Accurate Fetal Variant Calling in the Presence of Maternal Cell Contamination

Elena Nabieva[1, 2, ✉, *], **Satyarth Mishra Sharma**[1, *], **Yermek Kapushev**[1], **Sofya K. Garushyants**[1, 2], **Anna V. Fedotova**[1, 3], **Viktoria N. Moskalenko**[3], **Tatyana Serebrenikova**[4], **Eugene Glazyrina**[4], **Ilya V. Kanivets**[5], **Denis V. Pyankov**[5], **Tatyana V. Neretina**[2, 3], **Maria D. Logacheva**[1, 2, 3], **Georgii A. Bazykin**[1, 2], **and Dmitry Yarotsky**[1, 2]

[1]Skolkovo Institute of Science and Technology, Skolkovo, Russia,
[2]Institute for Information Transmission Problems (Kharkevich Institute), Russian Academy of Sciences, Moscow, Russia,
[3]Lomonosov Moscow State University, Moscow, Russia,
[4]Progen LLC, Moscow, Russia
[5]Genomed LLC, Moscow, Russia
[*]These authors contributed equally to this work.

Correspondence: *e.nabieva@skoltech.ru*

High-throughput sequencing of fetal DNA is a promising and increasingly common method for the discovery of all (or all coding) genetic variants in the fetus, either as part of prenatal screening or diagnosis, or for genetic diagnosis of spontaneous abortions. In many cases, the fetal DNA (from chorionic villi, amniotic fluid, or abortive tissue) can be contaminated with maternal cells, resulting in the mixture of fetal and maternal DNA. This maternal cell contamination (MCC) undermines the assumption, made by traditional variant callers, that each allele in a heterozygous site is covered, on average, by 50% of the reads, and therefore can lead to erroneous genotype calls. We present a panel of methods for reducing the genotyping error in the presence of MCC. All methods start with the output of GATK HaplotypeCaller on the sequencing data for the (contaminated) fetal sample and both of its parents, and additionally rely on information about the MCC fraction (which itself is readily estimated from the high-throughput sequencing data). The first of these methods uses maximum likelihood estimation to correct the fetal genotype calls produced by MCC-unaware HaplotypeCaller. The other two methods "learn" the genotype-correction model from examples. We use simulated contaminated fetal data to train and test the models. Using the test sets, we show that all three methods lead to substantially improved accuracy when compared with the original MCC-unaware HaplotypeCaller calls. We then apply the best-performing method to three chorionic villus samples from spontaneously terminated pregnancies.

1. Introduction

High-throughput sequencing of fetal DNA is increasingly being used in academic and clinical settings. It is a powerful tool with the potential for use in prenatal diagnosis based on chorionic villus or amniotic fluid sampling [16], or in the analysis of chorionic villi or products of conception for genetic diagnosis of an unsuccessful pregnancy. In prenatal diagnosis, whole-genome or whole-exome sequencing can discover novel clinically significant variants that are not present in SNP arrays or gene panels, resulting in higher diagnostic yield [1]. Prenatal sequencing can inform prenatal and postnatal care and counseling, and may lead to prenatal therapeutic interventions [1].

In standard practice, the DNA of both parents is sequenced together with fetal DNA ("trio sequencing") in order to establish patterns of inheritance and inform variant prioritization and interpretation. A technical difficulty that may arise in the analysis of fetal DNA is the contamination of the fetal sample with maternal cells. The prevalence of such *maternal cell contamination* (MCC) can be significant, depending on the experimental technique and quality of the sample; for example, one study reported 9.1% of amniotic fluid samples as having detectable MCC [15], while another found MCC fraction > 5% in as many as 26% of amniotic fluid samples under some practices [18]. High-level contamination (over 20%) was detected by one study [12] in a small, but non-zero number of samples (0.3% of cultured amniotic fluid samples and 1.3% of cultured chorionic villi samples; it must be noted that cultured amniotic fluid samples generally have less MCC than direct samples). In traditional prenatal analysis, such as that aimed at detecting chromosomal aberrations, maternal cell contamination is assayed by special tests, such as the Short Tandem Repeat analysis, and, if detected at a sufficient level, may nullify the analysis [13].

Meanwhile, standard variant calling software that is used to analyze next-generation sequencing data relies on the expectation that each allele is represented by half of the reads. MCC disrupts this assumption, leading to errors in variant calling. In this work, we propose and evaluate computational methods for reducing the MCC-caused error. All of these methods begin with variants called in the fetal specimen, the mother, and the father by a standard variant-calling pipeline and then "correct" the results from the fetal specimen. The first method uses a maximum-likelihood estimator to decide on the "true" fetal genotype based on the called genotypes of the trio. The other methods eschew making assumptions about the best way to uncover the "true" fetal genotype from the maternally-contaminated observed specimen data and instead solve this problem using machine learning. We train these methods on synthetic "mother-father-fetus" trios generated from real family trios by adding specified numbers of maternal reads to the child sample. We use these synthetic trios, where the child's genotype is known, to demonstrate that MCC correction significantly improves the accuracy of variant calling compared to contamination-naive calling, especially for higher fractions of MCC. As an intermediate technical step, we present a simple heuristic algorithm for estimating maternal cell contamination fraction in the fetal sample. We then apply the trained model to real sequencing data from a miscarried fetus with 40% estimated MCC, changing the fetal calls for a substantial number of SNPs.

