AMERICAN ACADEMY
OF OPHTHALMOLOGY®

Machine Learning Prediction of Adenovirus D8 Conjunctivitis Complications From Viral Whole-Genome Sequence

Kenji Nakamichi,[1,2] Lakshmi Akileswaran, PhD,[1,2] Thomas Meirick, MD,[1] Michele D. Lee, MD,[1] James Chodosh, MD,[3] Jaya Rajaiya, PhD,[3] David Stroman, PhD,[4] Alejandro Wolf-Yadlin, PhD,[5] Quinn Jackson,[5] W. Bradley Holtz,[5] Aaron Y. Lee, MD,[1,2] Cecilia S. Lee, MD,[1,2] Russell N. Van Gelder, MD, PhD,[1,2,6,7] for the BayNovation Study Group

Objective: To obtain complete DNA sequences of adenoviral (AdV) D8 genome from patients with conjunctivitis and determine the relation of sequence variation to clinical outcomes.

Design: This study is a post hoc analysis of banked conjunctival swab samples from the BAYnovation Study, a previously conducted, randomized controlled clinical trial for AdV conjunctivitis.

Participants: Ninety-six patients with AdV D8-positive conjunctivitis who received placebo treatment in the BAYnovation Study were included in the study.

Methods: DNA from conjunctival swabs was purified and subjected to whole-genome viral DNA sequencing. Adenovirus D8 variants were identified and correlated with clinical outcomes, including 2 machine learning methods.

Main Outcome Measures: Viral DNA sequence and development of subepithelial infiltrates (SEIs) were the main outcome measures.

Results: From initial sequencing of 80 AdV D8-positive samples, full adenoviral genome reconstructions were obtained for 71. A total of 630 single-nucleotide variants were identified, including 156 missense mutations. Sequence clustering revealed 3 previously unappreciated viral clades within the AdV D8 type. The likelihood of SEI development differed significantly between clades, ranging from 83% for Clade 1 to 46% for Clade 3. Genome-wide analysis of viral single-nucleotide polymorphisms failed to identify single-gene determinants of outcome. Two machine learning models were independently trained to predict clinical outcome using polymorphic sequences. Both machine learning models correctly predicted development of SEI outcomes in a newly sequenced validation set of 16 cases ($P = 1.5 \times 10^{-5}$). Prediction was dependent on ensemble groups of polymorphisms across multiple genes.

Conclusions: Adenovirus D8 has ≥ 3 prevalent molecular substrains, which differ in propensity to result in SEIs. Development of SEIs can be accurately predicted from knowledge of full viral sequence. These results suggest that development of SEIs in AdV D8 conjunctivitis is largely attributable to pathologic viral sequence variants within the D8 type and establishes machine learning paradigms as a powerful technique for understanding viral pathogenicity. *Ophthalmology Science 2022;2:100166 © 2022 Published by Elsevier Inc. on behalf of the American Academy of Ophthalmology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).*

Supplemental material available at www.ophthalmologyscience.org.

Conjunctivitis is among the most common infectious conditions worldwide.[1,2] In the United States alone, > 6 million cases of conjunctivitis are diagnosed annually with an attributed associated annual cost for diagnosis, treatment, and lost work of over $800 million.[1,3] Viruses are the most frequent cause of conjunctivitis. Human adenoviruses (AdV) are the most prevalent, accounting for approximately 75% of cases.[2,4–6] The most severe form of conjunctivitis is epidemic keratoconjunctivitis (EKC), a highly infectious form that frequently results in community outbreaks in health care settings, schools, and daycare. Epidemic keratoconjunctivitis may have a prolonged and severe course lasting from weeks to months and may lead to permanently decreased visual acuity.[7–10]

Subepithelial infiltrates (SEIs) are among the most serious sequela of AdV keratoconjunctivitis, with reported frequency ranging from 33% to 80% of patients with acute AdV conjunctivitis.[2,11,12] Histopathologically, SEIs are composed of lymphocytes, histiocytes, and fibroblasts accompanied by disruption of collagen fibers in the Bowman's layer of the cornea.[13] Subepithelial infiltrates are thought to be due to viral replication in corneal cells

https://doi.org/10.1016/j.xops.2022.100166
ISSN 2666-9145/22

and subsequent immunologic host reaction.[14–16] In addition to causing clinical symptoms including persistent photophobia and foreign-body sensation, these opacities may cause scarring of the cornea, optical aberrations, induced astigmatism, and a decrease in visual acuity.[17–20]

Traditionally, AdV were classified based on serum neutralization and hemagglutination assays into 6 species (A-F). More recently, classification has been based on genomics, with 7 species (A-G) and over 100 types identified based largely on variations in the immunogenic hexon, fiber, and penton base viral surface proteins.[21] Certain types of the B, D, and E species are capable of causing conjunctivitis, including types B3, E4, D8, D37, and D64 (previously D19).[10,22] The D species is most commonly associated with EKC, suggesting that virally encoded factors at least partially determine the severity of conjunctivitis.[6,14,23–25] However, the degree to which viral sequence variants determine clinical outcomes in adenoviral conjunctivitis has not been established to date.

Methods

Participants

This study was HIPAA-compliant, approved by institutional board review (US: Goodwin IRB [Cincinnati, OH]; India: Drug Controller General [DCGI] with local site ethics committees; Sri Lanka: Scientific and Ethical Review, Faculty of Medicine, University of Kelaniya, Colombo, Sri Lanka; Brazil: National Ethics Committee in Research [CONEP] plus local site ethic committees), and conducted in accord with the Declaration of Helsinki. The BAYnovation clinical trial was registered at clinicaltrials.gov (NCT01877694). Informed consent was obtained from all study subjects. Subjects were included if they were older than 18 years; had one of the following: recent upper respiratory tract infection, contact with an infected person, or a recent visit to an eye care provider; had at least 2 of 9 clinical signs indicative of conjunctivitis; had onset < 3 days prior to enrollment; and had a positive point-of-service AdV antigen screening test (Adeno Plus, Rapid Pathogen Screening, Inc, Saratosa, FL). Study subjects were recruited from centers in Brazil, Sri Lanka, and India.

Sequencing and Identification of Viral Clades

The samples from the placebo arm of the NVC-422 BAYnovation drug trial with quantitative polymerase chain reaction (PCR) quantitation of $> 1 \times 10^7$ copies/swab had DNA extracted and were sequenced using Illumina MiSeq. The resulting data were annotated, filtered, and aligned using the Scalable Metagenomics Alignment Research Tool-based pipeline.[26] Of the samples that yielded sufficient sequence for full-genome reconstruction, de novo reconstruction was accomplished using SPAdes 3.11.1.[27] Variant calling of the scaffolds was performed using Annovar to construct a sequence of nucleotides for each sample that described the genotype at all the genomic locations where any single-nucleotide variant was reported within the dataset with respect to reference sequence.[28] Maximum likelihood phylogenetic trees were constructed from the scaffolds using Clustal Ω for multiseq alignments and phylip for the phylogenetic analysis to determine monophyletic groups (clades).[29] In order to validate the findings of these clades, pairwise comparison of the single-nucleotide polymorphisms (SNPs) were completed. Principal component analysis (PCA) was performed, and the results from the clustering algorithms were compared in order to validate the clades. Subjects were only included for analysis if they were from the placebo arm and had clade data available.

