Pacific Symposium on Biocomputing 25:599-610(2020)

Coverage profile correction of shallow-depth circulating cell-free DNA sequencing via multi-distance learning

Nicholas B. Larson, Melissa C. Larson, Jie Na, Carlos P. Sosa,
Chen Wang, Jean-Pierre Kocher

Department of Health Sciences Research, Mayo Clinic College of Medicine and Sciences
200 1st Street SW, Rochester, MN 55901 USA
Email: Larson.nicholas@mayo.edu

Ross Rowsey

Department of Laboratory Medicine and Pathology, Mayo Clinic College of Medicine and Sciences
200 1st Street SW, Rochester, MN 55901 USA

Shallow-depth whole-genome sequencing (WGS) of circulating cell-free DNA (ccfDNA) is a popular approach for non-invasive genomic screening assays, including liquid biopsy for early detection of invasive tumors as well as non-invasive prenatal screening (NIPS) for common fetal trisomies. In contrast to nuclear DNA WGS, ccfDNA WGS exhibits extensive inter- and intra-sample coverage variability that is not fully explained by typical sources of variation in WGS, such as GC content. This variability may inflate false positive and false negative screening rates of copy-number alterations and aneuploidy, particularly if these features are present at a relatively low proportion of total sequenced content. Herein, we propose an empirically-driven coverage correction strategy that leverages prior annotation information in a multi-distance learning context to improve within-sample coverage profile correction. Specifically, we train a weighted k-nearest neighbors-style method on non-pregnant female donor ccfDNA WGS samples, and apply it to NIPS samples to evaluate coverage profile variability reduction. We additionally characterize improvement in the discrimination of positive fetal trisomy cases relative to normal controls, and compare our results against a more traditional regression-based approach to profile coverage correction based on GC content and mappability. Under cross-validation, performance measures indicated benefit to combining the two feature sets relative to either in isolation. We also observed substantial improvement in coverage profile variability reduction in leave-out clinical NIPS samples, with variability reduced by 26.5-53.5% relative to the standard regression-based method as quantified by median absolute deviation. Finally, we observed improvement discrimination for screening positive trisomy cases reducing ccfDNA WGS coverage variability while additionally improving NIPS trisomy screening assay performance. Overall, our results indicate that machine learning approaches can substantially improve ccfDNA WGS coverage profile correction and downstream analyses.

Keywords: cell-free DNA, kNN, distance, annotation, next-generation sequencing

1. Introduction

Circulating cell-free DNA (ccfDNA) is comprised of relatively short fragments of genomic material that naturally occur in bodily fluids and originate primarily from normal cell apoptosis[1]. A number of biomedical applications have been identified for shallow-depth whole-genome sequencing (WGS) of plasma ccfDNA, including liquid biopsy for early identification of invasive tumors[2] as well as non-invasive prenatal screening (NIPS) for fetal genetic and genomic abnormalities[3,4]. For downstream inference, these data are typically summarized by generating binned coverage profiles of the sequencing output, whereby the genome is uniformly partitioned into moderately sized contiguous regions (e.g., 10-50 kilobases (kb)) and count-based coverage is calculated by the enumerating the overlapping sequencing reads. Evidence of copy-number variants (CNVs) and aneuploidy may be detected from these profiles using standard CNV detection methods for coverage data[5].

Coverage profile patterns for plasma ccfDNA WGS exhibit a large degree of non-uniformity relative to standard nuclear DNA WGS, which is in part attributable to biased preservation of DNA originating from regions that are protected from degradation by nucleases in the blood stream, including nucleosome- and protein-bound DNA[6]. These patterns can in turn be exploited via deconvolution to identify evidence of tissue-of-origin admixture, and have been leveraged using machine learning methods to build fetal fraction predictors for NIPS[7]. However, ccfDNA exhibits a large deal of inter-sample heterogeneity and GC-content correction only explains a moderate proportion of this variability. In the context of NIPS and fetal trisomy detection, this may necessitate higher fetal-fraction quality control thresholds to achieve desired assay sensitivity and specificity, delaying recommended gestational age for the assay and leading to repeat maternal blood draws when estimated fetal fraction in too low.

Previous studies have recognized the value of large-scale empirical coverage correlations in ccfDNA coverage profile NIPS analyses to address inter- and intra-sample variability. For example, correlations of gross chromosomal read count proportions can be leveraged to improve trisomy detection[8,9]. Straver et al.[10] proposed WISECONDOR for coarsely binned (i.e., 1 Mb) micro-duplication and deletion detection using pair-wise empirical bin coverage similarity. Extension of these concepts to general profile bias correction under the small genomic bin sizes typical of ccfDNA coverage profiling is appealing, given the fine granularity of epigenomic variability that contributes to profile coverage heterogeneity. However, application of algorithms like WISECONDOR for these small bin sizes is computationally challenging, as this requires calculating pair-wise bin similarities under a much larger bin dimensionality.

