A hierarchical genotyping framework using DNA melting temperatures applied to adenovirus species typing

Ben Galvin^{1¶}, Jay Jones^{1¶*}, Michaela Powell¹, Katherine Olin¹, Matthew Jones¹, Tom Robbins¹

¹ present employees of bioMérieux Inc., Salt Lake City, Utah, US

* Corresponding author

E-mail: jay.jones@biomerieux.com (JJ)

[¶]These authors contributed equally to this work.

bioMérieux markets the BioFire® FilmArray® System, the BioFire Respiratory 2 Panel, the BioFire Respiratory 2.1 Panel, and BioFire Syndromic Trends.

Abstract

Known genetic variation for a pathogen, in conjunction with post-PCR melting curve analysis, can be leveraged to provide increased taxonomic detail for pathogen identification in commercial molecular diagnostic tests. The increased taxonomic detail may be used by clinicians and public health decision makers to observe circulation patterns, monitor for outbreaks, and inform local testing practices. We propose a method for expanding the taxonomic resolution of PCR diagnostic systems by incorporating *a priori* knowledge of the assay design and publicly available sequence information into a genotyping classification model. For multiplexed PCR systems, this framework is generalized to incorporate information from multiple assays that react with different gene targets of the same pathogen to increase classification accuracy. To illustrate the method, a hierarchical classification model is developed for the BioFire® Respiratory 2.1 Panel and the BioFire® Respiratory 2 Panel (collectively the BioFire Respiratory Panels – highly multiplexed PCR diagnostic tests) to predict the species of human adenovirus (HAdV) from Adenovirus Detected test results. Performance of the classification model was characterized via a 10-fold cross-validation on a labeled dataset and exhibited $95\% \pm 4\%$ accuracy. The model was then applied to the BioFire® Syndromic Trends dataset, which contains deidentified patient test data from BioFire Respiratory Panels collected at over 100 sites across the globe since 2015. Adenovirus Detected test results in BioFire Syndromic Trends were classified to produce predicted prevalence of each HAdV species within the United States from 2018 through 2021. These results show a marked change in both the predicted prevalence for HAdV and the species makeup with the onset of the COVID-19 pandemic. In particular, HAdV-B decreased from a pre-pandemic predicted prevalence of up to 40% to less than 5% in 2021, while HAdV-A and HAdV-F species both increased in predicted prevalence.

Author Summary

In the diagnosis of infectious disease there is often a trade-off between time to result, cost, and the amount of information gained about the pathogen. We develop a method that can provide increased genotypic information about a pathogen in a test result, with no additional time or cost. Our method is applied to the BioFire Respiratory® 2.1 Panel and the BioFire Respiratory® 2 Panel to speciate positive test results for adenovirus. The method shows robust performance in correctly identifying the adenovirus species from the test results. We estimate the prevalence of each adenovirus species in the United States by applying this method to adenovirus positive test results in the BioFire Syndromic Trends network. This information can be used in near real-time to augment the Centers for Disease Control and Prevention efforts to monitor adenovirus species circulation. Our mathematical framework can be generalized to other genotyping applications.

Introduction

Characterizing genetic variation in pathogen populations and expanding understanding of how they evolve between hosts can have important implications to public health [1]. Providing increased taxonomic information about a pathogen can inform medical decision making. For example, it has been shown that the specific strain of Enterovirus (D-68) presents an increased risk of severe respiratory infections for some at-risk populations (*e.g.*, pediatric and asthmatic patients) compared to other similar Enterovirus strains [2,3]. When this increased taxonomic resolution of a pathogen is available and provided in real-time, clinicians and health care professionals can make decisions for local and regional institutions to mitigate the effects of an outbreak [4].

At the present time, next generation sequencing (NGS) is the primary method of genetic characterization and is increasingly being integrated in the clinical microbiology laboratory, but there still exist many challenges in its routine use, namely cost and time to results [5]. Alternatively, lab developed tests (LDTs) to identify specific strains are common practice in diagnostics for the clinical microbiology laboratory. However, these tests cannot be distributed directly from lab to lab, and labs are not required to

standardize assays used for the identification and reporting of specific strains. Commercially available diagnostic tests for common pathogens are widely available due to their low cost, time to result, and/or clinical accuracy; however, most do not have the ability to provide speciation of the identified pathogen.

One commercially available platform that is in wide use is the BioFire® FilmArray® system, a multiplex, real-time PCR diagnostic system from bioMérieux. The BioFire system addresses many of the issues with NGS, LDTs, and many of the low-plex front-line testing options, in that the assays are standardized across labs using the BioFire system. The BioFire system is capable of providing a sample-to-answer result in about an hour, is in routine use in many labs, and is relatively low cost. Here, we propose a mathematical framework that takes advantage of the design of the molecular assays in the BioFire system and the known genetic variability within assay target regions to provide additional taxonomic resolution for a pathogen.