Clinical Predictors of Subepithelial Infiltrates

Visual acuity was determined using ETDRS vision charts.

Clinical signs that were collected and analyzed include the presence of lymphadenopathy, lid edema, lid erythema, bulbar conjunctival injection, palpebral conjunctival hyperemia, conjunctival discharge and lid crusting, presence of abnormal tear meniscus, corneal fluorescein staining, tear breakup time, and the presence of SEIs. Clinical symptoms included in the analysis were blurry vision, foreign-body sensation, tearing, itching, burning, and photophobia. To determine a clinical signs and clinical symptoms score, variables such as lymphadenopathy, tear breakup time, presence of SEIs, and blurry vision were scored as a binary 0 (not present) or 1 (present); all other variables were scored as 0 (absent), 1 (mild), 2 (moderate), or 3 (severe) based on study-defined criteria.

The primary outcome was to determine the presence or absence of an association between collected baseline factors and distinct monophyletic groups (clades). As a secondary outcome, an association between demographic factors, vision, clinical symptoms, and clinical signs and the development of SEIs was examined. Chi-squared and analysis of variance testing were performed to determine significance. Variables were considered statistically significant if $P < 0.05$.

Machine Learning Methods

Ensemble of Extra Trees. Initial testing of several methods including XGBoost, glm, and random forest classifiers identified a stacked ensemble of extra tree classifiers as well suited for this application. An extra trees model utilizing all sequence variants (coding and noncoding) was used to predict clade of the sample and serves as a meta-classifier gate in the stacked ensemble. A separate first level layer of ensembles of extra tree classifiers was built within each clade and was trained using only missense mutations for SEI outcomes.

Pattern Discovery Engine Analysis. The data were analyzed by Pattern Computer using its Pattern Discovery Engine™ methodology. Briefly, all SNP information was binarized, with 0 representing the reference allele and 1 the alternate. The data were then processed using its discovery platform leveraging the topology of directed acyclic graphs to discover hidden patterns within large-scale datasets without introducing bias into the data. The extracted patterns, including the most important SNPs and SNP interactions with respect to the different models' output, were used to reduce the dimensionality of the genomic space and create mathematically interpretable and testable predictive models for viral country, clade, and severity of eye infection.

Results

The BAYnovation trial was a randomized, masked controlled clinical trial of 500 patients with adenoviral keratoconjunctivitis from 4 countries on 3 continents to assess the efficacy of the non-specific antiviral compound auriclosene (NV-422) in improving clinical outcomes. While the agent did not meet the clinical endpoint, the placebo arm of this trial provides important natural history data.[11] In the course of this study, bilateral conjunctival swabs containing virus were obtained from 500 symptomatic subjects on days 1, 3, 8, 11, and 18, along

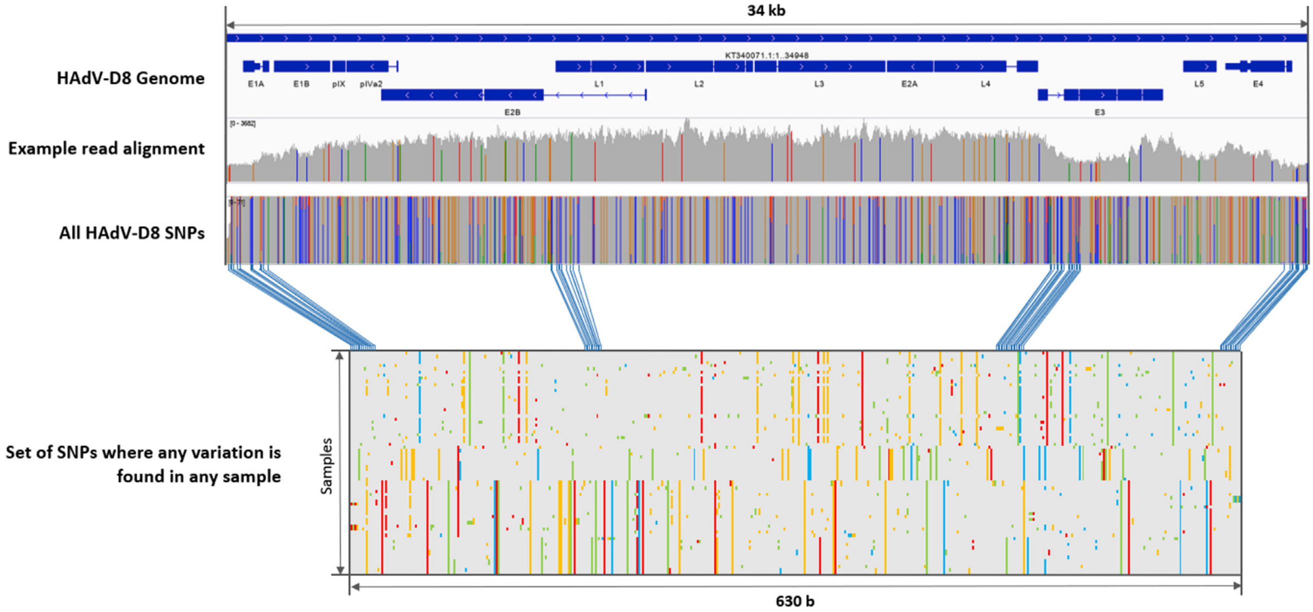


Figure 1. Adenovirus (AdV) 8 polymorphisms among 71 subjects with viral conjunctivitis. The top panel shows genetic map of AdVD8, with the middle panel showing sequencing depth and location of polymorphisms of a representative viral sequence compared with canonical sequence (KT340071.1) and location of all 630 single nucleotide polymorphisms (SNPs) identified in the training dataset of 71 samples in this study. The lower panel shows clustering of the 630 SNPs, identifying 3 potential clades of virus.

with prespecified clinical data. We analyzed the subset of samples typed as AdV D8 by direct PCR from the placebo arm of this trial, using shotgun metagenomic sequencing and viral genome reconstruction. Two hundred sixty-two of the 500 unique patients were PCR-positive for AdV D8, of which 127 were from the placebo arm of the trial. We randomly selected 80 of these with viral loads $> 1 \times 10^7$/ml for sequencing and successfully reconstructed the viral genome from 71 of these samples. In total, 47 of these 71 subjects (66%) developed SEIs during follow-up. Clinical features of subjects with and without SEIs are shown in Table S1.

Sequencing Identifies 3 Subtype Clades of AdV D8 with Differing Disease Severity

Metagenomic sequences from the 71 samples were generated from the Illumina platform and were annotated, filtered, and aligned using a Scalable Metagenomics Alignment Research Tool-based pipeline.[30] Of these, 43 samples were from India, 10 samples were from Sri Lanka, and 18 samples from Brazil (we have previously shown that AdV D8 was a rare cause of conjunctivitis in the United States for this cohort).[11] De novo reconstruction of these 71 sample genomes was performed using SPAdes 3.11.1.[26] In order to obtain high-quality scaffolds with minimal fragmentation given the large variance in genomic coverage, k-mer sizing was iterated until the average k-mer coverage was below 40. This resulted in an average k-mer coverage of the single largest scaffold node from each sample of 38.6x, with an average length of 34.5 kb, corresponding to ∼99% breadth of coverage of the 34.9 kb reference AdV-D8 genome (GenBank reference KT340071.1). The average coverage of the largest scaffold node from each sample was 330x, corresponding to an average genomic coverage of 326x (Fig 1, top). The remaining 1% of genomic sequence consisted of boundary inverted terminal repeats that were not fully resolved.

