposite may be true if there is no LD between QTLs and certain SNPs. Additionally, the SNPs in LD with QTL also contribute to the between-family information through genomic relationships, and thus will necessarily achieve some positive effect prediction relative to how it explains the relatedness in the dataset.

The above models deserve further investigations.

5.4 Conclusion

In this thesis, three methods are suggested for aiding in the following fields of study: parentage assignment, triploid genotyping and estimation of heritabilities and prediction of genomic breeding values using genotypes from just cases and their parents. To the best of the knowledge of the authors of the papers, all three methods are novel. The methods from papers 1 and 3, i.e. parentage assignment and TDGP, respectively, may be applied for the benefit of almost any modern breeding program which uses high-density SNP genotypes to perform selection of breeding candidates. The triploid genotype calling method developed in Paper 2 may be used to assist breeding programs when the species allows for triploidy, such as in aquaculture.

6 References

1. Bonetta, L., *Whole-genome sequencing breaks the cost barrier.* Cell, 2010. **141**(6): p. 917-9.
2. Schwarze, K., et al., *Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature.* Genet Med, 2018. **20**(10): p. 1122-1130.
3. O'Flynn, F.M., et al., *Comparisons of cultured triploid and diploid Atlantic salmon (Salmo salar L.)* ICES J. Mar. Sci, 1997. **54**.
4. Liu, S., et al., *Development and validation of a SNP panel for parentage assignment in rainbow trout.* Aquaculture, 2016. **452**: p. 178-182.
5. Miggiano, E., et al., *AFLP and microsatellites as genetic tags to identify cultured gilthead seabream escapees: data from a simulated floating cage breaking event.* Aquaculture International, 2005. **13**(1-2): p. 137.
6. Fleming, I., et al., *An experimental study of the reproductive behaviour and success of farmed and wild Atlantic salmon (Salmo salar).* Journal of Applied Ecology, 1996: p. 893-905.
7. Karlsson, S., et al., *Widespread genetic introgression of escaped farmed Atlantic salmon in wild salmon populations.* Ices Journal of Marine Science, 2016. **73**(10): p. 2488-2498.
8. Jacobsen, J.A. and E. Gaard, *Open-ocean infestation by salmon lice (Lepeophtheirus salmonis): Comparison of wild and escaped farmed Atlantic salmon (Salmo salar L.).* Ices Journal of Marine Science, 1997. **54**(6): p. 1113-1119.
9. Jensen, O., et al., *Escapes of fishes from Norwegian sea-cage aquaculture: causes, consequences and prevention.* Aquaculture Environment Interactions, 2010. **1**(1): p. 71-83.
10. Grashei, K.E., J. Odegard, and T.H.E. Meuwissen, *Genotype calling of triploid offspring from diploid parents.* Genet Sel Evol, 2020. **52**(1): p. 15.
11. Spielman, R.S., R.E. McGinnis, and W.J. Ewens, *Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM).* Am J Hum Genet, 1993. **52**(3): p. 506-16.
12. Jones, A.G., et al., *A practical guide to methods of parentage analysis.* Molecular Ecology Resources, 2010. **10**(1): p. 6-30.
13. Chakraborty, R., M. Shaw, and W.J. Schull, *Exclusion of paternity: the current state of the art.* Am J Hum Genet, 1974. **26**(4): p. 477-88.
14. Marshall, T.C., et al., *Statistical confidence for likelihood-based paternity inference in natural populations.* Molecular Ecology, 1998. **7**(5): p. 639-655.
15. Jobling, M.A. and P. Gill, *Encoded evidence: DNA in forensic analysis.* Nature Reviews Genetics, 2004. **5**(10): p. 739-751.

16. DeWoody, J.A., D. Walker, and J.C. Avise, *Genetic parentage in large half-sib clutches: Theoretical estimates and empirical appraisals.* Genetics, 2000. **154**(4): p. 1907-1912.
17. Norris, A.T., D.G. Bradley, and E.P. Cunningham, *Parentage and relatedness determination in farmed Atlantic salmon (Salmo salar) using microsatellite markers.* Aquaculture, 2000. **182**(1-2): p. 73-83.
18. Dong, S.R., et al., *Parentage determination of Chinese shrimp (Fenneropenaeus chinensis) based on microsatellite DNA markers.* Aquaculture, 2006. **258**(1-4): p. 283-288.
19. Castro, J., et al., *A microsatellite marker tool for parentage analysis in Senegal sole (Solea senegalensis): Genotyping errors, null alleles and conformance to theoretical assumptions.* Aquaculture, 2006. **261**(4): p. 1194-1203.
20. Hauser, L., et al., *An empirical comparison of SNPs and microsatellites for parentage and kinship assignment in a wild sockeye salmon (Oncorhynchus nerka) population.* Mol Ecol Resour, 2011. **11 Suppl 1**: p. 150-61.
21. Ranade, K., et al., *High-throughput genotyping with single nucleotide polymorphisms.* Genome Research, 2001. **11**(7): p. 1262-1268.
22. Jones, B., et al., *Using blocks of linked single nucleotide polymorphisms as highly polymorphic genetic markers for parentage analysis.* Molecular Ecology Resources, 2009. **9**(2): p. 487-497.
23. Nielsen, R., et al., *Statistical approaches to paternity analysis in natural populations and applications to the North Atlantic humpback whale.* Genetics, 2001. **157**(4): p. 1673-1682.
24. Hadfield, J.D., D.S. Richardson, and T. Burke, *Towards unbiased parentage assignment: combining genetic, behavioural and spatial data in a Bayesian framework.* Molecular Ecology, 2006. **15**(12): p. 3715-3730.
25. Whalen, A., G. Gorjanc, and J.M. Hickey, *Parentage assignment with genotyping-by-sequencing data.* Journal of Animal Breeding and Genetics, 2019. **136**(2): p. 102-112.
26. Huisman, J., *Pedigree reconstruction from SNP data: parentage assignment, sibship clustering and beyond.* Molecular Ecology Resources, 2017. **17**(5): p. 1009-1024.
27. Vandeputte, M., S. Mauger, and M. Dupont-Nivet, *An evaluation of allowing for mismatches as a way to manage genotyping errors in parentage assignment by exclusion.* Molecular Ecology Notes, 2006. **6**(1): p. 265-267.
28. Buckleton, J. and J. Curran, *A discussion of the merits of random man not excluded and likelihood ratios.* Forensic Science International: Genetics, 2008. **2**(4): p. 343-348.

29. Hayes, B.J., *Efficient parentage assignment and pedigree reconstruction with dense single nucleotide polymorphism data.* J Dairy Sci, 2011. **94**(4): p. 2114-7.
30. Boerner, V., *On marker-based parentage verification via non-linear optimization.* Genetics Selection Evolution, 2017. **49**.
31. Chester-Jones, I., P.M. Ingelton, and J.G. Phillips, *Fundamentals of comparative vertebrate endocrinology*. 1987, New York: Plenum Press. xvi, 666 p.
32. Piferrer, F., et al., *Polyploid fish and shellfish: Production, biology and applications to aquaculture for performance improvement and genetic containment.* Aquaculture, 2009. **293**(3-4): p. 125-156.
33. Cotter, D., et al., *An evaluation of the use of triploid Atlantic salmon (Salmo salar L.) in minimising the impact of escaped farmed salmon on wild populations.* Aquaculture, 2000. **186**(1-2): p. 61-75.
34. Fjelldal, P.G., et al., *Triploid (sterile) farmed Atlantic salmon males attempt to spawn with wild females.* Aquaculture Environment Interactions, 2014. **5**(2): p. 155-162.
35. Everson, J.L., et al., *Polyploidy affects fillet yield, composition, and fatty acid profile in two-year old, female rainbow trout, Oncorhynchus mykiss.* Aquaculture, 2020: p. 735873.
36. Weber, G.M., et al., *Growth performance comparison of intercross-triploid, induced triploid, and diploid rainbow trout.* Aquaculture, 2014. **433**: p. 85-93.
37. Herath, T.K., et al., *Impact of Salmonid alphavirus infection in diploid and triploid Atlantic salmon (Salmo salar L.) fry.* Plos One, 2017. **12**(9).
38. Taylor, J.F., et al., *Ploidy and family effects on Atlantic salmon (Salmo salar) growth, deformity and harvest quality during a full commercial production cycle.* Aquaculture, 2013. **410**: p. 41-50.
39. Bonnet, S., et al., *Genetic variation in growth parameters until commercial size in diploid and triploid freshwater rainbow trout (Oncorhynchus mykiss) and seawater brown trout (Salmo trutta).* Aquaculture, 1999. **173**(1-4): p. 359-375.
40. Friars, G.W., et al., *Family differences in relative growth of diploid and triploid Atlantic salmon (Salmo salar L.).* Aquaculture, 2001. **192**(1): p. 23-29.
41. Johnson, R.M., et al., *Dosage effects on heritability and maternal effects in diploid and triploid Chinook salmon (Oncorhynchus tshawytscha).* Heredity, 2007. **98**(5): p. 303-310.
42. Shrimpton, J.M., et al., *Effect of triploidy on growth and ionoregulatory performance in ocean-type Chinook salmon: A quantitative genetics approach.* Aquaculture, 2012. **362**: p. 248-254.

43. Vera, L.M., et al., *Early nutritional programming affects liver transcriptome in diploid and triploid Atlantic salmon, Salmo salar.* Bmc Genomics, 2017. **18**.
44. Vera, L.M., et al., *Enhanced micronutrient supplementation in low marine diets reduced vertebral malformation in diploid and triploid Atlantic salmon (Salmo salar) parr, and increased vertebral expression of bone biomarker genes in diploids.* Comparative Biochemistry and Physiology B-Biochemistry & Molecular Biology, 2019. **237**.
45. Caron, N., et al., *Mass Selection for 45-Day Body-Weight in Japanese-Quail - Selection Response, Carcass Composition, Cooking Properties, and Sensory Characteristics.* Poultry Science, 1990. **69**(7): p. 1037-1045.
46. Lukefahr, S.D., H.B. Odi, and J.K.A. Atakora, *Mass selection for 70-day body weight in rabbits.* Journal of Animal Science, 1996. **74**(7): p. 1481-1489.
47. Collins, W.M., H. Abplanalp, and W.G. Hill, *Mass Selection for Body Weight in Quail.* Poultry Science, 1970. **49**(4): p. 926-+.
48. Henderson, C.R., *Estimation of Variance and Covariance Components.* Biometrics, 1953. **9**(2): p. 226-252.
49. Henderson, C.R., et al., *The Estimation of Environmental and Genetic Trends from Records Subject to Culling.* Biometrics, 1959. **15**(2): p. 192-218.
50. Searle, S.R., *Henderson,C.R., the Statistician - and His Contributions to Variance-Components Estimation.* Journal of Dairy Science, 1991. **74**(11): p. 4035-4044.
51. Crump, S.L., *The Estimation of Variance Components in Analysis of Variance.* Biometrics Bulletin, 1946. **2**(1): p. 7-11.
52. Crump, S.L., *The Present Status of Variance Component Analysis.* Biometrics, 1951. **7**(1): p. 1-16.
53. Eisenhart, C., *The Assumptions Underlying the Analysis of Variance.* Biometrics, 1947. **3**(1): p. 1-21.
54. Gjedrem, T. and M. Baranski, *Selective Breeding in Aquaculture: An Introduction.*
55. Houston, R.D., et al., *Major quantitative trait loci affect resistance to infectious pancreatic necrosis in Atlantic salmon (Salmo salar).* Genetics, 2008. **178**(2): p. 1109-1115.
56. Moen, T., et al., *Confirmation and fine-mapping of a major QTL for resistance to infectious pancreatic necrosis in Atlantic salmon (Salmo salar): population-level associations between markers and trait.* Bmc Genomics, 2009. **10**.
57. Yu, J.M., et al., *A unified mixed-model method for association mapping that accounts for multiple levels of relatedness.* Nature Genetics, 2006. **38**(2): p. 203-208.

58. Moen, T., *Methods for detecting DNA polymorphisms in Atlantic salmon*. 2018, Google Patents.
59. Santi, N., T. Moen, and J. Ødegård, *METHOD FOR PREDICTING INCREASED RESISTANCE OF A RAINBOW TROUT TO INFECTIOUS PANCREATIC NECROSIS (IPN)*, in *European Patent Office*, E.P. Office, Editor. 2015.
60. Sigbjørn, L., M. SODELAND, and T. Moen, *Predicting the ability of atlantic salmon to utilise dietary pigment based on the determination of polymorphisms*. 2015, Google Patents.
61. Cobb, J.N., P.S. Biswas, and J.D. Platten, *Back to the future: revisiting MAS as a tool for modern plant breeding.* Theoretical and Applied Genetics, 2019. **132**(3): p. 647-667.
62. Habier, D., R.L. Fernando, and J.C. Dekkers, *The impact of genetic relationship information on genome-assisted breeding values.* Genetics, 2007. **177**(4): p. 2389-97.
63. Bangera, R., et al., *Genomic predictions can accelerate selection for resistance against Piscirickettsia salmonis in Atlantic salmon (Salmo salar).* Bmc Genomics, 2017. **18**.
64. Daetwyler, H.D., et al., *Accuracy of pedigree and genomic predictions of carcass and novel meat quality traits in multi-breed sheep data assessed by cross-validation.* Genetics Selection Evolution, 2012. **44**.
65. Nadaf, J. and R. Pong-Wong, *Applying different genomic evaluation approaches on QTLMAS2010 dataset.* BMC Proc, 2011. **5 Suppl 3**: p. S9.
66. VanRaden, P.M., *Efficient Methods to Compute Genomic Predictions.* Journal of Dairy Science, 2008. **91**(11): p. 4414-4423.
67. Meuwissen, T.H.E., B.J. Hayes, and M.E. Goddard, *Prediction of total genetic value using genome-wide dense marker maps.* Genetics, 2001. **157**(4): p. 1819-1829.
68. Goddard, M., *Genomic selection: prediction of accuracy and maximisation of long term response.* Genetica, 2009. **136**(2): p. 245-257.
69. Stranden, I. and D.J. Garrick, *Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit.* J Dairy Sci, 2009. **92**(6): p. 2971-5.
70. Habier, D., et al., *Extension of the Bayesian alphabet for genomic selection.* BMC Bioinformatics, 2011. **12**.
71. Erbe, M., et al., *Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels.* Journal of Dairy Science, 2012. **95**(7): p. 4114-4129.
72. Usai, M.G., M.E. Goddard, and B.J. Hayes, *LASSO with cross-validation for genomic selection.* Genetics Research, 2009. **91**(6): p. 427-436.

73. Ogutu, J.O., T. Schulz-Streeck, and H.-P. Piepho. *Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions*. in *BMC proceedings*. 2012. Springer.
74. Tsai, H.Y., et al., *Genomic prediction and GWAS of yield, quality and disease-related traits in spring barley and winter wheat.* Scientific Reports, 2020. **10**(1).
75. Aslam, M.L., et al., *Genome-wide association mapping and accuracy of predictions for amoebic gill disease in Atlantic salmon (Salmo salar).* Scientific Reports, 2020. **10**(1).
76. Muqaddasi, Q.H., et al., *Prospects of GWAS and predictive breeding for European winter wheat's grain protein content, grain starch content, and grain hardness.* Scientific Reports, 2020. **10**(1).
77. Voorrips, R.E., G. Gort, and B. Vosman, *Genotype calling in tetraploid species from bi-allelic marker data using mixture models.* BMC Bioinformatics, 2011. **12**: p. 172.
78. Voorrips, R.E. and G. Gort. *fitPoly: genotype calling for bi-allelic marker assays*. 2020 [cited 2020 13 Mar]; Version 3.0.0]. Available from: https://github.com/cran/fitPoly.
79. Scrucca, L., et al., *mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models.* R J, 2016. **8**(1): p. 289-317.
80. Schwarz, G., *Estimating the dimension of a model.* The annals of statistics, 1978. **6**(2): p. 461-464.
81. Biernacki, C., G. Celeux, and G. Govaert, *Assessing a mixture model for clustering with the integrated completed likelihood.* IEEE transactions on pattern analysis and machine intelligence, 2000. **22**(7): p. 719-725.
82. Barria, A., et al., *Genomic Predictions and Genome-Wide Association Study of Resistance Against Piscirickettsia salmonis in Coho Salmon (Oncorhynchus kisutch) Using ddRAD Sequencing.* G3-Genes Genomes Genetics, 2018. **8**(4): p. 1183-1194.
83. Odegard, J., et al., *Evaluation of statistical models for genetic analysis of challenge-test data on ISA resistance in Atlantic salmon (Salmo salar): Prediction of progeny survival.* Aquaculture, 2007. **266**(1-4): p. 70-76.
84. Jones, O.R. and J.L. Wang, *COLONY: a program for parentage and sibship inference from multilocus genotype data.* Molecular Ecology Resources, 2010. **10**(3): p. 551-555.
85. Dempster, A.P., N.M. Laird, and D.B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm.* Journal of the Royal Statistical Society: Series B (Methodological), 1977. **39**(1): p. 1-22.

86. Kjøglum, S., et al. *MULTIVARIATE GENOMIC MODEL FOR DIPLOID AND TRIPLOID GROWTH PERFORMACE IN ATLANTIC SALMON (Salmo salar)*. 2019 [cited 2020 2nd Sep. 2020]; Poster presentation at AquaCulture Europe 2019]. Available from: https://www.aquaeas.eu/uncategorised/402-welcome-to-aquaculture-europe-2019.
87. Moen, T., et al., *Epithelial Cadherin Determines Resistance to Infectious Pancreatic Necrosis Virus in Atlantic Salmon.* Genetics, 2015. **200**(4): p. 1313-+.
88. Lien, S., et al., *A dense SNP-based linkage map for Atlantic salmon (Salmo salar) reveals extended chromosome homeologies and striking differences in sex-specific recombination patterns.* BMC Genomics, 2011. **12**: p. 615.
89. Marchini, J. and B. Howie, *Genotype imputation for genome-wide association studies.* Nature Reviews Genetics, 2010. **11**(7): p. 499-511.
90. Falconer, D. and T. Mackay, *Introduction to quantitative genetics. 1996.* Harlow, Essex, UK: Longmans Green, 1996. **3**.
91. Odegard, J., et al., *Genomic prediction in an admixed population of Atlantic salmon (Salmo salar).* Frontiers in Genetics, 2014. **5**.
92. Tsai, H.Y., et al., *Genome wide association and genomic prediction for growth traits in juvenile farmed Atlantic salmon using a high density SNP array.* Bmc Genomics, 2015. **16**.
93. Correa, K., et al., *The use of genomic information increases the accuracy of breeding value predictions for sea louse (Caligus rogercresseyi) resistance in Atlantic salmon (Salmo salar).* Genetics Selection Evolution, 2017. **49**.

7 Papers

I Using genomic relationship likelihood for parentage assignment

Grashei, K.E., Ødegård, J. & Meuwissen, T.H.E. Genet Sel Evol 50, 26 (2018).
https://doi.org/10.1186/s12711-018-0397-7

Grashei et al. Genet Sel Evol (2018) 50:26
https://doi.org/10.1186/s12711-018-0397-7

GSE Genetics Selection Evolution

RESEARCH ARTICLE

Open Access

Using genomic relationship likelihood for parentage assignment

Kim E. Grashei[1,2*], Jørgen Ødegård[1,2] and Theo H. E. Meuwissen[2]

Abstract

Background: Parentage assignment is usually based on a limited number of unlinked, independent genomic markers (microsatellites, low-density single nucleotide polymorphisms (SNPs), etc.). Classical methods for parentage assignment are exclusion-based (i.e. based on loci that violate Mendelian inheritance) or likelihood-based, assuming independent inheritance of loci. For true parent–offspring relations, genotyping errors cause apparent violations of Mendelian inheritance. Thus, the maximum proportion of such violations must be determined, which is complicated by variable call- and genotype error rates among loci and individuals. Recently, genotyping using high-density SNP chips has become available at lower cost and is increasingly used in genetics research and breeding programs. However, dense SNPs are not independently inherited, violating the assumptions of the likelihood-based methods. Hence, parentage assignment usually assumes a maximum proportion of exclusions, or applies likelihood-based methods on a smaller subset of independent markers. Our aim was to develop a fast and accurate trio parentage assignment method for dense SNP data without prior genotyping error- or call rate knowledge among loci and individuals. This genomic relationship likelihood (GRL) method infers parentage by using genomic relationships, which are typically used in genomic prediction models.

Results: Using 50 simulated datasets with 53,427 to 55,517 SNPs, genotyping error rates of 1–3% and call rates of ~ 80 to 98%, GRL was found to be fast and highly (~ 99%) accurate for parentage assignment. An iterative approach was developed for training using the evaluation data, giving similar accuracy. For comparison, we used the Colony2 software that assigns parentage and sibship simultaneously to increase the power of the likelihood-based method and found that it has considerably lower accuracy than GRL. We also compared GRL with an exclusion-based method in which one of the parameters was estimated using GRL assignments.This method was slightly more accurate than GRL.

Conclusions: We show that GRL is a fast and accurate method of parentage assignment that can use dense, non-independent SNPs, with variable call rates and unknown genotyping error rates. By offering an alternative way of assigning parents, GRL is also suitable for estimating the expected proportion of inconsistent parent–offspring genotypes for exclusion-based models.

Background

In the field of animal genetics, low-density single nucleotide polymorphisms (SNPs), microsatellites, and amplified fragment length polymorphisms (AFLP) have long been the preferred types of genomic data for parentage assignment due to their low cost [1–3]. In practice, the foundation of parentage assignment rests on exclusion- and likelihood-based methods [4]. Exclusion-based methods rely on their ability to exclude false parent–offspring combinations when the offspring's candidate parents' genotypes violate Mendel's laws. These methods are often used due to their ease of interpretation, but the number of expected exclusions depends on allele frequencies in the population and on genotype call rates and error rates [5]. Exclusion-based methods also require more loci than likelihood-based methods since only genotypes with Mendelian inconsistencies are used [6]. Likelihood-based methods often calculate

*Correspondence: kim.erik.grashei@aquagen.no; kim.grashei@nmbu.no
[1] AquaGen AS, P.O. Box 1240, NO-7462 Trondheim, Norway
Full list of author information is available at the end of the article

the likelihood ratio (LR) of the genotype of the offspring, which is the probability of the offspring's genotype given the genotypes of the candidate parents, relative to the probability of observing the genotype in the population by chance. The LR statistic effectively gives more weight to rare alleles. Different loci are typically assumed independent, such that total LR is multiplied over all loci. Likelihood-based methods have higher power than exclusion-based methods, but their interpretation is more complicated. Both likelihood- and exclusion-based models usually assume known and homogenous genotype error rates and independent loci, and do not account for variation in genotype call rates [5, 7, 8], which are all important assumptions when working with high-density SNP data. For dense SNP chip data, the assumption of independent inheritance among loci is not realistic (i.e., alleles are inherited on large DNA segments), which may lead to inflated LR values when using conventional likelihood-based methods.

Parentage can also be assigned and tested by using realized genomic relationships. The interrelationship between parents governs the expected inbreeding in offspring, as well as parent–offspring relationships. Realized genomic relationships assess the average genomic similarity across loci and do not assume independence of the loci. Increasing the number of markers in the calculations, increases the precision of the genomic relationships. Our aim was to study whether genomic relationships can be used to perform computationally fast and accurate parentage testing with high-density SNP data.

Methods

Residual genomic relationships

Estimates of genomic relationships require large numbers of loci [9], and their expectation is proportional to the genetic covariance between individuals. The proposed method for parentage testing is developed for trio parentage testing, i.e. using a single offspring and two parental candidates. The method uses genomic relationships estimated by VanRaden's first method [10], in which the genomic relationship between two individuals is calculated as follows:

$$r_{ij} = \frac{\sum_{t=1}^{c} (m_{it} - 2p_t)(m_{jt} - 2p_t)}{2\sum_{t=1}^{c} p_t(1 - p_t)}, \quad (1)$$

where r_{ij} is the genomic relationship between individuals i and j, m_{it} and m_{jt} are the genotypes (coded 0, 1 or 2 for the alternative homozygous, the heterozygous, and the homozygous reference genotypes, respectively) for individuals i and j at locus t, p_t is the allele frequency in the population at locus t, and c is the number of loci (i.e. SNPs). Genomic relationships can be calculated even for extremely dense genomic data (even up to full sequence), and do not assume independence of the loci. Figure 1 shows the relationships in a trio consisting of an offspring and two (candidate) parents.

We used Eq. (1) to estimate the genomic interrelationships between parents and offspring, i.e., the relationship of the offspring with itself ($r_{O,O}$), relationships of the two parent candidates with themselves (r_{P_1,P_1} and r_{P_2,P_2}), relationships of the offspring with both parent candidates (r_{O,P_1} and r_{O,P_2}), and relationships between the parent candidates (r_{P_1,P_2}), see Fig. 1.

Expected genomic relationships of an offspring with its true parents (TP) are [11]:

$$E(r_{O,P1}|TP) = 0.5(r_{P1,P1} + r_{P1,P2}),$$

$$E(r_{O,P2}|TP) = 0.5(r_{P2,P2} + r_{P1,P2}).$$

In other words, the relationship of an offspring with a parent is the average of the genomic relationship of the parent with itself and the relationship between the two parents. The expected relationship of the offspring with itself is [12]:

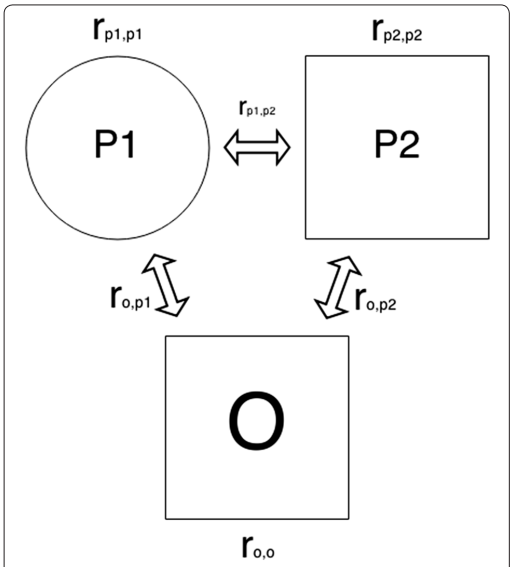


Fig. 1 A trio of offspring (O), first parent (P1) and second parent (P2). The variables near the arrows indicate genetic relationships between individuals, while the variables over P1 and P2, and below O, are the individuals' genetic relationships to themselves, respectively. Sexes are included in the figure but are not used by the GRL method

$$E(r_{O,O}|TP) = 1 + 0.5r_{P1,P2},$$

where $0.5r_{P1,P2}$ is the expected inbreeding coefficient of the offspring. Three residual relationships are defined as differences between actual and expected genomic relationships:

$$e_{O,P1} = r_{O,P1} - E(r_{O,P1}|TP),$$

$$e_{O,P2} = r_{O,P2} - E(r_{O,P2}|TP),$$

$$e_{O,O} = r_{O,O} - E(r_{O,O}|TP).$$

Inbreeding is accounted for when using the above residuals, as well as the direction of the relationships. For example, using the offspring as a candidate parent, and/or using a true parent as the offspring, will result in large residuals, i.e., realized relationships that deviate substantially from the expectations of a true parent–offspring trio.

