JMIR Bioinformatics and Biotechnology

Methods, devices, web-based platforms, open data and open software tools for big data analytics, understanding biological/medical data, and information retrieval in biology and medicine.
Volume 4 (2023) ISSN 2563-3570 Editor in Chief: Ece D. Uzun, MS, PhD, FAMIA

Contents

Original Papers

Editorial

Original Paper

Decision of the Optimal Rank of a Nonnegative Matrix Factorization Model for Gene Expression Data Sets Utilizing the Unit Invariant Knee Method: Development and Evaluation of the Elbow Method for Rank Selection

Emine Guven[1], MSc, PhD

Department of Biomedical Engineering, Düzce University, Düzce, Turkey

Corresponding Author:
Emine Guven, MSc, PhD
Department of Biomedical Engineering
Düzce University
College of Engineering, Main Campus, M-2 Building, #202
Düzce, 81620
Turkey
Phone: 90 5388733459
Email: emine.guven33@gmail.com

Abstract

Background: There is a great need to develop a computational approach to analyze and exploit the information contained in gene expression data. The recent utilization of nonnegative matrix factorization (NMF) in computational biology has demonstrated the capability to derive essential details from a high amount of data in particular gene expression microarrays. A common problem in NMF is finding the proper number rank (r) of factors of the degraded demonstration, but no agreement exists on which technique is most appropriate to utilize for this purpose. Thus, various techniques have been suggested to select the optimal value of rank factorization (r).

Objective: In this work, a new metric for rank selection is proposed based on the elbow method, which was methodically compared against the cophenetic metric.

Methods: To decide the optimum number rank (r), this study focused on the unit invariant knee (UIK) method of the NMF on gene expression data sets. Since the UIK method requires an extremum distance estimator that is eventually employed for inflection and identification of a knee point, the proposed method finds the first inflection point of the curvature of the residual sum of squares of the proposed algorithms using the UIK method on gene expression data sets as a target matrix.

Results: Computation was conducted for the UIK task using gene expression data of acute lymphoblastic leukemia and acute myeloid leukemia samples. Consequently, the distinct results of NMF were subjected to comparison on different algorithms. The proposed UIK method is easy to perform, fast, free of a priori rank value input, and does not require initial parameters that significantly influence the model's functionality.

Conclusions: This study demonstrates that the elbow method provides a credible prediction for both gene expression data and for precisely estimating simulated mutational processes data with known dimensions. The proposed UIK method is faster than conventional methods, including metrics utilizing the consensus matrix as a criterion for rank selection, while achieving significantly better computational efficiency without visual inspection on the curvatives. Finally, the suggested rank tuning method based on the elbow method for gene expression data is arguably theoretically superior to the cophenetic measure.

(JMIR Bioinform Biotech 2023;4:e43665) doi:10.2196/43665

KEYWORDS

gene expression data; nonnegative matrix factorization; rank factorization; optimal rank; unit invariant knee method; elbow method; consensus matrix

Introduction

Nonnegative matrix factorization (NMF) algorithms have been advanced for the application fields of bioinformatics, artificial intelligence [1], signal processing systems [2], and music signal processing systems [3]. Lee and Seung [4] formulated a parts-based illustrated algorithm to solve the problem of the NMF puzzle. Furthermore, various algorithms have been established to develop a solution to the NMF problem depending on the field [5-8].

Several approaches have been developed for clustering samples, mutational processes, and gene expression levels that draw similar expression motifs [4,9-11]. However, cancer analysis and classification based on genomic data offers a more powerful method that approach the sensitivity of advanced computational techniques to tackle certain problems such as modeling multiple, heterogeneous populations and reducing the number of variables (genes or mutations). Consequently, the choice of a trivial number of discriminatory features from thousands of features enhances crafting successful pinpointing classification systems [12-14]. Although neural networks are prone to overfitting, if the examined structure is noisy, as in the case of tumor expression profiling [15], Pal et al [12] suggested a variation of a multilayer perceptron network for biomarkers identification. Nevertheless, these approaches have severe constraints in capturing the entire framework essential in the data. Moreover, they generally highlight the dominant forms in a data set and cannot detect different signatures with a universal standard. Thus, an unbiased technique is needed for deciphering many clusters without visual inspection that is also capable of utilizing a computational program.