Variant calling of the scaffolds was performed using Annovar.[27] We found 630 SNPs (approximately 2% of the viral genome) among the 71 samples. Notably, the average per-sample pairwise distance relative to the reference genome was 62 SNPS, with 50 of these mapping to exonic polymorphisms. No nonsense mutations were found in the dataset. If only protein-level missense mutations are considered, the total variation within the dataset decreased to 156 SNPs across the 71 samples. Combined with the observation that almost 80% of the average pairwise distance between samples mapped to missense mutations, it appears there is significant clustering in the distribution of the missense mutations, while the remainder of the variation is more randomly distributed throughout the genome (Fig 1, bottom).

Maximum likelihood phylogenetic trees were constructed from the scaffolds using Clustal Ω for multiseq alignments[28] and phylip[21] for the phylogenetic analysis (Fig 2A). Three monophyletic groups were identified, one of which was shared by samples from India and Brazil (Clade 1, n = 30) and 2 of which arise from geographically distinct locations in Sri Lanka (Clade 2, n = 11, 10 of which were from Sri Lanka and one from India) and India (Clade 3, n = 30). Interestingly, within sites in India, there was substantial overlap of Clades 1 and 3 with several sites yielding viruses of both clades (Fig S1), suggesting these variants are in circulation

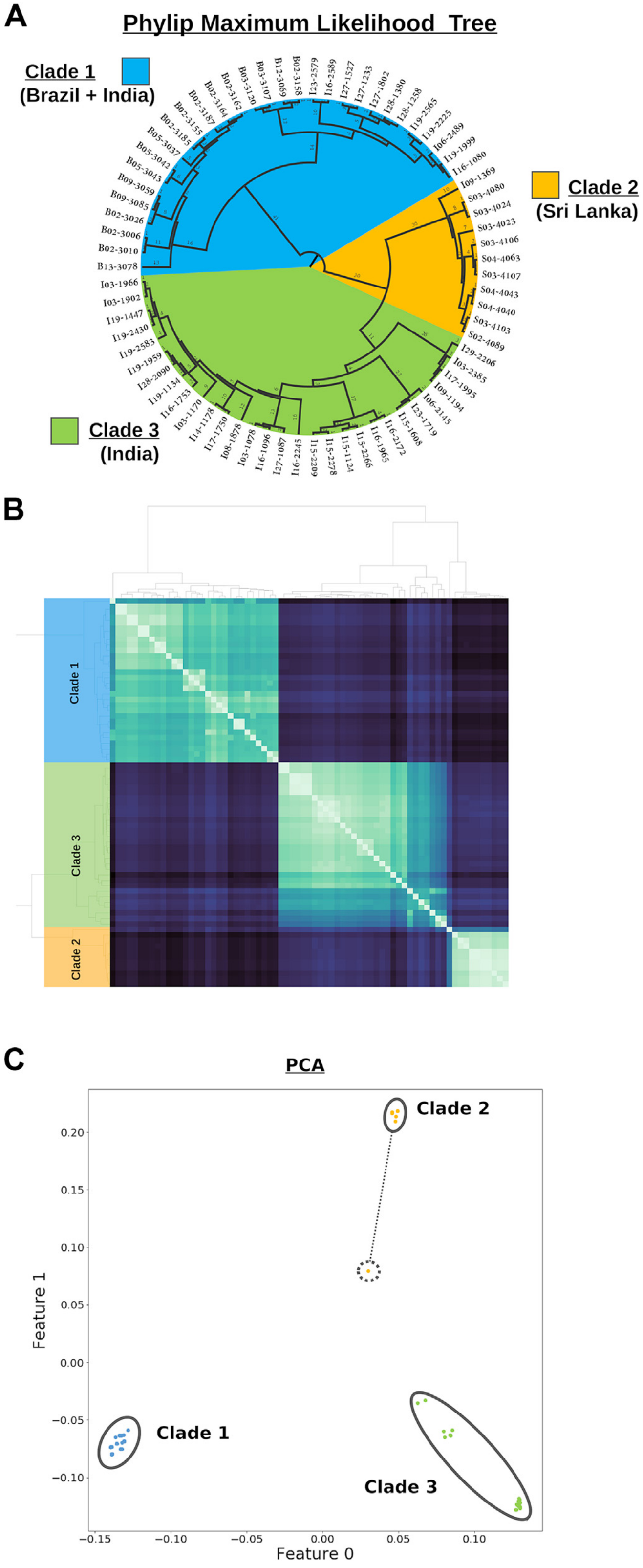


Figure 2. Clustering of Adenovirus D8 sequences. **A,** Phylip maximum likelihood tree showing geographic distribution of clusters. **B,** Jaccard nearest neighbor analysis demonstrating 3 clades of virus. **C,** Principal component analysis (PCA) of viral sequence demonstrating 3 clades.

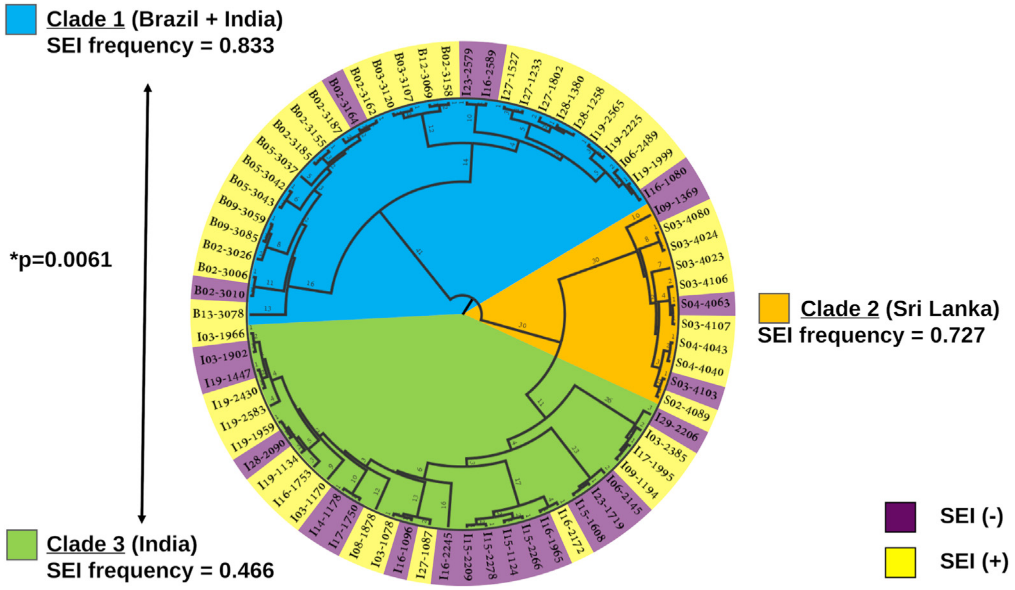


Figure 3. Distribution of subepithelial infiltrates (SEIs) among samples and clades.