The BioFire system combines multiple assays in a single diagnostic test to report the presence or absence of multiple pathogens in a single patient sample. For each pathogen there is at least one assay in the test that targets a specific region of its genome. However, there are some pathogens where multiple assays are necessary to capture strain diversity. The BioFire system uses PCR to amplify target regions of a pathogen genome for each assay and an end-point melting curve analysis (MCA) of the amplicon for the detection and identification of the target [6]. As part of MCA, the BioFire system calculates the temperature of the maximum rate of disassociation of the double-stranded DNA. This metric is referred to as the melting temperature or T_m [7,8]. A theoretical T_m value for each expected amplicon sequence can be computed with *in silico* analysis [9,10]. Here we propose a mathematical framework to provide predicted speciation based on genotypic differences using the BioFire test result and the observed T_m values from well characterized isolates, their theoretical T_m values, and the associated taxonomic lineage as training inputs.

As a practical application of this framework, we predict human mastadenovirus (HAdV) species from positive test results on the BioFire® Respiratory 2.1 Panel and BioFire® Respiratory 2 Panel, referred to

herein as BioFire Respiratory Pathogen Panels or RPP. Adenoviruses are DNA viruses from the family Adenoviridae, typically associated with respiratory and gastrointestinal infections [11,12]. In some cases, such as in young children or the immunocompromised, these infections can be serious [13,14]. There are seven known species of HAdV (A – G) [15,16]. Of these species HAdV A-F are most often associated with respiratory infections, with C resulting in more mild outcomes compared to B and E [17]. HAdV-G has only been observed to cause gastrointestinal infections. Outbreaks of HAdV are prone to occur in crowded community settings such as dormitories and hospitals [18]. While vaccines are effective at preventing outbreaks of infection cause by HAdV, they are currently only administered in the select settings and not widely available to the public [19]. HAdV typing in the United States is primarily motivated by the need to determine patterns of circulation, to detect outbreaks, and to inform vaccine and test design [20].

The RPP are designed to detect the six HAdV species known to cause respiratory infections in humans [21,22]. To ensure inclusivity, these products contain five assays that provide complementary coverage of this range of HAdV species. The RPP do not, however, distinguish between the HAdV species in the diagnostic result, instead returning a generalized Adenovirus Detected or Not Detected result.

In order to monitor circulation and detect outbreaks of HAdV, the model developed here is applied to the BioFire Syndromic Trends (Trends) system [23]. Trends is a network that collects and stores aggregated, deidentified patient test data from a subset of BioFire systems. More than 100 sites across the globe have contributed BioFire test data to Trends since 2015, with the aggregated dataset containing over 2 million RPP test results. By applying the HAdV speciation method to this dataset, we estimate changes in circulation patterns of HAdV species from 2018 through 2021.

In the following sections, we define a formal mathematical framework for predicting genotypic variability based on expected T_m values and tune a model to classify HAdV species using a labeled dataset of well characterized isolates. The performance of the classification model is detailed, and the model is then applied to RPP tests with Adenovirus Detected results in the Trends dataset to estimate HAdV species

prevalence. The use of this method to augment HAdV typing efforts are discussed. Finally, we discuss how the mathematical framework presented here could be generalized to other genotyping problems.

Methods

In this section, we develop a general mathematical framework for predicting genotypic differences and apply it to predict HAdV species from an unknown sample. To develop the framework, we first determine the HAdV species reactivity of each RPP HAdV assay by finding complementary alignment of the forward and reverse primer sequences for HAdV entries in the Nucleotide collection nr/nt sequence database [24]. The T_m values are predicted *in silico* for each expected resulting amplicon sequence. These estimated T_m values are used to form a probability density function for the T_m value of each species. We then employ Bayes' Theorem to estimate the posterior probabilities of each HAdV species given a T_m value. The results for all assays, with their associated HAdV species posterior probabilities, are the training input to a data-driven, hierarchical classification model to predict the species of an unknown sample. The classification model was optimized and the performance assessed and validated using a 10-fold cross validation on a labeled data set. The validated model was applied to Adenovirus Detected results from RPP test data in the Trends system to predict HAdV species prevalence in the United States from 2018 through 2021.

Estimating assay reactivity

For each RPP HAdV assay, forward and reverse primer sets were used as query inputs for a BLASTn search against the nr/nt database [24]. The results from these queries were filtered to HAdV hits. We identify predicted amplicons in HAdV genomes, where an amplicon is bounded by primers in the correct orientation and less than 1000 bp in length. The amplicons are linked to their respective taxonomic lineage, including the HAdV species. The set of HAdV species that have at least one predicted amplicon for an assay are considered to be reactive for a RPP HAdV assay.