A common problem in conventional multivariate data analysis methods such as factor analysis (FA), principal component analysis (PCA), cluster analysis, and NMF is to detect the proper number (r) of factors, principal components, clusters, and ranks, respectively. Item redundancy is common in long questionnaires such as those used in a pilot questionnaire study, arguing for the utilization of FA and the variance inflation factor on a lifestyle questionnaire. Staffini et al [16] concluded that both methods are acceptable for item reduction; however, both of these techniques might produce distinct features as an outcome.

The aim of this study was to utilize the unit invariant knee (UIK) method for obtaining related biological and molecular correlations in gene expression data. The UIK method is used to catch compositions essential for the data and to offer biological understanding by systematizing both the features and samples. The approach is based on a “knee point” and its unit invariant estimation using the extremum distance estimator method introduced by Christopoulous [17]. In this regard, NMF decomposes the gene expression data set into fragments of evocative features such as metagene and mutational signatures. When applying this method to conventional factorization techniques such as PCA or FA with World Values Survey Wave 5 United States data [18], certain factors (elements) clearly explained the questionnaire responses (1=“Not at all like me”...6=“Very much like me”) [19,20].

Therefore, given an NMF method and a data set (a target matrix), the tens of thousands of genes regarding a small number of signatures can be analyzed. Gene expression patterns of samples can then be studied to determine the expression motifs of the signatures. The signatures define an interesting decomposition of genes, analogous to the motifs of Hutchins et al [10] in which the first value is selected where the residual sum of squares (RSS) curvature presents an inflection point. The machinery of the UIK method can then be used to detect this inflection and expression motifs define a robust clustering of samples.

In this study, the elbow technique was considered for model selection utilizing alternative parsing and its robustness was evaluated [19,21]. The idea behind this approach is to develop an unbiased computable optimization point of the RSS curve that can then be used to select tuning parameters. The UIK method has proven to be useful for a variety of models, from classifying recordings of echolocation to a decision of predictive models for soil carbon at the field scale [22,23], but has not been used for NMF on genetic data to date. The advantage of the UIK method relative to the cophenetic measure method [24,25], as another NMF rank estimation measure, is that UIK yields a closed-form formula that can provide greater insight and computational speed in simulations, which can then be applied for selecting the rank of NMF for real high-dimensional hyperspectral data.

Finally, this study applies the combination of NMF and the UIK method (designated the uikNMF method) to simplify cancer classification tasks by clustering tumor samples and mutational signature data sets. This enables illustrating numerous sturdy decompositions of genetic and mutational signatures from experimental and simulated data sets.

Methods

NMF Approach

Given a target matrix $V^{m\times n}$, NMF identifies nonnegative matrices such that $N^{m\times r}$ and $M^{r\times n}$ (ie, with all entries≥0) to present the matrix decomposition as:

$$V \approx NM \quad \textbf{(1)}$$

In practice, N is typically viewed as a basis or metagenes matrix, and the mixture coefficient matrix and metagene expression profiles refer to the matrix N. The rank factorization is chosen such that r≤min(m,n). The goal behind this selection is to explain and split the details classified among V into r factors (ie, the columns of N). Given a matrix $V^{m\times n}$, NMF finds two nonnegative matrices, $N^{m\times r}$ and $M^{r\times n}$ (ie, with all elements≥0), to represent the decomposed matrix as

$$V \approx NM,$$

for instance by natural demanding of nonnegative N and M to minimize the reconstruction error:

$$||V - NM||_F, \text{ subject to } N \geq 0, M \geq 0 \quad \textbf{(2)}$$

In this case, we consider a gene expression data set characterized by the expression levels of *m* genes (probes) by *n* samples of unique tissues, cells, cell lines, time points, or experiments. The

number *m* of genes usually ranges from hundreds to thousands, and the *n* of experiments or patients is typically 100 for gene expression research. The gene expression data set is presented by a matrix of expression *V* of size N×M, whose rows consist of the expression levels of *m* genes and columns consist of *n* samples.