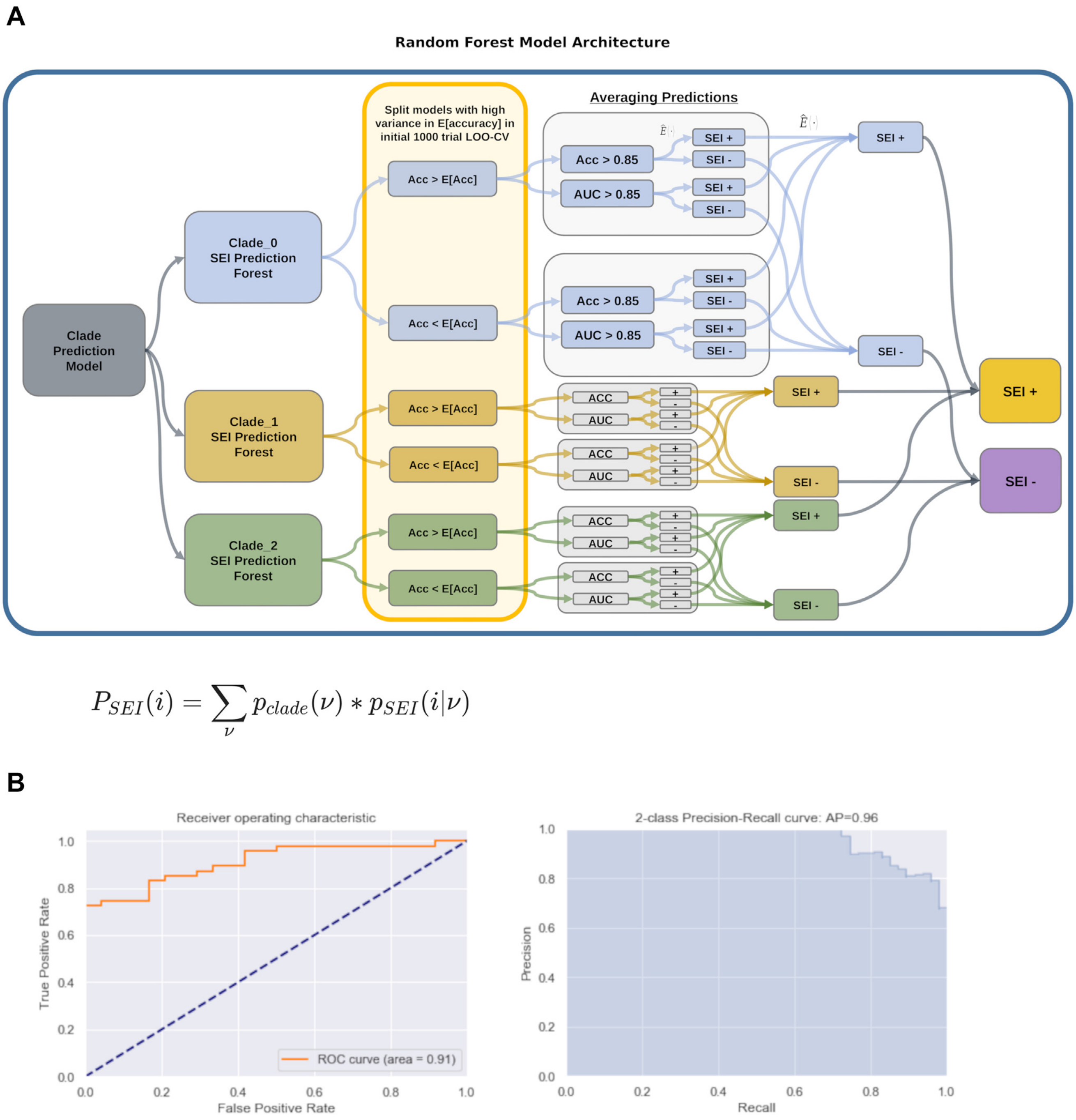
simultaneously in this population. To validate the finding of 3 clades of AdV D8, several independent clustering methods were employed. Pairwise comparison of the SNP sets was completed using multiple distance metrics including overlap coefficient (|X⋂Y|/min(|X|,|Y|)), Jaccard index (|X⋂Y|/|X⋃Y|), Manhattan (L1), Euclidean (L2), and Chebyshev (L¬∞) methods.[29,31,32] Group count, composition, and superstructure remained invariant with respect to metric by all clustering techniques. K-means clustering was run on the Jaccard index distance matrix to identify cluster boundaries and superstructure, followed by density-based spatial clustering of applications with noise (DBSCAN) to identify fine structure of each cluster (Fig 2B).[33] Principal component analysis was performed, and the results from the clustering algorithms were compared (Fig 2C). Finally, the cumulative 1D distribution function was used to compare diversity within and between subtypes. As expected, intragroup diversity was significantly reduced relative to intergroup diversity when all samples are considered, effectively showing a large reduction in the genomic entropy of each group and verifying structure of the clusters.

To determine if the identified 3-clade structure was unique to this dataset or encompassed previously sequenced AdV8 samples, the 54 AdV-D8 complete genomes present in the National Center for Biotechnology Information GenBank (as of January 1, 2021) were retrieved and added to the dataset. Identical clustering analysis again demonstrated 3 distinct monophyletic groups. All existing full-length sequences mapped onto the 3-clade structure (Fig S2) without outliers. Of note, historical samples were found in each of the 3 identified clades, and each clade contained samples originating from ≥ 2 continents.

The demographics and clinical outcomes of subjects infected with each of the 3 clades were compared (Table S2). Among 71 samples, there were 30 samples in clade 1, 11 samples in clade 2, and 30 samples in clade 3. The mean age of patients was 32.1 years, and 57.5% were male. Patients with Clade 2 virus tended to be older (mean age: 40.3 years) than in the other clades, and there was a trend toward female predominance in the Clade 3 samples. Visual acuity and presenting sign scores were indistinguishable between clades. Presenting viral load was similarly not correlated with clade. There was a statistically significant difference in the total presenting symptom score between clades, with subjects in Clade 1 presenting with a higher symptom score of 5.13 ± 3.53 than 2.36 ± 2.62 in Clade 2 subjects and 3.27 ± 2.77 in Clade 3 subjects (P=0.017). However, after multiple-comparison adjustment of statistics, none of 15 single presenting signs or symptoms in the composite score were significantly different between clades. With respect to outcomes, there were significant differences in the SEI frequency between Clades 1 (83% SEI) and 2 (73% SEI) and Clade 3 (46% SEI) (P=0.0061 by Fisher exact test, Fig 3).

Machine Learning Predicts Development of SEIs From Viral Sequence

We initially tested each of the 156 missense SNPs for prediction of SEI development using viral genome-wide association study (GWAS). No SNP was significantly predictive of outcomes. Having demonstrated that AdV D8 genomic variation produces clade-wise differential pathogenicity, we next sought to determine if a machine learning paradigm could be utilized to predict development of SEIs given genomic sequence variation across multiple loci. We initially utilized an ensemble of extra trees random decision tree classifier.[34–36] A 2-stage paradigm was utilized. First, a decision tree model utilizing all sequence variants (coding and noncoding) was used to predict clade of the sample.