In this paper, we explore computationally efficient machine learning strategies for improving within-sample coverage profile correction relative to standard regression-based methods commonly implemented for GC-content correction. Using a large set of plasma ccfDNA WGS coverage profiles from NIPS analyses, we reformulate the problem as a simple k-nearest neighbors (kNN) regression approach and further propose methods to integrate prior knowledge captured in genomic annotation sources via a supervised multi-distance learning framework. We compare coverage profile variability reduction in real NIPS maternal plasma ccfDNA WGS data, and additionally characterize potential improvement in discrimination of trisomy cases and negative controls for common fetal trisomies of 13, 18, and 21. Finally, we discuss further research directions in the area of ccfDNA WGS coverage data analysis.

2. Methods

2.1. *Data description*

2.1.1. Samples

De-identified samples from research and clinical NIPS results conducted by the Genomics Laboratory at Mayo Clinic were considered eligible for this study. Of these, we identified a total of 476 single fetus normal karyotype pregnancy samples, 145 positive trisomy samples (10 trisomy 13, 41 trisomy 18, 104 trisomy 21), and 790 non-pregnant female donor samples for our analyses. Plasma was obtained from blood and stored in a Streck Cell-Free DNA Blood Collection Tube (Streck, Omaha NE). Use of these data for research purposes was approved by the Institutional Review Board.

2.1.2. *Shallow-depth whole-genome sequencing*

DNA was extracted from plasma using the Qiagen Circulating Nucleic Acid Kit (Qiagen, Venlo Netherlands) and library preparation was conducted using the Illumina TruSeq® Nano DNA Sample Preparation Kit (Illumina, San Diego CA). Sequencing was performed on the Illumina HiSeq 2500 in Rapid Run mode to generate 50-cycle single-end reads, which were aligned to the hg19 human genome reference using Novoalign (Novocraft, Selangor Malaysia). Chromosomal coverage summaries for 50 kilobase (kb) contiguous genomic windows were generated from the resulting BAM files using the WANDY bioinformatics pipeline (http://bioinformaticstools.mayo.edu/research/wandy/), an in-house developed workflow for bin filtering, GC correction, and normalization of low-depth whole-genome sequencing output to identify copy-number variants and aneuploidy.

2.1.3. *Bin data preprocessing*

A total of 57,633 50 kb autosomal genomic bins were initially pre-filtered using an in-house defined set of bins that were previously classified as being unreliable (e.g., poor mappability, repeat regions), resulting in $B = 49{,}867$ bins (87%) under consideration. Raw coverage values for remaining bins were defined as the number of ccfDNA sequencing reads whose start overlaps each bin. We then normalized coverage values within sample by dividing by the mean coverage value across bins. We designate these intermediate coverage values as the $B \times N$ matrix $\boldsymbol{X}$ for some N set of observed coverage profiles.

2.2. *Neighborhood-based coverage correction*

GC-content-based coverage correction methods can be conceptualized as clustering bins by a shared annotation characteristic, such that bin strata defined by the same GC value serve as the basis for removing the bias induced by the sample-specific GC-coverage relationship. Similarly, empirical bin-to-bin similarity from retrospective data may improve coverage profile bias correction by identifying bins with similar empirical coverage patterns across multiple samples. WISECONDOR leverages a $B \times B$ dissimilarity matrix, $\boldsymbol{D}$, defined by the squared Euclidean distance (SED) of bin pairs across samples, which is then used to identify reference bins for each

given "target" bin. These reference bins then serve as the expected distribution of that bin in a new sample, such that the mean and standard deviations are used to compute bin-wise Z-scores. Such reference sets can alternatively be conceptualized as bin neighborhoods, such that the problem is alternatively posed as a type of k-nearest-neighbors (kNN) regression analysis.

2.2.1. *Weighted distance averaging*

Due to the high dimensionality of B under small bin partitioning, we may additionally wish to leverage a priori knowledge about fixed genomic annotation features (e.g., GC content) which contribute to a large proportion of bin coverage variability. Combining prior annotation dissimilarity with empirical dissimilarity measures in some supervised fashion is desirable, as the former could impart some form of regularization on empirical coverage dissimilarity measures and improve overall coverage correction performance, particularly if our training data suffers from small N dimensionality relative to B. This amounts to a supervised multi-distance learning problem, as we are seeking to optimally combine dissimilarities from two feature sets as a final input feature representation for model training.

Consider any distance function whereby the dissimilarity measure between bins i and j with corresponding feature vectors $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$ is the summation of pair-wise feature distances, such as the Manhattan or SED distance functions. For the latter, this is defined by

$$d(\boldsymbol{x_i}, \boldsymbol{x_j}) = \left\|\boldsymbol{x_i} - \boldsymbol{x_j}\right\|_2^2 = \sum_p \left(x_{i,p} - x_{j,p}\right)^2 \quad (1)$$