The aim is to identify a small number of rank factorizations, each defined as a positive linear combination of the *V* target matrix. The positive linear combination of metagenes is described by the gene expression motif of the samples. To obtain a dimensional reduction of the microarray data and evaluate the distinctions among samples, NMF was implemented utilizing R statistical environment version 3.6.3 with the "NMF" package [26].

Cophenetic Measure

In the framework of classification analyses, Brunet et al [9] suggested utilizing the *cophenetic correlation coefficient* as a metric asset of the clusters. Furthermore, a cophenetic measure was proposed as one of the metrics utilizing the consensus matrix as a criterion for rank selection [25]. Studying the values of the consensus matrix as a similarity metric, the cophenetic correlation coefficient is defined as the correlation between the sample distances induced by the consensus matrix and the cophenetic distances obtained by its hierarchical clustering.

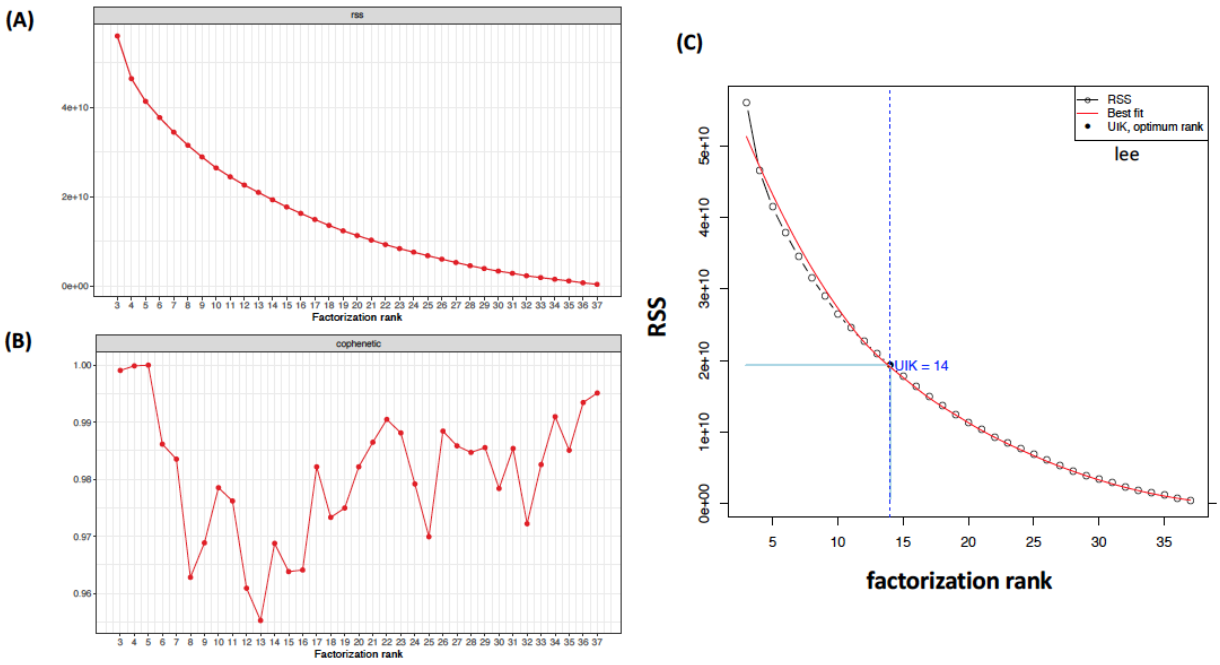
Proposed UIK Method

Hutchins et al [10] demonstrated how the variation in the RSS of the estimated matrix resulting from NMF analysis reveals a robust approximation of the proper number of elements (r). They employed Lee and Seung's [4] algorithm to select r, in which the plot of the RSS presents the first inflection point. In practice, the rank factorization r can be computed with a considerably smaller number of iterations, typically 20-30 runs for each value of r. In contrast, an optimal NMF interpretation requires a couple of hundred random restarts, which is computationally costly.