A
Random Forest Model Architecture
Clade Prediction Model
Clade_0 SEI Prediction Forest
Clade_1 SEI Prediction Forest
Clade_2 SEI Prediction Forest
Split models with high variance in E[accuracy] in initial 1000 trial LOO-CV
Acc > E[Acc]
Acc < E[Acc]
Averaging Predictions
Acc > 0.85
AUC > 0.85
ACC
AUC
SEI +
SEI -
$\hat{E}[\cdot]$
B
Receiver operating characteristic
True Positive Rate
False Positive Rate
ROC curve (area = 0.91)
0.0
0.2
0.4
0.6
0.8
1.0
2-class Precision-Recall curve: AP=0.96
Precision
Recall

$$P_{SEI}(i) = \sum_{\nu} p_{clade}(\nu) * p_{SEI}(i|\nu)$$

Figure 4. Machine learning architecture and results. **A,** Schematic of ensemble of trees model for machine learning. **B,** Receiver operating characteristic (ROC) and precision recall on 71 training samples using leave-one-out (LOO) methodology using extra random forest model. AUC = area under the curve; SEI = subepithelial infiltrate.

One hundred percent accuracy in assignment relative to clustering was observed. Second, within each clade, an extra tree random forest ensemble was trained using only missense mutations for SEI outcomes (Fig 4). Accuracy was assessed using multiple iterations with holdout set. After training, the engineered ensemble of extra random decision tree classifiers resulted in $>$ 97% accuracy with $>$ 99.5% receiver operator characteristic area under the curve (AUC) and 99.8% precision-recall (PR) score on leave-one-out cross-validation of the training set.

To evaluate potential overfitting, a 1000-trial bootstrap estimate of model performance with randomly permuted training set SEI outcomes was conducted. When the training set SEI outcome results were randomly permuted, both

Table 1. Polymorphisms and Linkages Used in Machine Learning Prediction of Outcomes

Prediction	Polymorphism	Gene	Amino Acid Substitution	Linked Mutations
Clade	C24044G	L4	Q433E	T20418C (L3), T21020C (L3), A22381G (E2A), A23710G (L4), A24202G (L4), C27475G (E3), C27969T (E3), C29037A (E3), A29393C (E3)
Clade	C28031G	E3	Q202E	C21209G, G27384A (E3), C27475G (E3), A28457G (E3), A30716G, G33373A (E4), G34504A
Country	G188A	Upstream of E1A	-	C2671T (E1B), C2682T (E1B), T3712C (pIX), T6873C (E2B), G10209A, C10510A, G13050C (L1), C13486T, C16591T (L2), C16801T (L1), G18405C (L3), C22963G (L4), C22973T (L4), C23032T (L4), G24908A (L4), C26536T, T26635G, C30531G, G30609A, C30841T (L5), G31354A (L5), C33901T (E4), C34315A
Country	G4470A	pIVa2	L251L	
Country	T20418C	L3	H889H	T21020C (L3), A22381G (E2A), A23710G (L4), C24044G (L4), A24202G (L4), C27475G (E3), C27969T (E3), C29037A (E3), A29393C (E3)
SEI 1	A3861G	Downstream of E1B, pIVa2, pIX	-	G7922A (E2B), A32968T (E4), C33208T (E4), C33236T (E4)
SEI 1	C4866T	pIVa2	L116L	C8814T (E2B), G9730A (E2B), C10958T (L1), C14645G (L2), G20199A (L3), C20729T (L3), C21715T (E2A), G25212C, C26751T, C26766A
SEI 1	C5187T	E2B, pIVa2	S1023N, Q12Q	G7175A (E2B), C20892T (L3), G28748C (E3)
SEI 1	A8962G	E2B	F411F	
SEI 1	G31042A	L5	G94G	
SEI 1	G31285A	L5	K175K	
SEI 1	G34825A	Upstream of E4	-	A34838C, G34840T, G34841C
SEI 2	C1707G	E1B	L54V	C6163T (E2B), C25946T (L4)
SEI 2	G31589C	L5	E277Q	
SEI 2	A33281T	E4	I5I	
SEI 2	T34747C	Upstream of E4	-	
SEI 3	G197A	Upstream of E1A	-	A1381G (E1A), T7980G (E2B), C25795G (L4)
SEI 3	C2401T	E1B	C183C	
SEI 3	C3630T	pIX	A69V	
SEI 3	G4566T	pIVa2	I219I	G17362A (L3), T28133A (E3)
SEI 3	C6155T	E2B	K700K	
SEI 3	G10295A	Upstream of E2B, L1	-	
SEI 3	G26961C	E3	L11L	
SEI 3	C31928T	Downstream of E4, L5	-	

Linked polymorphisms occurring within specific genes are denoted in parentheses. All polymorphisms are annotated with respect to reference adenovirus D8 sequence (GenBank KT340071.1). SEI = subepithelial infiltrate.

accuracy and AUC were indistinguishable from random choice (E[Acc] = 0.50 ± 0.12, E[AUC] = 0.50 ± 0.15), indicating that machine learning identified meaningful sequence variants in prediction of SEIs rather than polymorphisms with chance correlation to SEI outcomes.

To further validate the learning model, 16 previously unsequenced AdV D8 samples from the placebo arm of the trial, equally divided from India, Brazil, and Sri Lanka, were sequenced and genomes reconstructed. Eleven of these 16 samples were associated with development of SEIs. The extra random forest classifier trained on the previous 71 samples was applied to these new sequences. The machine learning paradigm successfully predicted the clade of origin of the samples (16/16 correct assignments with n = 3 possible outcomes, $P = 2.32 \times 10^{-8}$) and correctly predicted presence or absence of SEIs in all cases (16/16 correct assignments with n = 2 outcomes, $P = 0.000015$). A second machine pattern discovery method (Pattern Discovery Engine™ [PDE],[37,38] Pattern Computer®, Redmond, WA) was applied to the same dataset. Following training of the system on the 71-sample training set, this model was also able to predict clade of origin and clinical outcomes correctly in 16/16 samples.

Machine Learning Models Utilize Multiple Linked Polymorphisms for Prediction of SEIs

Extra random forest ensemble models have low explainability.[39] We utilized the PDE methodology to understand the basis for machine learning prediction of outcomes data. The discovery engine identifies, weights, and ranks the most informative covariates within a system and transforms that information into a model represented by a simple set of equations. Using this methodology, single equations were extracted that exactly predict viral country of origin and clade across the training and test datasets. A parsimonious set of equations built to predict SEIs across all clades successfully predicted SEI status for 93% and 100% of the training and test sets, respectively. Overall, the systems of equations built by the discovery engine utilized 21 SNPs.

$$\boldsymbol{Country = 2 - G4470A + T20418C \cdot G4470A + G188A \cdot T20418C} \tag{1}$$

$$\boldsymbol{Clade = 2 \cdot C28031G + C24044G} \tag{2}$$

$$\begin{aligned} \boldsymbol{SEI_0} &= \boldsymbol{(1 - G31042A) \cdot (1 - G31285C) \cdot (1 - G34825A) \cdot (1 - C5187T)} \\ &\quad \boldsymbol{\cdot (1 - C4866T) \cdot G31285A + A3861G + A8962G} \\ \delta_{SEI_0} &= \begin{cases} \textbf{\textit{Infiltrates,}} & \text{if } \boldsymbol{SEI_0 \geq 1,} \\ \textbf{\textit{Does Not Infiltrate,}} & \textbf{\textit{Otherwise.}} \end{cases} \end{aligned} \tag{3}$$

$$\begin{aligned} \boldsymbol{SEI_1} &= \boldsymbol{(1 - C1707G) \cdot (1 - G31589C) \cdot (1 - A33281T) \cdot T34747C} \\ \delta_{SEI_1} &= \begin{cases} \textbf{\textit{Infiltrates,}} & \text{if } \boldsymbol{SEI_1 \geq 1,} \\ \textbf{\textit{Does Not Infiltrate,}} & \textbf{\textit{Otherwise.}} \end{cases} \end{aligned} \tag{4}$$