The primer sequences for the RPP HAdV assays are not disclosed and the RPP HAdV assay names have been obfuscated for the commercial interest of bioMerieux, Inc.

Predicting distributions of Tm values

For a single HAdV assay on the RPP, let S_i represent one of the HAdV species (A – F). For each species S_i , there is a set of associated potential amplicon sequences denoted $R_{i(j)}$ for $j = 1, 2, 3, \dots, n_i$, with n_i being the number of distinct amplicon sequences that are associated with species S_i and amplified by the HAdV assay. Note, if the species is not expected to react with an assay then $n_i = 0$.

For each amplicon sequence $R_{i(j)}$, the predicted Tm value, based on the base-pair sequence of the amplicon, is denoted $\mu_{Tm, R_{i(j)}}$. For simplicity, we use the model proposed by Howley *et. al.* for all Tm value predictions [9]. To account for system variability due to slight changes in chemistry, instrument conditions, and sample makeup, the Tm value for each $R_{i(j)}$ is represented by a normally distributed random variable $T_{R_{i(j)}}$,

$$T_{R_{i(j)}} \sim N(\mu_{Tm, R_{i(j)}}, \sigma_{sys}^2). \#(1)$$

In Equation 1, the variance of $T_{R_{i(j)}}$, σ_{sys}^2 represents the measurement error of the PCR system, and is set to 0.5° C standard deviations for all sequences in the application based on empirical evidence [6]. The distribution of expected Tm values for all expected reactive sequences of species S_i can be represented as a weighted sum of all the individual sequence distributions $T_{R_{i(j)}}$,

$T_{S_i} \sim \sum_{j=1}^{n_i} w_j * T_{R_{i(j)}}. \#(2)$ Where $0 \leq w_j \leq 1$ for all w_j and $\sum_j w_j = 1$. For this application the weights w_j are naively set to $1/n_i$ for all j . For a given assay and species i , T_{S_i} is the random variable that describes the possible Tm values based on the reactivity of that assay with the individual sequences included from the species. The associated probability density function for T_{S_i} is $f_{T_{S_i}}$.

Posterior probabilities of a genotype group

To calculate the posterior probability of each species given a Tm value, x , we employ Bayes' Theorem for each S_i

$$P(x \in T_{S_i} | x) = \frac{P(x | x \in T_{S_i}) * P(S_i)}{P(x)}, \#(3)$$

where $P(x | x \in T_{S_i})$ is the likelihood of the Tm value, x , given it is from species S_i . For a Tm value, x , the conditional probability that the Tm value is associated with species S_i can be written,

$$P(x | x \in T_{S_i}) = \int_{x-\varepsilon}^{x+\varepsilon} f_{T_{S_i}}(x) dx, \#(4)$$

where ε is defined as the minimum resolution of the Tm detection algorithm, set to 0.01° C for the BioFire system. Note, to account for the possibility of a reactive sequence not included in R_i we incorporate a uniform distribution $S_{\text{unknown}} \sim U(Tm_{\text{low}}, Tm_{\text{high}})$, where Tm_{low} and Tm_{high} are the lower and upper bounds of the melting curve analysis range for all RPP HAdV assays, set to 75° C and 95° C, respectively. The final set of likelihood probabilities, S_{all} , is the union of the conditional probabilities for an observed Tm and the uniform distribution S_{unknown} .

The prior, $P(S_i)$, is a discrete distribution indicating the relative prevalence of HAdV species and for this application is set such that all species were equally likely. $P(x)$ is the marginal probability calculated using the Law of Total Probability.

Hierarchical classification model

The previous methods are applied to the five HAdV assays in the RPP independently to compute the posterior probabilities for the entire melting curve analysis range. These probabilities are combined into a data-driven hierarchical classification model to increase accuracy in predicting the species for an Adenovirus Detected result. Labeled data were generated by running isolates of known HAdV species using the RPP. A total of 347 RPP tests with an Adenovirus Detected result were sourced from three

internal BioFire studies used to characterize RPP performance: the limit of detection study, the inclusivity study, and the cross-reactivity study. The results of these studies contribute to the instructions for use of the RPP [21]. The number of replicates, serotype, and concentration of the samples for each study are summarized in Table 2. This set of varied serotypes for each species, with multiple replicates of each isolate, provides a diverse training dataset for optimizing a data-driven classification model.

Table 2. Training dataset summary.