Genomic relationship likelihood (GRL)

The above residual relationships are used to calculate a genomic relationship log-likelihood using a multivariate normal density function, assuming:

$$\mathbf{e} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where $\mathbf{e} = \begin{bmatrix} e_{O,P1} \\ e_{O,P2} \\ e_{O,O} \end{bmatrix}$ and $\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}$ is a vector of the overall means for the residuals for true parent–offspring trios. In the absence of genotyping errors, the residuals are expected to be approximately normally distributed around zero ($\mathbf{e} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, see [Additional file 1: Figure S1]. The central limit theorem states that the sum of many independently and identically distributed variates will be approximately normally distributed. The variates in Eq. (1) may be considered as originating from a common (albeit unknown) distribution, but not all are independent (i.e., the effective number of loci is lower than the actual number of loci). Still, given a substantial number of loci distributed over the entire genome (i.e., most of the loci are indeed independent), genomic relationships (summed over all variates) are still likely to approach a normal distribution (see [13], Theorem 27.4). Plotting the residual relationships for true parent–offspring trios revealed that they were approximately normally distributed [see Additional file 1: Figures S1, S2 and S3].

Since genotyping errors can occur in real data (and the expected residual relationship may thus deviate from 0), parameters of the distribution of residual relationships were estimated using an iterative method (see Section "Estimation of model parameters" below). Matrix $\boldsymbol{\Sigma}$ is the 3×3 (co)variance matrix of the three residual variates in true parent–offspring trios and was also estimated using the iterative method. The genomic relationship likelihood (GRL) was defined as:

$$\text{GRL} = -\frac{1}{2}(\mathbf{e} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{e} - \boldsymbol{\mu}),$$

which is proportional to the natural logarithm of a multivariate normal density function. Based on (iteratively assigned) parent–offspring trios, a threshold for acceptable GRL values can be defined. In this study, we assumed that a parent–offspring trio had to have a GRL value that was within the highest 99% of the known parent–offspring GRL values, thus accepting a false negative rate of 1%.

Difference between the top two trios (ΔGRL)

To reduce the false positive rate and increase the true negative rate, the value of ΔGRL was also assessed based on:

$$\Delta\text{GRL} = \text{GRL}_1 - \text{GRL}_2,$$

where GRL_1 (GRL_2) is the (second) highest GRL value achieved for an offspring across all candidate parent–offspring trios. This is analogous to the Δ statistic used in Marshall et al. [7], with more details in Appendix 1.

In datasets where both parents of an offspring are present and no other relatives are available, ΔGRL will typically be very high, since no other realistic trio exists. When other close relatives of the offspring are included among the candidate parents, ΔGRL may be lower due the potential existence of multiple "likely" false parent candidates, e.g. uncles, aunts, grandparents, siblings or descendants of the offspring. High relatedness to the offspring alone is not sufficient to obtain a high value for GRL_2 since the method accounts for interrelationships of the whole trio. For example, if the parent candidates consist of one true parent and one full-sib of the offspring, interrelationships of the trio will typically be inconsistent because of the high relationship between the two parental candidates, although the relationships of the offspring with itself and with the parent candidates may be "normal" (these should be elevated if the relationship among the two parent candidates is high). In cases where a parent is missing but many other close relatives of the offspring are present, GRL_1 can, in rare cases, exceed the threshold for GRL_1-values, but then ΔGRL will typically be low, since multiple highly-related candidate parents are present. Thus, thresholds for assignment must be set for both GRL_1 and ΔGRL.

Estimation of model parameters

Estimation of the GRL-parameters, i.e. $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and the GRL threshold, is undertaken with an iterative method which is briefly described below. The ΔGRL threshold was set to 6.9, which implies that the best parent pair should be at least 1000 ($=e^{6.9}$) times more likely than the second-best parent pair. See Section 2 in Additional file 2: for more details.

Step 1: allele dropping

Random matings between individuals from the dataset are performed in silico to produce simulated offspring. For simplicity, all loci are assumed to be inherited independently. The simulated trios are then used to obtain initial estimates of the GRL parameters. A smaller subset of the loci may be used in this step.

Step 2: assignment iteration

Trios are initially assigned using the GRL method based on the parameters estimated in Step 1. The method relies on the presence of true trios (albeit unknown) in the data. Parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are then re-estimated using the newly assigned trios from evaluation data, and then used as the basis of the next assignment iteration. Iteration stops when the number of assignments is smaller than in the previous iteration. Thus, the GRL training procedure iteratively assigns trios while (re-)estimating the GRL-parameters until no more trios can be assigned. See Section 1 in Additional file 2: for more information about the training procedure.The parameter estimates obtained in the second-to-last iteration are considered optimal. To limit the number of plausible trios to test, only individuals with a relationship larger than 0.25 with an offspring were considered as potential parents, i.e. $r_{O,P1} > 0.25$ and $r_{O,P2} > 0.25$. The GRL threshold is not re-estimated in this step.

When pre-defined parameter estimates are used, the assignment process starts without estimating parameters. This is equivalent to running only the second-to-last iteration of Step 2.

Simulation study

A simulation study was conducted to investigate the strengths and weaknesses of the GRL method. QMSim [14] was used to produce simulated datasets. The initial size of the historical population was set to 500 and remained constant for 5000 generations to achieve mutation/drift equilibrium. In generation 5001, the population size was reduced to 300, of which 100 were males and 200 were females. Twenty chromosomes were simulated, each 1 Morgan long, and the number of SNPs was set such that approximately 54,000 SNPs (53,427 to 55,517) with a minor allele frequency higher than 0.05 existed in the population. The SNP mutation rate was set to 0.00003, assuming a recurrent mutation model (i.e. only two possible alleles exist). After the historical population, a recent population was simulated over five generations, with 1000 individuals per generation (5000 individuals in total). These were produced by random mating of 100 sires and 200 dams per generation, with one sire mated with two dams and each mating resulting in five recorded offspring. Of these, the last two generations were used in the parentage assignment tests. Fifty repetitions of the QMSim simulations were performed to produce 50 datasets. Genotype errors (1 and 3%) and call rates (80–100%) were added using a custom script written in the Python programming language, allowing both erroneous and missing genotypes among individuals, see Section 2 in Additional file 2: for more information.

The GRL method was programmed in the C++ programming language that emphasizes parallel processing. The program was run in a Linux cluster environment using multiple CPU. Tests were run using the training procedure on all (evaluation) datasets. In addition, pre-estimated parameters were obtained from some of the runs with training. The datasets were not divided into offspring and parents, and thus all true offspring and parents had the potential to be assigned parents both correctly (offspring only) and incorrectly (parents and offspring).

There are three possible outcomes of the assignment process: (1) 'Correct', meaning correct assignment of true parents to the unknown offspring (parents must be present), (2) 'Incorrect', meaning wrong candidate parents were assigned and (3) 'No-assign', meaning no assignment was made. These were quantified for each analysis.

Comparison with a conventional likelihood-based method

To compare GRL with other methods, we analyzed five of the simulated datasets, arbitrarily chosen from all 50 datasets, by using the Colony2 software V2.0.6.3 [15]. Colony2 uses a likelihood-based method that jointly assigns both sibship and parentage based on a simulated annealing process [16, 17]. This increases the assignment power compared to methods that use a single unknown individual (the offspring) and one or two candidate parents. Colony2 was run using a 1% genotype error (true and assumed). In addition, the following settings were chosen: (1) do not update allele frequencies, (2) assume no inbreeding, (3) no sibship scaling, (4) no sibship prior, (5) short run length, (6) use the pairwise likelihood score (PLS) and (7) allelic dropout rate set to zero for all markers. The 'ParentPairs'-file produced by Colony2 was used to check accuracy of assignments. Any assignments for

which mother, father or both were missing, or for which the assignment probability reported by Colony2 was less than 0.5, were categorized as a "No-assign". Suggested parent pairs with at least one incorrect parent were categorized as "Incorrect" assignments and pairs with both parent candidates correct were categorized as "Correct" assignments.

Comparison with an exclusion-based method: the binomial exclusion method

We developed an exclusion-based method in which one of the parameters was estimated using GRL-assigned trios using custom scripts written in the R programming language. Exclusion ratios (ER) for the GRL-assigned trios were calculated as the ratio of the number of exclusions for a trio and the number of loci for which all three individuals in the trio had called genotypes. We used a binomial distribution as a basis for the new assignments, i.e. $E \sim Bin(n, p)$, where E is the number of trio exclusions, n (number of trials) is the number of calls for the trio, and p (success probability) is the median ER from the GRL assigned trios.

To limit the number of trios for binomial exclusion assignment, we used the same parent–offspring genomic relationship threshold that we used for the GRL assignments, i.e. $r_{O,P1} > 0.25$ and $r_{O,P2} > 0.25$. Assignment was done in a similar manner as with GRL, using both a confidence cutoff and a Δ-score. For more information, see Section 3 in Additional file 2:. We refer to this method as the binomial exclusion method (BEM) in the text.

Results

Assignment results using Colony2 are shown in Fig. 2, and the analogous GRL- and BEM results are shown in Figs. 3 and 4. The most noticeable differences in results between GRL- and BEM are shown in Figs. 5 and 6. Here, both methods used training estimates from a dataset with a 3% genotype error, while the true error was 1%. Results that were similar between GRL and BEM are shown in Figures S4, S5, S6, S7, S8, S9, S10 and S11 [see Additional file 3: Figures S4, S5, S6, S7, S8, S9, S10 and S11]. In Figures S4 (GRL) and S5 (BEM), parameters were pre-estimated at a 3% genotype error (true and assumed). Figures S6 (GRL) and S7 (BEM) show the results for a true error of 3% and an assumed error of 1%. Figures S8 (GRL) and S9 (BEM) show the results for training with a 1% error rate, and Figures S10 (GRL) and S11 (BEM) for training with a 3% error rate. Total results over all datasets are shown in Table S1 [see Additional file 4: Table S1].

The Colony2 software was tested using a 1% true genotype error rate (assumed and true). When parents are available, Colony2 had a correct assignment rate of 22.4%, a no-assign rate of 75.4% and an incorrect assignment rate of 2.2%. For individuals without parents, the incorrect assignment rate climbed to 14.7% and the (correct) no-assign rate climbs to 85.3% (see Fig. 2).

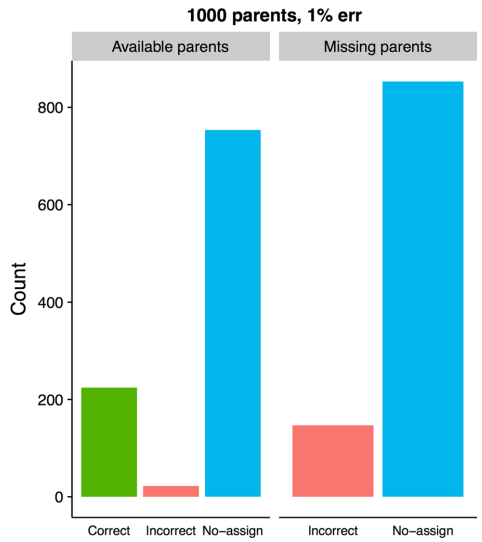


Fig. 2 Assignment results from Colony2 for individuals with (left panel) and without (right panel) available parents in the dataset. Results from five simulated datasets are averaged. The true and assumed genotype error rate was 1% for all datasets

Figures 3 and 4 show the comparison between GRL and BEM when parameter estimates from an arbitrarily chosen dataset were used. When parents were available in the dataset and the genotype error rate (true and assumed) was 1%, using GRL resulted in 99.5% of the individuals being correctly assigned both parents (Fig. 3), while 99.9% were assigned correctly with the (GRL-trained) BEM (Fig. 4). In both cases, no individuals with parents in the dataset were assigned incorrect parent pairs. When parents were not available, the incorrect assignment rate for GRL climbed to 0.01% for both 1% and 3% genotype error rates (Fig. 3 and Additional file 3: Figure S4).

The most notable difference in results between GRL and BEM was found for a true genotype error rate of 1% when parameter estimates were from a dataset with a 3% error rate (Figs. 5 and 6). Here, GRL did not assign any trios. However, BEM assigned all trios correctly when parents were available, but incorrectly assigned 1.0% of the trios when parents were not available. When the true and assumed genotype error rates were reversed (i.e. a true error rate of 3% and an incorrectly assumed error rate of 1%), neither method assigned any

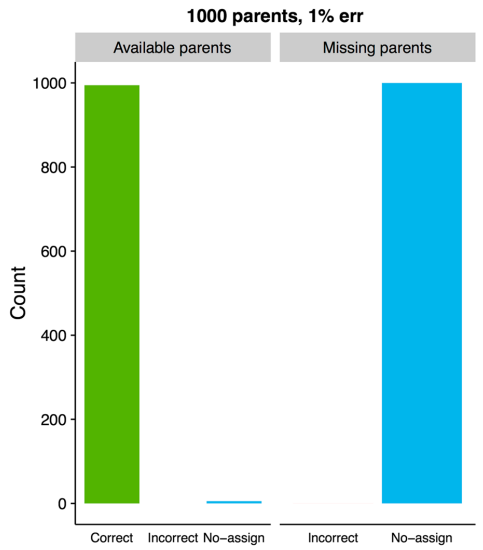


Fig. 3 Assignment results using GRL at a 1% genotype error rate (true and assumed) for individuals with (left panel) and without (right panel) available parents in the dataset. Results from 50 simulated datasets are averaged. Parameters were pre-estimated using one arbitrarily chosen dataset with a 1% genotype error

trios, while the GRL method incorrectly assigned 0.02% trios, both when parents were available and when they were missing [see Additional file 3: Figures S6 and S7] and [see Additional file 4: Table S1].

An alternative to assuming a set of predefined parameters is to estimate these by using the evaluation data directly. Averaged results for each dataset are shown in Figures S8 and S9 [see Additional file 3: Figures S8 and S9] (1% genotype error) and in Figures S10 and S11 [see Additional file 3: Figures S10 and S11] (3% genotype error). These results are very similar to the results shown in Figs. 3 and 4 (1% true and assumed error rates), and Figures S4 and S5 [see Additional file 3: Figures S4 and S5] (3% true and assumed error rates).

Discussion

Parentage assignment is mostly performed using likelihood-based models with microsatellites [2, 7], low-density SNPs [1] or exclusion-based models [18]. However, assignments methods often impose idealized assumptions, such as known age, generation and gender of all individuals, a limited number of known parental candidates, independent markers, little or no inbreeding, no stratification of the population or sample, no biased sampling of individuals, Hardy–Weinberg equilibrium (HWE) and little or no variation in genotype error or call rates within and between samples. For GRL and BEM, we perfomed assignments with unknown age, generation and gender, with no assumption as to independence of markers, HWE, inbreeding, family size or family composition, and with dense (SNP) markers, closely related individuals and varying genotype error and call rate. Colony2 assumes HWE, independent markers and no inbreeding.

GRL

Residual relationships were approximately normally distributed even when genotype errors were present [see Additional file 1: Figures S2 and S3], but with different expectations compared to genomic data without genotype errors [see Additional file 1: Figure S1].

It did not appear to be a problem that the parent and offspring generations were unknown when using GRL and BEM. High accuracies were achieved, although individuals had numerous close relatives that were eligible as parent candidates, such as the true parents, full- and half-sibs, own offspring, uncles/aunts and nieces/nephews. Similar results were obtained when the genotype error was increased to 3%, which was used to show that the GRL and BEM work even when the genotype error rate has quite extreme values. These properties may be useful for populations with large sibling groups, such as in fish, poultry or pigs, when generations cannot be clearly differentiated, or when the genotype error or call rates vary a lot.

A strength of the GRL training procedure is that no reference dataset with known pedigree is required for training and that the training is only partly done by simulation (allele-dropping). As long as there is a sufficient number of true (but unknown) trios present for assignment, the training can proceed. The method requires a pre-defined ΔGRL threshold (i.e. the minimum acceptable value). The ΔGRL is (the log of) the odds for correct assignment, given that the correct trio is among the two best trios (this is nearly always the case if true parents are present). In this study, the threshold was set to 6.9, i.e., the best trio should be at least $e^{6.9} = 1000$ times more likely than the second-best trio. Relaxing this assumption will increase both the true and false positive assignment rates of the model, while setting a stricter threshold will have the opposite effect.

In some cases, the iterative training method may fail because the initial iteration results in no assignments. This may be caused by two factors: (1) the number of loci used in the allele-dropping simulation step may be set too high (giving too idealized parent–offspring relationships compared with evaluation data), or (2) there are no

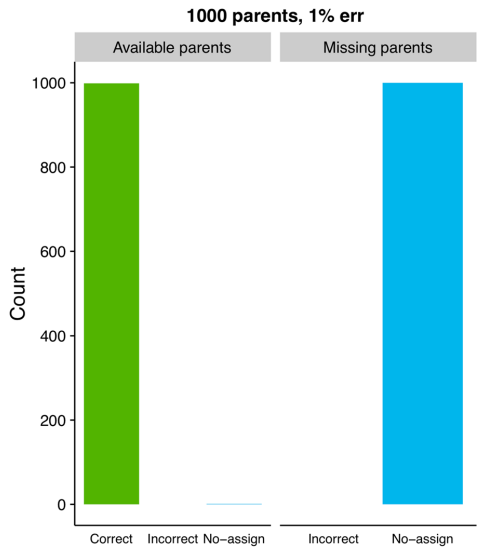


Fig. 4 Assignment results using BEM at a 1% genotype error rate (true and assumed) for individuals with (left panel) and without (right panel) available parents in the dataset. Results from 50 simulated datasets are averaged. Parameters were pre-estimated using the GRL-assignments from one arbitrarily chosen dataset with a 1% genotype error

true trios present in the evaluation dataset. If reducing the number of SNPs used in the allele-dropping step does not start the iteration process, the latter may be the case. During training, there is no need to estimate or assume a genotype error rate with the GRL method, as long as the training procedure is done using the evaluation dataset.

Exclusion using parent–offspring duos (i.e. offspring and a single candidate parent) or trios is a relatively simple method for parentage assignment, by identifying incorrect parents by genotypes that violate the laws of Mendelian inheritance ("exclusion genotypes"). The GRL method is a fundamentally different approach and can be used to estimate exclusion-based parameters in true parent–offspring trios (assigned by GRL). Assignment of a single parent to an offspring is also possible using a similar method as for trios, but this was not explored in this study. The training-based GRL has the advantage that it requires no prior assumption with respect to genotype error rate or expected number of exclusions.

Binomial exclusion method

Estimation of the p-parameter for the BEM was done using trios that were assigned using GRL. An alternative to using GRL-assignments is using a training dataset with genotyped trios and known pedigree. Such a training dataset would need to have a similar genotype error rate as the evaluation dataset since having a discrepancy between the true and assumed genotype error rate could lead to decreased accuracy [see Additional file 4: Table S1]. Since pedigree information is not always reliable, we prefer to use GRL assignments (preferably using a relatively big dataset) for parameter estimation.

Comparing GRL and the binomial exclusion method with Colony2

The GRL and BEM resulted in much more accurate assignments of parents than Colony2. Parameters for Colony2 were chosen to minimize running time, so assignment accuracy may be improved by adjusting the parameters, but at the expense of time and/or computing resources required to perform the analysis. Colony2 incorrectly assumes that marker loci are independently distributed, while GRL and BEM do not. This is likely the main reason for the poor results obtained with Colony2 on these relatively dense marker datasets.

Comparing GRL with the binomial exclusion method

Using BEM resulted in a slightly higher accuracy than GRL when the genotype error assumption was correct, or when GRL-parameters were estimated using the evaluation data (Figs. 3 and 4) and [see Additional file 3: Figures S4, S5, S8 and S9]. However, when pre-estimated model parameters are used, assuming a too high genotype error rate will lead to some false assignments with BEM (Fig. 6), and assignment failure for the GRL method (Fig. 5). Thus, GRL can be used when it is crucial to minimize the false-positive rate. Assuming a too low genotype error rate resulted in both methods failing to correctly assign any trios, but GRL had a small fraction (0.016%) of false assignments while BEM did not [see Additional file 4: Table S1]. Although the success parameter (p, see Methods) of BEM was estimated using already GRL-assigned trios, the results indicate that the two methods are somewhat complementary and can be used together to increase overall assignment accuracy.

When the assumed genotype error rate was correct (Figs. 3 and 4) and [see Additional file 3: Figures S4 and S5] or when the evaluation dataset was used to estimate parameters [see Additional file 3: Figures S8, S9, S10 and S11], nearly all the individuals were assigned correctly and there were hardly any false assignments with either method. Thus, parameters should be estimated using the available data whenever possible, which should be the case in most situations.

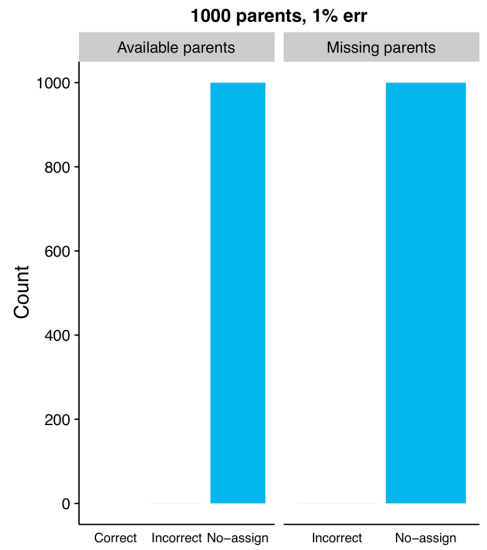


Fig. 5 Assignment results using GRL for a 1% true genotyping error rate but using parameter estimates from a dataset with 3% genotype errors. Individuals with (left panel) and without (right panel) available parents are present in the dataset. Results from 50 simulated datasets are averaged

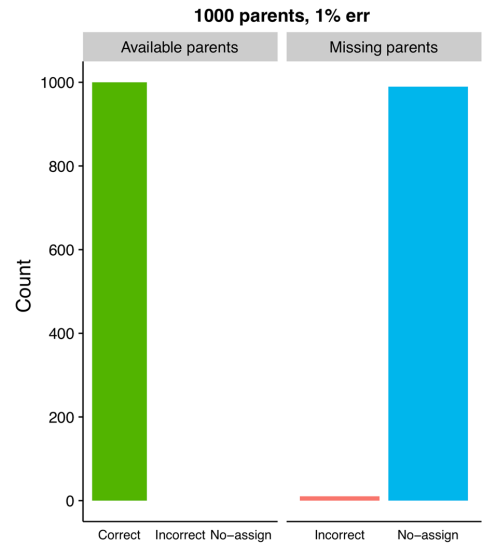


Fig. 6 Assignment results using BEM for a 1% true genotyping error rate but using parameter estimates using GRL-assignments from a dataset with 3% genotype errors. Individuals with (left panel) and without (right panel) available parents are present in the dataset. Results from 50 simulated datasets are averaged

Using GRL with clones or duplicated DNA

A possible novel use for the GRL method is analysis of genomic data that contain possibly duplicated genomes (e.g., by sampling of clones in plants or monozygotic twins in animals, or by duplicated sampling of DNA from the same individual). Using traditional likelihood-based or exclusion-based methods, duplicated samples/clones should be removed prior to the analysis, as these may be assigned as their own parents. For the GRL method, duplication of offspring genotypes is not a problem since GRL looks at patterns in parent–offspring relationships rather than the likelihood of each single genotype. For example, if clones of a non-inbred offspring are inserted as one or both putative parents, the GRL method would expect the offspring to be highly inbred, which will not match the observed relationship of the offspring with itself, and thus yields a low GRL value. However, duplication of parental genotypes will inevitably lead to assignment failure, since two or more trios will appear equally likely.

Conclusions

The GRL method is a promising trio parentage assignment method which is well suited to perform parentage assignment with high accuracy on high-density SNP datasets. GRL can be applied with success on datasets with high and/or unknown genotype error rates, highly dependent marker loci, closely-related individuals, inbreeding and in some cases clones. Estimation of the GRL parameters can be done without having a pre-existing reference dataset with known parent–offspring trio combinations. In addition, GRL can be used for training of exclusion-based methods.

Additional files

Additonal file 1: Figures S1, S2 and S3. Residual relationships plotted for all true trios from the 50 datasets. This file contains three figures (Figures S1, S2 and S3). Residual densities for offspring to itself (top panel), offspring to real mother (mid panel) and offspring to real father (bottom panel) are shown as a continuous line in all Figs. 50,000 values were sampled from the normal distribution using the means and variances of the residuals as parameters, shown as a dashed line in each panel. Figure S1 shows results in which there is no genotype error or call rate variance, Figure S2 in which there is 1% genotype error and a ~ 80 to 100% call rate and Figure S3 in which there is a 3% genotype error and a ~ 80 to 100% call rate.

Additonal file 2. Supplementary material. This file contains three sections with extended information about the GRL training procedure, call rate and genotype error simulation, and the binomial exclusion method (BEM), respectively.

Additonal file 3: Figures S4, S5, S6, S7, S8, S9, S10 and S11. Assignment results using GRL or BEM for individuals with (left panel) and without (right panel) available parents in the dataset. This file contains eight figures in which assignment results from 50 simulated datasets are averaged. Parameters were pre-estimated using one arbitrarily chosen dataset in Figures S4, S5, S6 and S7, while training was performed on each evaluation dataset in Figures S8, S9, S10 and S11. Figures S4, S6, S8 and S10 show results using GRL, while Figures S5, S7, S9 and S11 show results using BEM. Figures S4 and S5 show results when there is a 3% genotype error (true and assumed), Figures S6 and S7 have pre-esimated parameters from a dataset with a 1% genotype error, while the (true) evaluation genotype error is 3%. Figures S8, S9, S10 and S11 use training on each evaluation dataset, both at 1% (Figures S8 and S9) and 3% (Figures S10 and S11) genotype errors. In all figures, the call rates are ~ 80 to 100%.