$$\begin{aligned} \boldsymbol{SEI_2} &= \boldsymbol{((((1 - C6155T) \cdot (1 - G10295A) \cdot (1 - G197A) + (1 - C31928T))} \\ &\quad \boldsymbol{\cdot (1 - C6155T) + (1 - C2401T)) \cdot (1 - C3630T) + (1 - C31928T))} \\ &\quad \boldsymbol{\cdot (1 - C26961G) \cdot (1 - C26961G) \cdot (1 - C3630T) \cdot (1 - G4566T)} \\ &\quad \boldsymbol{+ (1 - C3630T) + (1 - C26961G) \cdot (1 - C3630T)} \\ \delta_{SEI_2} &= \begin{cases} \textbf{\textit{Infiltrates,}} & \text{if } \boldsymbol{SEI_2 \geq 6,} \\ \textbf{\textit{Does Not Infiltrate,}} & \textbf{\textit{Otherwise.}} \end{cases} \end{aligned} \tag{5}$$

Country Keys

- 1 = Brazil
- 2 = India
- 3 = Sri-Lanka

Possible clade values are 0, 1 and 2.

SEI_i, SEI model equation for clade i.

Figure 5. Predictive equations for determination of country, clade, and subepithelial infiltrate (SEI) development. Presence of a specific polymorphism (i.e., G31042A) equates to a binary value of 1, and all other values are 0. Clade is first established through equation (2), followed by calculation of SEI prediction through equations (3), (4), and (5).

Two SNPs were required to perfectly distinguish between clades, and 19 SNPs were used by the discovery engine to predict SEIs across viral clades (Table 1 and Fig 5). For clade prediction, the 2 utilized SNPs were perfectly linked to numerous other SNPs in the dataset (8 additional SNPs for C2044G and 7 additional SNPs for C28031G). These SNPs are located in the E2A, E3, E4, L3, and L4 genes. The 19 polymorphisms that are used in the machine learning paradigm to predict SEIs are in turn highly linked to numerous other genes, with polymorphisms used in clade 1 prediction linked to 17 additional polymorphisms, polymorphisms for clade 2 prediction linked to 2 additional SNPs, and clade 3 prediction polymorphisms linked to 5 additional SNPs (Table 1). As these SNPs are highly linked, it is impossible to attribute pathogenicity to any specific SNP in the linked group. Thus, information required to predict outcome is distributed broadly through the viral genome and represents either tightly linked polymorphisms across genes or high cooperativity among multiple genes in determining outcomes.

Discussion

We have found that whole-viral-genome sequencing of a total of 87 AdV D8 isolats causing adenoviral keratoconjunctivitis reveals previously unappreciated subtypes of this virus, with 3 well-delineated subtype clades. The finding that existing AdV D8 full-length sequences, including sequences from Europe, North America, and Japan, fall within the 3-clade structure suggests these are stable viral subtypes and that their evolution predated this study. It is possible that the relative geographic segregation of specific types in this study may be a consequence of specific localized outbreaks at the time of the study. It is also noteworthy that we did not see clear evidence of recombination of viral clades, even though multiple clades were in contemporaneous circulation in India during this study. This suggests that the linked polymorphisms which define clades (which constitute 19 SNPs in total in 2 linkage groups) may have achieved stability within each clade.

Risk of SEIs, a serious complication of viral conjunctivitis, varied significantly with clade, strongly indicating that viral sequence variants influence the likelihood of this complication. Further analysis of the dataset using 2 independent machine learning methods produced models capable of predicting development of SEIs with > 97% accuracy including perfect prediction of 16 novel samples, using solely viral sequence information. Taken together, these findings add to prior evidence that genetic variants beyond the hexon gene strongly influence pathogenesis[40] and that nearly all risk of SEIs is determined by complex factors identifiable within the viral sequence.

The concept that sequence variants may influence pathogenicity has a strong basis in other viruses. This has been best documented for influenza virus, in which the hemagglutinin and neuraminidase gene variants are associated with variable pathogenicity (for example, with H1N1 being the cause of the 1918 pandemic or the H3N2 associated with the 1968 Hong Kong outbreak).[41,42] Previous work in adenoviral conjunctivitis pathogenicity has linked hexon gene variants with disease severity. D-species variants (particularly types 8, 37, and 64[22]) are known to be associated with EKC, while B- and E-species variants such as B3 and E4 have been more commonly associated with milder follicular conjunctivitis or pharyngeal conjunctivitis. Within the D family, the D8 variant is the most common cause of EKC. Typing of adenoviruses has historically been based on immunogenicity, largely dependent on nucleocapsid proteins, particularly the hexon capsomere (L3) and fiber (L5) genes. The current study finds that noncapsid sequence variants appear to contribute to pathogenicity of this family of viruses.

The use of machine learning in identifying viral sequence pathogenicity appears to be a powerful approach for identification of pathogenic strains and associated genetic variants. The success of this technique in predicting outcomes in the validation samples demonstrates its power in identifying pathogenic variants of an infectious pathogen. The finding that machine learning could predict SEI outcomes without any knowledge of patient-dependent factors suggests that much (if not all) of the variation in outcomes in adenoviral conjunctivitis is driven by virus-specific factors. While a simple model for this might entail expression of a uniquely immunogenic protein causing SEIs in certain strains, traditional viral GWAS analysis, in which each individual SNP was assessed for contribution to outcome, failed to identify any SNPs that were uniquely predictive of SEIs. Only analysis of the ensemble genome by machine learning identified the patterns of SNPs associated with disease, suggesting that SEIs arise from conjoint action of several gene products.

The discovery engine methodology allows for identification of the viral features used to drive prediction. Determination of the clade relied on only 2 linked groups of SNPs but included linked variants in the E2A, E3, E4, L3, and L4 genes. Interestingly, among the factors identified by this approach, variants in the penton gene (L2), which codes for the protein that guides viral internalization, and the L5 capsid fiber gene did not appear to contribute substantially to clade prediction. However, the L5 gene, which has previously been shown to determine viral tropism for the corneal epithelium,[43] was among those genes with SNPs contributing to prediction of SEI for clades 1 and 2. It is remarkable that prediction of SEIs in the different clades used different sets of SNPs, with no single gene's SNPs required for all 3 models. The high degree of linkage between polymorphisms used by the machine learning approaches limits biological hypotheses as to pathogenesis of SEIs. For instance, the C4866T allele in the pIVa2 gene used in SEI clade 1 prediction is polymorphic with 7 other coding variants in 5 other genes as well as 2 noncoding variants, any one of which could be biologically relevant. In the absence of recombinants, it is impossible at present to assign roles for these alleles in pathogenesis. Individual alleles could be tested through production of viruses with high- and low-susceptibility alleles and subsequent in vitro and animal pathogenicity testing. For example, previous work has suggested a role for the E3-14.7 gene in viral pathogenesis via inactivation of host TNF-alpha,[44,45] and similarly, viruses with deleted E2B-pTP show altered pathogenicity due to attenuated host immune response.[46] It also remains to be determined if identical or similar polymorphisms in the same genes will affect outcomes for other conjunctivitis-causing species and strains such as E4, D37, or D64.