If we have two feature sets $\boldsymbol{X}$ and $\boldsymbol{Y}$ from which respective distance matrices $\boldsymbol{D_X}$ and $\boldsymbol{D_Y}$ may be calculated using $d(\cdot,\cdot)$, these can be efficiently combined under such a distance function by appropriately concatenating and weighting the feature vectors[11]. Define $D_\alpha(i,j)$ to be element (i,j) in a $B \times B$ distance matrix defined as $\boldsymbol{D_\alpha} = (1-\alpha)\boldsymbol{D_X} + \alpha \boldsymbol{D_Y}$, where α is a mixing parameter that defines the weighted contribution of each component distance matrix. It is simple to show that $D_\alpha(i,j)$ is equivalent to $d(\boldsymbol{z_i}, \boldsymbol{z_j})$, where $\boldsymbol{Z} = [\sqrt{1-\alpha}\boldsymbol{X} \quad \sqrt{\alpha}\boldsymbol{Y}]$ is the weighted column concatenation of matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$, since

$$d(z_i, z_j) = \sum_p \left(z_{i,p} - z_{j,p}\right)^2 = (1-\alpha)\sum_{p\in\mathcal{X}} \left(z_{i,p} - z_{j,p}\right)^2 + \alpha \sum_{q\in\mathcal{Y}} \left(z_{i,q} - z_{j,q}\right)^2 \quad (2)$$

where $\mathcal{X}$ and $\mathcal{Y}$ denote the respective feature sets unique to $\boldsymbol{X}$ and $\boldsymbol{Y}$. This relationship is computationally advantageous, as being able to concatenate the feature spaces in this manner facilitates the use of rapid distance calculation algorithms that identify the leading k neighbors and their corresponding distances within a single input feature set, rather than computing the complete distance matrices $\boldsymbol{D_X}$ and $\boldsymbol{D_Y}$ prior to weighted combination, which is computationally and memory intensive. Additionally, optimization over α provides information on relative contributions of "prior" and "observed" data from $\boldsymbol{Y}$ and $\boldsymbol{X}$, respectively.

To implement this approach, we applied the kd-tree searching functions for nearest neighbor indices and distances as implemented by the FNN R package, which support fast Euclidean distance calculations for the k nearest neighbors via the approximate near neighbors C++ library[12]. Since the relationship between SED and Euclidean distance is monotone, this is equivalent for identifying nearest neighbors under SED (although the output distances can be squared to maintain SED distance values). To also ensure distances are comparable across the feature sets prior to

combining, we adopted the double-scaled Euclidean approach for distance normalization, whereby column p in $\boldsymbol{Z}$ is further divided by $\sqrt{v_S \cdot d_p}$, where v_s is the column dimensionality of the respective feature set $S \in \{\mathcal{X}, \mathcal{Y}\}$ to which p originally belongs, and d_p is the maximum potential distance for feature p as defined by $\left(\max_i(z_{i,p}) - \min_i(z_{i,p})\right)^2$.

2.2.2. *Genomic annotation*

For this analysis, we designate the prior annotation information $\boldsymbol{Y}$ to be comprised of two specific genomic bin annotation sources: (1) GC content and (2) mappability scores, such that $\boldsymbol{Y}$ is $B \times 2$. Both of these have strong *a priori* relationships with coverage profile variability, as depicted in Figure 1.

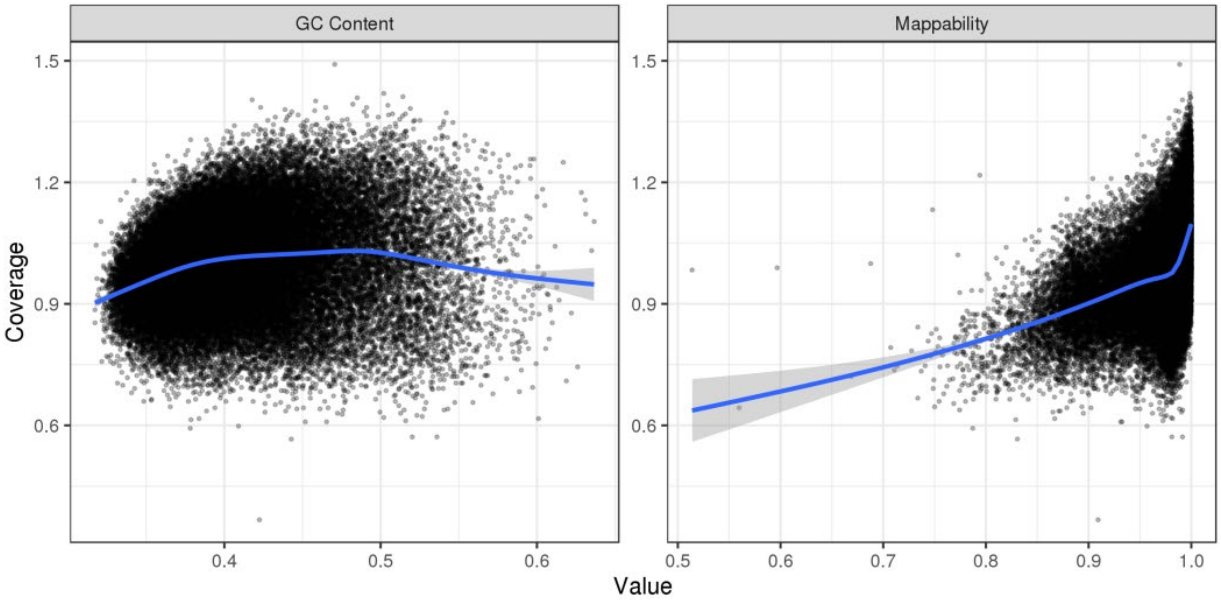

Fig. 1. Example coverage relationships with GC content and mappability for a given sample.