Additonal file 4: Table S1. Summary table of total number of correct, incorrect and non-assigned trios with or without parents and genotype errors for all 50 datasets. Genotype error: either 1% or 3%, and with assumption of genotype error in parenthesis (only applicable for models that are pre-trained). Available parents: all individuals with parents available for assignment in the dataset (Yes) or where all parents are missing (No). Correct: Number of correctly assigned individuals over all 50 datasets (only applicable when parents are available). Incorrect: Number of incorrectly assigned individuals over all 50 datasets. No-assign: Number of individuals that could not be assigned parents over all 50 datasets.

Authors' contributions
KEG wrote the software, performed the study and drafted the manuscript. JO conceived the GRL method, coordinated the whole study and contributed in writing and revising the manuscript. THEM helped finalize the theory behind the training portion of the GRL method as well as revising the manuscript critically. All authors read and approved the final manuscript.

Author details
[1] AquaGen AS, P.O. Box 1240, NO-7462 Trondheim, Norway. [2] Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, P.O. Box 5003, NO-1432 Ås, Norway.

Acknowledgements
The research leading to these results has received funding from The Research Council of Norway through both the NAERINGSPHD (Project No. 251664) and the HAVBRUK2 (Project No. 245519) programs, as well as the breeding company AquaGen AS. The authors thank Thore Egeland for helpful comments on an early version of the draft and Jinliang Wang for providing support for the Colony2 software. We also wish to thank the editors and reviewers, especially reviewer 2 whose comments lead to a significant increase in the quality of the end result.

Competing interests
KEG and JO are employed by AquaGen AS. AquaGen has applied for a patent regarding the use of the GRL methodology in parentage assignment.

Availability of data and materials
The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Consent for publication
Not applicable.

Ethics approval and consent to participate
Not applicable.

Funding
The research leading to these results has received funding from The Research Council of Norway through the research programs NAERINGSPHD (Project No. 251664) and the HAVBRUK2 (Project No. 245519), and AquaGen AS.

Appendix

Mathematical foundation for the GRL method

In this article, only the hypothesis of true parents is used for the GRL method:

H_1 : *Both candidate parents are the true parents of the child*.

We assume $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\mathbf{x}$ is the vector of residual genomic relationships, i.e. it holds the residual values for trio assignments. We define x_1 as being the most probable trio, while x_2 is the second most probable trio, that is $P(x_1|H_1) \geq P(x_2|H_1)$.

The difference $\Delta_{\text{GRL}} = \text{GRL}_1 - \text{GRL}_2$, where GRL_1 and GRL_2 refer to the best and the second best trio candidates, respectively, can be shown to be identical to the natural logarithm of the probability of observing x_1 given H_1 divided by the probability of observing x_2 given H_1. Since $\mathbf{x}$ is assumed to be normally distributed, the multivariate normal probability density function used is:

$$f(\mathbf{x}|H_1) = \frac{1}{\sqrt{(2\pi)^3|\Sigma|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})},$$

where $\mathbf{x}$ is the 3×1 vector of genomic residuals, $\boldsymbol{\mu}$ is 3x1 vector of expected residuals and $\boldsymbol{\Sigma}$ is the 3x3 covariance matrix. If we define $\frac{L_1}{L_2} = \frac{f(x_1|H_1)}{f(x_2|H_1)}$ (i.e. how many times more likely is x_1 given H_1 compared to x_2 given H_1), we find that:

$$\begin{aligned}\frac{f(\boldsymbol{x}_1|H_1)}{f(\boldsymbol{x}_2|H_1)} &= \frac{\frac{1}{\sqrt{(2\pi)^3|\Sigma|}} e^{-\frac{1}{2}(\mathbf{x}_1-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_1-\boldsymbol{\mu})}}{\frac{1}{\sqrt{(2\pi)^3|\Sigma|}} e^{-\frac{1}{2}(\mathbf{x}_2-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_2-\boldsymbol{\mu})}} \\ &= \frac{e^{-\frac{1}{2}(\mathbf{x}_1-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_1-\boldsymbol{\mu})}}{e^{-\frac{1}{2}(\mathbf{x}_2-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_2-\boldsymbol{\mu})}}.\end{aligned}$$

If we take the natural logarithm of this ratio we get:

$$\begin{aligned}\ln\left[\frac{f(x_1|H_1)}{f(x_2|H_1)}\right] &= \ln\left[\frac{e^{-\frac{1}{2}(\mathbf{x}_1-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_1-\boldsymbol{\mu})}}{e^{-\frac{1}{2}(\mathbf{x}_2-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_2-\boldsymbol{\mu})}}\right] \\ &= \ln\left(e^{-\frac{1}{2}(\mathbf{x}_1-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_1-\boldsymbol{\mu})}\right) \\ &\quad - \ln\left(e^{-\frac{1}{2}(\mathbf{x}_2-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_2-\boldsymbol{\mu})}\right) \\ &= -\frac{1}{2}(\mathbf{x}_1-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_1-\boldsymbol{\mu}) \\ &\quad -\left(-\frac{1}{2}(\mathbf{x}_2-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_2-\boldsymbol{\mu})\right) \\ &= \text{GRL}_1 - \text{GRL}_2 = \Delta_{\text{GRL}}.\end{aligned}$$

The above formula shows that Δ_{GRL} has a logarithmic point probability ratio expectation. We can compare this

to the $\Delta_{Marshall}$ test statistic which is defined as in [7], that is:

$$\mathrm{LOD} = \ln(LR) = \ln\left(\frac{P(data|H_1)}{P(data|H_2)}\right),$$

where H_2 is defined as:

H_2 : *Two random individuals are the true parents of the child*.

LOD_1 is defined to be the LOD-score of the most likely trio, while LOD_2 is the second most likely trio. Then:

$$\Delta_{\mathrm{Marshall}} = \mathrm{LOD}_1 - \mathrm{LOD}_2 = \ln(LR_1) - \ln(LR_2).$$

Both $P(data|H_1)$ and $P(data|H_2)$ can be written as follows:

$$P(data|H_1) = \prod_{t=1}^{c} P_t\left(g_C|g_F, g_M, H_1\right) * P_t\left(g_F\right) * P_t\left(g_M\right),$$
$$P(data|H_2) = \prod_{t=1}^{c} P_t\left(g_C|H_2\right) * P_t\left(g_F\right) * P_t\left(g_M\right),$$

where $P_t\left(g_C|g_F, g_M, H_1\right)$ is the probability of observing the offspring genotype given the father and mother genotypes under H_1 at locus t, $P_t\left(g_F\right)$ is the probability of observing the father genotype at locus t, $P_t\left(g_M\right)$ is the probability of observing the mother genotype at locus t, $P_t\left(g_C|H_2\right)$ is the probability of observing the offspring genotype under H_2 at locus t and c is the number of loci. Since $LR = \frac{P(data|H_1)}{P(data|H_2)}$, we can simplify LR to be:

$$LR = \frac{P(data|H_1)}{P(data|H_2)} = \prod_{t=1}^{c} \frac{P_t\left(g_C|g_F, g_M, H_1\right) * P_t\left(g_F\right) * P_t\left(g_M\right)}{P_t\left(g_C|H_2\right) * P_t\left(g_F\right) * P_t\left(g_M\right)} = \prod_{t=1}^{c} \frac{P_t\left(g_C|g_F, g_M, H_1\right)}{P_t\left(g_C|H_2\right)}.$$

Since LR_1 is the likelihood ratio of the most likely trio and LR_2 is the likelihood ratio of the second most likely trio (defined above), we can write LR_1 and LR_2 as:

$$LR_1 = \frac{P(data_1|H_1)}{P(data_1|H_2)} = \prod_{t=1}^{c} \frac{P_t\left(g_C|g_{F_1}, g_{M_1}, H_1\right)}{P_t\left(g_C|H_2\right)},$$

and

$$LR_2 = \frac{P(data_2|H_1)}{P(data_2|H_2)} = \prod_{t=1}^{c} \frac{P_t\left(g_C|g_{F_2}, g_{M_2}, H_1\right)}{P_t\left(g_C|H_2\right)},$$

where g_{F_1} and g_{M_1} are the genotypes of the father and mother in the most likely trio at locus t, respectively, and g_{F_2} and g_{M_2} are the genotypes of the father and mother at locus t in the second most likely trio, respectively. Since the same offspring is used in both trios, g_C is the same for both LR_1 and LR_2 for locus t.

Inserting LR_1 and LR_2 into the $\Delta_{Marshall}$-formula above we get:

$$\begin{aligned}\Delta_{Marshall} &= \ln(LR_1) - \ln(LR_2) = \ln\left(\frac{LR_1}{LR_2}\right) \\ &= \ln\left(\frac{\prod_{t=1}^{n} \frac{P_t\left(\mathrm{g}_C|\mathrm{g}_{F_1}, \mathrm{g}_{M_1}, H_1\right)}{P_t\left(\mathrm{g}_C|H_2\right)}}{\prod_{t=1}^{n} \frac{P_t\left(\mathrm{g}_C|\mathrm{g}_{F_2}, \mathrm{g}_{M_2}, H_1\right)}{P_t\left(\mathrm{g}_C|H_2\right)}}\right) \\ &= \ln\left(\prod_{t=1}^{n} \frac{P_t\left(\mathrm{g}_C|\mathrm{g}_{F_1}, \mathrm{g}_{M_1}, H_1\right)}{P_t\left(\mathrm{g}_C|\mathrm{g}_{F_2}, \mathrm{g}_{M_2}, H_1\right)}\right) \\ &= \ln\left(\frac{P\left(\mathbf{g}_C|\mathbf{g}_{F_1}, \mathbf{g}_{M_1}, H_1\right)}{P\left(\mathbf{g}_C|\mathbf{g}_{F_2}, \mathbf{g}_{M_2}, H_1\right)}\right),\end{aligned}$$

where the explanation for $\mathrm{g}_C, \mathrm{g}_{F_1}, \mathrm{g}_{M_1}, \mathrm{g}_{F_2}, \mathrm{g}_{M_2}$ is the same as above, while $\mathbf{g}_C$, $\mathbf{g}_{F_1}$, $\mathbf{g}_{M_1}$, $\mathbf{g}_{F_2}$ and $\mathbf{g}_{M_2}$ are the genotypes for the offspring (or child), for most probable father and mother and for the second most probable father and mother, respectively, over all loci in vector-notation. The $\Delta_{Marshall}$ method only uses the probability of observing the child genotypes given that F_1 and M_1, or F_2 and M_2 are the true parents. The fact that the information in the H_1 hypothesis is not used makes the $\Delta_{Marshall}$ method similar to Δ_{GRL}, we see this when the two method definitions are compared:

$$GRL\ method : \Delta_{GRL} = \ln\left[\frac{f(\boldsymbol{x}_1|H_1)}{f(\boldsymbol{x}_2|H_1)}\right],$$

$$Marshall\ method : \Delta_{Marshall} = \ln\left(\frac{P\left(\mathbf{g}_C|\mathbf{g}_{F_1}, \mathbf{g}_{M_1}, H_1\right)}{P\left(\mathbf{g}_C|\mathbf{g}_{F_2}, \mathbf{g}_{M_2}, H_1\right)}\right).$$

Both methods produce an estimated logarithmic ratio of the probability that C is the child of the two most probable parent candidates versus the probability that C is the child of the two second most probable parent candidates, hence the results produced by the two methods can be considered analogous.

Publisher's Note

Received: 16 October 2017 Accepted: 4 May 2018
Published online: 18 May 2018

References

1. Heaton MP, Leymaster KA, Kalbfleisch TS, Kijas JW, Clarke SM, McEwan J, et al. SNPs for parentage testing and traceability in globally diverse breeds of sheep. PLoS One. 2014;9:e94851.
2. Waldbieser GC, Bosworth BG. A standardized microsatellite marker panel for parentage and kinship analyses in channel catfish, *Ictalurus punctatus*. Anim Genet. 2013;44:476–9.
3. Campbell D, Duchesne P, Bernatchez L. AFLP utility for population assignment studies: analytical investigation and empirical comparison with microsatellites. Mol Ecol. 2003;12:1979–91.
4. Jones AG, Small CM, Paczolt KA, Ratterman NL. A practical guide to methods of parentage analysis. Mol Ecol Resour. 2010;10:6–30.
5. Morrissey MB, Wilson AJ. The potential costs of accounting for genotypic errors in molecular parentage analyses. Mol Ecol. 2005;14:4111–21.
6. Strucken EM, Lee SH, Lee HK, Song KD, Gibson JP, Gondro C. How many markers are enough? Factors influencing parentage testing in different livestock populations. J Anim Breed Genet. 2016;133:13–23.
7. Marshall TC, Slate J, Kruuk LE, Pemberton JM. Statistical confidence for likelihood-based paternity inference in natural populations. Mol Ecol. 1998;7:639–55.
8. Purfield DC, McClure M, Berry DP. Justification for setting the individual animal genotype call rate threshold at eighty-five percent. J Anim Sci. 2016;94:4558–69.
9. Goddard ME, Hayes BJ, Meuwissen TH. Using the genomic relationship matrix to predict the accuracy of genomic selection. J Anim Breed Genet. 2011;128:409–21.
10. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91:4414–23.
11. Falconer DS, Mackay TFC. Introduction to quantitative genetics. 4th ed. Harlow: Longman group Ltd; 1996.
12. Malécot G. Les mathématiques de l'hérédité. Paris: Masson; 1948.
13. Billingsley P. Probability and measure. 3rd ed. New York: Wiley; 1995.
14. Sargolzaei M, Schenkel FS. QMSim: a large-scale genome simulator for livestock. Bioinformatics. 2009;25:680–1.
15. Jones OR, Wang J. COLONY: a program for parentage and sibship inference from multilocus genotype data. Mol Ecol Resour. 2010;10:551–5.
16. Wang J, Santure AW. Parentage and sibship inference from multilocus genotype data under polygamy. Genetics. 2009;181:1579–94.
17. Wang J. Computationally efficient sibship and parentage assignment from multilocus marker data. Genetics. 2012;191:183–94.
18. Hayes BJ. Efficient parentage assignment and pedigree reconstruction with dense single nucleotide polymorphism data. J Dairy Sci. 2011;94:2114–7.

II Genotype calling of triploid offspring from diploid parents

Grashei, K.E., Ødegård, J. & Meuwissen, T.H.E. Genet Sel Evol 52, 15 (2020).
https://doi.org/10.1186/s12711-020-00534-w

Grashei *et al. Genet Sel Evol* (2020) 52:15
https://doi.org/10.1186/s12711-020-00534-w

GSE Genetics Selection Evolution

RESEARCH ARTICLE Open Access

Genotype calling of triploid offspring from diploid parents

Kim Erik Grashei[1,2*], Jørgen Ødegård[1,2] and Theo H. E. Meuwissen[2]

Abstract

Background: Polyploidy is widespread in animals and especially in plants. Different kinds of ploidies exist, for example, hexaploidy in wheat, octaploidy in strawberries, and diploidy, triploidy, tetraploidy, and pseudo-tetraploidy (partly tetraploid) in fish. Triploid offspring from diploid parents occur frequently in the wild in Atlantic salmon (*Salmo salar*) and, as with triploidy in general, the triploid individuals are sterile. Induced triploidy in Atlantic salmon is common practice to produce sterile fish. In Norwegian aquaculture, production of sterile triploid fish is an attempt by government and industry to limit genetic introgression between wild and farmed fish. However, triploid fish may have traits and properties that differ from those of diploids. Investigating the genetics behind traits in triploids has proved challenging because genotype calling of genetic markers in triploids is not supported by standard software. Our aim was to develop a method that can be used for genotype calling of genetic markers in triploid individuals.

Results: Allele signals were produced for 381 triploid Atlantic salmon offspring using a 56 K Thermo Fisher GeneTitan genotyping platform. Genotypes were successfully called by applying finite normal mixture models to the (transformed) allele signals. Subsets of markers were filtered by quality control statistics for use with downstream analyses. The quality of the called genotypes was sufficient to allow for assignment of diploid parents to the triploid offspring and to discriminate between maternal and paternal parents from autosomal inheritance patterns. In addition, as the maternal inheritance in triploid offspring is identical to gynogenetic inheritance, the maternal recombination pattern for each chromosome could be mapped by using a similar approach as that used in gene-centromere mapping.

Conclusions: We show that calling of dense marker genotypes for triploid individuals is feasible. The resulting genotypes can be used in parentage assignment of triploid offspring to diploid parents, to discriminate between maternal and paternal parents using autosomal inheritance patterns, and to map the maternal recombination pattern using an approach similar to gene-centromere mapping. Genotyping of triploid individuals is important both for selective breeding programs and unravelling the underlying genetics of phenotypes recorded in triploids. In principle, the developed method can be used for genotype calling of other polyploid organisms.

Background

Polyploidy is widespread in plants and exists both in vertebrate and invertebrate animals [1, 2]. In aquaculture species, triploidy can be induced by pressure-shocking newly fertilized eggs, resulting in unreduced gametes in the females [3]. In such induced triploids, the shocking of eggs prevents the second polar body from leaving the secondary oocyte during meiosis [4]. This results in a triploid cell, in which two sets of chromosomes are inherited from the mother, and one set from the father. This practice is commonly used by the aquaculture industry to produce sterile fish for farming, and by wildlife management for stocking of sterile game fish for recreational purposes.

The Thermo Fisher GeneTitan platform is commonly used to genotype Atlantic salmon using high-density single nucleotide polymorphism (SNP) chips [5]. However, genotyping of triploids is currently not possible using the

*Correspondence: kim.erik.grashei@aquagen.no; kim.grashei@nmbu.no
[1] AquaGen AS, P.O. Box 1240, 7462 Trondheim, Norway
Full list of author information is available at the end of the article

supplied Thermo Fisher software, which limits research projects for triploids of any species. The goal of this study was to develop a method for calling genotypes for triploid individuals using the output from the Thermo Fisher GeneTitan instrument. Secondary aims were to develop methods for dam- and sire-specific assignment of induced triploid offspring, and to use maternal inheritance to triploid offspring to map the maternal recombination pattern. Although the genotype calling method was tested only by using diploids and triploids, in principle, it can be extended to other ploidies as well.

Methods

Data

DNA was sampled from 381 triploid Atlantic salmon and genotyped on a Thermo Fisher SNP chip array with 56,177 SNPs (56 k chip). Triploidy was verified by visually identifying four distinct clusters of transformed allele strengths per individual, namely 'contrast' and 'size' (see [6]). Two individuals showed a correlation of genotypes higher than 0.99 (likely, duplicated or contaminated samples) and were removed, leaving 379 triploid individuals. In total, 3158 diploid individuals from the parent generation were also genotyped, using either the same 56 k chip or a 220,000 SNP chip (220 k chip), which had 52,458 (52 k) SNPs in common with the 56 k chip. Downstream analyses were performed using the 52 k common SNPs, or a subset of these. Two candidate parents had apparently duplicated genotypes, likely due to duplicated samples. After removal of duplicates, the number of individuals in the candidate parent dataset was equal to 3156.

In addition, 914 diploid Atlantic salmon and 116 of their previously parentage assigned parents were used to compare genotype calling methods. In total, 853 of the 914 diploid offspring had both parents assigned, i.e. trios. The 914 diploids were called using both the method developed here and using Thermo Fisher's APT software, the Affymetrix power tools (APT) [6]. The parents were genotyped using one of the two SNP chips described above and, thus, the same 52,458 SNPs were used in all downstream analyses. Additional details are provided in the subsection 'Calling genotypes with Affymetrix power tools' below.

New genotype calling method

Observations of $contrast = \log_2\left(A_{signal}/B_{signal}\right)$ (also known as 'Delta') were obtained for each DNA sample from the file 'AxiomGT1.normalized-summary.txt', which is produced by the APT software [6, 9]. The A_{signal} and B_{signal} are the signal strengths observed by the GeneTitan instrument for the two possible alleles (called A and B) for each SNP. Thus, the possible genotypes for a given SNP are AA, AB and BB for diploid and AAA, AAB, ABB and BBB for triploid individuals.

The R package "mclust" [7] was used for calling both diploid and triploid genotypes, fitting up to three and four genotype clusters, respectively, in a single dimension (the contrast). The clustering models assumed that the contrast is a mix of normally distributed variables, one for each genotype cluster, allowing for different expectations and variances for each cluster, depending on the model. The mclust package attempts to identify the underlying distributions by choosing the most likely out of two possible models for each genotype cluster. The two models are: (1) the 'E' model, in which each genotype cluster is assumed to have equal variances, and (2) the 'V' model, in which the genotype clusters can have different variances (see [7]). Not all markers will have all biologically possible clusters represented; e.g. markers of low minor allele frequency may only show the most common cluster(s). Thus, the two models are tested with the assumption that there are one, two, three, or four (for triploids only) genotype clusters in the data for a given locus. That is, for all biologically possible numbers of genotype clusters, both models ('E' and 'V') were fitted. This means that for triploid individuals, a marker could have up to four clusters (AAA, AAB, ABB and/or BBB), resulting in $4*2=8$ models being tested, while for diploid individuals up to three clusters are possible (AA, AB and/or BB), resulting in $3*2=6$ models. The integrated complete likelihood (ICL) for all models, defined as in [8] (a higher ICL is favorable), was calculated using the mclust package. ICL was chosen rather than the Bayesian information criterion (BIC) because of its tendency to favor well-separated clusters (see "Discussion" for more information). For each number of clusters G, the model with the highest ICL was saved, i.e. $\max(ICL(G))$, where $G \in \{1,2,3,4\}$ for triploids and $G \in \{1,2,3\}$ for diploids. Then, the model with the highest ICL (ICL_1) was assumed to produce genotypes with the lowest genotype error rate and, thus, was chosen to classify the genotypes.

Mclust uses the iterative expectation maximization (EM) algorithm for all models, which adjusts the parameters until the most likely set of parameters is found for each model. When no starting parameter values are set, mclust uses the mean contrast of each marker as a starting point for all possible genotype clusters in the first EM-iteration. In some cases, this may result in mclust choosing a local optimum for the parameter estimates due to, e.g. uneven numbers of individuals in the different genotype classes or DNA sample bias due to differences in DNA quality (see "Discussion"). To obtain better starting values, the initial numbers of individuals in each genotype group ($n_1, \ldots n_G$, where G is the chosen number of clusters) were predicted by using a rough estimate

related to the SNP allele frequency (see Appendix). Then, initial clustering was done by sorting the individuals by contrast values and initially assigning the first n_1 individuals to the first (left-wise) cluster, n_2 to the second cluster, etc. The contrast means of the initial clusters were set as starting values in the EM-algorithm and used as priors for the cluster means. Further details are in Appendix.

In cases where all four triploid clusters are found (i.e. all genotype groups are represented for the locus in question), the lowest cluster (with respect to contrast) is assumed to correspond to genotype *BBB*, the second to genotype *ABB*, etc. The same logic applies to diploids, except that there are up to three possible clusters. If three or fewer clusters are identified for triploids, the correspondence between left-to-right cluster number and genotype value is less obvious, similar to the case of diploids having two or fewer clusters. In such cases, the genotype calls are determined by the mean contrast of each cluster. Distributions of estimated mean contrast values for all markers that are predicted to have three or four clusters for diploids and triploids, respectively, are in Fig. 1. The estimated mean of each cluster in Fig. 1 is used to call genotypes for markers with less than the maximum possible number of clusters, i.e. markers with less than four clusters for triploids and less than three clusters for diploids. These estimated reference contrast means were approximately -1.76 ($=BB$), 0.16 ($=AB$), and 1.94 ($=AA$) for diploids, and -1.97 ($=BBB$), -0.50 ($=ABB$), 0.75 ($=AAB$), and 2.14 ($=AAA$) for triploids. These contrast means were estimated using the entire 56 k chip, where 44,431 and 38,792 of the 56,177 SNPs had a maximum number of clusters for diploids and triploids, respectively.

Some SNPs will not have all four (triploids) or all three (diploids) clusters because, for example, they might be fixed or have very high or very low allele frequencies. For such markers, the following approach was used: (1) retrieve the estimated mean contrast of each genotype cluster, and (2) find the closest reference contrast mean from the markers that had the maximum number of clusters (see Fig. 1) and set the genotypes to be the same as for these reference clusters. However, if two or more cluster contrast means are closest to the same reference cluster, the locus will not be used (defined as no-calls). This was the case for 950 of 11,746 SNPs in the diploid group with less than three clusters and for 4041 out of 17,385 markers in the triploid group with less than four clusters.

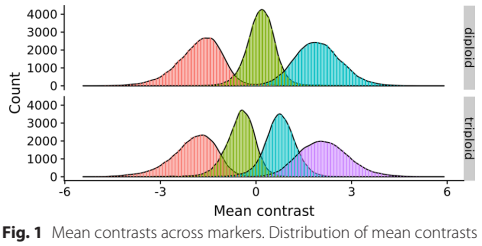


Fig. 1 Mean contrasts across markers. Distribution of mean contrasts across all 56,177 markers with three (diploids, top panel) or four (triploids, bottom panel) predicted genotype clusters, as estimated by the expectation maximization method

After choosing a model, the probabilities of belonging to each of the possible clusters are calculated for each contrast, and the cluster with the highest probability is chosen. The uncertainty is then the probability of the genotype belonging to any of the other clusters (1 minus the probability of belonging to the most likely cluster). If the uncertainty exceeds 0.15, the genotype value is defined as a no-call (i.e. a missing genotype). The threshold of 0.15 was chosen as this is the default threshold used by the APT software and, thus, provides a good comparison between the methods. Varying this threshold will result in different marker call rates, however we have not investigated the effects of varying this threshold on downstream analyses.

Calling genotypes with Affymetrix power tools

In addition to our mclust implementation, genotypes of the 914 diploid individuals were also called by using standard Thermo Fisher APT software based on the following three-step procedure: step 1: DQC-step: generate dish quality check (DQC) values for each sample and exclude samples below a chosen threshold, step 2: call genotypes for all remaining samples and calculate sample call rates, and step 3: call genotypes again using only the individuals from step 2 with call rates above a chosen threshold (see [9] for more background and information). Thus, individuals from step 3 have higher call rates than those from step 2. On a general basis, Thermo Fisher recommends setting the DQC threshold at 0.82 and the sample call rate threshold at 97%. However, we visually inspected the curves of ordered DQC- and call rate values, and set the threshold manually. An uncertainty value ('confidence' in Thermo Fisher terms) is estimated by APT for each call from each sample, which is equivalent to the uncertainty calculated by mclust. The recommended and default threshold for this uncertainty is 0.15, which is what we used both for calling with APT and mclust to provide a fair comparison of the two methods (see [9] for more information). All 914 diploids had genotypes from step 3, while of the 116 known parents, 104 had genotypes from step 3 and 12 from step 2.