For instance, a fundamental step for any unsupervised algorithm is to determine the optimal number of clusters (k) into which the data may be clustered [27]. The *elbow method* is one of the most popular methods to determine the optimal value of such components of optimum features [17,18]. The utilization of UIK methodology for identification of the knee (elbow) point of a curve has consistently proven to be immensely advantageous in a wide variety of studies to locate the optimal number of "components" on a scree plot of k-means, PCA, FA, and NMF [27-32].

In many cases, utilization is referred to as uik(x,y), where x is the vector of ranks, components, clusters, or factors and y is the related vector of the RSS curve [10,22,33]. In regression analysis, the term mean squared error (MSE) is sometimes used to refer to the unbiased estimate of error variance (ie, the RSS divided by the degrees of freedom). Ulfarsson and Solo [34] proposed a metric for rank selection in NMF by selecting the tuning parameters of an unbiased computable estimator of the MSE [25]. Thus, as illustrated in Figure 1, the aim is to find an inflection where r meets the proper number of the factorization ranks utilizing the "elbow point," which is virtually the point where a severely decreasing or increasing curve begins to turn "flat enough" [19,20,22,33,35]. Furthermore, this study considered the function of the rank factorization curve and used the function uik() from the R package *inflection* to select the optimal rank [33,36,37]. The uik() function detects the factorization rank when the curve begins to climb faster (start point) and the point beyond which the curve flattens out (ending point), which are generally known as the *knee points* of a curve (Figure 1). In Figure 1, the emergence of factorization rank for the Golub et al [38,39] gene expression data set is shown on the rank survey plot. The optimal rank of the RSS plot is in between knee points detected by the uik() function of the R package *inflection* at the curve to which the cumulative rank factorization belongs.

Figure 1. (A) Rank survey plots for residual sum of squares (RSS) and (B) cophenetic coefficient curves factorization rank. The factorization rank ranges from 3 to 37. The aim is to decide whether the optimal rank factorization is very rigid by simple visual inspection. (C) The function of factorization rank is selected as the emergence rank of the RSS survey. The rank range between knee points is detected by the uik() function of the R package "inflection" at the curve of the cumulative rank units. The best fit is determined using a linear regression model.



Cross-validation

This study used cross-validation to select an optimal number of implicit elements in NMF. The goal of NMF is to obtain low-dimensional N and M with all nonnegative elements by minimizing the reconstruction error $|V - NM|^2$. Leaving out a single entry of V (eg, $\mathbf{V_{ab}}$) and implementing NMF of the resulting matrix may produce a different result than the actual result. In other words, finding N and M while minimizing reconstruction error over all nonmissing entries results in:

$$\sum_{ij \neq ab}(V_{ij} - [NM]_{ij})^2 \textbf{(3)}$$

Consequently, the left-out element V_{ab} can be predicted by calculating $[\mathbf{WH}]_{ab}$ and then determining the prediction error as:

$$E(ab) = (V_{ab} - [\mathbf{WH}]_{ab})^2 \textbf{(4)}$$

One can repeat this process by crossing out all entries of V_{ab} one at a time and adding up the error of prediction overall, a_a and b_b. This will lead to the predicted residual sum of squares (PRESS) value. The PRESS value is defined as $E(r) = \sum_{ab}E(ab)$, which will strongly depend on the rank r. The prediction error, E(r), will have a minimum defined as an "optimal rank" r.

Since the NMF must be reiterated for each crossed-out value and might also be difficult to code (depending on the target matrix entries and how smooth it is to implement NMF with missing values), this can be a computationally expensive procedure. For instance, in PCA, one can avoid this by crossing out entire rows of V, which eventually speeds up the computing [40]. All the traditional cross-validation rules can apply here. Therefore, by not including multiple entries instead of a single entry and iterating the computation process by bootstrapping the entries instead of looping over all the entries, both techniques can help speed up the procedure.

Note that various techniques have been developed to select the optimal rank factorization. For example, Brunet et al [9] suggested seizing the first value of r for which the cophenetic coefficient value was decreasing, whereas Frigyesi et al [11] considered the smallest value at which the decrease in the RSS is lower than the decay of the RSS simulated from random data. The aim of this study was to decide how and which approach performs better on an estimation of the latent factors given different algorithms of NMF.