2.3. *Standard annotation-based correction*

GC content can heavily influence coverage profiles derived from WGS[13], and a standard pre-processing step for WGS coverage profile data prior to CNV analysis is within-sample GC-content correction. For a given sample, genomic bins are typically stratified by a shared GC content value and the mean or median stratum-specific coverage value (often with additional lowess smoothing) is used for further normalization. Herein, we fit generalized additive models (GAMs) using the *gam* R package, where smooth functions are fit using GC content and mappability values as predictors for observed bin coverage for a given sample. Observed raw coverage values were then divided by the predicted coverage values produced by the fitted model, such that corrected coverage values are non-negative with an expectation of 1.

2.4. *Proposed coverage correction approach*

Since bins within the physical neighborhood of a given bin (i.e., in *cis*) may also have a higher likelihood of having highly correlated coverage, including these bins in the correction procedure may inadvertently over-correct true copy-number alterations that overlap those bins. This is particularly true in the instances of aneuploidy, where abnormal numbers of chromosomes could be less detectable if neighborhoods were largely comprised of *cis* bins. Thus, we restrict the potential neighboring bins to be those in $trans$ with the candidate bin b, such that they occur on different chromosomes than b. To facilitate this, we split the data by individual chromosomes and their complement, such that bins in the chromosome of interest are queried against the complement for nearest neighbors in the combined feature space. This also allows for parallelization and improves overall training computational efficiency.

For coverage correction, we considered both an unweighted and dissimilarity-weighted kNN (wkNN) strategy for within-sample coverage correction, such that raw coverage values of the k nearest *trans* bins are used to generate a mean prediction for said bin. That is, for a $B \times 1$ vector $\boldsymbol{x}$ of raw coverage profile data for a new sample, we define the predicted coverage for bin i as $\hat{x}_i = \frac{1}{\sum_{j \in \mathcal{K}_i} w_{i,j}} \sum_{j \in \mathcal{K}_i} w_{i,j} x_j$ where $\mathcal{K}_i$ is the size k neighboring set for bin i and $w_{i,j}$ is defined as $1/D_\alpha(i,j)$ for the distance-weighted approach and $w_{i,j} = 1 \forall i,j$ in the unweighted version. To perform coverage correction, we again define $x_i^{corr} = \frac{x_i}{\hat{x}_i}$, such that the observed value is divided by the predicted value. More generally, the "model" itself can be represented simply by a $B \times B$ sparse weight matrix $\boldsymbol{W}$ where $W_{i,j} = w_{i,j}$ for $(i,j) \in \{\mathcal{K}_1^*, \dots, \mathcal{K}_B^*\}$ where $\mathcal{K}_i^*$ are of tuples corresponding to the query bin i and its neighboring bins $\mathcal{K}_i$, and 0 otherwise, and $\hat{\boldsymbol{x}} = \boldsymbol{W}\boldsymbol{x}$.

2.5. *Performance Evaluation*

2.5.1. *Model Fitting*

To tune the model parameters (α, k), we adopted a simple grid search in combination with 5-fold cross-validation in our training set of N = 790 non-pregnant female donor samples. We elected to use only these non-pregnant samples for model training and validation because contaminating fetal ccfDNA in maternal plasma has a different coverage profile due to underlying epigenomic differences[7]. Training a model using data from pregnant female samples could lead to neighborhoods of bins that highly correlate with fetal fraction of ccfDNA and inadvertently over-correct true fetal genomic signal in the coverage data.

Under the cross-validation framework, we considered the samples within $\boldsymbol{X}$ to be the mode of cross-validation, such that the columns of $\boldsymbol{X}$ were split into folds for purposes of contributing to $\boldsymbol{D_\alpha}$ or to performance evaluation. We set the potential tuning parameter values to be $k \in (5, \dots, 300)$ at increments of 5 and $\alpha \in (0, 0.001, 0.01, 0.1, 0.25, 0.5, 0.75, 1)$, such that $\alpha = 0,1$ respectively denote models based entirely on $\boldsymbol{X}$ or $\boldsymbol{Y}$, respectively. We defined our loss function for tuning as the mean absolute error (MAE) of the individual out-of-fold sample coverage profiles across samples, such that $MAE = \frac{1}{N \times B} \sum_{l=1}^{N} \sum_{b=1}^{B} |x_{l,b} - \hat{x}_{l,b}|$. A final model parameterization was then selected based on best cross-validated performance and fit on the complete training data.