2. Materials and Methods

In the rest of the paper, we use the term *specimen* to denote the obtained fetal sample which may be contaminated with maternal cells. In practice, it can be a chorionic villus or amniotic fluid sample or an abortus sample.

We assume that DNA samples are available for all three members of the trio. We expect the trio to be sequenced (we gear our work towards and test it on Illumina whole-exome sequencing, currently the most widely used protocol in biomedical practice) and the reads mapped and variant-called according to a standard bioinformatics pipeline, e.g., one involving the Genome Analysis Toolkit (GATK, [5]), although our approach can work with other calling pipelines too, as we show in Section 3.3. We expect that the called genotypes are available along with accompanying quality information, such as allelic depths, genotype likelihoods, and so on. The VCF file produced by the pipeline is the starting point for all our analyses (see Figure 1).

2.1. Estimating the contamination fraction. We define the maternal cell contamination (MCC) fraction, which we denote by $\alpha \in [0,1]$, to be the share of maternal DNA in the DNA of the fetal specimen. MCC accounts for the discrepancy between the results of variant calling on the contaminated specimen sample and the true fetal genotype. The MCC-aware genotype correction methods presented below all rely on knowing the value of α, and therefore it is vital to be able to estimate it accurately.

To estimate α in a fetal sample, we look at the results of variant calling by GATK HaplotypeCaller [5] and consider positions in the genome where the mother and the father are homozygous for different alleles (i.e., one of the parents is homozygous for the reference allele, and the other for the alternative allele); we only consider biallelic sites. The fetus then should be heterozygous at these sites, with equal read coverage for the two parental alleles.

In the presence of MCC, however, the mother's allele coverage fraction m should be higher, namely $\frac{1+\alpha}{2}$. Then, we can use the value of m at the position obtained from the VCF file to compute α as $2m-1$. Since the actual coverage fluctuates, we average this ratio over all relevant sites to get the MCC fraction estimate. Let m_i be the fraction of maternal reads at site i, Then we compute $\widehat{m}$ to be the average of m_i over all sites i where the mother and the father are homozygous for different alleles, and estimate $\widehat{\alpha} = 2\widehat{m} - 1$. Only calls that pass basic filtering criteria (GQ > 30 in both parents, depth at site > 10 in all three samples) are considered in the calculation.

Although this procedure should be symmetric with respect to which of the parents is homozygous for the reference allele at a site, we observe that estimating α when the mother is homozygous for the reference allele ("mo00") systematically gives larger values than when the mother is homozygous for the alternative allele ("mo11"). We attribute this phenomenon to the reference bias of the variant calling pipeline [6]. We, therefore, estimate α separately over the two alternatives ("fa00_mo11" and "fa11_mo00") and compute the average of the two estimates as the final result. We found this aggregated estimate to be more accurate than either alternative.

Using simulated data (see Section 2.4), we found that the MCC estimation error made by this method does not exceed 2% and in the vast majority of the cases is below 1% Figure S1.

While there are methods for estimating sample contamination that are geared towards detecting contamination during sequencing with unrelated samples and make use of population allele frequencies (VerifyBamID [9] and the CalcualteContamination module of the cancer-sequencing workflow in Genome Analysis Toolkit v.4 (https://software.broadinstitute.org/gatk/), they are not well suited for the particular case of maternal contamination; in fact, the authors of VerifyBamID predicted that their method would underestimate contamination with maternal sample by half [9]. Our evaluation of these methods on simulated contaminated data generally agrees with this prediction, although we find it to be something of an underestimate itself, especially for large MCC fractions (Figure S1).

2.2. Maximum likelihood-based genotype correction. As our first method, we consider a maximum likelihood-based approach that carries out a fetal genotype call adjustment based on a simple explicit mathematical model. To this end, we first observe that the disagreement between the true variants of the child and the called variants of the fetal specimen is in the vast majority of cases due to the mislabeling of homozygous positions in the fetus as heterozygous in the contaminated fetal specimen. This happens when the maternal genotype, and accordingly the mixture of fetal and maternal reads, is heterozygous.

Following this observation, we implement a fetal genotype correction procedure in which the candidates for readjustment are the positions where both the contaminated specimen's genotype and the maternal genotype are called heterozygous, say 01. We assume for simplicity that the maternal genotype 01 is known with high certainty (this can be ensured by filtering out positions with low genotype quality). We consider three hypotheses, that the child's genotype is 00, 01, or 11. We then determine the child's genotype using the maximum likelihood estimate based on the fetal allelic depths and the previously estimated MCC fraction α. Specifically, given child's true genotype 00, 01 or 11 (and remembering that the maternal genotype is 01), the respective conditional probabilities of observing the allelic depths $AD0, AD1$ in the