Comparing APT and mclust using exclusion ratios in known trios

The 914 diploid individuals were genotyped using both APT and mclust with two goals in mind: (1) to compare the genotype calling accuracy between the two methods, and (2) to investigate if a threshold on absolute *ICL* or ΔICL could be used to identify high-quality markers with reduced genotyping errors, where $\Delta ICL = ICL_1 - ICL_2$, i.e. ΔICL is the difference in *ICL* between the two most likely numbers of genotype clusters (for the best model of each cluster number). Exclusion ratios (ER) between offspring and known parents were used as the main statistic for benchmarking the methods, in addition to some other support statistics (see "Results"). Exclusions are Mendelian mismatches between offspring and parent(s), and the ER is the number of exclusions between an offspring and its parent(s) divided by the number of SNPs where the genotypes are called for both offspring and parent(s). See Tables 1 and 2 for an overview of the combinations of genotypes between trios that were regarded as exclusions when offspring are diploid and triploid, respectively. Given the sex of triploid parents, additional erroneous genotypes can be identified using the maternal-specific inheritance to triploid offspring (see Table 2).

Of the 914 diploid offspring, 853 had both parents known. Thus, genotypes were called for 914 offspring, but comparisons using exclusion ratios were based on 853 known trios. For the triploid offspring, 304 known trios were used (see text regarding parentage assignment and parental sex prediction of triploids in "Methods" and "Results" sections).

Table 1 Possible Mendelian exclusions between a diploid offspring ("O exclusion") and its diploid parents ("P1" and "P2")

P1	P2	O exclusion	Comment
NA	*AA*	*BB*	One parent has NoCall
NA	*BB*	*AA*	One parent has NoCall
AA	*NA*	*BB*	One parent has NoCall
BB	*NA*	*AA*	One parent has NoCall
AA	*BB*	*AA, BB*	Oppositely homozygous parents
BB	*AA*	*AA, BB*	Oppositely homozygous parents
AA	*AA*	*AB, BB*	Identically homozygous parents
BB	*BB*	*AA, AB*	Identically homozygous parents
AB	*AA*	*BB*	One parent heterozygous and the other homozygous
AB	*BB*	*AA*	One parent heterozygous and the other homozygous
AA	*AB*	*BB*	One parent heterozygous and the other homozygous
BB	*AB*	*AA*	One parent heterozygous and the other homozygous

NA indicates missing genotype

Parentage assignment of diploid parents to triploid offspring

The fraction of parents to triploids represented in the data was unknown. An exclusion-based approach was used to assign diploid parents to triploid offspring. Neither triploid nor diploid offspring can be opposite homozygotes (i.e. have exclusions) relative to their true parents through Mendelian inheritance of alleles. For example, a parent with genotype *BB* at a given SNP cannot have offspring with genotype *AA* (diploid)/*AAA* (triploid) at that SNP. However, since genotype errors can occur, a relatively small number of exclusions should be expected even in true parent–offspring pairs. The

Table 2 Possible Mendelian exclusions between a triploid offspring ("O exclusion") and its diploid mother ("M") and father ("F")

M	F	O exclusion	Comment
BB	*NA*	*AAA, AAB*	Father has missing genotype and mother is homozygous
AA	*NA*	*ABB, BBB*	Father has missing genotype and mother is homozygous
NA	*AA*	*BBB*	Mother has missing genotype and father is homozygous
NA	*BB*	*AAA*	Mother has missing genotype and father is homozygous
AA	*BB*	*AAA, ABB, BBB*	Oppositely homozygous parents
BB	*AA*	*AAA, AAB, BBB*	Oppositely homozygous parents
AA	*AA*	*AAB, ABB, BBB*	Identically homozygous parents
BB	*BB*	*AAA, AAB, ABB*	Identically homozygous parents
AA	*AB*	*ABB, BBB*	Mother homozygous, father heterozygous
BB	*AB*	*AAA, AAB*	Mother homozygous, father heterozygous
AB	*AA*	*BBB*	Mother heterozygous, father homozygous
AB	*BB*	*AAA*	Mother heterozygous, father homozygous

NA indicates missing genotype

expected number of exclusions between an offspring and a non-parent individual depends on their genomic relationship, i.e. greater relatedness between individuals usually means a smaller number of exclusions. ER was used for parentage assignment instead of the number of exclusions to account for variation in individual call rates due to differences in DNA quality. Only markers for which triploid $\Delta ICL > 150$ and with parent call rates > 95% were used in the parentage assignment. ER were calculated for each pair of offspring and candidate parent. An assignment ER-threshold of 0.002 for offspring–candidate duos was applied, which means that all candidate parents with ER below this threshold for an offspring were assumed the true parents. See "Results", for more detailed information regarding the choice of ER threshold.

Parent sex prediction for triploid offspring

When using pressure-shock induced triploidy in fish, the offspring receives two sets of chromosomes from the mother and a single set from the father (see "Background" section). As a result, certain genotypes are not possible for a true mother of the offspring but are possible for the true father. For example, a triploid offspring with marker genotype *AAB* implies that the true mother should have at least one allele *A*, i.e. the true mother cannot have genotype *BB* at that marker. Likewise, if the offspring has a marker genotype of *ABB*, the true mother cannot have genotype *AA*. In contrast, true father and offspring can have any genotype combination, except opposing homozygotes. Using this information, true mothers and fathers can be distinguished. The "mother-specific exclusions" were used along with opposing homozygotes to construct mother exclusion ratios, coined 'mother.ER', which was calculated by dividing the number of mother exclusions by the number of markers for which both the offspring and the candidate mother had called genotypes. In addition, ER from non-mother-specific exclusions were also used when constructing the 'mother.ER' shown in our results. The same markers were used to calculate candidate 'mother.ER', as was used in parentage assignment (see above).

Maternal recombinations

By pressure induced triploidy, the second polar body is not extruded during Meiosis II [4]. This implies that the sister chromatids formed during Meiosis I in the mother are still found within the ovum, along with the alleles passed down by the father, making the cell triploid. The sister chromatids passed down from the mother are identical, except for any recombinations that might have occurred during prophase I [10]. For markers for which the father is homozygous (*AA* or *BB*), the paternal allele state (i.e. a single *A* or *B*) of the offspring can be deduced, implying that maternal inheritance at that locus can also be deduced. Thus, markers for which the father is homozygous and the mother heterozygous can be used to map maternal crossovers with high accuracy given a relatively high density of such markers for multiple known offspring–mother–father trios. At the centromere, recombination is suppressed, and two identical alleles are thus inherited from the mother. On each chromosome arm, the maternally inherited alleles shift from homozygotes to heterozygotes at the location of the first crossover. A second maternal crossover (further away from the centromere) will cause the maternal alleles to shift back from heterozygous to homozygous. Hence, the triploid offspring genotypes can be used to study the recombination patterns of different chromosomes, simply by comparing genotypes of diploid parents and triploid offspring, without phasing of genotypes. This method of recombination mapping is essentially the same as gene-centromere mapping in gynogenetic diploids [11], except that induced triploids do have paternal inheritance, which is lacking in gynogenetic diploids.

In total, 304 trios of triploid offspring with assigned mothers and fathers were used to estimate maternal recombination rates. To retain the majority of the markers and still have high enough marker quality to interpret the downstream results, we chose a ΔICL threshold of 50 and marker call rate thresholds of 0.80 for the triploid offspring and 0.95 for diploid parents. Each marker had to be mapped to a given chromosome and have at least 50 trios with informative genotypes (i.e. homozygous father and heterozygous mother). This resulted in 27,130 informative markers with a maternal recombination estimate (see Fig. 7). The markers used were placed on the ICSASG_v2 Atlantic salmon genome reference assembly [12, 13].

Results

Calling genotypes with mclust and APT

Without filtering SNPs, the numbers of SNPs for diploid offspring predicted to have one, two, or three clusters were 155, 3970 and 48,333, respectively, when using APT and 2008, 7534 and 42,916, respectively, when using mclust. For each SNP called by mclust, there were two possible models: heterogenous and homogenous cluster variance. The heterogenous cluster variance model was chosen for 68% of the SNPs for diploids and for 49% of the SNPs for triploids. Using mclust, the numbers of SNPs for triploid offspring predicted to have a single, two, three, or four clusters were 3149, 3528, 10,708, and 38,792, respectively. Of the 9542 SNPs with less than three clusters called by mclust for diploids and that were used in downstream analyses, 724 were given 100% no-calls, due to insufficient

separation of clusters (see subsection "New genotype calling method" in "Methods"). For triploids, there were 3620 such markers. Figure 2 shows indicator statistics of mclust marker calling quality in diploids for different thresholds of Δ*ICL*. Increasing Δ*ICL* resulted in a lower ratio of SNPs with one or two clusters for diploids, while the ratio of SNPs with three clusters increased. Furthermore, the ratio of Mendelian errors (ER) decreased as Δ*ICL* increased, indicating that increasing the threshold for Δ*ICL* improves calling quality. This was supported by the decreasing ratio of missing genotypes ('NoCalls'), which indicates that higher thresholds for Δ*ICL* results in retaining SNPs that have good separation of genotype clusters, i.e. SNPs with a low call uncertainty. The red horizontal lines in Fig. 2 show the values achieved by using the genotype calls from APT with all SNPs included in the analyses. APT achieved fewer Mendelian errors (ER) and fewer missing genotypes ('NoCalls') than our mclust implementation when all SNPs were used. Mclust needs a Δ*ICL* threshold of ~ 110 to obtain similar levels of ER and no-calls as APT, which resulted in the use of ~ 7500 fewer SNPs.

Figure 3 shows the same statistics as Fig. 2 after removing SNPs below an *ICL* threshold (note: not Δ*ICL*), i.e. a threshold on the *ICL* of the most likely model. All investigated *ICL* thresholds result in higher ER and NoCalls than was achieved by APT without SNP quality filtering. Because the variability of ER and no-calls seemed erratic, we decided not to use *ICL* as a marker quality filtering statistic in downstream analyses.

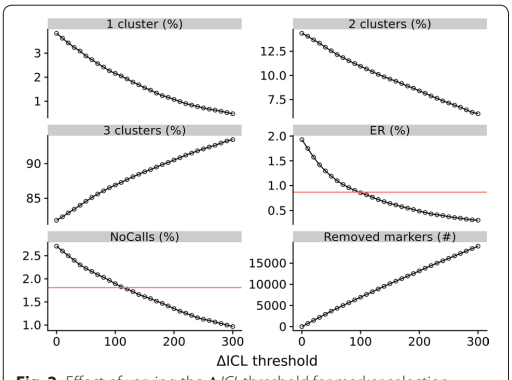


Fig. 2 Effect of varying the Δ*ICL* threshold for marker selection. Statistics for diploid offspring/parent trios when varying the Δ*ICL* threshold for marker selection when genotypes are called by the mclust algorithm. '1 cluster', '2 clusters' and '3 clusters' show the percentage of markers predicted to have one, two, and three clusters. 'ER' is the exclusion ratio shown in percent for trios. 'NoCalls' is the percentage of missing genotypes and 'Removed markers' shows the number of markers which are removed. The horizontal red lines show the values found for 'ER' and 'NoCalls' when using genotype calls from APT

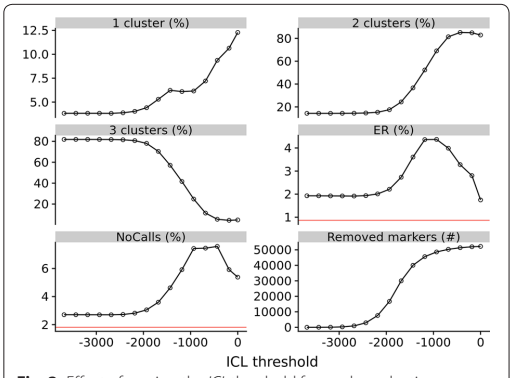


Fig. 3 Effect of varying the *ICL* threshold for marker selection. Statistics for diploid offspring/parent trios when varying the *ICL* threshold for marker selection when genotypes are called by the mclust algorithm. '1 cluster', '2 clusters' and '3 clusters' show the percentage of markers predicted to have one, two and three clusters. 'ER' is the exclusion ratio shown in percent for trios. 'NoCalls' is the percentage of missing genotypes and 'Removed markers' shows the number of markers which are removed. The horizontal red lines show the values found for 'ER' and 'NoCalls' when using genotype calls from APT

The histograms in Fig. 4 show that the Δ*ICL* achieved for one, two, and three clusters were roughly the same in triploids, while higher Δ*ICL* could be achieved for four clusters.

Parentage assignment of diploid parents to triploid offspring

The QC filtering of SNPs for parentage assignment based on triploid Δ*ICL* > 150 and parent call rates > 95% resulted in retaining 35, 375, 238, and 13,258 SNPs with one, two, three, and four genotype clusters, respectively, which resulted in the use of 13,906 SNPs for this parentage assignment. The ER between offspring and their first, second, third, and less likely candidate parents are shown in the top panel of Fig. 5, while the bottom panel zooms in on the best fitting parent candidates. The lowest ER between all triploid offspring and their third most likely candidate parent (i.e. the closest-fitting non-parent) was ~ 0.003. Consequently, a 0.002 assignment threshold for offspring–candidate duos was applied, which also fitted well, based on visual inspection of Fig. 5. In other words, the candidate parent in any duo with an ER < 0.002 was assigned and assumed to be a true parent. At least one parent was assigned to all 379 triploids, and 304 were assigned both parents. Lacking assignments were likely

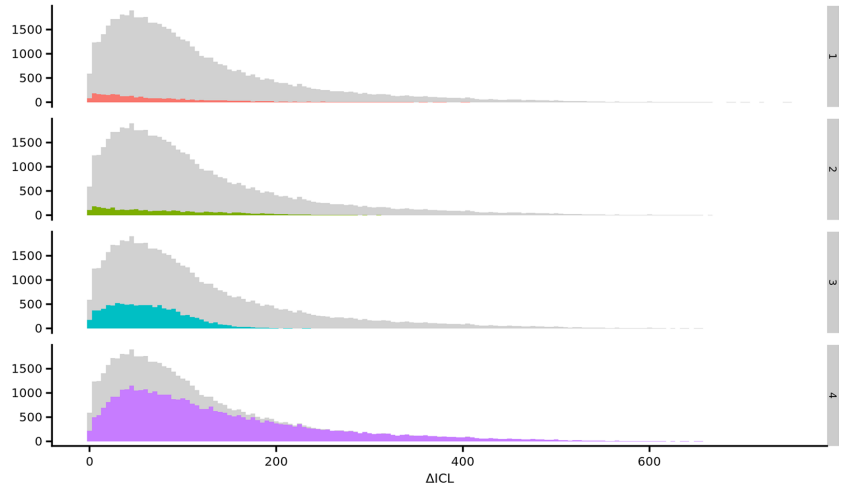


Fig. 4 Δ*ICL* distributions for markers genotypes in triploids. Distribution of Δ*ICL* values for markers predicted to have one (top panel) to four (bottom panel) genotype clusters in triploids. Markers with red, green, cyan and purple indicate 1, 2, 3 and 4 predicted genotype clusters, respectively, while the gray bars shows the total number of markers for each value of ΔICL

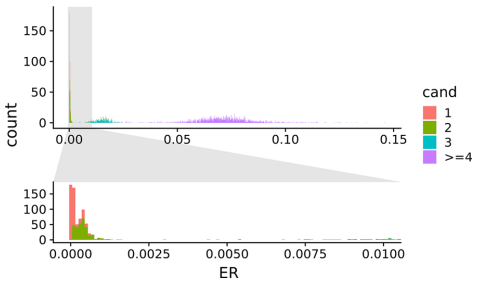


Fig. 5 Triploid exclusion ratios. Exclusion ratios (ER) for duos of triploid offspring and diploid parents are shown for the duos with lowest ER (red), second-lowest ER (green), third-lowest ER (turquoise) and 2000 randomly sampled ER from fourth lowest and above (purple)

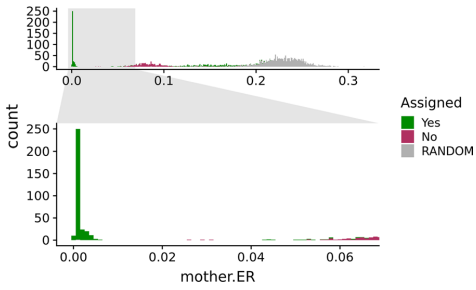


Fig. 6 Triploid maternal exclusion ratios. Maternal exclusion ratios (mother.ER) for duos of triploid offspring and diploid candidate mothers are shown for the assigned (green) and unassigned (maroon) duos and 2000 generated 'mother.ER' values from randomly sampled parent candidates, i.e. both sampling males and females which can possibly be true mothers (grey)

due to genotypes of some parents being absent in the dataset.

Parent sex prediction for triploid offspring

A similar procedure as for parentage assignment was used for assignment of mothers of triploid offspring, using the mother exclusion ratios ('mother.ER'). Figure 6 shows the 'mother.ER' between assigned, unassigned, and random pairs of offspring–parent candidates (note that 'Assigned' in Fig. 6 is for parentage assignment, not mother assignment). The minimum 'mother.ER' of the third-best parental candidates was ~0.022, thus we set the 'mother.ER' assignment threshold at 0.02. Any assigned parent with a 'mother.ER' < 0.02 was assigned as mother, and any assigned parent with 'mother.ER' ≥ 0.02 was assumed to be the true father. This resulted in 58 assigned mothers and 65 assigned fathers. No mothers were assigned as fathers, or vice versa. In total, 304 offspring were assigned both their mother and father (the same as the two parents assigned above), 14 were assigned a mother only, and 61 were assigned a father

only (i.e. as above, all individuals were assigned a father, a mother, or both). No offspring were assigned two apparent mothers or two apparent fathers.

Investigating different thresholds for Δ*ICL* and marker call rate

In addition to the Δ*ICL* threshold statistic explored above, marker call rate is another marker quality statistic that is often employed when analyzing genotype datasets. Marker call rate is related to ICL, as both call rate and ICL use call uncertainty as a measure of marker quality. Δ*ICL* provides a probabilistic penalization of the mixture model likelihood [8], whereas marker call rate is the fraction of genotypes that fall below a pre-defined uncertainty threshold. We chose the uncertainty threshold of 0.15, i.e. all genotype calls above this threshold were defined as no-calls (missing genotypes) for both triploid offspring and diploid parents. Table 3 shows that increasing either the marker call rate threshold or the Δ*ICL* threshold tended to decrease Mendelian errors, i.e. decrease the ER, but also increased the number of removed markers. Note that, for the parents, we always used a marker call rate threshold of 95%.

Maternal recombination rates

Figure 7 shows the estimated maternal recombination fraction along each of the 29 chromosomes in the Atlantic salmon genome by looking at where the triploid offspring inherited the homozygous (*AA*/*BB*) or heterozygous (*AB*) allele from the mother (see "Methods"). The region with the lowest maternal recombination fraction on each chromosome was at the centromere, where recombination is known to be suppressed. In [13], chromosome 8 was reported to be metacentric, but in Fig. 7 it appears as acrocentric or telocentric. However, the p-arm of chromosome 8 contains highly repetitive regions and, therefore, few or no markers from this region may be represented on the SNP chip (personal communication with S Lien, see also [14]).

Discussion

Sterile triploid Atlantic salmon have been produced for decades and differences in traits between the triploid and diploid Atlantic salmon have been observed [15]. To assign parentage, to identify population background, or to perform any kind of genetic analysis of triploids with genotype data requires methods for genotype calling in triploid individuals.

Calling SNP genotypes using sequencing data relies on the number of alleles that is called at a certain locus, and to know how this pattern varies for the two homozygous and the different heterozygous genotype groups [16, 17]. Conceptually, this differs from calling genotypes based on the aggregated light signal created by the Thermo Fisher GeneTitan instrument, as investigated here, because each allele has already been called in the sequence data. To the best of our knowledge, no official software for calling triploid genotypes using output from the Thermo Fisher GeneTitan instrument currently exists. However, software for polyploid genotyping has been created by other groups, such as the R package fitPoly [18, 19]. We chose to use mclust due to our familiarity with its functionality and its substantial documentation, frequent updates, and extensive use. Both mclust and fitPoly use mixture models and the EM algorithm to estimate parameters. Thus, fitPoly and mclust implementations may give similar results but significant differences cannot be ruled out. See "Calling genotypes with mclust and APT" subsection below for more discussion on this. However, a comparison between mclust and fitPoly was outside the scope of this study. In [20], Serang et al. use graphical Bayesian modelling to incorporate information on population allele frequencies or parental genotypes into the model to achieve increased genotype calling accuracy.

Table 3 Marker quality filtering using different thresholds for call rate and Δ*ICL*

Δ*ICL* threshold	Marker call rate threshold				
	0.80	**0.85**	**0.90**	**0.95**	**1.00**
0	0.02604 (3668)	0.02573 (3774)	0.02505 (4505)	0.02338 (7411)	0.0263 (29,905)
50	0.01582 (16,329)	0.01571 (16,340)	0.01545 (16,423)	0.01471 (17,526)	0.01461 (34,111)
100	0.00924 (29,888)	0.00916 (29,892)	0.00906 (29,902)	0.0089 (30,076)	0.00723 (39,535)
150	0.00433 (38,521)	0.00433 (38,521)	0.00433 (38,521)	0.0043 (38,535)	0.00287 (43,563)
200	0.00206 (43,557)	0.00206 (43,557)	0.00206 (43,557)	0.002 (43,561)	0.00114 (46,168)
250	0.0007 (46,679)	0.0007 (46,679)	0.0007 (46,679)	0.0007 (46,679)	0.00051 (47,936)
300	0.00037 (48,705)	0.00037 (48,705)	0.00037 (48,705)	0.00037 (48,705)	0.00029 (49,286)

Filtering markers using trios of triploid offspring and diploid parents with predicted sexes using marker call rate thresholds (top row in bold) and thresholds for ΔICL (left column in italic). The first number in each internal cell is the overall ER for all markers and all trios where there are informative genotypes (i.e. trios where offspring and at least one parent has called genotypes, see Table 2). The number of removed markers is shown in parenthesis.

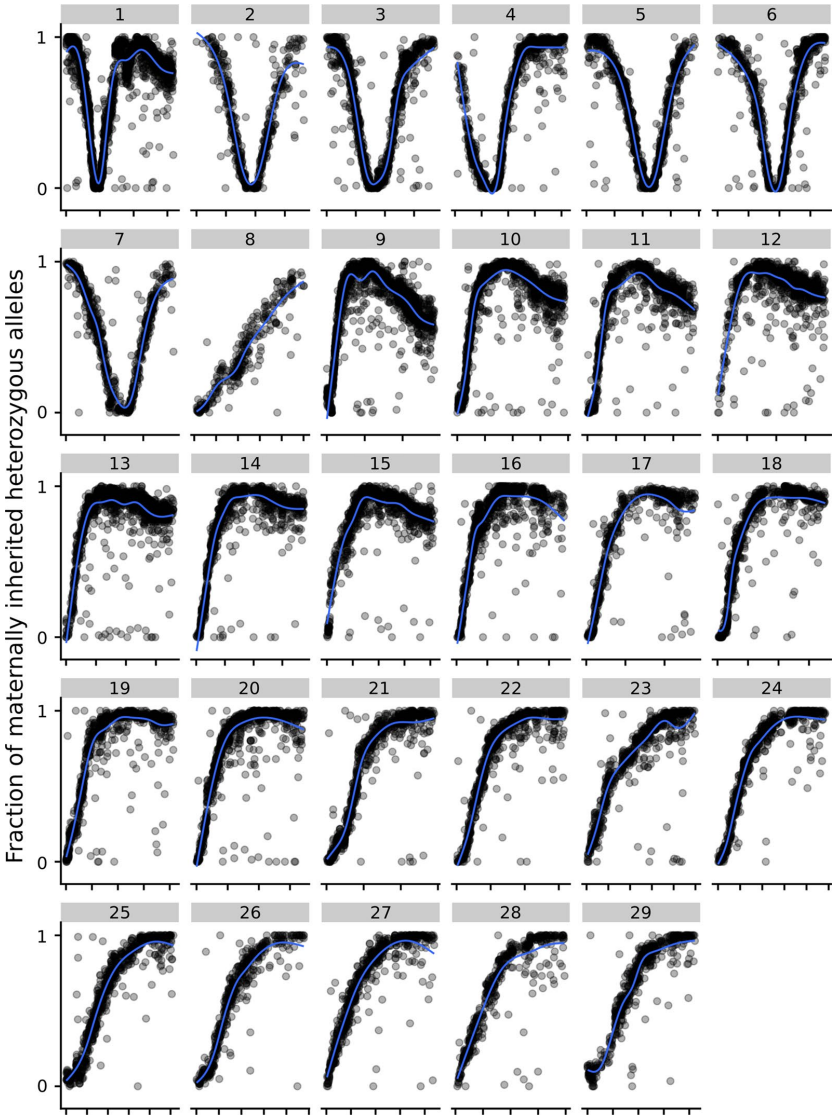


Fig. 7 Triploid maternal recombination events. The fraction of heterozygous (*A* and *B*) alleles inherited from mothers for informative loci along each of the 29 chromosomes in Atlantic salmon. The x-axis is marker position on each chromosome and is scaled by chromosome size. All markers are required to have at least 50 trios with informative genotypes

Although we did not use information on allele frequencies or parental genotypes, this information could be used to increase the accuracy of the genotype calling, as shown in [19]. However, we estimated starting values for the EM algorithm and priors for the cluster means based on a rough estimate of allele frequencies (see Appendix). Another approach could be to correct genotypes by using allele frequency information after genotype calling has

been performed. Unlike in [21], offspring and parents were not called using the same method in our dataset. If any bias in the calling procedures used by APT or mclust exists, it can lead to incorrectly assuming that there are fewer Mendelian inconsistencies when calling genotypes using APT in both parents and offspring, as opposed to calling parents with APT and triploid offspring with mclust. We have not investigated whether this is the case in our dataset. Although we focused on Atlantic salmon, the method can be extended to other polyploid species, e.g. in plants [1].