2.5.2. *Profile Correction of NIPS Samples*

To characterize maternal ccfDNA coverage profile bias correction performance, we considered N=476 available clinical NIPS samples ostensibly free of trisomies per previously reported screening results. The applied modes were based on training of all available NPP data. We considered the within-sample median absolute deviation about the median (MAD), a commonly used metric to characterize sample coverage variability in CNV analysis of WGS data, as a sample-level performance metric for coverage profile correction. As a baseline comparator method for coverage correction, we also performed standard within-sample annotation-based coverage correction described in Section 2.3.

2.5.3. *Fetal trisomy detection*

In addition to variability reduction, we want to ensure our approach preserves and potentially improves true positive signals relative to standard coverage profile correction methods. In NIPS, we typically derive chromosome-wise proportions of coverage profiles to determine the presence or absence of common fetal trisomies (i.e., 13, 18, and 21). That is, for a given sample bin coverage vector $\boldsymbol{x}$ and chromosome c, we define proportion $\pi_c = \frac{\sum_{b \in \mathcal{C}_c} x_b}{\sum_k x_k}$ where $\mathcal{C}_c$ is the set of bins that correspond to chromosome c and $\sum_{c=1}^{22} \pi_c = 1$. We considered the simple Z-score approach for trisomy detection[9], such that a large fixed set of reference normal NIPS samples free of trisomies is used to characterize population distributional properties about π_c (i.e., mean μ_c and standard deviation SD_c) for trisomy-prone chromosomes. Then, the screening test statistic is defined as $Z_c = \frac{\pi_c - \hat{\mu}_c}{\widehat{SD}_c}$. For our purposes, we randomly selected 300 negative NIPS samples to serve as the reference set. Trisomy screening Z-scores were then calculated for the remaining 176 negative samples along with 145 positive trisomy samples (10 Trisomy 13, 41 Trisomy 18, 104 Trisomy 21). This contrasts WISECONDOR's strictly within-sample inference for trisomy, which is based on a Stouffer's combined Z approach and requires explicit tuning of a decision threshold.

Screening performance for our available NIPS samples already achieves near perfect discrimination using standard coverage profile correction methods due to imposed quality control standards (e.g., minimum sufficient fetal fraction). To assess significant signal improvement (i.e., larger Z-scores for positive cases), we alternatively performed a Wilcoxon signed rank test of the paired Z-scores under each coverage correction approach by trisomy. Evidence of increased Z-scores for positive cases (but not in controls) would indicate that existing thresholds could be relaxed, reducing the number of quality control failures and improving overall assay sensitivity/specificity.

2.6. *Code availability*

All analyses were performed using R version 3.5.2 (R Core Team, Vienna, Austria). Relevant R code is made publicly available at https://github.com/nblarson/ccfdna_coverage.git.

3. Results

3.1. *Simulation Results*

The MAE performance measures for the non-pregnant donor data are presented in Figure 2 for the unweighted kNN approach across the considered tuning parameter values. Results were highly comparable to the weighted approach, with a median difference in MAE of 1.0E-05 in favor of the weighted method (range: 0 – 1.8E-04). The optimal tuning parameter settings were also the same for both types of approaches ($\alpha = 0.01$ and $k = 150$), which also led to nearly identical cross-validation performance results (MAE = 0.04134 for both methods). In contrast, using the same $k = 150$, performance for $\alpha = 0$ and $\alpha = 1$ were respectively 0.04137 and 0.06816. Overall, these results indicate that the bins were very densely arranged in the feature space, as the selected k is quite large and the weighted kNN demonstrated modest performance gains. Moreover, the amount of available training data was largely sufficient to capture bin-to-bin dissimilarities, with the kNN approach benefitting from a small degree of "regularization" afforded by the annotation-based dissimilarity in our multi-distance learning approach, as indicated by the small value of α selected for the final models and the modest difference in MAE relative to $\alpha = 0$.

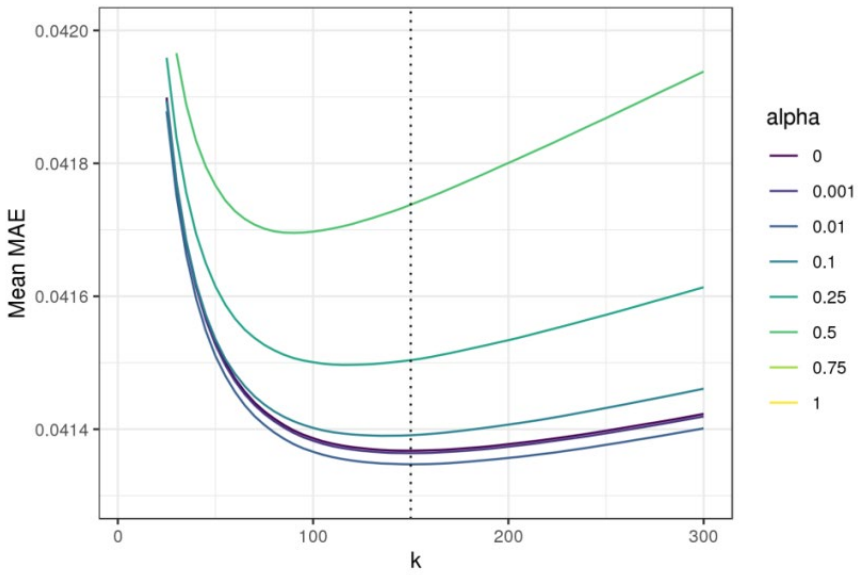

Fig. 2. Cross-validated MAE measures for unweighted kNN approach. Figure is zoomed in to demonstrate separation across alpha values for best performing settings. Vertical dotted line indicates optimal k for $\alpha = 0.01$.