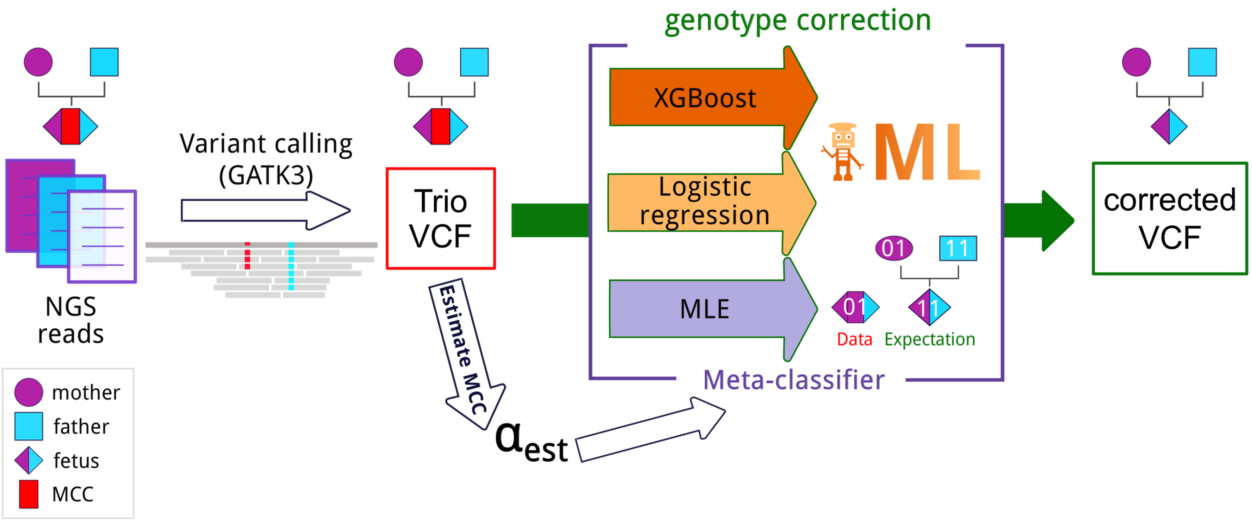

Fig. 1. Pipeline for accurate fetal variant calling in the presence of maternal cell contamination (MCC). NGS reads for each sample in the trio are mapped to the reference genome, and then variants are called with GATK v3. In the presence of MCC, the resulting VCF file contains incorrect calls for the fetus. To overcome this, we estimate MCC from this VCF and utilize the estimated value to correct the genotype calls for the fetus. Three different approaches for genotype correction were utilized (see Materials and Methods section): an explicit maximum likelihood method (MLE) and two machine learning-based approaches, namely, logistic regression and Gradient Boosted Decision Trees (XGBoost). The meta-classifier combines outputs of all previous methods. ML - machine learning.

mixture are:

$$\mathrm{P}(AD0,AD1|00) = C(\tfrac{\alpha}{2})^{AD1}(1-\tfrac{\alpha}{2})^{AD0},$$
$$\mathrm{P}(AD0,AD1|01) = C(\tfrac{1}{2})^{AD1}(\tfrac{1}{2})^{AD0},$$
$$\mathrm{P}(AD0,AD1|11) = C(1-\tfrac{\alpha}{2})^{AD1}(\tfrac{\alpha}{2})^{AD0},$$

with the binomial coefficient $C = \frac{(AD0+AD1)!}{AD0!AD1!}$ Then, assuming for simplicity that the three possible fetal genotypes have the same prior probability, we just pick the genotype with the maximum probability.

2.3. Machine learning-based genotype correction. In this approach, the predictive model for genotype readjustment is not based on heuristics, but is trained using a machine learning algorithm. The input to the model consists of two components. The first is the estimated fraction of maternal DNA in the fetal specimen, $\widehat{\alpha}_{glob}$. The second is the vector of features characterizing the variants called in the fetal specimen and the parents at a particular position in the genome (practically, this is the line in the VCF file for that position; it includes the genotype likelihoods, genotype qualities, read depths, etc. for all three samples). The output of the model is the adjusted fetal genotype at this position in the genome.

The approach assumes that we either consistently use a particular mapping and variant-calling pipeline or train predictive models using a sufficiently rich training dataset combining data obtained using several pipelines, since a single mapping and calling pipeline may have bias dependent on the specifics of the read mapper and variant caller. In practice, it is necessary that the fields describing the variant (genotype likelihood, read depth, etc.) be the same in the training VCFs and the VCF to be genotype-corrected; therefore, the same, or closely related variant callers should be used to produce all VCFs. The "ground truth" to which the predictive model is fit is the genotype calls on the pure fetal sample made by the same pipeline as was used on the fetal specimen sample. To train and test the model, we simulate "virtual specimens" from a number of publicly available father-mother-child trios by randomly mixing mother and child reads at various MCC fractions, as described in Section 2.4, and then calling the variants in the simulated trios. Since none of the trios on which the "virtual specimens" were based have MCC, we can use them to obtain the true child genotypes. The predictive model is then trained and tested on the data set that maps, genomic position-wise, father-mother-virtual specimen variant data to the child variant data of the corresponding father-mother-child sample. The predictive models may overfit to particular variant-calling pipelines or underlying trios, and care must be taken to ensure sufficient diversity of training samples.