Several caveats apply to the machine learning approach. First, as noted previously, there are numerous linked polymorphisms across clades. While the ensemble of trees approach will weight each polymorphism, it is possible that only one of the linked sequence changes is actually the driving outcome, while the others serve as linkage markers. Indeed, performance was slightly better on the test set than on the training set, which may be attributable to divergent outcomes arising from nearly identical viral variants in the latter. Additionally, the validation set consisted of 16 subjects taken from the same clinical protocol as the training set. These cases may have been in temporal or spatial proximity to each other, which may limit the range of detected polymorphisms. The generalizability of the machine learning algorithm for prediction remains untested for other populations. Despite these provisos, it appears that utilization of machine learning approaches to analysis of viral variants and their clinical courses allows for identification of complex genetic interactions that determine outcome. Such an approach may be applicable to many other questions in viral pathogenesis such as determination of oncogenic potential of human papilloma viruses,[47] understanding risks for reactivation of varicella zoster causing shingles,[48] or understanding determinants of outcomes from SARS-CoV2.[49]

Footnotes and Disclosures

Originally received: November 23, 2021.
Final revision: April 24, 2022.
Accepted: April 25, 2022.
Available online: May 10, 2022.

[1] Department of Ophthalmology, University of Washington School of Medicine, Seattle, Washington.

[2] Roger and Angie Karalis Johnson Retina Center, University of Washington School of Medicine, Seattle, Washington.

[3] Department of Ophthalmology, Massachusetts Eye and Ear Infirmary, Harvard Medical School, Boston, Massachusetts.

[4] NovaBay, Emeryville, California.

[5] Pattern Computer, Redmond, Washington.

[6] Department of Biological Structure, University of Washington School of Medicine, Seattle, Washington.

[7] Department of Laboratory Medicine and Pathology, University of Washington School of Medicine, Seattle, Washington.

Disclosure(s):

All authors have completed and submitted the ICMJE disclosures form.

The authors made the following disclosures: A.W.-Y.: Paid employee – Pattern Computer, Inc. Q.J.: Paid employee – Pattern Computer, Inc. W.B.H.: Paid employee – Pattern Computer, Inc. D.S.: Employee – NovaBay, LLC.

Aaron Y. Lee and Cecilia S. Lee, an Associate Editor and Editor of this journal, were recused from the peer-review process of this article and had no access to information regarding its peer-review. Co-author David Stroman passed away during the review of this manuscript; the authors dedicate this work to his memory.

HUMAN SUBJECTS: Human Subjects were used in this study. This study was HIPAA-compliant, approved by institutional board review (Multiple IRBs approved this study: US: Goodwin IRB [Cincinnati, OH]; India: Drug Controller General [DCGI] with local site ethics committees; Sri Lanka: Scientific and Ethical Review, Faculty of Medicine, University of Kelaniya, Colombo, Sri Lanka; Brazil: National Ethics Committee in Research [CONEP] plus local site ethic committees), and conducted in accord with the Declaration of Helsinki. The BAYnovation clinical trial was registered at clinicaltrials.gov (NCT01877694). Informed consent was obtained from all study subjects.

Supported by the National Institutes of Health R21EY027453, R21EY033174, and P30EY0001730 (R.V.G.), R01EY013124 and R01EY021558 (J.C. and J.R.), an unrestricted grant from Research to Prevent Blindness (R.V.G.), and the Mark J. Daily, MD Research Fund.

Author Contributions:

Research Design: Nakamichi, Stroman, Wolf-Yadlin, Jackson, Holz, A.Y.Lee, C.S.Lee, Van Gelder

Data Acquisition: Nakamichi, Akileswaran, Stroman, C.S. Lee, Van Gelder

Data Analysis: Nakamichi, Meirick, M. Lee, Chodosh, Rajaiya, Stroman, Wolf-Yadlin, Jackson, Holz, A.Y. Lee, C.S. Lee, Van Gelder

Obtained funding: Stroman, Chodosh, Rajaiya, Van Gelder

Manuscript preparation: Nakamichi, Akileswaran, Stroman, Van Gelder

Abbreviations and Acronyms:
AdV = adenovirus; **AUC** = area under the curve; **EKC** = epidemic keratoconjunctivitis; **GWAS** = genome-wide association study; **PCA** = Principal component analysis; **PCR** = polymerase chain reaction; **PR** = precision-recall; **SEI** = subepithelial infiltrate; **SNP** = single-nucleotide polymorphism.

Keywords:
Adenovirus D8, Conjunctivitis, Machine learning, Subepithelial infiltrates.

Correspondence:
Russell N. Van Gelder, MD, PhD, Department of Ophthalmology, University of Washington School of Medicine, Campus Box 359608, 325 9th Avenue, Seattle, WA 98104. E-mail: russvg@uw.edu.