To explore conditions of substantially reduced availability of training data, we additionally performed the same kNN model training with 10 random subsets of 50 NPP samples. The mean (range) of optimal tuning parameters was $k = 155$ (100-195) and $\alpha = 0.14$ (0.1-0.5), with mean MAE = 0.0419 (range: 0.0420-0.491). These results follow intuition that a larger value of α would be selected in model training when more limited information is available in $\boldsymbol{X}$, and furthermore illustrate that good performance can be achieved with relatively few training samples.

3.2. *Coverage correction*

There were 476 negative trisomy NIPS samples available for comparative performance analysis. Using the wkNN with tuning parameters selected via the cross-validation results in Section 3.1, we derived corrected coverage profiles and compared them to results based on the annotation-based GAM model. Overall, we observed a median single-sample MAD reduction of 38.7% relative to the GAM regression approach, with the sample MAD consistently lower using our proposed wkNN method (range in MAD reduction: 26.5-53.5%). Further summary statistics of these results are presented in Table 1.

Table 1. Profile coverage correction MAD summaries for GAM and wkNN methods.

MAD (N = 476)	GAM Model	wkNN Model	% Reduction
Minimum	0.072	0.041	26.5
1st Quartile	0.081	0.049	36.3
Mean	0.085	0.051	38.7
Median	0.085	0.052	38.7
3rd Quartile	0.089	0.055	41.4
Maximum	0.118	0.075	53.5

Comparison of individual coverage profiles also demonstrated substantial smoothing effects of both variability and bias relative to the GAM approach. An illustrated example of typical coverage profile improvement (MAD reduction: 34.2%) is presented in Figure 3 for the trisomy prone chromosomes 13, 18 and 21.

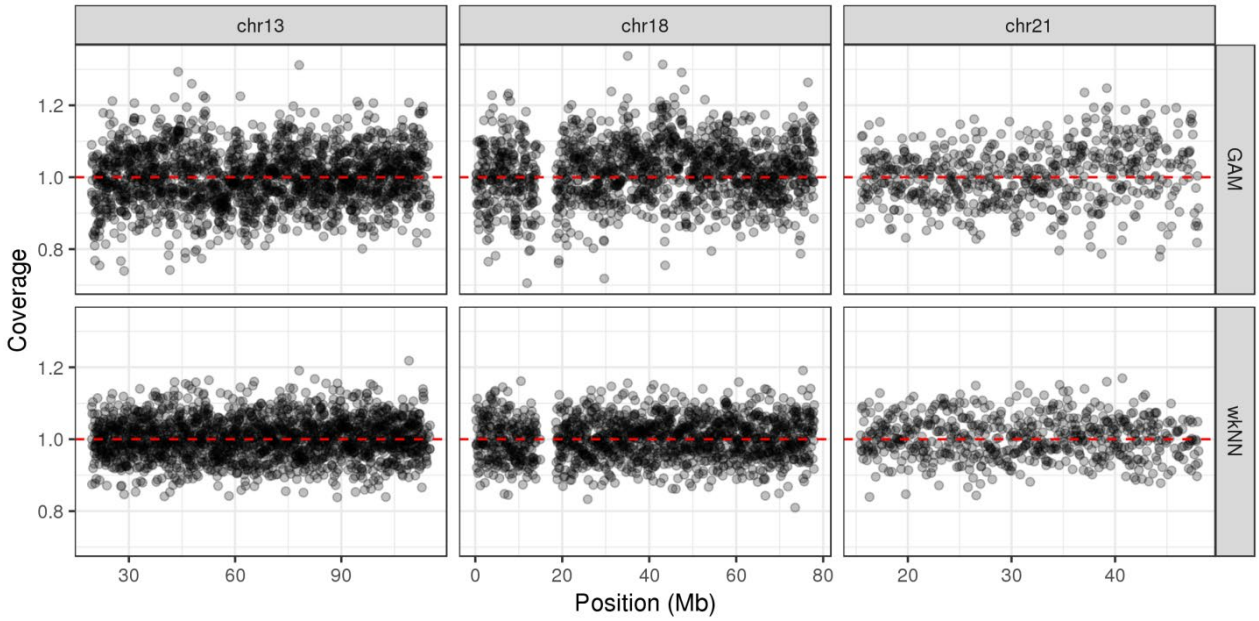

Fig. 3. Corrected 50 kb bin coverage profiles using the GAM approach (top) vs. the proposed wkNN model (bottom) for an example negative trisomy NIPS test sample for chromosomes 13, 18, and 21 (columns).