Two machine learning algorithms were considered: logistic regression (as implemented in `scikit-learn`[1]) and Gradient Boosted Decision Trees (XGBoost implementation[2] [2]). The inputs to the models are read from a VCF file and standardized, with categorical features being one-hot encoded (split into binary columns for each unique value the feature can take). L2 regularization is applied to both algorithms, with early stopping based on performance on a validation set used to prevent XGBoost from overfitting. We found both the models to be fairly robust to choice of hyperparameters, and while these can be further tuned for a particular dataset, the gains in performance are marginal and do not warrant the loss in generalizability this incurs.

Additionally, an ensemble meta-classifier combining the re-

[1] `http://scikit-learn.org/`
[2] `https://xgboost.readthedocs.io/en/latest/`

sults of all of the above methods (maximum-likelihood estimator, logistic regression, XGBoost) was constructed. An ensemble makes predictions by combining 'votes' from multiple classifiers. Here we use soft-voting, i.e. voting based on the predicted probabilities of each class for the different classifiers, rather than a simple class prediction. This combined prediction helps mitigate the shortcomings of each individual classifier.

2.4. Contaminated trio generation. We obtained testing and training datasets by using publicly available exome reads from father-mother-child trios ("real-world trios") to produce "virtual specimen trios" with predetermined maternal contamination fraction α. Namely, we mixed randomly selected reads from the "real-world child" and the "real-world mother" in such a proportion that the fraction of maternal reads would be α, and thus obtained "virtual specimen" reads. We used the remaining reads from the "real-world mother" to create the "virtual specimen mother", while the "virtual specimen father's" reads were identical to the original "real-world father's" reads. For each "virtual specimen" trio, we also generated a corresponding "pure child" trio, which was identical to the "virtual specimen" trio except that no mother's reads were added to the child's reads (thus, the "virtual child" and the "virtual mother" reads are subsets of the corresponding child and mother reads from the original family exome). Keeping everything but the presence of maternal contamination the same between the "virtual" and the "real" trios ensures that the genotypes called with and without contamination would be compared fairly.

We performed this procedure for four "real-world" trios with publicly available exome data: the Ashkenazim trio HG002_NA24385 from the Genome in the Bottle project [19] ("AJT"), the YRI NA19240 trio from the 1000 Genomes project [3] ("YRI"), the CHD trio [8] ("CHD"), and the Corpas family daughter trio [4] ("Corpas"). Since the child in the "real-world" trios was either a living individual (AJT, YRI, Corpas) or stillborn (CHD), there was no maternal cell contamination in the "real-world" reads. If the data were available only as an alignment, we first obtained raw reads using the bam2FastQ program of the bamUtil package [10]. The CHD trio had a large number of read duplicates, which distorted the contamination fraction, so we deduplicated the aligned reads using the biobambam package[3], and used the reads from the deduplicated file to generate the "contaminated" trio.

For each "real-world trio", we repeated this procedure with $\alpha = 0.01, 0.03, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5$.

We aligned the "virtual specimen" and the "pure child" trio reads to the GRCh38DH reference genome following the 1000 Genomes pipeline [14]. Variants were called using Genome Analysis Toolkit 3.8 HaplotypeCaller with the option `-dontUseSoftClippedBases` and restricting the sequence considered to Gencode v.24 protein-coding exons.

2.5. Miscarriage samples. We analyzed DNA from spontaneously miscarried euploid abortuses. Internal Review Board approval was obtained from the Ethics Committee of the Institute of Information Transmission Problems (document 11616-2116/726, 27/11/2017) and the Institutional Review Board of the Skolkovo Institute of Science and Technology (20/06/2019). Families who suffered a miscarriage and requested chromosomal microarray analysis (CMA) of the embryo were offered the choice to participate additionally in the whole-exome sequencing study. Informed consent was obtained from the parents. Only cases where no copy-number anomalies were detected by the CMA were considered for the WES study. Chorionic villi were used for the abortus DNA sample, while blood was drawn from both parents for sequencing. Parenthood and the presence of fetal DNA in the chorionic villus sample were verified using short tandem repeat (STR) analysis with the COrDIS Plus system (GORDIZ, Russian Federation).

Libraries were prepared using the TruSeq DNA Library Prep for Enrichment kit (Illumina, USA). Exome capture was performed using the xGen Exome Research Panel v.1.0 (IDT, USA). Sequencing was performed on the HiSeq4000 instrument (Illumina, USA) in paired-end mode The reads were aligned to the hg19 reference genome. Calling was restricted to the Gencode v.2.7 protein-coding exons with 50-nt flanks. In other respects, the mapping and calling procedure was the same as for the "virtual specimen" trios (see Section 2.4).

3. Results

We compared the genotypes corrected by each of the MCC-aware methods on the "virtual specimen trios" (see Section 2.4) to the genotypes produced by the GATK HaplotypeCaller on the corresponding "pure child" trios, the latter being considered the ground truth. As a baseline, we also included the comparison to initial genotypes called by the GATK HaplotypeCaller on the "virtual specimen" trios ("no genotype correction"). To control for possible artifacts that may result from experimental differences in the original sequencing of the real-world trios, including differences in coverage, we present the results separately for the four sets of "virtual specimen" trios generated from each original family. The machine learning-based models were trained with a "leave-one-out" strategy, where every trio excluding the test trio was used for training. This ensures that we test on data derived from individuals unseen by the trained classifier.