References

1. Udeh BL, Schneider JE, Ohsfeldt RL. Cost effectiveness of a point-of-care test for adenoviral conjunctivitis. *Am J Med Sci*. 2008;336:254–264.
2. Jhanji V, Chan TC, Li EY, et al. Adenoviral keratoconjunctivitis. *Surv Ophthalmol*. 2015;60:435–443.
3. Smith AF, Waycaster C. Estimate of the direct and indirect annual cost of bacterial conjunctivitis in the United States. *BMC Ophthalmol*. 2009;9:13.
4. Centers for Disease Control and Prevention (CDC). Adenovirus-associated epidemic keratoconjunctivitis outbreaks–four states, 2008-2010. *MMWR Morb Mortal Wkly Rep*. 2013;62:637–641.
5. Cheung D, Bremner J, Chan JT. Epidemic keratoconjunctivitis–do outbreaks have to be epidemic? *Eye (Lond)*. 2003;17:356–363.
6. Ismail AM, Zhou X, Dyer DW, et al. Genomic foundations of evolution and ocular pathogenesis in human adenovirus species D. *FEBS Lett*. 2019;593:3583–3608.
7. Montessori V, Scharf S, Holland S, et al. Epidemic keratoconjunctivitis outbreak at a tertiary referral eye care clinic. *Am J Infect Control*. 1998;26:399–405.
8. Paparello SF, Rickman LS, Mesbahi HN, et al. Epidemic keratoconjunctivitis at a U.S. military base: Republic of the Philippines. *Mil Med*. 1991;156:256–259.
9. Ohnsman CM. Exclusion of students with conjunctivitis from school: policies of state departments of health. *J Pediatr Ophthalmol Strabismus*. 2007;44:101–105.
10. Zhang L, Zhao N, Sha J, et al. Virology and epidemiology analyses of global adenovirus-associated conjunctivitis outbreaks, 1953-2013. *Epidemiol Infect*. 2016;144:1661–1672.
11. Lee CS, Lee AY, Akileswaran L, et al. Determinants of outcomes of adenoviral keratoconjunctivitis. *Ophthalmology*. 2018;125:1344–1353.
12. Levinger E, Trivizki O, Shachar Y, et al. Topical 0.03% tacrolimus for subepithelial infiltrates secondary to adenoviral keratoconjunctivitis. *Graefes Arch Clin Exp Ophthalmol*. 2014;252:811–816.
13. Lund OE, Stefani FH. Corneal histology after epidemic keratoconjunctivitis. *Arch Ophthalmol*. 1978;96:2085–2088.
14. Pennington MR, Saha A, Painter DF, et al. Disparate entry of adenoviruses dictates differential innate immune responses on the ocular surface. *Microorganisms*. 2019;7:351.
15. Chintakuntlawar AV, Zhou X, Rajaiya J, Chodosh J. Viral capsid is a pathogen-associated molecular pattern in adenovirus keratitis. *PLoS Pathog*. 2010;6:e1000841.
16. Alsuhaibani AH, Sutphin JE, Wagoner MD. Confocal microscopy of subepithelial infiltrates occurring after epidemic keratoconjunctivitis. *Cornea*. 2006;25:1102–1104.
17. Aydin Kurna S, Altun A, Oflaz A, Karatay Arsan A. Evaluation of the impact of persistent subepithelial corneal infiltrations on the visual performance and corneal optical quality after epidemic keratoconjunctivitis. *Acta Ophthalmol*. 2015;93:377–382.
18. Butt AL, Chodosh J. Adenoviral keratoconjunctivitis in a tertiary care eye clinic. *Cornea*. 2006;25:199–202.
19. Kepez Yildiz B, Urvasizoglu S, Yildirim Y, et al. Changes in higher-order aberrations after phototherapeutic keratectomy for subepithelial corneal infiltrates after epidemic keratoconjunctivitis. *Cornea*. 2017;36:1233–1236.
20. Tekin K, Kiziltoprak H, Koc M, et al. The effect of corneal infiltrates on densitometry and higher-order aberrations. *Clin Exp Optom*. 2019;102:140–146.
21. Bailey A, Mautner V. Phylogenetic relationships among adenovirus serotypes. *Virology*. 1994;205:438–452.
22. Zhou X, Robinson CM, Rajaiya J, et al. Analysis of human adenovirus type 19 associated with epidemic keratoconjunctivitis and its reclassification as adenovirus type 64. *Invest Ophthalmol Vis Sci*. 2012;53:2804–2811.
23. Robinson CM, Shariati F, Zaitshik J, et al. Human adenovirus type 19: genomic and bioinformatics analysis of a keratoconjunctivitis isolate. *Virus Res*. 2009;139:122–126.
24. Robinson CM, Shariati F, Gillaspy AF, et al. Genomic and bioinformatics analysis of human adenovirus type 37: new insights into corneal tropism. *BMC Genomics*. 2008;9:213.
25. Lee JS, Mukherjee S, Lee JY, et al. Entry of epidemic keratoconjunctivitis-associated human adenovirus type 37 in human corneal epithelial cells. *Invest Ophthalmol Vis Sci*. 2020;61:50.
26. Bankevich A, Nurk S, Antipov D, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol*. 2012;19:455–477.
27. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38:e164.
28. Sievers F, Wilm A, Dineen D, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol*. 2011;7:539.
29. Bastien O, Aude JC, Roy S, Marechal E. Fundamentals of massive automatic pairwise alignments of protein sequences: theoretical significance of Z-value statistics. *Bioinformatics*. 2004;20:534–537.
30. Lee AY, Lee CS, Van Gelder RN. Scalable metagenomics alignment research tool (SMART): a scalable, rapid, and complete search heuristic for the classification of metagenomic sequences from complex sequence populations. *BMC Bioinformatics*. 2016;17:292.
31. Hou L, Wang L, Berg A, et al. Comparison and evaluation of network clustering algorithms applied to genetic interaction networks. *Front Biosci (Elite Ed)*. 2012;4:2150–2161.
32. Jay JJ, Eblen JD, Zhang Y, et al. A systematic comparison of genome-scale clustering algorithms. *BMC Bioinformatics*. 2012;13 Suppl 10:S7.
33. Zhao Y, Liu X, Li X. An improved DBSCAN algorithm based on cell-like P systems with promoters and inhibitors. *PLoS One*. 2018;13:e0200751.

34. Mishra G, Sehgal D, Valadi JK. Quantitative structure activity relationship study of the anti-hepatitis peptides employing random forests and extra-trees regressors. *Bioinformation*. 2017;13:60−62.
35. Zhu R, Zeng D, Kosorok MR. Reinforcement learning trees. *J Am Stat Assoc*. 2015;110:1770−1784.
36. Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. Inferring regulatory networks from expression data using tree-based methods. *PLoS One*. 2010;5:e12776.
37. Mohammed I, Singh N, Venkatasubramanian M. Computer-assisted detection and diagnosis of pediatric pneumonia in chest X-ray images. https://www.patterncomputer.com/wp-content/uploads/2022/02/Computer-Assisted-Detection-and-Diagnosis-of-Pediatric-Pneumonia-in-Chest-X-ray-Images.pdf; 2019. Accessed April 24, 2022.
38. Singh N, Venkatasubramanian M, Mohammed I, et al. On the road to personalized medicine: discovery of prognostic combinatorial high-order interactions in breast cancer. https://www.patterncomputer.com/wp-content/uploads/Discovery-of-Hidden-Patterns-in-Complex-Data.pdf; 2018. Accessed April 24, 2022.
39. Vandewiele G, Steenwinckel B, Turck F, Ongenae F. MINDWALC: mining interpretable, discriminative walks for classification of nodes in a knowledge graph. *BMC Med Inform Decis Mak*. 2020;20:191.
40. Singh G, Zhou X, Lee JY, et al. Recombination of the epsilon determinant and corneal tropism: human adenovirus species D types 15, 29, 56, and 69. *Virology*. 2015;485:452−459.
41. Kileng H, Kjellin M, Akaberi D, et al. Personalized treatment of hepatitis C genotype 1a in Norway and Sweden 2014-2016: a study of treatment outcome in patients with or without resistance-based DAA-therapy. *Scand J Gastroenterol*. 2018;53:1347−1353.
42. Yan Z, Wang Y. Viral and host factors associated with outcomes of hepatitis C virus infection (review). *Mol Med Rep*. 2017;15:2909−2924.
43. Ismail AM, Lee JS, Dyer DW, et al. Selection pressure in the human adenovirus fiber knob drives cell specificity in epidemic keratoconjunctivitis. *J Virol*. 2016;90:9598−9607.
44. Robinson CM, Rajaiya J, Zhou X, et al. The E3 CR1-gamma gene in human adenoviruses associated with epidemic keratoconjunctivitis. *Virus Res*. 2011;160:120−127.
45. Klingseisen L, Ehrenschwender M, Heigl U, et al. E3-14.7K is recruited to TNF-receptor 1 and blocks TNF cytolysis independent from interaction with optineurin. *PLoS One*. 2012;7:e3 8348.
46. Osada T, Yang XY, Hartman ZC, et al. Optimization of vaccine responses with an E1, E2b and E3-deleted Ad5 vector circumvents pre-existing anti-vector immunity. *Cancer Gene Ther*. 2009;16:673−682.
47. Ou Z, Chen Z, Zhao Y, et al. Genetic signatures for lineage/sublineage classification of HPV16, 18, 52 and 58 variants. *Virology*. 2020;553:62−69.
48. Chow VT, Tipples GA, Grose C. Bioinformatics of varicella-zoster virus: single nucleotide polymorphisms define clades and attenuated vaccine genotypes. *Infect Genet Evol*. 2013;18: 351−356.
49. Nakamichi K, Shen JZ, Lee CS, et al. Hospitalization and mortality associated with SARS-CoV-2 viral clades in COVID-19. *Sci Rep*. 2021;11:4802.