3.3. *Trisomy detection*

A total of 300 negative trisomy NIPS samples were randomly selected to serve as a reference set for deriving estimates of μ_c and SD_c for $c = 13,18,21$. Corresponding Z-scores were generated for 176 negative trisomy samples and the 141 positive trisomy samples using chromosome proportions derived from corrected coverage profiles under each method (GAM vs wkNN). Boxplots of the Z-scores for the three chromosomes by trisomy status and the method of coverage profile used are depicted in Figure 4. Trisomy Z-scores among negative samples were highly correlated across method (Pearson ρ>0.69 for each chromosome), while signed-rank testing for increased positive case Z-scores were significant when assessed separately for all three chromosomes (all p-values < 0.0005). Results for trisomy 13 demonstrated the most substantial improvement, with a median Z-score increase of 7.08 among positive cases, and only four positive trisomy case Z-scores were higher for the GAM corrected coverage profiles relative to the wkNN approach. None of these four cases would have resulted in different NIPS results (all Z>6.0).

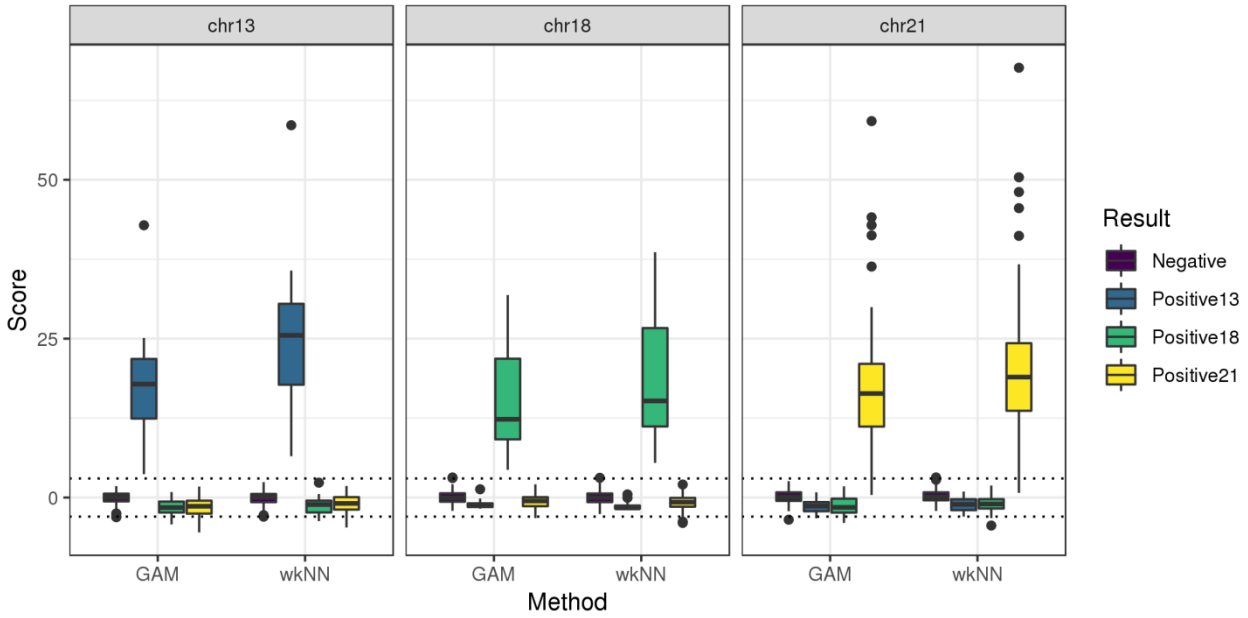

Fig. 4. Boxplots of Z-scores used for trisomy screening, separated by method (GAM vs wkNN), chromosome, and trisomy status. Typical screening threshold values of (-3,3) are depicted by dotted horizontal lines.

4. Discussion

In this paper, we proposed a computationally efficient approach to within-sample coverage profile bias correction for shallow-depth ccfDNA WGS. The noted inter- and intra-sample heterogeneity of these coverage profiles make the identification of even large structural alterations challenging in applications where the mutation frequency is anticipated to be low. Adopting similar principles to those proposed for large-scale micro-duplication/deletion detection, we defined a straight-forward kNN-type implementation toward leveraging empirical and annotation-based measures of genomic bin dissimilarity data under a multi-distance learning framework. In contrast to standard GC-content correction procedures implemented via GAM methods, this approach allows for the

control of known and latent sources of coverage variability, using genomic annotation as a manner of regularizing empirical measures of dissimilarity in retrospective data.

Relative to the regression-based coverage profile correction method based solely on genomic annotation, the wkNN approach demonstrated substantial improvement in both coverage profile variability reduction and improved detection of positive fetal trisomies. Although overall screening performance was already nearly perfect, the increased discrimination distance between positive and negative cases indicates improved trisomy sensitivity and specificity at lower fetal fractions. Thus, adopting these types of coverage correction procedures should render the NIPS assay more robust to random fluctuations in sample fetal fractions and improve reference sample-based NIPS analyses.