HAdV Species	Limit of Detection Study		Inclusivity Study		Cross-Reactivity Study		Species Run Count
	Isolate serotype	Run Count	Isolate serotype	Run Count	Isolate serotype	Run Count	
A	A18	39	A12	3	A31	3	48
			A31	3			
B	B7a	40	B3	3	B3	3	83
			B7d/d2	3			
			B7h	3	7a	3	
			B11	4			
			B14	3	B14	3	
			B16	3			
			B21	3	B21	3	
			B34	3			
			B35	3			
			B50	3			
C	C2	39	C1	3	C2	3	57
			C5	3	C5	3	
			C6	3	C6	3	
D	D37	40	D8	4	D8	3	60
					D20	3	
			D20	3	D26	3	
					D37	4	
E	E4a	38	E4	3	E4	3	44
F	F41	40	F40	6	F40	3	55
			F41	3	F41	3	
Total Run Count							347

A description of the labeled data included in the logistic regression classification model optimization.

These studies contain different isolates of each HAdV species tested at varied concentrations.

For each sample in the labeled dataset, the posterior probabilities for each assay were computed from the

T_m value observed from each assay. Note, the T_m values for all data in this application have been

normalized against the internal controls of the RPP [25]. The posterior probabilities are combined to create a feature matrix X , and the known species from each isolate form a response vector y . X and y are used to optimize a multiclass elastic-net regularized logistic regression model to predict HAdV species. This model was optimized using an L1 ratio of 0.6 with balanced class weights, and all other parameters set to the default values in the scikit-learn v1.0.1 package with Python v3.8 [26]. Performance of the optimized classification model is assessed and validated using 10-fold cross validation and reported as the mean and standard deviation of the accuracy across all folds. The mean of the precision, recall, and F1-scoring across the ten cross-validation folds are also used to assess overall and class-wise performance [27].

In addition to the optimized logistic regression model, we developed another HAdV species classification model as a sensitivity analysis to ensure that the data-driven model provided increased performance over a naive approach. This model is referred to as the assay design informed classification model and sets predetermined weights for each of the posterior probabilities based on *a priori* knowledge of assay reactivity. The weights for the posterior probabilities are set equally across all species that an assay has expected reactivity to, as shown in Table 1. For each sample, the prediction is determined by computing the Argmax of the weighted sum of posterior probabilities [26]. Because this model was not optimized using the labeled dataset, no cross validation was performed.

Application to Syndromic Trends

The optimized classification model was elected to be used for application to Trends due to its increased performance over the assay design informed classification model developed for the sensitivity analysis. HAdV species prevalence was predicted from Adenovirus Detected results from RPP test data in the Trends system in the United States from 2018 through 2021.

Results

Estimated assay reactivity

The BLAST search identified 2,249 HAdV amplicons that met our inclusion criteria. Each predicted amplicon's species was determined from the BLAST search. An assay's predicted reactivity to a given HAdV species was then based on the amplicons that could be produced by the assay's primer sets. Table 1 contains the predicted reactivity of HAdV species for the five HAdV assays on the RPP.

Table 1. RPP HAdV assay predicted reactivity summary.

RPP HAdV Assay	HAdV Reactive Species
RPP HAdV-1	B, D, E
RPP HAdV-2	A, F
RPP HAdV-3	C
RPP HAdV-4	A, E, F
RPP HAdV-5	B, C, D, E

The predicted HAdV species reactivity of the five RPP HAdV assays.

Calculated distributions of T_m values and posterior probabilities

The predicted T_m values for the expected reactive sequences, $R_{i(j)}$, for the RPP HAdV-5 assay are shown in Fig 1A. The likelihood distribution $P(x|x \in T_{S_i})$ for each species i are shown in Fig 1B. The posterior probability distributions are shown in Fig 1C. For a T_m value, each posterior probability describes the probability that the sample is of a particular species. Note that for each T_m value in the melting curve analysis range, the posterior probabilities for all species, S_{all} , sum to one. Analogous figures for the other four HAdV assays (RPP HAdV-1, RPP HAdV-2, RPP HAdV-3, and RPP HAdV-4) can be found in the supporting information S1 – S4 Figs.

Fig 1. Predicted T_m values of reactive sequences, likelihood distributions, and resulting posterior probability distributions for reactive species with the RPP HAdV-5 assay. The assay RPP HAdV-5 had predicted reactivity with HAdV serotypes from the HAdV-B, HAdV-C, HAdV-D, and HAdV-E species. RPP HAdV-5 shows the most predicted reactivity in the range of 85-93° C. A) Predicted T_m values for the expected reactive species of the RPP HAdV-5 assay. B) Predicted Likelihood distributions

$P(x|x \in T_{S_i})$ of Tm values for each expected reactive species of the RPP HAdV-5 assay. C) Posterior probabilities across the temperature range. Species HAdV-B, HAdV-C, HAdV-D, and HAdV-E have peak posterior probabilities at 85.8, 87.9, 92.1, and 90.5° C respectively.