Calling genotypes with mclust and APT

Genotypes were called in both triploid and diploid offspring through Gaussian finite mixture modelling using the EM algorithm, implemented in the R package mclust. The R packages fitTetra 1.0 [19] and fitTetra 2.0 [21] were developed by the same group that developed the fitPoly package [18, 19], and they all use the same underlying EM-based algorithm for genotype calling. However, fitTetra 1.0/2.0 are limited to calling tetraploid genotypes only. Both fitTetra 1.0/2.0 [19, 21] and fitPoly [18, 19] assume a common variance for all genotype clusters for a marker by transforming the intensity signal ratios to obtain approximately constant variance. Our implementation runs two models for each marker, one model that assumes equal variance and another model that allows for heterogeneous variance of the clusters, where the model with the highest ICL value is assumed to be the best. With our implementation, it seems that both models are useful, as 68 and 49% of the markers had heterogenous cluster variance for diploids and triploids, respectively. The reason for the difference between rate of markers with heterogenous cluster variance between diploids and triploids is unknown, but it can be hypothesized that the increased overlap of clusters in triploids causes mclust to prefer the heterogenous cluster variance model. Both fitTetra 1.0 and 2.0 use the Bayesian information criterion (BIC) to score models. However, genotype clusters are well separated for markers with high quality, which harmonizes well with the ability of ICL to penalize models with a low degree of cluster separation [8]. Variation in DNA quality in a sample dataset can affect the signal intensities from DNA hybridizing with the probes on the SNP chip [22–24]. This can result in clusters being distributed non-Gaussian. The algorithm can fit additional Gaussian clusters to account for violations of model assumptions [25]. Hence, the number of fitted clusters can exceed the number of biological genotype groups. Biernacki et al. [8] showed that BIC tends to overestimate the number of clusters when the model fits the data poorly. Since genotype clusters are expected to have different contrast means (i.e. separated clusters), and since factors such as heterogeneous sample quality can result in model assumptions to be violated, we chose ICL over BIC as a model selection criterion.

All genotype calls were given an estimate of uncertainty based on the probability of the genotype belonging to another cluster than that with the highest probability. The threshold for this was set to 0.15, which means that all genotypes with an uncertainty >0.15 were no-calls (i.e. missing genotypes). Decreasing the threshold for uncertainty would decrease the call rate of each marker and, thus, the number of markers used in downstream analyses. In our opinion, results from the downstream analysis indicate that the genotype calling method was appropriate and gave reliable and trustworthy results.

Salmonids have been through several genome duplication events and their genomes are in a state of rediploidization from the last genome duplication event, which resulted in a tetraploid genome [13]. That is, different regions of the Atlantic salmon genome are still in a tetraploid state. When creating SNP chips for such a genome, it is necessary to ensure that the SNP is in a region of the genome that is not duplicated, or that the SNP is in a tetraploid region where only one of the homologues is polymorphic for the SNP ("semi-fixed"). Having markers targeting "semi-fixed" SNPs can complicate the calling procedure because shifts in contrasts can be observed. For example, for a "semi-fixed" SNP with possible genotypes *AAAA*/*AAAB*/*AABB*, the contrast for the marker targeting this SNP can be shifted towards the right (i.e. towards the '*A*'-allele). Furthermore, hybridization affinity between the probes on the SNP chip may not be equal for the *A*- and *B*-alleles, which can also result in shifts of contrasts for the genotype. The problem of duplicated regions and/or differences in allele affinity for hybridizing with the probe is expected to be worse in triploids. Tetraploid regions in diploids become hexaploid regions in triploids, potentially resulting in more severe shifts in allele affinity compared to normal diploids (or tetraploids). In addition, since triploids have two heterozygote groups, there is an elevated risk of overlap between the clusters (see Fig. 1). We did not investigate to what degree any of the SNPs on our chip were affected by such semi-fixed SNPs.

APT uses the BRLMM-P algorithm, which was developed by Affymetrix (now Thermo Fisher) [26]. Many elements of the BRLMM-P algorithm are similar to the current implementation of mclust, e.g. estimation of contrast cluster means and variances and use of priors. However, one key aspect that differentiates APT from the mclust implementation is the use of covariances between different cluster means [26]. Currently, software limitations prevent this from being implemented with mclust. Another difference is that APT provides

uncertainty estimates for each genotype call from markers that have only one genotype class (e.g. monomorphic markers within dataset), while mclust does not. Thus, no-calls are produced for such markers by APT but not by mclust. Although possible, we did not investigate implementing this in mclust. This could account for some of the increase in ER when the ΔICL threshold is low, since the fraction of markers with one cluster was increased (Fig. 2).

Figure 1 shows the distribution of estimated mean contrasts in each of the four genotype clusters for markers with three/four (for diploids/triploids) predicted clusters. Note that these distributions are across all markers (containing many different marker clusters skewed in different directions), and therefore there is more overlap between clusters than would be the case for individual markers. There was also some evidence for a widening of the contrast space from diploids to triploids, i.e. the contrast cluster means ranged from -1.76 ($=BB$) to 1.94 ($=AA$) for diploids and from -1.97 ($=BBB$) to 2.14 ($=AAA$) in triploids. All DNA extractions were normalized to the same concentration, so the reason for this difference in range is not known. However, normalization of DNA concentration is not only based on the initial concentration of DNA, but also on the concentration of other components such as protein, which may result in a higher final DNA concentration for the triploids. In any case, the contrast means, as expected, overlapped more in triploids than in diploids because of the two heterozygote clusters, making it more difficult to distinguish between clusters.

Figure 4 shows the distribution of ΔICL for markers with one to four predicted clusters for triploid individuals. The markers that had four predicted clusters seemed to be able to achieve higher ΔICL than what was possible with fewer predicted clusters. Markers with three or less biological clusters are expected to have low minor allele frequencies (MAF). For such markers, the number of observations within some of the clusters is likely very small, making estimates of cluster parameters less precise, and thus limiting ICL due to uncertain clustering.

APT uses pre-determined priors for means and variances, with equal priors for all markers as default. These priors were also used in the current study.

A small fraction of the induced triploids is expected to have failed triploidization. If any of the individuals assumed to be triploid are in fact diploid, calling accuracy is expected to decrease. However, the presence of diploids in the triploid dataset was deemed unlikely in this study, as inspection of (transformed) allele strength distributions revealed four distinct clusters for all individuals that were assumed to be triploids (see "Methods").

Comparing APT and mclust using exclusion ratios in known and assigned offspring–parent configurations

To compare our mclust implementation with APT, we called the same diploid offspring with both mclust and APT. In Fig. 2, it is clear that, without marker quality filtering, APT achieved lower Mendelian error rates compared to our implementation of mclust. When markers were filtered based on ΔICL, around 7 to 8000 markers had to be removed before Mendelian error rates achieved with mclust and APT were comparable. Figure 2 also shows that, without marker quality filtering, the percentage of missing genotypes (no-calls) was higher for mclust than for APT.

Knowing the parents' sex enabled us to identify more Mendelian exclusions for triploid offspring (see Table 2) than for diploid offspring (see Table 1). As a result, exclusion rates for diploid and triploid offspring could not be directly compared. Higher error rates are expected in triploids due to more overlapping genotype clusters (e.g. Fig. 1). Because of this, the genotype error rate (or e.g. ER) should be estimated separately for triploids.

We used APT to call the parents of both triploid and diploid offspring. Consequently, there may exist bias in favor of APT. Hence, mclust may appear to give more mismatches than APT between genotypes of offspring and parents (always called with APT), which would affect the estimated Mendelian error rate (ER) for both diploid- and triploid offspring.

Parentage assignment of diploid parents to triploid offspring

A threshold of $\Delta ICL > 150$ was chosen to retain high-quality markers for parentage assignment of diploid parents to triploid offspring. This arbitrarily large number was chosen to ensure that accurate parentage assignment was used in downstream analyses. Note that the threshold of $\Delta ICL > 150$ was only used for parentage assignment of the triploids, not in downstream analyses, where other thresholds were investigated and chosen. The fact that parentage could be assigned to a substantial number of triploid offspring with clear differences in exclusion rates for assigned parents compared to non-assigned parents is an indication that the calling of triploid genotypes was successful. Parent sex prediction, comparisons between our implementation and APT, and mapping of maternal recombinations, all depend on correct parentage assignment. This is another indication that both the triploid genotyping and parentage assignment were successful and accurate. Applying an ER-threshold of 0.002 worked well in this dataset but may not be applicable in all situations (it may depend on, e.g. the SNP chip, genotype errors, or relatedness between individuals in the sample). The ER-threshold should be set lower than

the minimum ER of the third most likely candidate for all duos (assuming that duplicates or clones of parental DNA are not present in the data). Furthermore, (visual) inspection of the ER distribution is required to locate the probable region of true parental ER's. Parentage assignment using high-density SNP genotypes and exclusions (opposing homozygotes) is frequently used for parentage assignment of diploid offspring (e.g. [27–29]). Parentage assignment in triploid Pacific oyster (*Crassostrea gigas*) offspring with diploid mothers and autotetraploid fathers was performed by Miller et al. [30] using microsatellite markers. Nonetheless, we are not aware of any case where triploid offspring have been assigned diploid parents using high-density SNP data.

Parent sex prediction

Accurately identifying sex in salmonids using genotypes is not trivial [31, 32]. In pressure-induced triploids, the fact that mothers contribute two alleles to their offspring and fathers one allele can be used to separate the already assigned parents into mothers and fathers. Two assigned "mothers" or "fathers" indicate a false assignment, either by incorrectly assuming triploidy in diploid offspring or by duplicated parental samples. In our analysis, all assigned parents were consistently assigned as either fathers or mothers across all triploid offspring.

Since the parent candidate dataset included closely-related individuals, several candidates were likely closely related with the true parents. Close relatives of the mother will have a high fraction of genotypes that resemble the genotypes of the true mother, which gives such candidates relatively low 'mother.ER', even compared with the true father (Fig. 6).

In the ER-based parentage assignment (Fig. 5), sex of the parent was not considered. Still, we observed some differences in ER between the sexes, with lower average ER for mother–offspring pairs. This may be explained by the fact that the mother contributes two alleles and the father one allele. For example, if the true mother has genotype *AA*, the triploid offspring can have genotypes *AAA* and *AAB*. In contrast, a true *AA* father can have triploid offspring with genotypes *AAA*, *AAB* and *ABB*. The latter genotype is more likely to be misinterpreted as *BBB* (see Fig. 1), generating a false exclusion genotype.

Maternal recombinations

Figure 7 shows an increase in maternal recombination rates when moving away from the predicted centromeric region for all 29 chromosomes [14]. By visual inspection, the centromeric regions for the most part aligned well with what was reported by Lien et al. [14]. However, chromosome 8 was reported to be metacentric in [14], while we observed it to be acrocentric or telocentric probably due to a lack of markers on the p-arm of chromosome 8, see "Results". Figure 7 shows that inheritance of maternal alleles was highly dependent on the distance between the locus and the centromere. For loci that are heterozygous in the mother and close to the centromere, the offspring usually inherited two identical maternal alleles, while for loci far from the centromere the offspring usually inherited two different maternal alleles. Thus, the inherited alleles are not expected to be in Hardy–Weinberg equilibrium. Estimating the number and position of recombinations is possible for each individual mother by searching for transitions from homozygous to heterozygous maternal inheritance.

Figure 7 shows the fraction of offspring that inherited heterozygous alleles from the mother at different positions along each chromosome (only informative genotypes were included, i.e. heterozygous mother and homozygous father). There were signs of interference for all chromosomes in Fig. 7. Under a model of no interference, secondary recombinations on the chromosome arms would frequently occur. Instead, all chromosomes showed a rapid increase in the fraction of heterozygous maternal alleles when moving away from the centromere, with little indication of secondary recombination (which would result in homozygous inheritance of maternal alleles). For some of the bigger acrocentric chromosomes, the maternally inherited heterozygous fraction approached 1 before it started to decline. This was most prominent for chromosome 9 and might suggest that interference was affected by distance from the last recombination event. Since induction of triploidy by pressurization occurs after prophase, the pattern of recombination should not be different when ordinary oocytes are formed for haploid inheritance of alleles.

Because this study focused on genotyping polyploids, and specifically triploids, further investigations on the implications of maternal recombinations were deemed outside the scope of the current study.

Application to other methods for creating triploid offspring

Other ways of producing triploid individuals are possible, such as mating tetraploids with diploids [33, 34]. In such cases, the methods used here for genotype calling and ER-based parentage assignment can still be used (given that genotypes can be called for the tetraploid parent), but the methods used here for parent sex

detection and mapping of recombination events are not necessarily applicable.

Conclusions

We have developed a technique for genotyping triploid individuals using allele signals from the Thermo Fisher GeneTitan genotyping platform, or other platforms that use light intensity for estimating the allele hybridization ratio. Using the called triploid genotypes, diploid parents could be assigned to induced triploid offspring and sex of the assigned parents could be predicted. No a priori information about the parents was needed, except their genotype information (not including any sex-linked markers). Furthermore, the genotypes of triploid offspring and their assigned parents were used to map maternal recombination events along the chromosomes. The methods and results of this study can be used for further genetic analyses (genomic prediction, genome-wide association studies) of phenotypic traits recorded in triploids as well as their genetic covariance with phenotypic traits recorded in diploids.

Acknowledgements
We wish to thank The Research Council of Norway, which through both the NAERINGSPHD (Project no. 251664) and the HAVBRUK2 (Project no. 245519) programs provided funding, as well as the breeding company AquaGen AS for funding and other support. We also want to thank the reviewers and the editor for helpful comments resulting in better algorithms and manuscript.

Authors' contributions
KEG wrote the software, performed the study and drafted the manuscript. JO and KEG conceived the methods for genotyping, estimation of starting parameters and parentage assignment of the triploids. JO conceived the methods for parent sex prediction and estimation of maternally inherited recombinations, coordinated the study and contributed in writing and revising the manuscript. THEM helped in revising the manuscript critically and contributed to the discussions regarding the methods and the underlying statistical theory. All authors read and approved the final manuscript.

Funding
The research leading to these results has received funding from The Research Council of Norway through the research programs NAERINGSPHD (Project no. 251664) and HAVBRUK2 (Project no. 245519). Additional funding was provided by AquaGen AS.

Availability of data and materials
The data that support the findings of this study are available from AquaGen AS but restrictions apply to the availability of these data, which were used under license for the current study, and thus are not publicly available. Data are, however, available from the authors upon reasonable request and with permission of AquaGen AS.

Ethics approval and consent to participate
Treatment of animals used in this study is in concordance with Norwegian laws and regulations.

Consent for publication
Not applicable.

Competing interests
KEG and JO are employed with AquaGen AS. The authors declare that they have no competing interests.

Author details
[1] AquaGen AS, P.O. Box 1240, 7462 Trondheim, Norway. [2] Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, P.O. Box 5003, 1432 Ås, Norway.

Appendix

Providing the mclust package informative starting proportions and prior mean parameter estimates

In the first iteration of the expectation maximization (EM) procedure, the overall mean for each variable is normally used as starting parameters for μ_k by mclust [7]. In our experience, this may lead to incorrect classifications, and informative starting parameters and priors can be used to increase classification accuracy. As with APT, the contrast variable was deemed the most informative with respect to genotype classification, and informative starting values and prior parameters were thus estimated for this variable.

Assuming that DNA is sampled from random individuals in a population, the number and size of clusters depend on the allele frequency of the marker. For example, if the allele frequency of allele A is 50%, in diploids, we would expect to see equally-sized clusters for the two opposing homozygotes (AA and BB) and a heterozygote AB cluster with twice the size of either the AA- or the BB-cluster. If the minor allele frequency is low, some genotype classes may be absent from the data, leading to a reduced number of clusters. As discussed in the main text, the maternal inheritance of alleles for induced triploid offspring depends on the maternal recombination rate. This may result in genotype groups not strictly adhering to Hardy–Weinberg equilibrium. However, it should still be better to have slightly imprecise starting estimates and priors compared to using the overall contrast mean as the starting point for all genotype clusters.

When estimating informative priors and starting parameters for mclust for a defined number of possible cluster classes (e.g. $G \in \{1, 2, 3, 4\}$ in triploids), the number of individuals in cluster class C (i.e. having genotype C) approximately follows a binomial distribution:

$$C \sim bin(G - 1, p), \tag{1}$$

where p is the success probability (affected by the allele frequency, but not necessarily equivalent to it). A priori, the p parameter is unknown, but can be roughly estimated from normalized contrast values (d_{norm}):

$$d_{norm} = \frac{d - d_{min}}{d_{max} - d_{min}}, \tag{2}$$

$$\hat{p} = mean(d_{norm}), \tag{3}$$

where d is the contrast value for the marker obtained from the Thermo Fisher genotyping platform for an

individual, d_{max} is the maximum and d_{min} is the minimum contrast values for the marker. Each SNP on the chip is a collection of probes. The light signal produced when the different alleles are hybridized with probes for a single marker on the SNP chip can reach maximum intensity when 100% of the probes are hybridized. However, it is not certain that 100% of the probes hybridize with alleles, and thus the contrast range may vary. By transforming the contrast range such that the normalized contrasts are in the range between 0 and 1, we are able to use $\hat{p}$ as the success parameter in a binomial distribution. From this, we can roughly estimate the number of genotypes in each genotype class as:

$$\hat{n}_c = Pr\left(C = c|G-1,\hat{p}\right)\cdot n, \tag{4}$$

where n is total number of individuals genotyped for this marker and $Pr\left(C = c|G-1,\hat{p}\right)$ is the binomial probability of observing c successes in $G-1$ trials with success probability $\hat{p}$.

For the marker in question, contrast values were then sorted from smallest to largest, and the first $\hat{n}_0$ observations were used to estimate μ_0, the next $\hat{n}_1$ observations were used to estimate μ_1, etc. Finally, $\hat{\mu}_d = \begin{bmatrix} \hat{\mu}_0 \\ \dots \\ \hat{\mu}_{G-1} \end{bmatrix}$ was used as both an input starting parameter vector and a prior by mclust for the contrast cluster means (corresponds to μ_k in [7]). A vector of $\left[\hat{n}_1/n,\dots,\hat{n}_c/n\right]$ was used as starting proportions. All initial and prior variance parameters (see text regarding Σ in [7]) were set to 0.06 since this is also the prior variance used by the APT software and should provide a good comparison of the methods. Both the means and the variances are updated in every iteration of the EM-algorithm, ending with the maximum likelihood estimates of the parameters. The models assume a normal distribution for each cluster with different means. For the models 'E' (equal) and 'V' (varying), the variances are either equal or different, respectively, for all clusters.

Received: 29 November 2018 Accepted: 11 March 2020
Published online: 18 March 2020

References

1. Bourke PM, Voorrips RE, Visser RGF, Maliepaard C. Tools for genetic studies in experimental populations of polyploids. Front Plant Sci. 2018;9:513.
2. Song C, Liu S, Xiao J, He W, Zhou Y, Qin Q, et al. Polyploid organisms. Sci China Life Sci. 2012;55:301–11.
3. Chester-Jones I, Ingelton PM, Phillips JG. Fundamentals of comparative vertebrate endocrinology, vol. xvi. 1st ed. New York: Plenum Press; 1987. p. 666.
4. Piferrer F, Beaumont A, Falguière JC, Flajšhans M, Haffray P, Colombo L. Polyploid fish and shellfish: production, biology and applications to aquaculture for performance improvement and genetic containment. Aquaculture. 2009;293:125–56.
5. Fisher T. GeneTitan multi-channel (MC) instrument. 2018. https://www.thermofisher.com/no/en/home/life-science/microarray-analysis/microarray-analysis-instruments-software-services/microarray-analysis-instruments/genetitan-multi-channel-instrument.html. Accessed 13 Mar 2020.
6. Fisher T. Affymetrix power tools. 2018. https://www.thermofisher.com/no/en/home/life-science/microarray-analysis/microarray-analysis-partners-programs/affymetrix-developers-network/affymetrix-power-tools.html. Accessed 13 Mar 2020.
7. Scrucca L, Fop M, Murphy TB, Raftery AE. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. R J. 2016;8:289–317.
8. Biernacki C, Celeux G, Govaert G. Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Trans Pattern Anal. 2000;22:719–25.
9. Fisher T. Axiom genotyping solution data analysis guide. https://assets.thermofisher.com/TFS-Assets/LSG/manuals/axiom_genotyping_solution_analysis_guide.pdf.
10. Brooker RJ, Widmaier EP, Graham LE, Stiling PD. Biology. 2nd ed. New York: McGraw-Hill Education; 2011.
11. Thorgaard GH, Allendorf FW, Knudsen KL. Gene-centromere mapping in Rainbow trout: high interference over long map distances. Genetics. 1983;103:771–83.
12. ICSASG_v2 NCBI Assembly: International cooperation to sequence the Atlantic salmon genome. 2015. https://www.ncbi.nlm.nih.gov/assembly/GCF_000233375.1/#/def_asm_Primary_Assembly. Accessed 13 Mar 2020.
13. Lien S, Koop BF, Sandve SR, Miller JR, Kent MP, Nome T, et al. The Atlantic salmon genome provides insights into rediploidization. Nature. 2016;533:200–5.
14. Lien S, Gidskehaug L, Moen T, Hayes BJ, Berg PR, Davidson WS, et al. A dense SNP-based linkage map for Atlantic salmon (*Salmo salar*) reveals extended chromosome homeologies and striking differences in sex-specific recombination patterns. BMC Genomics. 2011;12:615.
15. O'Flynn FM, McGeachy SA, Friars GW, Benfey TJ, Bailey JK. Comparisons of cultured triploid and diploid Atlantic salmon (*Salmo salar* L.). ICES J Mar Sci. 1997;54:1160–5.
16. Pereira GS, Garcia AAF, Margarido GRA. A fully automated pipeline for quantitative genotype calling from next generation sequencing data in autopolyploids. BMC Bioinformatics. 2018;19:398.
17. Clark LV, Lipka AE, Sacks EJ. polyRAD: genotype calling with uncertainty from sequencing data in polyploids and diploids. G3 (Bethesda). 2019;9:663–73.
18. Voorrips RE, Gort G. fitPoly: genotype calling for bi-allelic marker assays. Version 3.0.0. 2018. https://github.com/cran/fitPoly. Accessed 13 Mar 2020.
19. Voorrips RE, Gort G, Vosman B. Genotype calling in tetraploid species from bi-allelic marker data using mixture models. BMC Bioinformatics. 2011;12:172.
20. Serang O, Mollinari M, Garcia AA. Efficient exact maximum a posteriori computation for bayesian SNP genotyping in polyploids. PLoS One. 2012;7:e30906.
21. Zych K, Gort G, Maliepaard CA, Jansen RC, Voorrips RE. FitTetra 2.0—improved genotype calling for tetraploids with multiple population and parental data support. BMC Bioinformatics. 2019;20:148.
22. Miclaus K, Wolfinger R, Vega S, Chierici M, Furlanello C, Lambert C, et al. Batch effects in the BRLMM genotype calling algorithm influence GWAS results for the Affymetrix 500 K array. Pharmacogenomics J. 2010;10:336–46.
23. Clayton DG, Walker NM, Smyth DJ, Pask R, Cooper JD, Maier LM, et al. Population structure, differential bias and genomic control in a large-scale, case–control association study. Nat Genet. 2005;37:1243–6.
24. Hong H, Su Z, Ge W, Shi L, Perkins R, Fang H, et al. Assessing batch effects of genotype calling algorithm BRLMM for the Affymetrix GeneChip Human Mapping 500 K array set using 270 HapMap samples. BMC Bioinformatics. 2008;9:S17.
25. McLachlan GJ, Peel D. Finite mixture models. New York: Wiley; 2000.
26. Thermo Fisher. BRLMM-P: a genotype calling method for the SNP 5.0 array. 2007. https://www.google.fr/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&ved=2ahUKEwiqjLeUlannAhWQmBQKHSlbCvsQFjABegQ

lChAF&url=http%3A%2F%2Ftools.thermofisher.com%2Fcontent%2Fsfs%2Fbrochures%2Fbrlmmp_whitepaper.pdf&usg=AOvVaw1F_EnjOHCE1r6JCcGbFvbR. Accessed 13 Mar 2020.
27. Hayes BJ. Efficient parentage assignment and pedigree reconstruction with dense single nucleotide polymorphism data. J Dairy Sci. 2011;94:2114–7.
28. Grashei KE, Odegard J, Meuwissen THE. Using genomic relationship likelihood for parentage assignment. Genet Sel Evol. 2018;50:26.
29. Strucken EM, Lee SH, Lee HK, Song KD, Gibson JP, Gondro C. How many markers are enough? Factors influencing parentage testing in different livestock populations. J Anim Breed Genet. 2016;133:13–23.
30. Miller PA, Elliott NG, Vaillancourt RE, Koutoulis A, Henshall JM. Assignment of parentage in triploid species using microsatellite markers with null alleles, an example from Pacific oysters (*Crassostrea gigas*). Aquacult Res. 2016;47:1288–98.
31. Eisbrenner WD, Botwright N, Cook M, Davidson EA, Dominik S, Elliott NG, et al. Evidence for multiple sex-determining loci in Tasmanian Atlantic salmon (*Salmo salar*). Heredity. 2014;113:86–92.
32. Kijas J, McWilliam S, Naval Sanchez M, Kube P, King H, Evans B, et al. Evolution of sex determination loci in Atlantic salmon. Sci Rep. 2018;8:5664.
33. Chourrout D, Chevassus B, Krieg F, Happe A, Burger G, Renard P. Production of second generation triploid and tetraploid rainbow trout by mating tetraploid males and diploid females—potential of tetraploid fish. Theor Appl Genet. 1986;72:193–206.
34. Liu S. Distant hybridization leads to different ploidy fishes. Sci China Life Sci. 2010;53:416–25.

Publisher's Note

III Genomic prediction using a case-parental-control model

Grashei, K.E., Ødegård, J. & Meuwissen, T.H.E. Manuscript

Genomic prediction using a case-parental-control model

Kim E Grashei*[1,2], Jørgen Ødegård[1,2], Theo HE Meuwissen[2]

[1]AquaGen AS, P.O. Box 1240, NO-7462 Trondheim, Norway

[2]Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, P.O. Box 5003, NO-1432 Ås, Norway

*Corresponding author

E-mail addresses:

KEG: kim.erik.grashei@aquagen.no; kim.grashei@nmbu.no

JO: Jorgen.Odegard@aquagen.no; jorgen.odegard@nmbu.no

THEM: theo.meuwissen@nmbu.no

Abstract

Background

Performing genomic prediction has traditionally been restricted to using continuous or binary phenotypes. However, when searching for single QTLs across a genome, the transmission disequilibrium test (TDT) allows the use of individuals of a single category, i.e. only “cases”, in genome-wide association analysis. The TDT method is a case-parental-control method where deviations from neutral inheritance of alleles between parents and their offspring are used to construct a test statistic. The aim of this study was to develop a case-parental-control model that could be used for genomic prediction and to estimate heritability.