As noted above, our methods are conceptually similar to those proposed in WISECONDOR, albeit differing in large part by the approach to bin size and the perspective by which inference about how trisomy detection in NIPS is conducted. Our focus here was to adopt these methods in the context of coverage profile bias correction for downstream analyses using established NIPS inference methods. Additional work is necessary to discern which approach ultimately performs better, as our strategy can be coupled with a variety of other methods that are based on normal sample reference sets[9]. We are additionally exploring how valuable our methods are in improving regional fetal signal for fetal fraction prediction training[14], which necessitates these smaller bin sizes. Finally, kd-tree querying of nearest neighbors is generally most effective under relatively small feature dimensionality. Although our methods were feasibly implemented even under training sample sizes >700, more advanced parameter tuning and repeated cross-validation could be computationally burdensome. Column-wise data dimensionality reduction (e.g., PCA) could significantly alleviate this by identifying leading "eigen-samples" with large coverage variability.

A number of limitations and potential extensions warrant mention. The data used to derive our models were from healthy female donor plasma, which negates the ability to provide coverage profile correction for chromosome Y. Male donor data would be useful for improving chromosome Y correction as well as sex aneuploidy detection. Alternative machine learning strategies may also provide more accurate coverage correction performance if they were trained for each individual bin, although this would require fitting and validating B separate models. Sophisticated data dimensionality reduction methods, such as autoencoders, could also prove useful and warrant investigation. The grid search over values of α was fairly crude, and additional research improving how to tune α could lead to improved coverage profile correction. Selected genomic bin size may heavily influence overall performance, particularly if smaller bins (e.g., 5-10 kb) are used instead of 50 kb. Finally, it is not clear how generalizable trained models are across external labs, where differences in sequencing conditions may yield different inter-bin correlation patterns. Further investigation into genomic annotations shared by neighboring bins may elucidate characteristics that contribute to the observed coverage biases of ccfDNA WGS.

5. Acknowledgments

Funding for this project was supported by the Mayo Clinic Center for Individualized Medicine.

References

1 El Messaoudi, S., Rolet, F., Mouliere, F. & Thierry, A. R. Circulating cell free DNA: preanalytical considerations. *Clinica Chimica Acta* **424**, 222-230 (2013).

2 Jung, K., Fleischhacker, M. & Rabien, A. Cell-free DNA in the blood as a solid tumor biomarker—a critical appraisal of the literature. *Clinica chimica acta* **411**, 1611-1624 (2010).

3 Norton, M. E. *et al.* Cell-free DNA analysis for noninvasive examination of trisomy. *New England Journal of Medicine* **372**, 1589-1597 (2015).

4 Jiang, F. *et al.* Noninvasive Fetal Trisomy (NIFTY) test: an advanced noninvasive prenatal diagnosis methodology for fetal autosomal and sex chromosomal aneuploidies. *BMC Med Genomics* **5**, 57, doi:10.1186/1755-8794-5-57 (2012).

5 Zhou, B. *et al.* Whole-genome sequencing analysis of CNV using low-coverage and paired-end strategies is efficient and outperforms array-based CNV analysis. *Journal of medical genetics* **55**, 735-743 (2018).

6 Snyder, M. W., Kircher, M., Hill, A. J., Daza, R. M. & Shendure, J. Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin. *Cell* **164**, 57-68 (2016).

7 Kim, S. K. *et al.* Determination of fetal DNA fraction from the plasma of pregnant women using sequence read counts. *Prenat Diagn* **35**, 810-815, doi:10.1002/pd.4615 (2015).

8 Lau, T. K. *et al.* Noninvasive prenatal diagnosis of common fetal chromosomal aneuploidies by maternal plasma DNA sequencing. *J Matern Fetal Neonatal Med* **25**, 1370-1374, doi:10.3109/14767058.2011.635730 (2012).

9 Johansson, L. F. *et al.* Novel algorithms for improved sensitivity in non-invasive prenatal testing. *Scientific Reports* **7**, 1838 (2017).

10 Straver, R. *et al.* WISECONDOR: detection of fetal aberrations from shallow sequencing maternal plasma based on a within-sample comparison scheme. *Nucleic acids research* **42**, e31-e31 (2013).

11 Lee, W.-J., Duin, R. P., Ibba, A. & Loog, M. in *2010 2nd International Workshop on Cognitive Information Processing.* 304-309 (IEEE).

12 Arya, S., Mount, D. M., Netanyahu, N. S., Silverman, R. & Wu, A. Y. An optimal algorithm for approximate nearest neighbor searching fixed dimensions. *Journal of the ACM (JACM)* **45**, 891-923 (1998).

13 Benjamini, Y. & Speed, T. P. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic acids research* **40**, e72-e72 (2012).

14 Raman, L. *et al.* PREFACE: In silico pipeline for accurate cell-free fetal DNA fraction prediction. *Prenat Diagn* **39**, 925-933, doi:10.1002/pd.5508 (2019).