Hierarchical classification model performance

Table 3 contains a confusion matrix of the predicted HAdV species against known HAdV species using the optimized classification model. Values along the diagonal indicate that the prediction agreed with the known genotype for the sample, and off-diagonal elements indicate misclassifications. Table 4 summarizes the performance results of the model on the labeled dataset with the mean of the precision, recall, and F1-score across the ten cross validation folds, along with support by HAdV species. The mean class-wise precision for HAdV-A of 0.88 indicates that when HAdV-A is predicted, it is correct 88% of the time. The mean class-wise recall for HAdV-A of 0.96 indicates that when the true classification for a sample is HAdV-A, the species is correctly predicted 96% of the time. The mean class-wise F1-score for HAdV-A of 0.92 indicates balanced accuracy as the harmonic mean of precision and recall. Support indicates the number of samples of each species included in the analysis (per Table 2). The macro average for each metric is the average across all classes with equal class weights and the weighted average is the average across all classes weighted by the number of samples in each class [27].

Table 3. Confusion matrix for the optimized classification model of HAdV species.

HAdV Species		Predicted Label					
		A	B	C	D	E	F
Known Label	A	46	0	0	0	0	2
	B	0	81	0	0	2	0
	C	0	0	57	0	0	0
	D	0	0	0	53	4	3
	E	0	0	0	0	43	1
	F	6	1	0	0	0	48

Confusion matrix of the predicted HAdV species by the optimized classification model against the true HAdV species.

Table 4. Optimized HAdV species classification model performance.

HAdV Species	Precision	Recall	F1-score	Support
A	0.88	0.96	0.92	48
B	0.99	0.98	0.98	83
C	1.00	1.00	1.00	57
D	1.00	0.88	0.94	60
E	0.88	0.98	0.92	44
F	0.89	0.87	0.88	55
Macro average	0.94	0.94	0.94	
Weighted average	0.95	0.95	0.95	
Accuracy ($\pm$ s.d.)	0.95 ($\pm$ 0.04)			

Detailed performance characteristics for the optimized classification model.

The logistic regression model finds coefficients of each assay's posterior probabilities that optimize model classification performance of the training dataset in Table 2. In contrast, the model used in the sensitivity analysis sets the coefficients from the heuristics of the intended assay design and demonstrated 75% overall class accuracy for the same training dataset. The detailed performance results and confusion matrix for this method are shown in S1 and S2 Tables, respectively, in the supporting information.

Application to Syndromic Trends

Fig 2A shows the optimized classification model applied to the RPP runs in Trends with an Adenovirus Detected result. The rate of Adenovirus Detected results are displayed as a rate of total RPP runs in the database. Subtypes are displayed in stacked area under the total detections indicating the percentage of RPP runs with an Adenovirus Detected results attributed to each species. Fig 2B shows the normalized percentage of each HAdV species from Fig 2A.

Fig 2. Estimated HAdV Species Prevalence from Trends data. HAdV species predicted from the optimized classification model applied to RPP Adenovirus Detected results in the Trends database from 2018 through 2021. Rates are calculated on a weekly basis and smoothed with a six-week centered rolling window. A) Estimated prevalence of each HAdV species shown as a proportion of total runs RPP. B) Estimated prevalence of each HAdV species shown as a proportion of total Adenovirus Detected results.

Discussion

Hierarchical classification model performance

The mathematical framework for predicting genotypic differences applied to HAdV species typing showed robust cross-validation performance in assigning correct species labels to samples tested with the RPP, with an overall class accuracy of 95% and a standard deviation of 4% across the 10-fold cross validation. However, the optimized classification model performance was not equal across all HAdV species with the lowest species class accuracy equal to 87% (see Table 4). Higher class precision and recall were observed for HAdV species B and C compared to A, D, E, and F. This may be attributed to the assay coverage of these species, as A and F, and D and E often shared high posterior probabilities in similar temperature ranges. The model also demonstrates performance in line with the Xu, McDonough, and Erdman HAdV species typing assay that show an overall class accuracy of 96% with the lowest species class accuracy of 83% [28].

Compared to the optimized classification model, the assay design informed classification model had two primary deficiencies in performance. The most discrepant results were observed with HAdV-A and HAdV-F samples, often mistaking one for the other. It also misclassified many HAdV-E samples as HAdV-D. While the optimized classification model exhibits the worst performance with HAdV-A and HAdV-F, it avoids most of the specific biases of the assay design informed model by using the labeled data to inform the posterior probability weights. Although the training dataset used for these species

classification models was comprised of a diverse set of serotypes for each HAdV species, we would expect lower accuracy than what is demonstrated here when applied to unseen serotypes.