Results

A method called Transmission Disequilibrium Genomic Prediction (TDGP) was developed to estimate heritability and predict genomic breeding values for a binary trait where only the cases and their parents were genotyped. TDGP estimated highly accurate heritabilities for the schemes simulated in this study. The prediction accuracy of genomic breeding values was lower than for an ordinary case-control SNP-BLUP model when predicting individuals closely related with the training population. However, TDGP proved superior to ordinary SNP-BLUP when predicting less related individuals.

Conclusions

TDGP can be a useful method when genotyping is restricted to individuals from a single phenotypic category and their parents. Additionally, TDGP may be used when the training population is distantly related to breeding candidates.

Background

Genomic prediction methods have been successfully applied to binary traits (case-control data), e.g. [1-3]. In general, genomic prediction is typically based on a training data set consisting of genotyped and phenotyped individuals, where the trait analyzed must have measurable phenotypic variation as well as a non-zero heritability. Hence, for a binary trait, both categories (cases and non-cases/controls) must be present within the training data set. However, for some binary traits, targeted genotyping of cases may be more practical and cost-effective. To the best of our knowledge, no genomic prediction model has so far been proposed for case-only data. However, such data has for a long time been used in single-locus genome-wide association studies, through transmission disequilibrium testing (TDT), comparing (case) offspring genotypes to the genotypes of their respective parents [4].

For analysis of binary traits using ordinary statistical models (i.e. linear or generalized linear models), intermediate incidence is considered the most informative. However, some naturally occurring diseases or conditions may have extreme frequencies, e.g. very low (or in some cases very high) case incidence. Ideally, the proportion of cases included in training data should equal the actual case incidence, e.g. at 1% case incidence the training sample should consist of 1% cases and 99% non-cases (controls). This would imply that a very large number of genotyped individuals would be needed to achieve a sufficient number of cases. Other approaches include case-control models where similar numbers of cases and controls are targeted for genotyping, artificially increasing the incidence within the training data set. However, for disease outbreaks, it is not always straightforward to obtain representative/informative control samples. For infectious diseases, it may be easier to identify susceptible than non-susceptible individuals, as affected individuals (cases) are

obviously susceptible, while controls may be unexposed to the pathogen rather than non-susceptible. Consequently, controls may be a mix of truly non-susceptible individuals (exposed, but able to resist the pathogen), individuals in the early disease phase (may die/show symptoms at a later stage) and unexposed individuals. A sampling strategy targeting only identified cases may enable more cost-effective genomic prediction.

The aim of the current study was to develop and validate a novel genomic prediction (GP) method, called transmission disequilibrium genomic prediction (TDGP), using training information from cases and their parents only (i.e. no sampling of phenotypic controls). The results were compared with traditional GP using case-control data with a SNP-BLUP model.

Methods

Transmission disequilibrium null hypothesis (H_0)

A model for the transmission of an allele between a heterozygous parent and its offspring at a locus not affected by selection can be formulated as:

$$t_{ij} \sim Bernoulli(f = 0.5)$$

where t_{ij} is a binary indicator for whether or not the reference allele is transmitted for the parent - offspring duo i at locus j, and the success parameter $f = 0.5$ indicates that there is an equal probability of transmitting the reference allele and the alternative allele. Assuming this null-hypothesis distribution, $E(t_{ij}) = f = 0.5$ and $Var(t_{ij}) = f(1-f) = 0.25$. Note that under the null hypothesis, the maximum possible variance of t_{ij} is obtained since any other success parameter than $f = 0.5$ results in t_{ij} having a reduced variance. Over N_j duos the sum at locus j is $\sum_i^{N_j} t_{ij}$ where t_{ij} is a transmission for duo i at locus j. Consequently,

$E\left(\sum_i^{N_j} t_{ij}\right) = \sum_i^{N_j} E(t_{ij}) = \sum_i^{N_j} f = 0.5N_j$ and $Var(\sum_i^{N_j} t_{ij}) = \sum_i^{N_j} Var(t_{ij}) =$ $\sum_i^{N_j} f(1-f) = 0.25N_j$.

Transmission disequilibrium under selection

Assume in the following that genotyped offspring are those that have a certain binary phenotype, i.e. genotyping is restricted to offspring affected by some condition or disease (cases). When a locus is in linkage disequilibrium (LD) with a quantitative trait locus (QTL) affecting the phenotype (risk of condition or disease), the null hypothesis of neutral inheritance is violated and $f \neq 0.5$ within the case-offspring group. Under such circumstances, the variance of t_{ij} is reduced. Consider the following model which includes selection pressure:

$$t_{ij} \sim Bernoulli(f_j = 0.5 + s_j)$$

where s_j is the expected deviation from neutral inheritance at locus *j*. Thus, offspring inheriting the reference allele at locus *j* may have a higher/lower probability of ending up as a "case", causing a deviation in probability of observing the reference allele in case-offspring. As an example, assume that the locus is in high LD with a QTL such that $s_j = -0.2$. Then, $E(t_{ij}) = 0.5 - 0.2 = 0.3$. In the following we show how such deviations from neutral inheritance from parents to case offspring can be used to predict marker effects.

SNP-BLUP for a binary trait

Consider first the ideal situation where all offspring (cases and non-cases) in a training population are genotyped. The genotypes are centered by training population frequencies as:

$$\mathbf{M} = \mathbf{Q} - 2 \cdot \mathbf{1}\mathbf{p}'$$

where $\mathbf{Q}$ is the genotype matrix consisting of values 0, 1 or 2, i.e. the number of reference alleles for individual *i* at locus *j*, $\mathbf{p}$ is a vector of allele frequencies and $\mathbf{1}$ is a vector of ones matching the dimension of $\mathbf{p}$. Then, a linear SNP-BLUP model is:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{M}\mathbf{m} + \mathbf{e}$$

where $\mathbf{y}$ is a vector of binary phenotypes (1 = case, 0 = non-case), $\mathbf{b}$ is a vector of fixed effects, $\mathbf{m}$ is a vector of random marker effects, $\mathbf{e}$ is a vector of random residuals, $\mathbf{X}$ is an incidence matrix associating phenotypes with fixed effects (e.g. overall mean, temperature, location, person sampling etc.) and $\mathbf{M} = \begin{bmatrix} \mathbf{M_c} \\ \mathbf{M_{nc}} \end{bmatrix}$ is a matrix of genotypes, centered by the population allele frequencies, where $\mathbf{M_c}$ is the centered genotype matrix for the cases and $\mathbf{M_{nc}}$ is the centered genotypes for the non-cases. Then, the corresponding mixed-model equation system is:

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{M} \\ \mathbf{M}'\mathbf{X} & \mathbf{M}'\mathbf{M} + \mathbf{I}\lambda \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \widehat{\mathbf{m}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\hat{\mathbf{y}} \\ \mathbf{M}'\hat{\mathbf{y}} \end{bmatrix}$$

where $\lambda = \frac{\sigma_e^2}{\frac{\sigma_g^2}{\rho}} = \rho\frac{\sigma_e^2}{\sigma_g^2}$, $\rho = 2\sum p_j(1 - p_j)$, p_j is the allele frequency at locus *j*, σ_e^2 is the residual variance and σ_g^2 is the total genetic variance (i.e. $\frac{\sigma_g^2}{\rho}$ is the marker effect variance). Since genotypes are centered within the training population, we may assume $E(\mathbf{X}'\mathbf{M}) = \mathbf{0}'$ and $E(\mathbf{M}'\mathbf{X}) = \mathbf{0}$, which assumes independence between the fixed effect covariates and the marker genotypes. If a simple fixed effect structure is fitted such as just the overall mean, resulting in $\mathbf{X} = \mathbf{1}$, the matrix products are $\mathbf{1}'\mathbf{M} = \mathbf{0}'$ and $\mathbf{M}'\mathbf{1} = \mathbf{0}$. Under the assumption of such, or a similar simple fixed effect structure, the marker effects can be predicted

independently from fixed effects. Further, because non-cases have phenotypes 0 and cases phenotypes 1, the following applies:

$$\mathbf{M}'\mathbf{y} = \begin{bmatrix}\mathbf{M_c}\\ \mathbf{M_{nc}}\end{bmatrix}' \begin{bmatrix}\mathbf{y}_{\mathrm{c}}\\ \mathbf{y}_{\mathrm{nc}}\end{bmatrix} = \mathbf{M_c'1} + \mathbf{M_{nc}'0} = \mathbf{M_c'1} = \begin{bmatrix}\sum_{i=1}^{N_c} m_{i1}\\ \vdots \\ \sum_{i=1}^{N_c} m_{iL}\end{bmatrix} = \begin{bmatrix}\sum_{i=1}^{N_c} q_{i1} - 2\sum_{i=1}^{N_c} p_1\\ \vdots \\ \sum_{i=1}^{N_c} q_{iL} - 2\sum_{i=1}^{N_c} p_L\end{bmatrix}$$

$$= \begin{bmatrix}N_c\bar{q}_1 - 2N_c p_1\\ \vdots \\ N_c\bar{q}_L - 2N_c p_L\end{bmatrix} = \begin{bmatrix}N_c 2p_{1_c} - 2N_c p_1\\ \vdots \\ N_c 2p_{L_c} - 2N_c p_L\end{bmatrix} = 2N_c \begin{bmatrix}p_{1_c} - p_1\\ \vdots \\ p_{L_c} - p_L\end{bmatrix} = 2N_c(\mathbf{p_c} - \mathbf{p})$$

where q_{ij} is the genotype (either 0, 1 or 2) for individual i at locus j, $m_{ij} = q_{ij} - 2p_j$ is the genotype adjusted for the training population allele frequency, p_j, at locus j, $\bar{q}_j$ is the mean genotype among cases at locus j, N_c is the number of cases, L is the number of loci, $\mathbf{p_c}$ is a vector of allele frequencies of cases, while $\mathbf{p}$ is a vector of allele frequencies for the entire training population (cases and non-cases). A simplified equation system may thus be used to compute the marker effects for a binary trait when assuming independence between fixed effects and the marker genotypes:

$$[\mathbf{M}'\mathbf{M} + \mathbf{I}\lambda]\hat{\mathbf{m}} = 2N_c(\mathbf{p_c} - \mathbf{p})$$

Thus, by defining case-phenotypes as 1 and non-case phenotypes as 0, the non-case genotypes do not contribute to the right-hand side of the equation system. They do, however, contribute to the left-hand-side through $\mathbf{M}'\mathbf{M}$, which has expectation (see proof 1 in Appendix):

$$E(\mathbf{M}'\mathbf{M}) = 2N \cdot \mathbf{JCJ}$$

where N is the number of individuals in the training population (cases + non-cases), $\mathbf{C}$ is a genotype correlation matrix between different loci and $\mathbf{J}$ is a diagonal matrix with elements (see proof 1 in Appendix):

$$J_{jj} = \sqrt{p_j(1-p_j)}$$

where p_j is the corresponding allele frequency from the vector $\mathbf{p}$. Hence, given $\mathbf{p}$ and $\mathbf{C}$, the entire equation system above may thus be approximated as:

$$[2N \cdot \mathbf{JCJ} + \mathbf{I}\lambda]\widehat{\mathbf{m}} \approx 2N_c(\mathbf{p}_c - \mathbf{p})$$

Consequently, genomic prediction may be performed based on genotyped cases only (must be genotyped to compute $\mathbf{p}_c$), given that $\mathbf{p}$ (for the entire training population) and the correlation structure among genotypes for different loci can be assumed known. $\mathbf{X}'\mathbf{M}$ cannot be calculated when only cases are genotyped (since the non-cases are missing), and thus equations for fixed and markers effects must be assumed independent. This is only expected to be true if $\mathbf{p}$ is correct and under simple fixed effect structures with independence between the fixed effect classification and the marker genotypes. The computed marker effects are functions of $(\mathbf{p}_c - \mathbf{p})$. Any error in $\mathbf{p}$ will thus have a substantial effect on both the sign and the size of the estimated marker effect. Consequently, the most critical assumption for predicting genomic breeding values when genotyping only the cases and their parents is that $\mathbf{p}$ can be properly estimated, e.g. by using parental allele frequencies adjusted for the different family sizes. In contrast, an ordinary genomic prediction model is more robust as it estimates fixed- and marker effects jointly, i.e. the left-hand-side mixed model matrix blocks $\mathbf{X}'\mathbf{M}$ and $\mathbf{M}'\mathbf{X}$ are computed rather than being assumed zero. However,

the latter model requires more widespread genotyping (both cases and non-cases) to calculate **X'M**.

Transmission Disequilibrium Genomic Prediction (TDGP)

When just the cases are genotyped, the training population allele frequencies in $\mathbf{p}$ (including cases and non-cases) cannot be correctly estimated without prior knowledge of the size of each family in the training population (not only among cases). Hence, the centered genotype matrix cannot be computed. Here we assume parents are genotyped and can be assigned (or their genotypes are inferred from their offspring genotypes). In this case, genotypes can still be centered by their parental expectations:

$$\mathbf{T} = \mathbf{Q} - \frac{1}{2}(\mathbf{Z}_\mathrm{s} + \mathbf{Z}_\mathrm{d})\mathbf{Q}_\mathrm{p}$$

where $\mathbf{Q}$ (offspring x SNPs) and $\mathbf{Q}_\mathrm{p}$ (parents x SNPs) are matrices of genotypes for offspring (cases and non-cases) and their parents, respectively, $\mathbf{Z}_\mathrm{s}$ (offspring x parents) and $\mathbf{Z}_\mathrm{d}$ (offspring x parents) are appropriate incidence matrices for sires and dams, respectively, and $\mathbf{T}$ (offspring x SNPs) is a matrix where each element is the deviation in number of offspring reference alleles from what is expected given the parent genotypes. For simplification of the notation, we use 'sire' and 'dam'. Although both parents must be represented, knowing the sex of the parents is not required by the method. Using $\mathbf{T}$ instead of $\mathbf{M}$ implies that between-family genetic variation is ignored, and thus enters the residual, while within-family genetic variation is captured by the marker genotypes (centered by parental means). Consider first an ideal situation where all offspring (and their parents) are genotyped:

$$\mathbf{y} = \mathbf{Xb} + \frac{1}{2}(\mathbf{Z}_\mathrm{s} + \mathbf{Z}_\mathrm{d})\mathbf{u} + \mathbf{Tm} + \mathbf{e}$$

where $\mathbf{u} \sim N(\mathbf{0}, \mathbf{G}_{\mathrm{sd}}\sigma_g^2)$ is a vector of parental genetic effect values (breeding values), $\mathbf{G}_{\mathrm{sd}}$ is a genomic relationship matrix for the parents and the other parameters are as defined above.

The covariance in breeding values between individuals in a family is equal to the between-family variance [5]. Thus, for two full-sibs $BV_1 = 0.5(BV_s + BV_d) + W_1$ and $BV_2 = 0.5(BV_s + BV_d) + W_2$ where BV_1 and BV_2 are true breeding values for the two full-sibs, BV_s and BV_d are, respectively, the true breeding values of their sire and dam, and W_1 and W_2 are the (within-family) Mendelian sampling deviations. The Mendelian sampling deviations are independent both within and across families. Assuming that the parents are not related, i.e. $Cov(BV_s, BV_d) = 0$, the covariance between the full sib breeding values reduces to $Cov(BV_1, BV_2) = \sigma_{bf}^2 = 0.5\sigma_g^2$, where σ_g^2 is the total additive genetic variance, and σ_{bf}^2 is the between-family variance. The Mendelian sampling deviations in the above equations are responsible for the within-family variance, i.e. the random segregation of alleles from heterozygous parents to their offspring. So, in absence of selection, $\sigma_{wf}^2 = \sigma_{bf}^2 = \frac{1}{2}\sigma_g^2$, where σ_{wf}^2 and σ_{bf}^2 are the within-family and between-family genetic variance, respectively. The equation system can be written as:

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\frac{1}{2}(\mathbf{Z}_{\mathrm{s}} + \mathbf{Z}_{\mathrm{d}}) & \mathbf{X}'\mathbf{T} \\ \frac{1}{2}(\mathbf{Z}_{\mathrm{s}} + \mathbf{Z}_{\mathrm{d}})'\mathbf{X} & \frac{1}{4}(\mathbf{Z}_{\mathrm{s}} + \mathbf{Z}_{\mathrm{d}})'(\mathbf{Z}_{\mathrm{s}} + \mathbf{Z}_{\mathrm{d}}) + \mathbf{G}_{\mathrm{sd}}^{-1}\lambda\rho^{-1} & \frac{1}{2}(\mathbf{Z}_{\mathrm{s}} + \mathbf{Z}_{\mathrm{d}})'\mathbf{T} \\ \mathbf{T}'\mathbf{X} & \mathbf{T}'\frac{1}{2}(\mathbf{Z}_{\mathrm{s}} + \mathbf{Z}_{\mathrm{d}}) & \mathbf{T}'\mathbf{T} + \mathbf{I}\lambda \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \\ \hat{\mathbf{m}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \frac{1}{2}(\mathbf{Z}_{\mathrm{s}} + \mathbf{Z}_{\mathrm{d}})'\mathbf{y} \\ \mathbf{T}'\mathbf{y} \end{bmatrix}$$

where λ and ρ are as defined for the ordinary SNP-BLUP model above. Marker variance $\left(\frac{\sigma_g^2}{\rho}\right)$ is still the same, since the allele substitution effects within and across families are assumed to be equal (but genotypes in **T** has less variance than genotypes in **M**). If the individuals included in $\mathbf{T}$ (i.e. cases and non-cases) are an unselected sample, centering genotypes using

parental means imply that $E(\mathbf{T}'\mathbf{X}) = \mathbf{0}$ and $E\left(\mathbf{T}'\frac{1}{2}(\mathbf{Z}_\mathrm{s} + \mathbf{Z}_\mathrm{d})\right) = \mathbf{0}$, even though the allele frequencies in $\mathbf{p}$ are unknown. The marker effect equations above may be approximated as:

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\frac{1}{2}(\mathbf{Z}_\mathrm{s} + \mathbf{Z}_\mathrm{d}) & \mathbf{X}'\mathbf{T} \\ \frac{1}{2}(\mathbf{Z}_\mathrm{s} + \mathbf{Z}_\mathrm{d})'\mathbf{X} & \frac{1}{4}(\mathbf{Z}_\mathrm{s} + \mathbf{Z}_\mathrm{d})'(\mathbf{Z}_\mathrm{s} + \mathbf{Z}_\mathrm{d}) + \mathbf{G}_\mathrm{sd}^{-1}\lambda\rho^{-1} & \frac{1}{2}(\mathbf{Z}_\mathrm{s} + \mathbf{Z}_\mathrm{d})'\mathbf{T} \\ \mathbf{T}'\mathbf{X} & \mathbf{T}'\frac{1}{2}(\mathbf{Z}_\mathrm{s} + \mathbf{Z}_\mathrm{d}) & \mathbf{T}'\mathbf{T} + \mathbf{I}\lambda \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \\ \hat{\mathbf{m}} \end{bmatrix}$$

$$\xrightarrow{E(\mathbf{X}'\mathbf{T})=\mathbf{0},\ E\left(\frac{1}{2}(\mathbf{Z}_\mathrm{s}+\mathbf{Z}_\mathrm{d})'\mathbf{T}\right)=\mathbf{0}} \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\frac{1}{2}(\mathbf{Z}_\mathrm{s} + \mathbf{Z}_\mathrm{d}) & \mathbf{0} \\ \frac{1}{2}(\mathbf{Z}_\mathrm{s} + \mathbf{Z}_\mathrm{d})'\mathbf{X} & \frac{1}{4}(\mathbf{Z}_\mathrm{s} + \mathbf{Z}_\mathrm{d})'(\mathbf{Z}_\mathrm{s} + \mathbf{Z}_\mathrm{d}) + \mathbf{G}_\mathrm{sd}^{-1}\lambda\rho^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{T}'\mathbf{T} + \mathbf{I}\lambda \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \\ \hat{\mathbf{m}} \end{bmatrix}$$

Assuming that marker effect equations are independent of both fixed and parental equations, the marker effects can thus be estimated with the following simplified equation system:

$$[\mathbf{T}'\mathbf{T} + \mathbf{I}\lambda]\hat{\mathbf{m}} = \mathbf{T}'\mathbf{y}$$

The right-hand side of the equation system is now:

$$\mathbf{T}'\mathbf{y} = \begin{bmatrix} \mathbf{T_c} \\ \mathbf{T_{nc}} \end{bmatrix}' \begin{bmatrix} \mathbf{y}_\mathrm{c} \\ \mathbf{y}_\mathrm{nc} \end{bmatrix} = \mathbf{T_c}'\mathbf{1} + \mathbf{T_{nc}}'\mathbf{0} = \mathbf{T_c}'\mathbf{1} = \begin{bmatrix} \sum_{i=1}^{N_c} T_{c_{i1}} \\ \vdots \\ \sum_{i=1}^{N_c} T_{c_{iL}} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{N_c} \left[q_{i1} - \frac{1}{2}\left(q_{d_i1} + q_{s_i1}\right)\right] \\ \vdots \\ \sum_{i=1}^{N_c} \left[q_{iL} - \frac{1}{2}\left(q_{d_iL} + q_{s_iL}\right)\right] \end{bmatrix}$$

where $\mathbf{T}_\mathrm{c}$ and $\mathbf{T}_\mathrm{nc}$ are matrices with genotypes adjusted for parental expectation, $\mathbf{y}_\mathrm{c} = \mathbf{1}$ and $\mathbf{y}_\mathrm{nc} = \mathbf{0}$ are vectors of ones and zeros, respectively, with the number of rows equal to the number of cases and non-cases, respectively, q_{ij} is genotype (0, 1 or 2) for offspring (case) i at locus j, q_{d_ij} and q_{s_ij} are, respectively, the genotypes of the dam and sire of

offspring i at locus j, L is total number of loci and N_c is number of cases. The right-hand side of the equation system can thus be set up exactly, solely from case-genotypes.

At this point, it is important to recognize that the (original, uncentered) genotype for offspring i at locus j, i.e. q_{ij}, depends on the genotypes of its dam and sire, q_{d_ij} and q_{s_ij}, respectively:

$$q_{ij} = h_{s_ij} + h_{d_ij}$$

where h_{s_ij} and h_{d_ij} are the haploid genotypes of the sire and dam gametes of offspring *i* at locus *j*, i.e. the number of reference alleles being transmitted from the dam and sire are:

$$h_{d_ij} \sim Bernoulli\left(\frac{q_{d_ij}}{2}\right)$$

$$h_{s_ij} \sim Bernoulli\left(\frac{q_{s_ij}}{2}\right)$$

Hence, for parental genotypes being 1, the Bernoulli probability of the gamete genotype is ½, while for parental genotypes being 0 or 2, Bernoulli probabilities are 0 and 1, respectively (i.e. the resulting gamete genotype is given). The expectation of the offspring genotype is the average of the parental genotypes:

$$E\left(q_{ij} \middle| q_{d_ij}, q_{s_ij}\right) = \frac{1}{2}\left(q_{d_ij} + q_{s_ij}\right)$$

Hence, the realized offspring genotype can also be written as:

$$q_{ij} = \frac{1}{2}\left(q_{d_ij} + q_{s_ij}\right) + T_{ij}$$

where T_{ij} is a centered random Binomial variate:

$$T_{ij} \sim Binomial\left(trials = q_{d_ij}\left(2 - q_{d_ij}\right) + q_{s_ij}\left(2 - q_{s_ij}\right), prob = \frac{1}{2}\right) - \frac{1}{2}\left(q_{d_ij}\left(2 - q_{d_ij}\right) + q_{s_ij}\left(2 - q_{s_ij}\right)\right)$$

The T_{ij} equals element (*i, j)* of the **T** matrix. Note that $q_{d_ij}(2 - q_{d_ij}) + q_{s_ij}(2 - q_{s_ij})$ is the number of heterozygote parents for offspring *i* at locus *j*. To make an analogy to the TDT-method [4], we may also express T_{ij} as:

$$T_{ij} = \frac{1}{2}(t_{ij} - n_{ij})$$

where t_{ij} equals number of reference alleles being transferred from heterozygous parent(s) to offspring *i* at locus *j*, while n_{ij} is the corresponding number for the alternative alleles (see Supplementary Material for a comparison of TDT and TDGP). For offspring of two homozygous parents, $t_{ij} = n_{ij} = 0$. Thus:

$$\mathbf{T}'\mathbf{y} = \mathbf{T}_\mathbf{c}'\mathbf{1} = \begin{bmatrix} \sum_{i=1}^{N_c} T_{i1} \\ \vdots \\ \sum_{i=1}^{N_c} T_{iL} \end{bmatrix} = \frac{1}{2}\begin{bmatrix} \sum_{i=1}^{N_c} (t_{i1} - n_{i1}) \\ \vdots \\ \sum_{i=1}^{N_c} (t_{iL} - n_{iL}) \end{bmatrix} = \tfrac{1}{2}(\mathbf{t} - \mathbf{n})$$

where $\mathbf{t}$ and $\mathbf{n}$ are, respectively vectors (one entry per locus) of the total numbers of reference and alternative alleles being transferred from heterozygous parents to case-offspring. Assuming that elements of **T** are centered, the matrix product $\mathbf{T}'\mathbf{T}$ is expected to be proportional to a scaled covariance matrix, which has an expectation (see proof 2 in Appendix):

$$E(\mathbf{T}'\mathbf{T}) = \frac{1}{2}E(\mathbf{M}'\mathbf{M}) = N \cdot \mathbf{JCJ}$$

where N is number of individuals (cases + non-cases), $\mathbf{C}$ is a correlation matrix (across-loci genotype correlations) and $\mathbf{J}$ is a diagonal matrix with elements:

$$J_{jj} = \sqrt{p_j(1 - p_j)}$$

where J_{jj} is the expected (parent-corrected) genotype standard deviation at locus *j*. Since only a fraction of the rows in $\mathbf{T}$ may be known (cases only), we use the approximation: $\mathbf{T}'\mathbf{T} \approx N \cdot \mathbf{JCJ}$. The equation system is then further simplified:

$$[N \cdot \mathbf{JCJ} + \mathbf{I}\lambda]\widehat{\mathbf{m}} = \tfrac{1}{2}(\mathbf{t} - \mathbf{n})$$

This requires that the elements of **J** (standard deviations) and correlation structure among loci (**C**) can be estimated. Each diagonal element of **J** is:

$$J_{jj} = \sqrt{p_j(1 - p_j)} = \sqrt{\tfrac{1}{2}\eta_j}$$

i.e. J_{jj} is a function of the expected fraction of heterozygous parent-offspring duos under Hardy-Weinberg equilibrium:

$$\eta_j = 2p_j(1 - p_j)$$

If parental allele frequency p_j is unknown, then η_j is also unknown. Furthermore, family size and families not represented among the cases may be unknown, and the fraction of parent-offspring duos (among cases and non-cases) where the parent is heterozygous may also be unknown. However, assuming that case-offspring have approximately the same fraction of heterozygous parents as the entire training population, η_j may be estimated from the cases only:

$$\hat{\eta}_j = \frac{t_j + n_j}{2N_c} = \frac{\#duos\ among\ cases\ where\ parent\ is\ heterozygous\ at\ locus\ j}{\#duos\ among\ cases}$$

where t_j and n_j are, as above, the number of reference and alternative alleles transmitted from a heterozygous parent to case-offspring at locus *j*, N_c is the number of case-offspring

and $2N_c$ is thus the number of parent-offspring duos among case-offspring. The diagonal elements in **J** can then be estimated by the following equation:

$$J_{jj} \approx \sqrt{\frac{1}{2}\hat{\eta}_j} = \sqrt{\frac{t_j + n_j}{4N_c}}$$

Because 'cases' is a selected sample, the average heterozygosity of their parents may to some extent deviate from the heterozygosity of the parent population as a whole, especially around loci having a major effect on the trait. However, for complex traits, this effect is likely rather small.