Figure 2 summarizes the results separately for each publicly available family that served as the basis for the "virtual" contaminated trios, with the accuracy achieved by each method plotted against the contamination fraction. All three individual methods fare well, reducing the number of miscalled positions by 40-80% over the baseline of using the original MCC-unaware calls (the "no genotype correction" approach). Both machine learning approaches outperform the maximum likelihood-based method and overcome some of its problems. XGBoost is the best-performing individual method, while the meta-classifier outperforms all individual methods. Restricting analysis just to indels similarly shows

[3] https://www.sanger.ac.uk/science/tools/biobambam

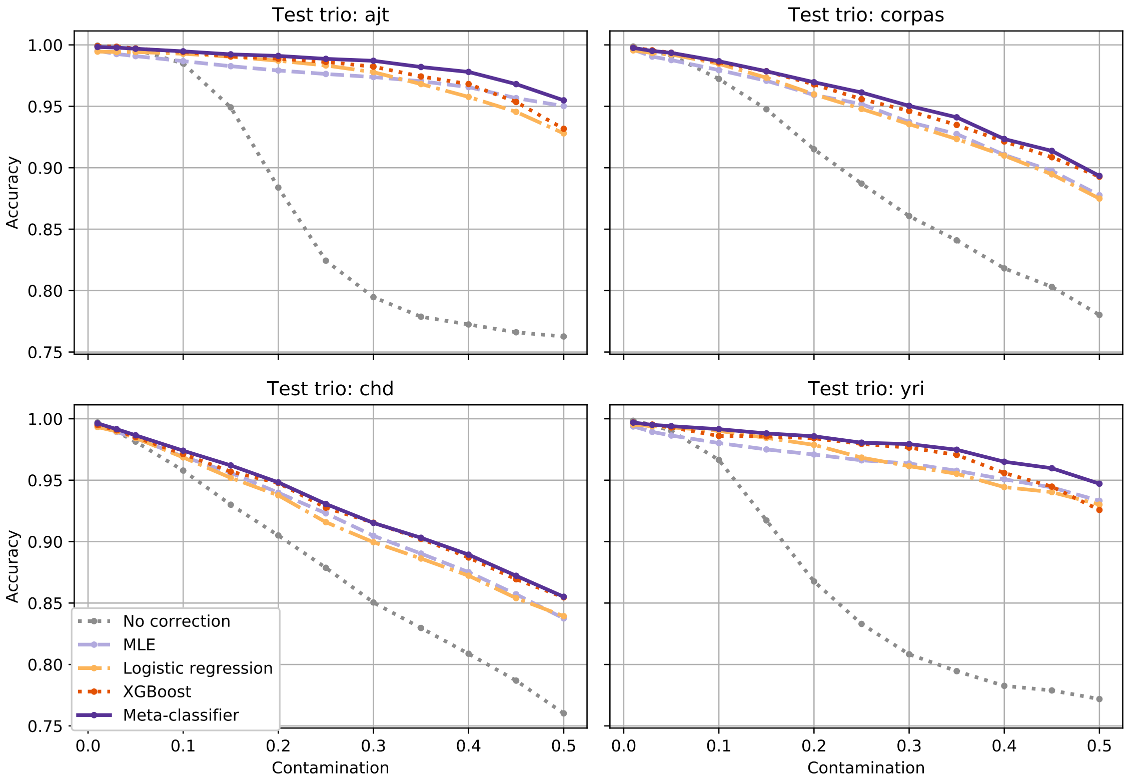

Fig. 2. Accuracy of the genotype correction methods at various MCC fractions. Machine learning methods were trained with a "leave one out" strategy. Each curve consists of twelve data points corresponding to accuracy of the method applied to "contaminated specimens" at various MCCs (the contamination values used are as given in Section 2.4). The accuracy is calculated on the intersection of the calls made on the virtual fetal specimen and the calls made on the real-world child.

Table 1. Training and running times of our methods. Training includes training a model on the entire synthetic dataset excluding the test trio (Section 2.4). Running involves genotype-correcting a single VCF file with a pre-trained model.

Training time	
Logistic Regression	180 s
XGBoost	750 s
Meta-classifier	940 s
Running time	
Logistic Regression	0.0025 s
XGBoost	0.34 s
Maximum Likelihood	0.0021 s
Meta-classifier	0.35 s

Table 2. Top 10 most important features for machine learning methods in descending order. The parameters learned by the trained models (the coefficients of the hyperplanes separating the different classes for logistic regression; the degree of decision tree branching on each feature for XGBoost) can give an estimate of relative feature importance. Parentheses indicate the member of the trio (specimen, mother, father).

	Logistic regression	**XGBoost**
1	Phred likelihood (0/0) (S)	Phred likelihood (0/0) (S)
2	Allelic depth (ref) (S)	Phred likelihood (1/1) (S)
3	Phred likelihood (1/1) (M)	MCC %
4	Allelic depth (alt) (S)	Phred likelihood (0/0) (M)
5	Phred likelihood (1/1) (S)	Phred likelihood (1/1) (M)
6	Phred likelihood (1/1) (M)	Allelic depth (ref) (S)
7	Genotype (0/1) (M)	Phred likelihood (1/1) (F)
8	Phred likelihood (0/1) (S)	Phred likelihood (0/0) (F)
9	Genotype (0/0) (M)	Allelic depth (alt) (S)
10	Allelic depth (alt) (M)	Read depth (F)

improvement in accuracy from genotype correction (Figure S2).