Generalization of the genotyping framework

The mathematical framework presented here was well-suited for the RPP HAdV assays. These assays were designed specifically to capture the diversity of the HAdV genome across the different species. However, it could be generalized to classifying genotypic diversity using other assays, provided there is variability in T_m values inside the relevant temperature range due to sequence differences of the amplicon sequences associated with genotypic diversity. The variation in T_m values associated with genotypic diversity must be greater than the temperature resolution of the instrumentation and the run-to-run variability for a reactive sequence. The error of the predicted T_m values for a given set of reaction conditions and amplicon sequence must also be less than the differences in T_m values between the genotypic groups. If assays were designed to maximize the variation of T_m values from different genotypic groups, performance of this method could be further improved.

Adenovirus surveillance

The relative prevalence of HAdV species in specific subpopulations has been studied extensively [11,12,18,19,29–32]. Surveillance of HAdV species in the greater population is currently addressed by the CDC’s National Adenovirus Type Reporting System (NATRS) [33]. This program collects and summarizes typing results from labs and hospitals and publishes a static prevalence estimation for several time periods. NATRS has published prevalence data from 2003-2016 collected from 11 sites and from 2014-2017 collected from at least 33 sites. We applied the HAdV species classification model to the RPP patient test data collected by the Trends system from 2018 through 2021. Because HAdV species prevalence varies from year to year, and the data collection timeframe does not overlap, no direct comparison of prevalence estimated here to that of NATRS is made [14,32]. However, the relative species proportions between NATRS (2014-2017) and the model predictions of the Trends data (2018-2020) are similar with HAdV-B and HAdV-C combined consisting of greater than 80% of HAdV relative

prevalence in both estimates. The application of the HAdV species classification model to Trends data supports and augments the surveillance efforts of NATRS. The Trends database contains a large pool of Adenovirus Detected results (>25,000) reported from over 100 testing sites across the United States. The application of the model to an active reporting system, such as Trends, confers several benefits over traditional passive reporting. First, this application requires minimal additional effort beyond running the samples on the RPP. Second, because all results are produced by the same testing platform, they share the same sample preparation protocol, reagents, and assays, which improves data consistency. Finally, because the data are automatically processed and categorized by the system and available in near real time, it can be used to detect species-specific outbreaks and monitor seasonality and changes in circulation patterns. The HAdV species classification model can be continuously applied to Trends data moving forward, and as more sites are added to the platform, more detailed patterns of circulation can be discerned.

Trends data was utilized in this study to investigate United States HAdV species prevalence. While Trends data has been used to suggest pathogen circulation dynamics [4,34,35], it should be noted that the testing practices and demographics of institutions contributing to Trends are unknown. Biases in sites contributing to the network may exist that could give Trends a skewed representation of population pathogen prevalence. These limitations need to be kept in mind when making inferences about HAdV species prevalence based on the classification model results applied to Trends data.

Investigating the predicted prevalence data produced by the model shown in Fig 2A reveals several interesting results. Immediately apparent is the large drop in positivity in the spring of 2020 due to the COVID-19 pandemic. This pattern was not limited to HAdV, however, and was observed across all positivity in the Trends network [35]. As social practices return to normal over the course of 2021, rates of Adenovirus Detected results increase in Trends to levels seen prior to the onset of the pandemic. It is interesting to note in Fig 2B that the estimated proportion of HAdV-B does not recover to its previous state of about 30% of Adenovirus Detected results after the onset of the pandemic; in 2021 it has

consisted of less than 5% of total Adenovirus Detected results. In the same timeframe HAdV-C increased in relative prevalence to make up approximately 70% of all Trends Adenovirus Detected results. Additionally, two species (HAdV-A and HAdV-F) appear to be increasing in prevalence in the absence of HAdV-B. HAdV-A has increased from a low in 2020 of 4.5% relative prevalence to a high in 2021 of 11.6%, and HAdV-F increased from a low of 1.8% in 2020 to a high in 2021 of 17.9%. As stated above, this data may not reflect the actual circulating patterns of HAdV in the general population. Although this is an interesting observation on potential changes in the prevalence of several HAdV species, further study and monitoring is needed to understand the dynamics of circulating virus populations after a major disruption in social practices and competition between similar species.