In large populations, only a fraction of the cases may be genotyped. If so, we may estimate: $N = \frac{N_c}{p_{cr}}$, which is required in the above mixed model equations, and where N_c is the actual number of genotyped cases, and p_{cr} is the fraction of cases in the population. If the latter is incorrectly assessed, the estimated λ is likely to change. Thus, for proper interpretation of heritability from the estimated λ, appropriate scaling of N is essential. For example, if N is set to half the real value (i.e. $N \cdot \mathbf{JCJ} \approx \frac{1}{2}\mathbf{T}'\mathbf{T}$), we may estimate $\hat{\lambda} \approx \frac{1}{2}\lambda$. Hence, heritability (=1/(λ/ρ+1)) will be overestimated, resulting in a general downscaling of the left hand-side of the equation system, which is expected to have a scaling effect on the solutions, while ranking of individuals remains the same. This also applies in ordinary case-control models, where controls (non-cases) may be massively under-sampled, leading to downscaling of the entire left-hand side of the equation system (including λ, implying overestimated heritability unless corrected for).

Among cases, loci in LD with QTL affecting the underlying trait may differ in frequency compared with the original parental genotypes. This does not, however, imply that these loci

292 are linked or in LD with each other in the overall population. The Pearson correlation
293 coefficients between genotypes of cases was used to estimate the elements in $\mathbf{C}$. General
294 changes in allele frequency from parents to case-offspring will be corrected for and thus not
295 contribute to the estimated correlation structure.

296 **Estimating heritability for case-only data using the Method R algorithm**

297 We are not aware of any software packages that can use case-only data to estimate
298 heritability for binary traits. However, marker effects can be estimated using the model
299 above, which combined with the "method R" algorithm [6] can be used for estimating the
300 variance ratio (λ), and thus the heritability. This is done by regressing the predicted random
301 effects from a dataset on the predicted random effects from a subsample of the same
302 dataset. The method relies on the fact that, when the variance ratio is correct, the expected
303 covariance between SNP effects using all individuals and a reduced subset of individuals
304 equals the variance of the SNP effects when only using the smaller subset of individuals [7]:

305 $$E\left(Var\left(\widehat{\boldsymbol{m}}_s|\hat{\lambda}=\lambda\right)\right)=E\left(Cov\left(\widehat{\boldsymbol{m}},\widehat{\boldsymbol{m}}_s|\hat{\lambda}=\lambda\right)\right)$$

306 where $\widehat{\boldsymbol{m}}$ and $\widehat{\boldsymbol{m}}_s$ are vectors of predicted SNP effects using all- and a subset of the
307 individuals, respectively, λ is the true variance ratio and $\hat{\lambda}$ is the variance ratio assumed in
308 the equation system. The Method R statistic is:

309 $$R_c=\frac{Cov(\widehat{\boldsymbol{m}},\widehat{\boldsymbol{m}}_s)}{Var(\widehat{\boldsymbol{m}}_s)}$$

310 where R_c is a statistic indicating whether the variance ratio ($\hat{\lambda}$) is set too high or too low.
311 More specifically, $R_c>1$ indicates that $\hat{\lambda}$ should be decreased, while $R_c<1$ indicates that
312 $\hat{\lambda}$ should be increased. An $R_c=1$ indicates that $\hat{\lambda}$ is correct. Using Method R, we estimate λ

iteratively by building the equation system using all case-individuals and a random subsample of case-individuals and predicting $\widehat{\boldsymbol{m}}$ and $\widehat{\boldsymbol{m}}_s$ for each iteration. As noted by Reverter *et al.* in [6] (Eq. 4), $Cov(\widehat{\boldsymbol{m}}, \widehat{\boldsymbol{m}}_s)$ achieves the same mathematical form as Restricted Maximum Likelihood (REML) uses to solve for variance components. Thus, the most likely estimate of λ, and the maximum accuracy, is expected when $R_c = 1$.

Define, as above, $\rho = 2\sum p_j(1 - p_j)$, where p_j is the allele frequency at locus j (e.g., among unselected selection candidates, or among case-parents as used in this study, see below). The total additive genetic variance can then be defined as:

$$\sigma_g^2 = \rho\sigma_m^2$$

where σ_m^2 is the variance of marker effects, which is the same under both ordinary SNP-BLUP and TDGP. The heritability is defined as:

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2} = \frac{\rho\sigma_m^2}{\rho\sigma_m^2 + \sigma_e^2}$$

For an ordinary GBLUP model, the lambda value is:

$$\lambda_a = \frac{(1 - h^2)}{h^2}$$

For the ordinary SNP-BLUP model and TDGP the lambda value is:

$$\lambda = \rho\lambda_a = \frac{\rho(1 - h^2)}{h^2}$$

$$h^2 = \frac{\rho}{\lambda + \rho}$$

An iteration strategy using a starting overall additive heritability of 0.5 was used to test different $\hat{\lambda}$, i.e. we start at $\hat{h}^2 = 0.5$, with $\hat{\lambda} = \frac{\rho(1-\hat{h}^2)}{\hat{h}^2} = \frac{\rho(1-0.5)}{0.5} = \rho$. If $R_c < 1$, the next

iteration will solve for marker effects using the $\hat{\lambda}$ where the current $\hat{h}^2$ is decremented by $\frac{1}{2}\hat{h}^2$, and if $R_c > 1$, the next iteration will solve for marker effects using the $\hat{\lambda}$ where the current $\hat{h}^2$ is incremented by $\frac{1}{2}\hat{h}^2$. When $R_c = 1 \pm 0.001$ the iteration stops, and convergence is achieved. This iteration strategy was repeated 10 times using a subset of 4k SNPs to reduce compute time, and for each repetition a random half of the cases were removed to form the subset to calculate SNP effects $\widehat{\boldsymbol{m}}_{s1}$. In addition, the other half of the cases were, in the same iteration, used as the subset for calculating another set of SNP effects, $\widehat{\boldsymbol{m}}_{s2}$. Subsequently, we calculated $R_{c1} = \frac{Cov(\widehat{\boldsymbol{m}},\widehat{\boldsymbol{m}}_{s1})}{Var(\widehat{\boldsymbol{m}}_{s1})}$, $R_{c2} = \frac{Cov(\widehat{\boldsymbol{m}},\widehat{\boldsymbol{m}}_{s2})}{Var(\widehat{\boldsymbol{m}}_{s2})}$ and $R_c = \frac{R_{c1}+R_{c2}}{2}$. Thus, we utilize more information from each repetition compared to just using half the individuals as a subsample. Lastly, the mean heritability across the 10 repetitions was used as the estimated heritability in our study. Thus, the $\hat{\lambda}$ we estimate using 4k SNPs had to be rescaled to be used with a 60k SNP dataset:

$$\hat{\lambda}_{60k} = \frac{\hat{\lambda}}{2\boldsymbol{p}'_{\boldsymbol{4k}}(\mathbf{1}_{\boldsymbol{4k}} - \boldsymbol{p}_{\boldsymbol{4k}})} 2\boldsymbol{p}'_{\boldsymbol{60k}}(\mathbf{1}_{\boldsymbol{60k}} - \boldsymbol{p}_{\boldsymbol{60k}})$$

where $\hat{\lambda}$ is the variance component ratio estimated using a random 4k subset of SNPs, $\boldsymbol{p}_{\boldsymbol{4k}}$ and $\boldsymbol{p}_{\boldsymbol{60k}}$ are vectors of allele frequencies from the 4k random subset of SNPs and the full 60k SNPs (estimated from case-parents) and $\mathbf{1}_{\boldsymbol{4k}}$ and $\mathbf{1}_{\boldsymbol{60k}}$ are vectors of ones with length 4k and 60k, respectively. Note that since we are repeating the whole simulation 10 times, a total of 100 (=10*10) heritability estimations using Method R were done across all simulations (see below).

Simulated data – genomic structure

We used AlphaSimR [8] and the R programming language [9] to simulate all data used in this study. Ten replications were run, each with a generic non-inbred founder population consisting of 1000 individuals with 30 chromosomes simulated by the MaCS simulation software included in AlphaSimR[10]. The length of each chromosome was set to 10^8 base pairs (bp), with genetic length of 1 Morgan, resulting in a genome size of $3 \cdot 10^9$ bp. After the founder haplotypes were established, the 1000 founders were used as the parent generation. When genotyping, a simulated SNP chip with 60 000 (60k) SNP markers was used. In each replication we simulated a complex trait with 1500 QTLs which were not included on the 60k SNP chip by design. However, some of the QTLs may be included among the 60k marker SNPs due to random chance as potential SNPs and potential QTLs are the same in AlphaSimR.

Simulated data – family and population structure

We randomly mated 100 mothers and 100 fathers from the founder population (see above) and set the number of offspring across all families to be either 2000, 5000, 10 000, 50 000 or 100 000. Thus, the number of full-sib families were always approximately 100. However, as sires and dams are randomly sampled with replacement, the same parents can potentially be mated more than once, randomly creating a few half-sib families and, more rarely, duplicated full-sib families. These offspring are called the "Training+Sibs" population. The cases were the 1000 or 5000 offspring with the largest phenotype value on the continuous scale when population size allowed for it, that is, when population size is 2000 or 5000, only 1000 cases are simulated. The latter restriction resulted in a total of 8 simulation schemes per replication. When comparing TDGP and case-control models, the accuracy of the models were compared under the same number of genotyped training individuals, either all cases (TDGP) or 50% cases and 50% non-cases (case-control).

A "Halfsibs" population consisting of 1000 individuals was simulated by randomly crossing the same females and males used as parents for the "Training+Sibs" population with family size of 1 (i.e. 1000 randomly picked mating pairs). Thus, the "Half-sibs" population consists mostly of individuals having half-sib (maternal or paternal) relationships to the "Training+Sibs" population. The "Halfsibs" population was used to check if breeding values could be accurately predicted using TDGP for individuals with relatively high relatedness with the training population.

In addition to the "Training+Sibs" and "Halfsibs" populations, a "Validation" population was created by randomly mating the 500 females and 500 males (=all individuals) from the founder population. The validation population is thus regarded as generally distantly related to the training (i.e. "Training+Sibs") population and is used to test how accurately TDGP can predict breeding values on distantly related individuals.

Continuous phenotypes were simulated with a heritability of 0.4. By applying a percentile threshold to the continuous data, cases and non-cases were determined, see Table 1.

Expected heritability of a binary trait

A common assumption in analysis of binary traits is that the binary outcomes are determined by whether or not an underlying continuous liability exceeds a threshold value (logit and probit models). The observed phenotype is thus a binary categorization of the unknown underlying liability. As the categorization of a continuous trait implies loss of information, the realized heritability using the binary phenotypes is expected to be lower than the heritability from the continuous phenotypes. For all simulations in this study we have simulated binary phenotypes using a threshold on the continuous phenotypes, where all individuals above the threshold are given the value 1, while all others are given the value 0. While the heritability for the continuous phenotypes of $h_x^2 = 0.4$ remains unchanged across all simulations, the

expected observed-scale (binary) heritabilities under the different case incidence thresholds used in this study are computed as [11]:

$$h^2 = \frac{\bar{z}^2 h_x^2}{\bar{p}\bar{q}} \qquad (1)$$

where h^2 is the additive heritability on the binary scale, $\bar{z}$ is height at the threshold percentile in the standard normal distribution, $h_x^2 = 0.4$ is the heritability of the continuous trait, $\bar{p}$ is the case incidence and $\bar{q} = 1 - \bar{p}$. The heritabilities on the binary scale used in this study are shown in Table 1.

Table 1: The top row shows the different case incidences used in our study, while the bottom row shows the corresponding expected heritabilities, computed using formula (1). The heritability of the continuous phenotypes is always 0.4.

Case incidence ($\bar{p}$)	0.01	0.02	0.05	0.10	0.20	0.50
h^2	0.029	0.048	0.090	0.137	0.196	0.255

Results

Figure 1 shows 95% confidence intervals (CI) of heritability estimates using either SNP-REML on continuous or binary phenotypes (both cases and controls), or Method R using cases only across eight sampling schemes. Note that the same number of individuals are genotyped when comparing SNP-REML and Method R (see Methods). Training population sizes were set to be 2000, 5000, 10 000, 50 000 or 100 000 with either 1000 (all population sizes) or 5000 (population sizes of 10 000 or larger) individuals being sampled for genotyping. The dashed horizontal line indicates how much each method over- or underestimates the heritability compared with the true heritability of 0.4 for continuous phenotypes, or the expected observed-scale heritability of the binary phenotypes (see Table 1), across sampling schemes. When the true case incidence is 50%, the SNP-REML-based methods are accurate (sampling schemes “H” and “G”), while TDGP Method R underestimated the heritability. For all the other

sampling schemes (i.e. schemes A-F), TDGP Method R delivers as-good or more precise heritability estimates compared with SNP-REML. However, for sampling schemes A-F, all methods over-estimate the heritability. As the case incidence decreases, the overestimation of the heritability estimates of the SNP-REML methods increases dramatically. The latter is probably due to the artificially increase of the case incidence due to over-sampling of cases vs. non-cases. The estimates from the TDGP Method R seem, however, much less affected.

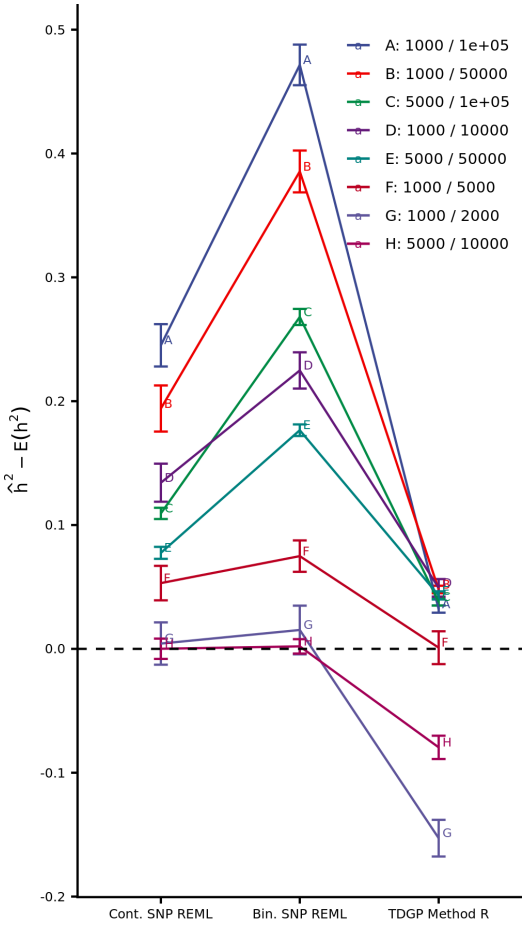


Figure 1: 95% confidence intervals (CI) of each heritability estimate minus the expected heritability ± the standard error of the mean. The letter and the colorization of each sampling scheme is shown in the upper right part of the figure on the form "number of cases" / "total number of individuals" in the training population. A line is drawn between the estimation methods to further indicate the sampling scheme. The methods for estimating the heritability is either SNP-REML for continuous-(underlying) or binary phenotypes, or Method R for TDGP. The expected heritabilities using binary SNP-REML and TDGP Method R are shown in Table 1, while the expected heritability for continuous SNP-REML is always 0.4. All cases are used for

estimating the heritability using TDGP Method R, while half of the cases and an equivalent number of non-cases are used when estimating heritability using SNP-REML.

Figure 2 shows 95% CI of accuracies of genomic breeding value predictions for the different methods, sampling schemes and training population sizes, as described above. Accuracies are computed for three different test-populations (given on the x-axis) with decreasing relatedness to the training population. The 'Training+Sibs' population includes the individuals sampled for genotyping and used for training as well as their non-sampled siblings from the training population (i.e. all cases and non-cases). The 'Halfsibs' population was created by randomly mating the 200 parents used to create the 'Training+Sibs' population, i.e. consisting of mostly paternal or maternal half siblings of the 'Training+Sibs' population. Finally, the 'Validation' population was created by crossing all 1000 founders randomly, i.e. having little or no relatedness with the individuals used for training.

The SNP-BLUP results indicate, as expected, reduced accuracy with reduced relatedness between training population and the individuals used to evaluate the accuracy (Figure 2). However, this tendency was much less profound for the TDGP model, and for sampling schemes having large 'Training+Sibs' populations (50 000 or 100 000 individuals), no reduction in accuracy by more distant relationships to the training population was observed. In fact, for the largest schemes, i.e. "C" and "E", sampling 5000 training individuals out of 'Training+Sibs' populations of, respectively, 100 000 or 50 000, the accuracy rather seemed to slightly increase with decreasing relatedness to the training population. When SNP effects are applied to the 'Training+Sibs' population itself, the SNP-BLUP model was consistently more accurate than TDGP. All three methods achieve the highest accuracies when there are many genotyped individuals, and large 'Training+Sibs' population (i.e., when the cases are numerous and phenotypically more extreme). For the low case incidence of 5% used in sampling scheme "C", TDGP outperformed both SNP-BLUP using underlying continuous

phenotypes and binary phenotypes when predicting genomic breeding values in the most distant validation population.

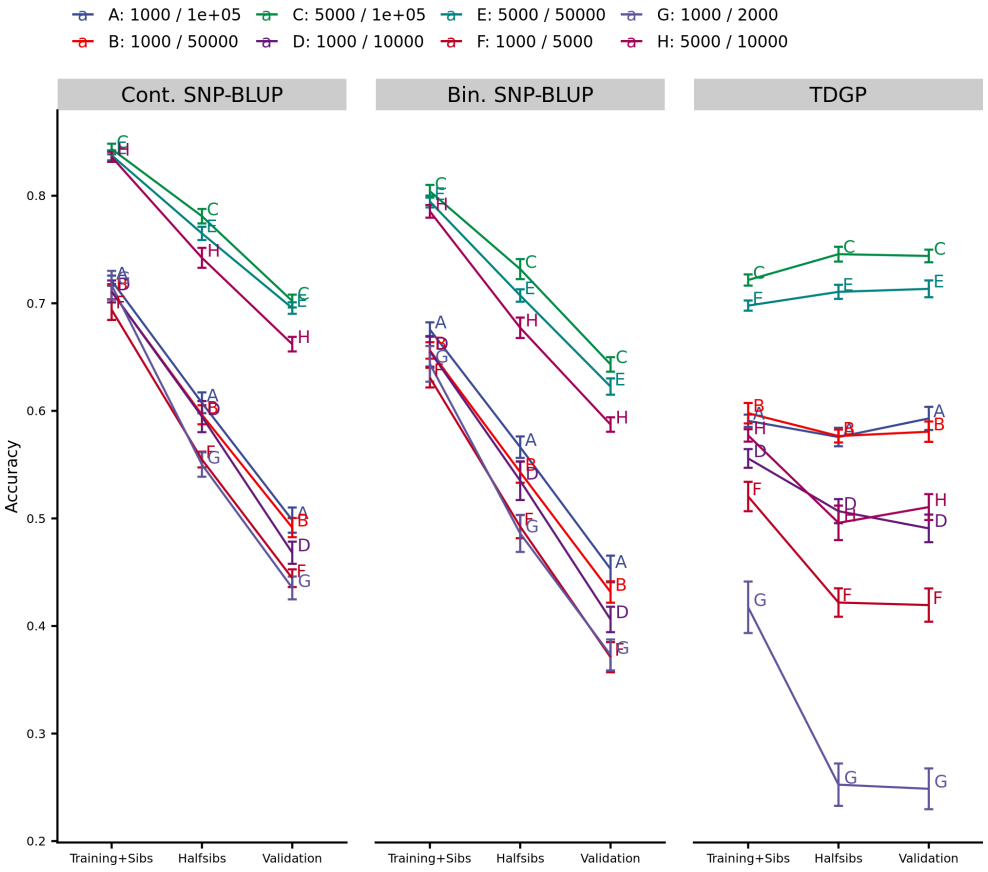


Figure 2: 95% confidence intervals of mean accuracy ± standard error of the mean for 10 replicates. Each CI is colorized for its group on the format "number of cases" / "total number of individuals". SNP effects are predicted using all cases in the training population for TDGP Method R, while half of the cases and an equivalent number of non-cases from the training population were used to predict SNP effects using SNP-BLUP. For each population, the methods continuous SNP-BLUP ('Cont.SNP-BLUP'), binary SNP-BLUP ('Bin. SNP-BLUP') and TDGP are used to calculate accuracies using the SNP effects predicted using the chosen individuals in the training population.

Discussion

The simulation study has shown that the TDGP model can provide accurate GEBVs, which can outperform ordinary case-control SNP-BLUP for distantly related validation populations.

If the case incidence is low, non-case individuals are expected to approach a random sample of the population, which will not provide much information with respect to marker effects.

As an example of a possible use for TDGP, Brun *et al.* [12] reported a mortality of 6.1% in groups of Atlantic Salmon (*Salmo salar*) in Norway with outbreaks of cardiomyopathy syndrome (CMS), compared to 2.5% mortality in non-infected groups. Thus, if ~3.6% (=6.1% - 2.5%) of the fish die due to CMS in a sea cage containing 100 000 individuals, there will be a total of ~3600 "cases" available. Such circumstances are advantageous for TDGP, namely restricting genotyping to the phenotypically most extreme offspring individuals and their parents without having to genotype a suitable contrast group from the remaining offspring.

We are not aware of any method where estimation of heritability and prediction of genetic breeding values can be performed by genotyping affected offspring and their parents only. However, as noted in Background, the TDT method has been successfully used to perform genome wide association studies (GWAS) using such data. TDGP is a generalization of TDT such that genome-wide breeding values (GEBVs) can be predicted and heritabilities estimated. In the special case where the TDGP model is used to fit a single fixed marker effect ($\lambda = 0$), the significance of the marker can be tested using an appropriate t-test (as shown in Supplementary Material). This t-test statistic is asymptotically equivalent with the χ^2-statistic used by TDT [4, 13] when case incidence approaches zero. Using TDGP with ridge regression, including a covariance structure to adjust for correlation structure (LD) among loci, enables us to predict individual GEBVs using cases only as training data combined with genotypes of their parents. TDGP resembles SNP-BLUP, while the latter also needs a contrast group (i.e. controls) to be included in training data.

In this study, we use mortality during a disease outbreak in a large population as a typical example where TDGP can be used. However, there are many categorical traits, not necessarily involving mortality, where cases can be separated from non-cases. In livestock, such traits may include reproductive failure, inadequate performance, locomotor problems, milking problems and non-fatal diseases in swine [14], infertility and mastitis in rabbits [15] and failure to conceive, feet problems and abortion in cows [16]. Thus, for any trait where cases can be clearly defined and has a low incidence, the case-individuals may contribute relatively more information compared to a largely random sample of non-cases.

TDGP Method R for estimating heritability

Figure 1 shows that when using TDGP Method R to estimate heritability, the estimates are underestimated when the case incidence is 50%. This may be due to the fact that, because of time constraints, we used a random subset of 4k SNPs when running TDGP Method R for heritability estimation. Even though we re-sampled the 4k SNPs ten times and used the mean heritability as the final heritability estimate, each 4k subset of SNPs may not capture the full genetic variation explained by the 1500 simulated QTLs, and thus the heritability may be underestimated. However, the TDGP Method R-estimated heritability seems to be slightly upward-biased when the case incidence decreases below 50% (see Figure 1). A potential explanation is that the left-hand side of the equation system is approximated using data from cases and their parents only. For example, the **J** matrix, used to set up the left-hand side of the equation system, is estimated based on the observed heterozygosity of case-parents, as an estimate for heterozygosity of the entire parental population of both cases and non-cases. Because the cases are phenotypically more extreme, highly susceptible families are likely overrepresented among them, potentially resulting in a somewhat lower

heterozygosity among case-parents for loci in LD with QTL. Consequently, the elements of the **J** matrix are therefore slightly underestimated, resulting in scaling down the left-hand side of the equation system. In contrast, the right-hand side is computed exactly, and the predicted marker effects may therefore be slightly inflated, causing a corresponding inflation of the estimated heritability. However, the heritability estimate from TDGP Method R is still closer to the expectation (see Table 1) than the ordinary SNP-REML estimates. Figure 1 shows that SNP-REML overestimates the heritability for sampling schemes where the case incidence is below 50%. The reason for this is probably that SNP-REML utilizes a genotyped subset of the population, of which cases are typically over-sampled, meaning that the heritability in the data sampled for SNP-REML is higher than in the actual data. The MME is set up using the genotyped individuals only, for which cases are the only ones contributing to the right-hand side MME (as $\mathbf{M}'\mathbf{y} = \mathbf{M_c}\mathbf{1}$), while all genotyped individuals contribute to the left-hand side ($\mathbf{M}'\mathbf{M}$). Consequently, for a smaller subset of the population where the cases are over-sampled, the left-hand side of the MME ($\mathbf{M}'\mathbf{M}$) will be relatively more reduced than the right-hand side ($\mathbf{M}'\mathbf{y}$), resulting in inflated marker effects and thus also inflated estimated heritability. In any case, the highest accuracy is expected to be achieved at the estimated heritability. The latter may not necessarily be true for TDGP with the current implementation of TDGP Method R using a random 4k subset of SNPs to estimate heritabilities rather than all SNPs. Consequently, if we use all the 60k SNPs using TDGP Method R to achieve a more accurate estimate of the heritability, we may also achieve increased prediction accuracy.