We also compared the running times of the methods for both training and testing (Table 1). In terms of speed, the maximum-likelihood method, which requires no training, is the clear winner. The machine learning methods take on the order of minutes to train. Once the models are trained, though, the running times for genotype-correcting a specimen are comparable.

3.1. Factors affecting the quality of machine learning-based genotype correction.

We further explored the effect of various settings on the performance of the machine learning approaches.

As noted earlier, the training dataset needs to be sufficiently rich for machine learning to to succeed at genotype correction. To demonstrate this, the two machine learning algorithms were trained and tested on different pairs of families (a 'one versus one' approach), at all contamination fractions. The results are summarized in Figure S3. While all the methods lead to an improvement in accuracy, the gains are much more modest than when training with a 'leave-one-out' approach (Figure 2), due to overfitting to the features of a particular trio. To combat overfitting, we ultimately adopted the approach where we train on all available trios except for the test one.

The trained classifiers can tell us about the discriminative power of our input features. Table 2 shows the top 10 features ordered by discriminative power for both classifiers. The letter in the last parentheses represents which sample the feature belongs to (Specimen, Mother, Father).

It is interesting to note that the degree of contamination is not found among the top features for logistic regression. Additionally, the father's features are prominently towards the bottom of the table for logistic regression, yet are included in the top 10 for XGBoost, suggesting some complex relationship that the linear model does not capture. Finally, the linear model gives importance to the one-hot encoded genotypes as well as to the likelihood scores, whereas XGBoost completely ignores the genotypes, implying that the information in the genotypes is redundant when the likelihood scores are known (which is indeed true).

3.2. Robustness of the genotype correction method.

We explored the robustness of the meta-classifier performance to variations in the setup.

First, since in practice, the MCC fraction is not known in advance, we investigated the robustness of the model to errors in MCC fraction estimate. We found that for the range of errors in the MCC estimate we obtained for the simulated specimens (Figure S1), the decrease in performance is close to negligible (Figure S4).

We then investigated the dependence of the genotype correction method performance on read coverage and found that for positions covered by more than 80 reads, which is typical coverage for exome sequencing, the genotype correction accuracy is close to maximal (Figure S5).

Finally, we investigated the effect of the choice of the reference genome on the performance of the meta-classifier. We tested the performance of the classifier with GRCh38 as the reference genome, and in settings where the classifier was trained on hg19-mapped data and tested on GRCh38-mapped data, and vice versa. As seen in Figure S6, the machine learning approach is robust not only to the choice of the reference genome build, but also to changing the reference genome build between training and evaluation. Therefore, a model trained on one reference genome can be used to correct variants called with a different reference.

3.3. Strelka2 genotype correction. To test the performance of the genotype-correction framework on a variant caller other than the GATK HaplotypeCaller, we repeated the tests with the vcf files obtained using the recently published caller Strelka2 [11] using its standard germline calling pipeline and the same VCF features as were available from GATK HaplotypeCaller. We observed the same trends, namely, that genotype correction increases the concordance between the calls on the "pure child" and the "contaminated specimen" trios (Figure S7).

3.4. Application to clinical miscarriage data. We applied the MCC estimation algorithm from Section 2.1 to 33 abortus samples from a larger study aimed at finding genetic causes of spontaneous miscarriages of euploid fetuses (data to be published separately). In five of the 33 cases, the estimated MCC in the abortus exceeded 5%, with one case of 8% MCC and one of 40% MCC. We will focus on the one abortus sample with the 40% MCC (miscarried at 9 weeks). We applied the recommended GATK hard filters [17] to the calls and further required that GQ be at least 30 for both parents. We did not apply the GQ filter to the abortus sample, as we expect maternal contamination to lower the quality of the calls in the abortus and the GQ filter may therefore exclude legitimate variants. As a result, we obtained 65431 calls by MCC-naive GATK HaplotypeCaller before genotype correction, of which 11796 were genotype-corrected by the meta-classifier (trained on the "virtual specimens"). As we expected (see Section 2.2), all but one corrections affected the heterozygous genotype call. Also not surprisingly, correction predominantly affected heterozygous variants with skewed allelic depths and favored the “elimination” of the allele with the lower allelic depth (Figure S8).

Contamination-aware genotype correction had several effects, both on the global variant statistics and on the particular list of candidate mutations that could explain the pregnancy loss. Genotype correction rectified the highly skewed ratio of heterozygous to homozygous variants called in the abortus and brought it in line with that for the uncontaminated parent samples. Thus, the parents' heterozygous to homozygous call ratio was 3.1, while for the MCC-naive abortus call set, it was 6.7. After genotype correction, that number for the abortus was brought down to 3.8, much closer to the parents' value. Regarding the possible genetic explanation of pregnancy loss, the uncorrected call set contained six candidate compound heterozygote variants (Supplementary Methods; Table S1). Of these, genotype correction eliminated three variant calls, which reduced the number of candidate compound heterozygotes to three. Thus, in this case, MCC-aware variant (re-)calling eliminated potential false positive candidate variants.