Conclusion

Using *a priori* information and *in silico* analysis, we developed a mathematical framework for providing additional taxonomic resolution for pathogen targets of the highly multiplexed BioFire system. This framework is flexible enough to provide increased genotypic classification for any PCR-based test with end point melting analysis, so long as the genotypic variability being classified results in predicted T_m value variability. We applied this framework to the problem of predicting HAdV species from RPP Adenovirus Detected results. The HAdV species classification accuracy is enhanced by combining the results from multiple assays in a data-driven regularized logistic regression model. The model exhibited high accuracy on the labeled dataset. As a result, the optimized classification model was applied to unlabeled data from the Trends network. These results show a marked change in both the predicted prevalence for HAdV and the species makeup with the onset of the COVID-19 pandemic. In particular, HAdV-B decrease from a pre-pandemic predicted prevalence of approximately 30% to less than 5% in 2021, which led to relative increases in predicted prevalence for HAdV-A and HAdV-F during that same timeframe. Further development to apply this HAdV species classification model in real time to the Trend system could augment CDC efforts to surveil the circulation of HAdV species and provide outbreak detection.

Acknowledgements

The authors would like to thank the bioMerieux research and development teams of the BioFire Respiratory 2 Panel and the BioFire Respiratory 2.1 Panel, who generated the training datasets and made it available for this work.

References

1. Alizon S, Luciani F, Regoes RR. Epidemiological and clinical consequences of within-host evolution. *Trends Microbiol.* 2011;19(1):24-32. doi:10.1016/J.TIM.2010.09.005
2. Holm-Hansen CC, Midgley SE, Fischer TK. Global emergence of enterovirus D68: a systematic review. *Lancet Infect Dis.* 2016;16(5):e64-e75. doi:10.1016/S1473-3099(15)00543-5
3. Oermann CM, Schuster JE, Connors GP, Newland JG, Selvarangan R, Jackson MA. Enterovirus D68: A focused review and clinical highlights from the 2014 U.S. outbreak. *Ann Am Thorac Soc.* 2015;12(5):775-781. doi:10.1513/ANNALSATS.201412-592FR/SUPPL_FILE/DISCLOSURES.PDF
4. Meyers L, Dien Bard J, Galvin B, et al. Enterovirus D68 outbreak detection through a syndromic disease epidemiology network. *J Clin Virol.* 2020;124. doi:10.1016/J.JCV.2020.104262
5. Rossen JWA, Friedrich AW, Moran-Gilad J. Practical issues in implementing whole-genome-sequencing in routine diagnostic microbiology. *Clin Microbiol Infect.* 2018;24(4):355-360. doi:10.1016/J.CMI.2017.11.001
6. Poritz MA, Blaschke AJ, Byington CL, et al. FilmArray, an Automated Nested Multiplex PCR System for Multi-Pathogen Detection: Development and Application to Respiratory Tract Infection. *PLoS One.* 2011;6(10):e26047. doi:10.1371/JOURNAL.PONE.0026047
7. Ririe KM, Rasmussen RP, Wittwer CT. Product Differentiation by Analysis of DNA Melting Curves during the Polymerase Chain Reaction. *Anal Biochem.* 1997;245(2):154-160. doi:10.1006/ABIO.1996.9916
8. Wittwer CT, Herrmann MG, Gundry CN, Elenitoba-Johnson KSJ. Real-Time Multiplex PCR Assays. *METHODS.* 2001;25:430-442. doi:10.1006/meth.2001.1265
9. Howley PM, Israel MA, Law MF, Martin MA. A Rapid Method for Detecting and Mapping Homology between. Heterologous DNAs EVALUATION OF POLYOMAVIRUS GENOMES*. *J Biol Chem.* 1979;254(11):4876-4883. doi:10.1016/S0021-9258(17)30093-5
10. SantaLucia J. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc Natl Acad Sci.* 1998;95(4):1460-1465. doi:10.1073/PNAS.95.4.1460
11. Lion T. Adenovirus infections in immunocompetent and immunocompromised patients. *Clin Microbiol Rev.* 2014;27(3):441-462. doi:10.1128/CMR.00116-13
12. Uhnoo I, Wadell G, Svensson L, Johansson ME. Importance of enteric adenoviruses 40 and 41 in acute gastroenteritis in infants and young children. *J Clin Microbiol.* 1984;20(3):365-372. doi:10.1128/JCM.20.3.365-372.1984
13. Cui X, Wen L, Wu Z, et al. Human adenovirus type 7 infection associated with severe and fatal