It may be possible to modify SNP-REML so that it can estimate heritability directly using TDGP MME. However, since the majority of software does not support estimation of

heritabilities using customized MME, we found it more convenient to implement Method R instead of SNP-REML for estimating TDGP-based heritabilities.

Genomic predictions of TDGP vs. SNP-BLUP

If SNP-BLUP is used, normally both cases and non-cases/controls must be sampled and contrasted. In some situations, sampling non-cases from a disease outbreak may not be trivial. For example, CMS in Atlantic salmon have long-lasting, moderately elevated mortality, making it hard to identify when the outbreak has passed and non-cases can be safely sampled [12]. Consequently, clear cases (e.g. deceased) may thus be more easily identified than clear non-cases, making such traits prime candidates for use with TDGP.

Even though we mostly describe applications of the TDGP model where the phenotype is naturally categorical, the model may also be extended to the extreme(s) of continuous phenotypes. For example, for traits like growth, one may genotype e.g. the 2% fastest-growing individuals in a large population as “cases”. However, categorization of continuous traits necessarily implies loss of information. For very large populations (e.g. net-cages with 100k fish or more), where only a fraction of the individuals can be genotyped, genotyping the most extreme individuals may still be cost-effective compared with genotyping a random sample.

SNP-BLUP uses both between- and within-family information, while TDGP is restricted to within-family information for prediction of marker effects. Since TDGP adjusts all genotypes by parent expectation, the predicted SNP effects depend on the deviation from neutral inheritance seen within case offspring of heterozygous parents. SNPs in LD with QTLs will generally tend to deviate from their parents’ genotype average. It is clear from Figure 2 that modelling marker effects solely from within-family variation has both advantages and

disadvantages. Predicting genomic breeding values (GEBVs) using SNP-BLUP (i.e. using both within- and between-family information) results in higher accuracies when predicting the training population itself. However, when predicting GEBVs for individuals with low relatedness with the training individuals, the accuracy of SNP-BLUP is considerably reduced, while TDGP (when applied to large populations with low case incidences) is less affected. For SNP-BLUP, marker alleles being more common in high-ranking vs. low-ranking families may be used to model family differences, despite having no clear LD to the QTLs underlying the trait, i.e. the marker data captures family structure. In the current study, the training families were mostly created by mating a single sire to a single dam (producing many offspring), and sire and dam genetic effects are thus likely heavily confounded. Hence, genomic prediction may be accurate when applied to the specific families included in the training population. However, the accuracy is considerably lower when applied to other families, both when applied to new matings of the same parents (half-sib population) and, even more, when applied to offspring of different parents (validation population). In contrast, the TDGP marker estimates are based on whether the markers show a consistent deviation from parent expectation (neutral inheritance) within all segregating families represented in the case-data. Consequently, for large populations of low case incidence, TDGP shows consistent accuracies irrespective of their relatedness with the training population (training, half-sib and validation). When applied to binary data in our simulations, the TDGP model appears superior to the SNP-BLUP model when applied outside the training population itself, both when applied to half-sibs and, even more, when applied to the lowly related validation population (see Figure 2).

Importance of SNP density for TDGP

As the TDGP is largely LD-based, it is essential to include markers in close LD with underlying QTLs. Thus, if some of the underlying QTLs have little or no LD with the SNPs included in the analysis, the TDGP model will have limited possibilities to capture their effects. In contrast, such QTL-effects can be partly captured through the family structure, which is implicitly used in SNP-BLUP marker models. For SNP-BLUP/GBLUP it has been shown that increasing the SNP density gives only a minor increase in accuracy except when the initial SNP density is very low [17-21]. The 60k SNPs used in this study can be regarded as medium density, probably capturing most, but not all, of the genetic variation from the 1500 QTLs through LD. Because higher densities of SNPs could possibly be beneficial when using TDGP, we performed an additional simulation study where we varied the number of SNPs from 30 to 61 440, for a binary trait of 5% case incidence and 100 000 individuals in total (results not shown). Here, the accuracy plateaued at ~15 000 SNPs, indicating that a further increase in SNP density would not significantly improve TDGP accuracy of the simulated population. Although denser SNPs imply a higher probability of having SNPs in close LD with underlying QTLs, both the TDGP and SNP-BLUP approaches assume *a priori* that total genetic variance is distributed over all SNPs fitted. Consequently, adding more SNPs imply that predicted SNP effects would be more restricted, which potentially can explain the lacking improvement in accuracy when approaching higher SNP densities. Variable selection models may be more capable of utilizing higher marker densities (e.g., BayesB- and BayesC-like implementations of TDGP).

Effect of preselection on TDGP

A downside of using TDGP is that all selection (artificial or natural) performed from fertilization until the individual is identified as a case may affect the prediction. For example, pre-selection for fast growth rate in early life-stages may bias TDGP marker effect estimates

610 from a subsequent disease outbreak. In this example, a “case” is not only characterized by
611 high susceptibility to the disease/condition in question, but also by fast growth prior to the
612 disease outbreak. Genomic selection for reduced case incidence, based on marker effects
613 from a TDGP model applied to such data may thus partly imply selection for reduced growth
614 rate and partly selection for reduced susceptibility to the disease. Still, such pre-selection
615 may have low selection intensity and poor accuracy, and the effect may thus be limited. In
616 any case, we advise caution when predicting SNP effects using TDGP on individuals which
617 have been involved in some form of pre-selection. The individuals, prior to the disease
618 outbreak, should be a random sample from their respective families.

619 # Conclusion

620 Genotyping affected cases, e.g. from a disease outbreak, and their parents can be used to
621 predict SNP effects and individual genomic breeding values through the use of Transmission
622 Disequilibrium Genomic Prediction (TDGP). Even though only cases and their parents are
623 included in the analysis, we estimate heritability for the TDGP model using Method R. The
624 bias in the estimated heritability from a SNP-BLUP model using a selected sample and REML
625 was significantly higher than the bias from TDGP using Method R, except when the case
626 incidence was 50%. However, the TDGP model adjusted for the case incidence, while case-
627 control SNP-BLUP had no such adjustment, which is probably the main reason for the
628 differences with respect to bias in heritability estimates. For the simulated individuals used
629 in this study, the accuracy of the TDGP model is best when the training population is large
630 and has a low case incidence, i.e. when a more extreme training sample from a large
631 population is genotyped. In such situations, the accuracy of TDGP is not sensitive to degree
632 of relationship between training and validation populations (e.g., siblings or distantly

related), having similar accuracy as the training population itself. In contrast, predictive ability of the case-control SNP-BLUP model clearly diminished with diminishing relationships between training population and the validation population. For large populations of low case incidences, the accuracy of SNP-BLUP GEBVs was superior to TDGP only when used to predict the training population itself. However, TDGP was superior both when applied to closely related validation populations (siblings) and even more so when applied to a distantly related validation population.

Appendix – Proofs

Proof 1: $E(\mathbf{M}'\mathbf{M}) = 2N \cdot \mathbf{JCJ}$

If $\mathbf{M}$ is a matrix of genotypes adjusted for 2 x allele frequency, then:

$$E(\mathbf{M}'\mathbf{M}) = E\left(\begin{bmatrix} \sum_{i=1}^{N}(q_{i1}-2p_1)^2 & \cdots & \sum_{i=1}^{N}(q_{i1}-2p_1)(q_{iL}-2p_L) \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^{N}(q_{iL}-2p_L)(q_{i1}-2p_1) & \cdots & \sum_{i=1}^{N}(q_{iL}-2p_L)^2 \end{bmatrix}\right)$$

$$= E\left(\begin{bmatrix} \sum_{i=1}^{N} m_{i1}m_{i1} & \cdots & \sum_{i=1}^{N} m_{i1}m_{iL} \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^{N} m_{iL}m_{i1} & \cdots & \sum_{i=1}^{N} m_{iL}m_{iL} \end{bmatrix}\right) = E\left(\begin{bmatrix} N\cdot\widehat{Var}(m_1) & \cdots & N\cdot\widehat{Cov}(m_1,m_L) \\ \vdots & \ddots & \vdots \\ N\cdot\widehat{Cov}(m_L,m_1) & \cdots & N\cdot\widehat{Var}(m_L) \end{bmatrix}\right)$$

$$= E\left(\begin{bmatrix} N\cdot 1\cdot\hat{\sigma}_1^2 & \cdots & N\cdot\hat{r}_{1L}\cdot\hat{\sigma}_1\hat{\sigma}_L \\ \vdots & \ddots & \vdots \\ N\cdot\hat{r}_{L1}\cdot\hat{\sigma}_L\hat{\sigma}_1 & \cdots & N\cdot 1\cdot\hat{\sigma}_L^2 \end{bmatrix}\right) = E\left(N\begin{bmatrix} 1\cdot\hat{\sigma}_1^2 & \cdots & \hat{r}_{1L}\cdot\hat{\sigma}_1\hat{\sigma}_L \\ \vdots & \ddots & \vdots \\ \hat{r}_{L1}\cdot\hat{\sigma}_L\hat{\sigma}_1 & \cdots & 1\cdot\hat{\sigma}_L^2 \end{bmatrix}\right)$$

$$= E\left(N\begin{bmatrix} \hat{\sigma}_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \hat{\sigma}_L \end{bmatrix}\begin{bmatrix} 1 & \cdots & \hat{r}_{1L} \\ \vdots & \ddots & \vdots \\ \hat{r}_{L1} & \cdots & 1 \end{bmatrix}\begin{bmatrix} \hat{\sigma}_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \hat{\sigma}_L \end{bmatrix}\right)$$

$$= N\begin{bmatrix} \sqrt{2p_1(1-p_1)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{2p_L(1-p_L)} \end{bmatrix}\begin{bmatrix} 1 & \cdots & r_{1L} \\ \vdots & \ddots & \vdots \\ r_{L1} & \cdots & 1 \end{bmatrix}\begin{bmatrix} \sqrt{2p_1(1-p_1)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{2p_L(1-p_L)} \end{bmatrix}$$

$$= 2N\begin{bmatrix} \sqrt{p_1(1-p_1)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{p_L(1-p_L)} \end{bmatrix}\begin{bmatrix} 1 & \cdots & r_{1L} \\ \vdots & \ddots & \vdots \\ r_{L1} & \cdots & 1 \end{bmatrix}\begin{bmatrix} \sqrt{p_1(1-p_1)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{p_L(1-p_L)} \end{bmatrix} = 2N\cdot\mathbf{JCJ}$$

where $q_{ij} \in \{0,1,2\}$ is the genotype for individual *i* at locus *j*, p_j is allele frequency at locus *j*, $m_{ij} = q_{ij} - 2p_j$ is genotype adjusted for expected genotype ($2p_j$), N is the number of individuals, L is the number of loci, σ_j is the standard deviation at locus *j*, $r_{j_1 j_2}$ is the genotype correlation between locus j_1 and locus j_2, $\boldsymbol{J}$ is a diagonal matrix of standard deviations for the markers and $\boldsymbol{C}$ is a matrix of genotypic correlations between loci.

Proof 2: $E(\mathbf{T}'\mathbf{T}) = \frac{1}{2}E(\mathbf{M}'\mathbf{M}) = N \cdot \mathbf{JCJ}$

$E(\mathbf{T}'\mathbf{T})$

$$= E\left(\begin{bmatrix} \sum_{i=1}^{N}\left(q_{i1} - \frac{1}{2}(q_{d_i1} + q_{s_i1})\right)^2 & \cdots & \sum_{i=1}^{N}\left(q_{i1} - \frac{1}{2}(q_{d_i1} + q_{s_i1})\right)\left(q_{iL} - \frac{1}{2}(q_{d_iL} + q_{s_iL})\right) \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^{N}\left(q_{iL} - \frac{1}{2}(q_{d_i1} + q_{s_i1})\right)\left(q_{i1} - \frac{1}{2}(q_{d_iL} + q_{s_iL})\right) & \cdots & \sum_{i=1}^{N}\left(q_{iL} - \frac{1}{2}(q_{d_iL} + q_{s_iL})\right)^2 \end{bmatrix}\right)$$

$$= E\left(N\begin{bmatrix} \widehat{Var}(t_1) & \cdots & \widehat{Cov}(t_1, t_L) \\ \vdots & \ddots & \vdots \\ \widehat{Cov}(t_L, t_1) & \cdots & \widehat{Var}(t_L) \end{bmatrix}\right) = E\left(N\begin{bmatrix} 1 \cdot \hat{\sigma}_1^2 & \cdots & \hat{r}_{1L} \cdot \hat{\sigma}_1\hat{\sigma}_L \\ \vdots & \ddots & \vdots \\ \hat{r}_{L1} \cdot \hat{\sigma}_L\hat{\sigma}_1 & \cdots & 1 \cdot \hat{\sigma}_L^2 \end{bmatrix}\right)$$

$$= E\left(N\begin{bmatrix} \hat{\sigma}_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \hat{\sigma}_L \end{bmatrix}\begin{bmatrix} 1 & \cdots & \hat{r}_{1L} \\ \vdots & \ddots & \vdots \\ \hat{r}_{L1} & \cdots & 1 \end{bmatrix}\begin{bmatrix} \hat{\sigma}_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \hat{\sigma}_L \end{bmatrix}\right)$$

$$\xrightarrow{\sigma_{T_{ij}} = \sqrt{p_j(1-p_j)}} N\begin{bmatrix} \sqrt{p_1(1-p_1)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{p_L(1-p_L)} \end{bmatrix}\begin{bmatrix} 1 & \cdots & r_{1L} \\ \vdots & \ddots & \vdots \\ r_{L1} & \cdots & 1 \end{bmatrix}\begin{bmatrix} \sqrt{p_1(1-p_1)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{p_L(1-p_L)} \end{bmatrix}$$

$$= N\begin{bmatrix} \sqrt{p_1(1-p_1)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{p_L(1-p_L)} \end{bmatrix}\begin{bmatrix} 1 & \cdots & r_{1L} \\ \vdots & \ddots & \vdots \\ r_{L1} & \cdots & 1 \end{bmatrix}\begin{bmatrix} \sqrt{p_1(1-p_1)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{p_L(1-p_L)} \end{bmatrix} = N \cdot \mathbf{JCJ}$$

$$= \frac{1}{2}E(\mathbf{M}'\mathbf{M})$$

where $q_{ij} \in \{0,1,2\}$ is the genotype for individual *i* at locus *j*, q_{d_ij} and q_{s_ij} are the genotypes of the dam and sire of individual *i*, respectively, N is the number of individuals (cases and non-cases), L is the number of loci, σ_j is the standard deviation at locus *j*, $r_{j_1j_2}$ is the genotype correlation between locus j_1 and locus j_2, $\boldsymbol{J}$ is a diagonal matrix of standard deviations for the markers and $\boldsymbol{C}$ is a matrix of genotypic correlations between loci. The proof that $\sigma_{T_{ij}} = \sqrt{0.5p_j(1-p_j)}$ is shown in the following.

$Var(T_{ij}) = Var\left(q_{ij} - \frac{1}{2}(q_{d_ij} + q_{s_ij})\right)$ is the variance of the sum of two binary variates h_{s_ij} and h_{d_ij}, i.e. the gamete genotypes inherited from the paternal and maternal parent at locus *j*, respectively.

$$h_{s_ij} \sim Bernoulli\left(success = 0.5 * q_{s_ij}\right)$$

$$h_{d_ij} \sim Bernoulli\left(success = 0.5 * q_{d_ij}\right)$$

where q_{s_ij} and q_{d_ij} is the sire and dam genotype, respectively. Thus, the variance in offspring genotype adjusted for single-parent expectation where the parent is heterozygous is:

$$P\left(q_{s_ij} = 1\right) Var\left(h_{s_ij} | q_{s_ij} = 1\right) = P\left(q_{d_ij} = 1\right) Var\left(h_{d_ij} | q_{d_ij} = 1\right)$$

$$= 2p_j\left(1 - p_j\right) 0.5 q_{s_ij}\left(1 - 0.5 q_{s_ij}\right) = 2p_j\left(1 - p_j\right) 0.5 q_{d_ij}\left(1 - 0.5 q_{d_ij}\right)$$

$$= 2p_j\left(1 - p_j\right) 0.5(1 - 0.5) = 0.5 p_j\left(1 - p_j\right)$$

Note that parents with homozygous genotypes do not contribute to the variance because $Var\left(h_{s_ij} | q_{s_ij} \neq 1\right) = Var\left(h_{d_ij} | q_{d_ij} \neq 1\right) = 0$. Consequently, assuming that the parent genotypes are in Hardy-Weinberg Equilibrium, the variance of the offspring genotype adjusted for single-parent expectation at a single locus is:

$$Var\left(h_{s_ij}\right) = Var\left(h_{d_ij}\right) = 0.5 p_j\left(1 - p_j\right)$$

Because h_{s_ij} and h_{d_ij} are assumed independent (i.e. no relatedness between the parents) their sum has no covariance:

$$Var\left(T_{ij}\right) = Var\left(q_{ij} - \frac{1}{2}\left(q_{d_ij} + q_{s_ij}\right)\right) = Var\left(h_{s_ij} + h_{d_ij}\right) = Var\left(h_{s_ij}\right) + Var\left(h_{d_ij}\right)$$

$$= 0.5 p_j\left(1 - p_j\right) + 0.5 p_j\left(1 - p_j\right) = p_j\left(1 - p_j\right)$$

Thus, the standard deviation of the offspring genotype adjusted for parental expectation is:

$$\sigma_{T_{ij}} = \sqrt{Var\left(T_{ij}\right)} = \sqrt{p_j\left(1 - p_j\right)}$$

Competing interests

KEG and JO are employed with AquaGen AS. The authors declare that they have no competing interests.

Funding

The research leading to these results has received funding from The Research Council of Norway through the research programs NAERINGSPHD (project no. 251664) and HAVBRUK2 (project no. 245519). Additional funding was provided by AquaGen AS.

Authors' contributions

KEG: helped conceive the idea for TDGP and helped with some of the mathematical foundations and proofs. KEG Drafted the manuscript, wrote all the code and performed the simulations.

JO: Conceived the idea for the TDGP method and formulated the majority of the quantitative genetic theory. Critically reviewed and improved the manuscript.

THEM: Suggested to use Method R for estimation of heritability and helped with quantitative genetic theory. Critically reviewed and improved the manuscript.

Acknowledgements

A big thank you to Thomas Moen who, through critical review, improved the readability of this manuscript.

We wish to thank The Research Council of Norway which through both the NAERINGSPHD (project no. 251664) and the HAVBRUK2 (project no. 245519) programs provided funding, as well as the breeding company AquaGen AS for funding and other support.

References

1. Vallejo, R.L., et al., *Evaluation of Genome-Enabled Selection for Bacterial Cold Water Disease Resistance Using Progeny Performance Data in Rainbow Trout: Insights on Genotyping Methods and Genomic Prediction Models.* Frontiers in Genetics, 2016. **7**.
2. Vallejo, R.L., et al., *Genomic selection models double the accuracy of predicted breeding values for bacterial cold water disease resistance compared to a traditional pedigree-based model in rainbow trout aquaculture.* Genetics Selection Evolution, 2017. **49**.
3. Bangera, R., et al., *Genomic predictions can accelerate selection for resistance against Piscirickettsia salmonis in Atlantic salmon (Salmo salar).* Bmc Genomics, 2017. **18**.
4. Spielman, R.S., R.E. McGinnis, and W.J. Ewens, *Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM).* Am J Hum Genet, 1993. **52**(3): p. 506-16.
5. Falconer, D. and T. Mackay, *Introduction to quantitative genetics. 1996.* Harlow, Essex, UK: Longmans Green, 1996. **3**.
6. Reverter, A., et al., *Method R variance components procedure: application on the simple breeding value model.* J Anim Sci, 1994. **72**(9): p. 2247-53.
7. Reverter, A., et al., *Technical note: detection of bias in genetic predictions.* J Anim Sci, 1994. **72**(1): p. 34-7.
8. Gaynor, C., *AlphaSimR: Breeding Program Simulations.* 2019.
9. *R: A Language and Environment for Statistical Computing*. 2018; Available from: https://www.R-project.org/.
10. Chen, G.K., P. Marjoram, and J.D. Wall, *Fast and flexible simulation of DNA sequence data.* Genome Research, 2009. **19**(1): p. 136-142.
11. Dempster, E.R. and I.M. Lerner, *Heritability of Threshold Characters.* Genetics, 1950. **35**(2): p. 212-36.
12. Brun, E., et al., *Cardiomyopathy syndrome in farmed Atlantic salmon Salmo salar: occurrence and direct financial losses for Norwegian aquaculture.* Dis Aquat Organ, 2003. **56**(3): p. 241-7.
13. Ruiz-Narvaez, E.A. and H. Campos, *Transmission disequilibrium test (TDT) for case-control studies.* Eur J Hum Genet, 2004. **12**(2): p. 105-14.
14. D'Allaire, S., T.E. Stein, and A.D. Leman, *Culling patterns in selected Minnesota swine breeding herds.* Can J Vet Res, 1987. **51**(4): p. 506-12.
15. Rosell, J.M. and L.F. de la Fuente, *Culling and mortality in breeding rabbits.* Prev Vet Med, 2009. **88**(2): p. 120-7.
16. Bascom, S.S. and A.J. Young, *A summary of the reasons why farmers cull cows.* J Dairy Sci, 1998. **81**(8): p. 2299-305.
17. VanRaden, P.M., et al., *Genomic evaluations with many more genotypes.* Genet Sel Evol, 2011. **43**: p. 10.

18. Kriaridou, C., et al., *Genomic Prediction Using Low Density Marker Panels in Aquaculture: Performance Across Species, Traits, and Genotyping Platforms.* Front Genet, 2020. **11**: p. 124.
19. Ni, G., et al., *Whole-genome sequence-based genomic prediction in laying chickens with different genomic relationship matrices to account for genetic architecture.* Genet Sel Evol, 2017. **49**(1): p. 8.
20. Su, G., et al., *Comparison of genomic predictions using medium-density (approximately 54,000) and high-density (approximately 777,000) single nucleotide polymorphism marker panels in Nordic Holstein and Red Dairy Cattle populations.* J Dairy Sci, 2012. **95**(8): p. 4657-65.
21. Kjetså, M.H., J. Ødegård, and T.H.E. Meuwissen, *Accuracy of genomic prediction of host resistance to salmon lice in Atlantic salmon (Salmo salar) using imputed high-density genotypes.* Aquaculture, 2020. **526**.

Supplementary Material

Similarity between TDGP and TDT

Single locus associations to binary traits can be tested using TDT [4]. The TDT provides a χ^2 statistic which can be used to test if the SNP is in transmission disequilibrium. That is, to test if alleles inherited in a non-random subset of offspring from heterozygous parents consistently deviates from the parent expectation. Instead of testing a single locus, TDGP predicts SNP effects from transmission disequilibria across the entire genome simultaneously, taking correlation structure among loci into account and predicts individual genomic breeding values across all SNPs. One noteworthy distinction between TDT and TDGP is that the latter assumes SNP effects to be "random" (for $\lambda > 0$), while TDT assumes marker effects to be "fixed". Note that in the special case where a single SNP *j* is fitted with TDGP as "fixed" (i.e. assuming $\lambda = 0$), $[N \cdot \mathbf{JCJ} + \mathbf{I}\lambda] = N \cdot J_{jj}^{\mathbf{2}}$ and the equation system of TDGP thus simplifies to:

$$N \cdot J_{jj}^{\mathbf{2}} \widehat{m}_{\mathbf{j}} = \tfrac{1}{2}\left(t_j - n_j\right)$$

And thus

$$\widehat{m}_{\mathbf{j}} = \frac{\frac{1}{2}\left(t_j - n_j\right)}{N J_{jj}^2} = \frac{\left(t_j - n_j\right)4N_c}{2N\left(t_j + n_j\right)} = 2p_{cr}\frac{t_j - n_j}{t_j + n_j}$$

where the notation is the same as in the Methods section. The standard error of a single, fixed (i.e. $\lambda = 0$) marker effect *j* is (assuming that the left-hand side is correctly approximated):

$$SE\left(\widehat{m}_j\right) = \sigma_e\sqrt{\left(N \cdot J_{jj}^{\mathbf{2}}\right)^{-1}} = \sigma_e\sqrt{\frac{1}{N} \cdot \frac{4N_c}{t_j + n_j}} = \sqrt{p_{cr}(1 - p_{cr})} \cdot 2\sqrt{\frac{p_{cr}}{t_j + n_j}} = 2p_{cr}\sqrt{\frac{1 - p_{cr}}{t_j + n_j}}$$

where σ_e is the residual standard deviation for the binary trait, here assumed to equal the total phenotypic variance (i.e. assuming that the single SNP fitted has no effect), $\sigma_e = \sqrt{p_{cr}(1-p_{cr})}$. Then, a Student's t statistic of marker *j* based on the simplified equation system above is:

$$\tau_j = \frac{\widehat{m}_j}{SE(\widehat{m}_j)} = \frac{2p_{cr} \cdot \frac{t_j - n_j}{t_j + n_j}}{2p_{cr}\sqrt{\frac{1-p_{cr}}{t_j + n_j}}} = \frac{1}{\sqrt{(1-p_{cr})}} \cdot \frac{t_j - n_j}{\sqrt{t_j + n_j}}$$

Note that:

$$\lim_{p_{cr}\to 0} \tau_j = \frac{t_j - n_j}{\sqrt{t_j + n_j}}$$

In the original TDT-article by Spielman *et al.* [4], the corresponding χ^2 testing statistic (with one degree of freedom) is:

$$\chi_j^2 = \frac{\left(t_j - n_j\right)^2}{t_j + n_j}$$

i.e. the square of the asymptotic TDGP τ_j. Hence, assuming a single ("fixed") SNP ($\lambda = 0$) and the case-frequency approaching zero, the TDGP and TDT testing statistics are equivalent.

ISBN: 978-82-575-1775-5
ISSN: 1894-6402

Norwegian University
of Life Sciences

Postboks 5003
NO-1432 Ås, Norway
+47 67 23 00 00
www.nmbu.no

109194 / ANDVORDGRAFISK.NO