4. Discussion and Conclusions

As its costs are dropping, high-throughput sequencing is becoming increasingly routine in clinical and academic practice, and is gaining ground in prenatal diagnosis and in the diagnostic study of miscarried fetuses. While the bioinformatics toolkit for analyzing the sequencing data of individuals is well-developed and mature, it may encounter difficulties in the analysis of products of conception, which may be mixed with maternal tissue. We have shown that the accuracy of standard variant-calling pipelines does indeed degenerate as the maternal cell contamination increases, but that much of this decline can be alleviated by MCC-aware variant genotype correction. We have demonstrated that the MCC fraction can be readily estimated from trio sequencing data and then used to inform the genotype correction. We have further shown, using synthetic contaminated samples obtained from real trio data, that even a simple method based on maximum likelihood estimation does much to correct for the MCC-related loss of calling accuracy, and that the machine-learning methods trained on simulated data show even greater improvement, with the more sophisticated XGBoost having the better performance of the individual methods. The best performance, meanwhile, was obtained with the ensemble method that incorporates the results of all three methods that we tried.

We note that all methods discussed in this paper only correct the variants that have been discovered by a general-purpose variant caller (we used GATK HaplotypeCaller which is commonly employed for human data analysis, but any variant caller which outputs sufficiently rich information about variants into a VCF file should work). Therefore, MCC-aware genotype correction is a simple add-on to a standard bioinformatics pipeline, needing only the VCF file produced by it. The user can then compare the original VCF to the corrected VCF to see which genotypes had been changed. On the other hand, the reliance on standard variant calling implies that if MCC had caused the true fetal variant to be missed in all members of the trio, then that variant will not be called by the genotype-correction pipeline, either. We expect such cases to be rare.

Among published methods, CleanCall [7] applies its own probabilistic model to correct for sample contamination, starting with read-level information, and it has an option for doing so with a known contaminating sample (https://github.com/hyunminkang/cleancall/). Since our method works with the output of a third-party variant caller, and comparing it to CleanCall cannot be decoupled from comparing CleanCall's model to HaplotypeCaller's, they are not directly comparable. Nevertheless, we believe that in the case of maternal cell contamination with maternal sequence data available, our method is preferable, both because it is designed for that particular situation and because it allows the user to utilize the caller of his or her choice.

To show the applicatiblity of our method to real-world data, we applied it to in-house sequencing data for a spontaneously miscarried euploid embryo with 40% maternal contamination and demonstrated that it can correct for artifacts caused by maternal contamination.

Currently, next-generation sequencing is applied to fetal DNA obtained by invasive methods which extract amniotic fluid or fetal or chorionic villi cells. Our work is aimed at the analysis of this sort of sequencing data, and we do not consider MCC fractions higher than 50%. As its costs drop, sequencing may become a practical option in non-invasive prenatal testing (NIPT) as well. NIPT analyzes fetal DNA circulating in maternal blood which constitutes a small fraction of the DNA sample obtained. Analyzing that sort of data would make MCC-aware variant calling a necessity, and may in fact give rise to more sophisticated and/or specialized algorithms, such as those that work directly with mapped reads.

Supplementary material

Supplementary material is available at *bioRxiv* online. We also present a repository with scripts and notebooks for reproducing our experiments as well as applying our methods: https://github.com/bazykinlab/ML-maternal-cell-contamination

Simulated trio data are available at https://archive.org/download/simulated_MCC/simulated_MCC.tar.gz

Funding

The work was funded by the Skoltech Biomedical Initiative grant to GAB and DY.

References

1. Sunayna Best, Karen Wou, Neeta Vora, Ignatia B. Van der Veyver, Ronald Wapner, and Lyn S. Chitty. Promises, pitfalls and practicalities of prenatal whole exome sequencing. *Prenatal diagnosis*, 38(1):10–19, January 2018.

2. Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. pages 785–794, 2016.

3. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571):68–74, October 2015.

4. Manuel Corpas, Willy Valdivia-Granda, Nazareth Torres, Bastian Greshake, Alain Coletta, Alexej Knaus, Andrew P. Harrison, Mike Cariaso, Federico Moran, Fiona Nielsen, Daniel Swan, David Y. Weiss Solís, Peter Krawitz, Frank Schacherer, Peter Schols, Huangming Yang, Pascal Borry, Gustavo Glusman, and Peter N. Robinson. Crowdsourced direct-to-consumer genomic analysis of a family quartet. *BMC Genomics*, 16(1):910, November 2015.

5. Christopher DeBoever, Matthew Aguirre, Yosuke Tanigawa, Chris C. A. Spencer, Timothy Poterba, Carlos D Bustamante, Mark J. Daly, Matti Pirinen, and Manuel A Rivas. Bayesian model comparison for rare variant association studies of multiple phenotypes. *bioRxiv*, 2018.

6. Jacob F. Degner, John C. Marioni, Athma A. Pai, Joseph K. Pickrell, Everlyne Nkadori, Yoav Gilad, and Jonathan K. Pritchard. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*, 25(24):3207–3212, December 2009.