- acute lower respiratory illness and nosocomial transmission. *J Clin Microbiol.* 2015;53(2):746-749. doi:10.1128/JCM.02517-14
14. Zhao S, Wan C, Ke C, et al. Re-emergent human adenovirus genome type 7d caused an acute respiratory disease outbreak in Southern China after a twenty-one year absence. *Sci Rep.* 2014;4. doi:10.1038/SREP07365
15. Tam CC, O'Brien SJ, Tompkins DS, et al. Changes in causes of acute gastroenteritis in the United Kingdom over 15 years: microbiologic findings from 2 prospective, population-based studies of infectious intestinal disease. *Clin Infect Dis.* 2012;54(9):1275-1286. doi:10.1093/CID/CIS028
16. Ismail AM, Lee JS, Lee JY, et al. Adenoviromics: Mining the Human Adenovirus Species D Genome. *Front Microbiol.* 2018;9(SEP). doi:10.3389/FMICB.2018.02178
17. Lynch JP, Kajon AE. Adenovirus: Epidemiology, Global Spread of Novel Serotypes, and Advances in Treatment and Prevention. *Semin Respir Crit Care Med.* 2016;37(4):586-602. doi:10.1055/S-0036-1584923
18. Kandel R, Srinivasan A, D'Agata EMC, Lu X, Erdman D, Jhung M. Outbreak of Adenovirus Type 4 Infection in a Long-Term Care Facility for the Elderly. *Infect Control Hosp Epidemiol.* 2010;31(7):755-757. doi:10.1086/653612
19. Gray GC, Callahan JD, Hawksworth AW, Fisher CA, Gaydos JC. Respiratory diseases among U.S. Military personnel: Countering emerging threats. *Emerg Infect Dis.* 1999;5(3):379-387. doi:10.3201/eid0503.990308
20. Binder AM, Biggs HM, Haynes AK, et al. Morbidity and Mortality Weekly Report Human Adenovirus Surveillance-United States, 2003-2016. 2017;6(39).
21. BioFire Diagnostics LLC. *BioFire® Respiratory Panel 2.1 (RP2.1) De Novo Instructions for Use.* Vol 1.; 2021.
22. Ghebremedhin B, Ghebremedhin B. HUMAN ADENOVIRUS: VIRAL PATHOGEN WITH INCREASING IMPORTANCE. *Eur J Microbiol Immunol.* 2014;4:26-33. doi:10.1556/EuJMI.4.2014.1.2
23. Meyers L, Ginocchio CC, Faucett AN, et al. Automated Real-Time Collection of Pathogen-Specific Diagnostic Data: Syndromic Infectious Disease Epidemiology. *JMIR public Heal Surveill.* 2018;4(3). doi:10.2196/PUBLICHEALTH.9876
24. NCBI. BLASTN (RRID:SCR_001598). Published online 2022. https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&BLAST_PROGRAMS=megaBlast&PAGE_TYPE=BlastSearch
25. Robbins T, Wilson T, Cox C, Jones J, Galvin B, Kiser A. (2021). System and Method for Identifying Analytes in Assay Using Normalized Tm Values. *International Patent WO2021/113554.* World International Property Organization.
26. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. *J Mach Learn Res.* 2011;12:2825-2830.
27. M H, M.N S. A Review on Evaluation Metrics for Data Classification Evaluations. *Int J Data Min Knowl Manag Process.* 2015;5(2):01-11. doi:10.5121/ijdkp.2015.5201
28. Xu W, McDonough MC, Erdman DD. Species-Specific Identification of Human Adenoviruses by a Multiplex PCR Assay. *J Clin Microbiol.* 2000;38(11):4114. doi:10.1128/jcm.38.11.4114-4120.2000
29. Nakamura H, Fujisawa T, Suga S, et al. Species differences in circulation and inflammatory responses in children with common respiratory adenovirus infections. *J Med Virol.*

- 2018;90(5):873-880. doi:10.1002/jmv.25032
30. Probst V, Datyner EK, Haddadin Z, et al. Human adenovirus species in children with acute respiratory illnesses. *J Clin Virol*. 2021;134. doi:10.1016/j.jcv.2020.104716
31. Yuhong DOU, Yuxia LI, Caifeng MA, et al. Rapid diagnosis of human adenovirus B, C and E in the respiratory tract using multiplex quantitative polymerase chain reaction. *Mol Med Rep*. 2018;18(3):2889-2897. doi:10.3892/mmr.2018.9253
32. Kendall Scott M, Chommanard C, Lu X, et al. Human adenovirus associated with severe respiratory infection, Oregon, USA, 2013-2014. *Emerg Infect Dis*. 2016;22(6):1044-1051. doi:10.3201/EID2206.151898
33. National Adenovirus Type Reporting System Publications | CDC.
34. Park SW, Farrar J, Messacar K, Meyers L, Pons-Salort M, Grenfell BT. Epidemiological dynamics of enterovirus D68 in the US: implications for acute flaccid myelitis. *medRxiv Prepr Serv Heal Sci*. Published online July 2021. doi:10.1101/2020.07.23.20069468
35. Nawrocki J, Olin K, Holdrege MC, et al. The Effects of Social Distancing Policies on Non-SARS-CoV-2 Respiratory Pathogens. *Open forum Infect Dis*. 2021;8(7). doi:10.1093/OFID/OFAB133

Figure 1

A **B**

Figure 2

Figure S1

Figure S2

Figure S3

Figure S4