7. Matthew Flickinger, Goo Jun, Gonçalo R. Abecasis, Michael Boehnke, and Hyun Min Kang. Correcting for Sample Contamination in Genotype Calling of DNA Sequence Data. *American Journal of Human Genetics*, 97(2):284–290, August 2015.

8. Zhen Jia, Mao Fengbiao, Lu Wang, Mingzhen Li, Yueyi Shi, Baorong Zhang, and Guolan Gao. Whole-exome sequencing identifies a de novo mutation in trpm4 involved in pleiotropic ventricular septal defect. 10:5092–5104, 05 2017.

9. Goo Jun, Matthew Flickinger, Kurt N. Hetrick, Jane M. Romm, Kimberly F. Doheny, Gonçalo R. Abecasis, Michaẽl Boehnke, and Hyun Min Kang. Detecting and Estimating Contamination of Human DNA Samples in Sequencing and Array-Based Genotype Data. *The American Journal of Human Genetics*, 91(5):839–848, November 2012.

10. Goo Jun, Mary Kate Wing, Gonçalo R. Abecasis, and Hyun Min Kang. An efficient and scalable analysis framework for variant extraction and refinement from population scale DNA sequence data. *Genome Research*, page gr.176552.114, April 2015.

11. Sangtae Kim, Konrad Scheffler, Aaron L. Halpern, Mitchell A. Bekritsky, Eunho Noh, Morten Källberg, Yeonbin Chen, Xiaoyu a nd Kim, Doruk Beyter, Peter Krusche, and Christopher T. Saunders. Strelka2: fast and accurate calling of germline and somatic variants. *Nature Methods*, 15(8):591, August 2018.

12. Allen N. Lamb, Jill A. Rosenfeld, Justine Coppinger, Erin T. Dodge, Mindy Preston Dabell, Beth S. Torchia, J. Britt Ravnan, Lisa G. Shaffer, and Blake C. Ballif. Defining the impact of maternal cell contamination on the interpretation of prenatal microarray analysis. *Genetics in Medicine*, 14(11):914–921, November 2012.

13. Narasimhan Nagan, Nicole E. Faulkner, Christine Curtis, and Iris Schrijver. Laboratory Guidelines for Detection, Interpretation, and Reporting of Maternal Cell Contamination in Prenatal Analyses. *The Journal of Molecular Diagnostics : JMD*, 13(1):7–11, January 2011.

14. 1000 Genomes Project. Grch38 alignment readme, 2015. https://github.com/igsr/1000Genomes_data_indexes/blob/master/data_collections/1000_genomes_project/README.1000genomes.GRCh38DH.alignment.

15. Taita Stojilkovic-Mikic, Kathy Mann, Zoe Docherty, and Caroline Mackie Ogilvie. Maternal cell contamination of prenatal samples assessed by QF-PCR genotyping. *Prenatal Diagnosis*, 25(1):79–83, 2005.

16. Ahmad N. Abou Tayoun, Nancy B. Spinner, Heidi L. Rehm, Robert C. Green, and Diana W. Bianchi. Prenatal DNA Sequencing: Clinical, Counseling, and Diagnostic Laboratory Considerations. *Prenatal Diagnosis*, 38(1):26–32, January 2018.

17. Geraldine Van der Auwera. (howto) apply hard filters to a call set, 2013. https://gatkforums.broadinstitute.org/gatk/discussion/2806/howto-apply-hard-filters-to-a-call-set, Edited November 2018.

18. Jennifer Weida, Avinash S. Patil, Frank P. Schubert, Gail Vance, Holli Drendel, Angela Reese, Stephen Dlouhy, Shaochun Bai, and Men-Jean Lee. Prevalence of maternal cell contamination in amniotic fluid samples. *The Journal of Maternal-Fetal & Neonatal Medicine*, 30(17):2133–2137, September 2017.

19. Justin M. Zook, David Catoe, Jennifer McDaniel, Lindsay Vang, Noah Spies, Arend Sidow, Ziming Weng, Yuling Liu, Christopher E. Mason, Noah Alexander, Elizabeth Henaff, Alexa B. R. McIntyre, Dhruva Chandramohan, Feng Chen, Erich Jaeger, Ali Moshrefi, Khoa Pham, William Stedman, Tiffany Liang, Michael Saghbini, Zeljko Dzakula, Alex Hastie, Han Cao, Gintaras Deikus, Eric Schadt, Robert Sebra, Ali Bashir, Rebecca M. Truty, Christopher C. Chang, Natali Gulbahce, Keyan Zhao, Srinka Ghosh, Fiona Hyland, Yutao Fu, Mark Chaisson, Chunlin Xiao, Jonathan Trow, Stephen T. Sherry, Alexander W. Zaranek, Madeleine Ball, Jason Bobe, Preston Estep, George M. Church, Patrick Marks, Sofia Kyriazopoulou-Panagiotopoulou, Grace X. Y. Zheng, Michael Schnall-Levin, Heather S. Ordonez, Patrice A. Mudivarti, Kristina Giorda, Ying Sheng, Karoline Bjarnesdatter Rypdal, and Marc Salit. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific Data*, 3:160025, June 2016.