Gene Expression Data Set

This study illustrates the utilization of NMF based on the UIK method to select the optimal rank on the RSS curve with a leukemia gene expression data set (esGolub) in simplifying cancer subtypes [38,41,42]. This data set has been used in several previous studies on NMF and is built in the NMF package's data [9,26,43], packed into an ExpressionSet object [39]. To achieve biologically meaningful results, we used the entire gene expression data set including 5000 features for 38 leukemia samples. The difference between acute myelogenous leukemia and acute lymphoblastic leukemia (ALL) has been noted. ALL is also separated into two subtypes: T-cell and B-cell ALL.

Furthermore, this data set has served as a touchstone in cancer classification at the molecule, histology, and stage levels [38,44]. In this study, this data set was reprocessed to compare several clustering techniques regarding their effectiveness and permanence in recuperating other differentially expressed genes (DEGs) and associated pathways. Before the NMF procedure, dimension reduction is recommended for larger gene expression

data sets by nonspecific criteria based on the characteristics of the expression estimates (ie, the mean threshold of variance and genes with the smallest average variances) [45].

For example, by looking at the NMF rank survey plot of RSS in Figure 1, we want to decide how many basis vectors we should keep to obtain the optimal rank of the target (original) matrix. To achieve such a task, an unbiased technique for deciding the number of clusters without visual interpretation that is simultaneously capable of utilizing a computational program is needed.

Simulated Mutational Processes Data

The simulated mutational process data obtained from Alexandrov et al [46] is publicly available as a MATLAB file on SigProfiler [47]. They identified the handful of functional processes for a group of 100 simulated cancer genomes based on the repeatability of their signatures and low error for reconstructing the novel catalogs. The data set was generated by employing 10 mutational processes with different signatures (motifs), each with 96 mutation types, and adding a Poisson noise. The data also correspond to the six subtypes: C:G to A:T, C:G to G:C, C:G to T:A, T:A to A:T, T:A to C:G, and T:A to G:C and their immediate 5′ and 3′ sequence background.

Analyses were performed utilizing the R programming language. Before the procedure, the low-quality genes with an inadequate number of reads were eliminated and gene expression values were converted to a logarithmic scale. The data set (Table 1) was then normalized by computing the averages of each sample in R. The *NMF* R package was used to draw plots of rank surveys using the plot() function [48]. Rank survey analysis was performed to compare the optimal rank with distinct methods using the *inflection* package's uik() and check_curve() functions [36]. The readMat() function of the R.matlab package [49] was used to import the simulated mutational processes data (Table 1) from the MATLAB file into the R environment (see Supplementary Data S1 in Multimedia Appendix 1).

Table 1. Gene expression and simulated mutational data sets.

Data set	Size	Samples
esGolub gene expression	5000×38	38
Mutational processes	100×96	96

Results

Applications of NMF Based on the UIK Method

Leukemia (esGolub) Data Set

The present results are based on the NMF package of Gaujoux and Seoighe [26] combined with the technique introduced by Hutchins et al [10] (Figure 1). However, as shown in Figure 2, this study also tested other algorithms taken from the "brunet" and "nsNMF" algorithms to illustrate remarkable differences. It is important to emphasize that there is no remarkable base in the experimental data examined herein. Consequently, it is not possible to demonstrate considerable doubt that the proposed approach operates effectively on the experimental data set. As indicated in Figure 2, the uik() function selects the optimal rank as the curve starts to decline faster (start point) and the point beyond that the curve flattens out (ending point), which are generally known as the knee points of a curve (Figure 1). The UIK method identified 15 components for the brunet algorithm, whereas the nsNMF algorithm detected 14 latent factors as the best representation for the whole esGolub data set.

By simply looking at the cophenetic correlation or RSS plots of rank factorization in Figure 3A, one can confirm that the optimum rank factorization is 3. For performance reasons, the submatrix esGolub (1:200) was initially performed with only 10 runs for each rank value. As demonstrated in Figure 3B, the UIK method of optimal rank factorization was validated by comparing with Gaujoux's estimates of the esGolub subdata set [50] (also see Supplementary Data S2 in Multimedia Appendix 1). Consensus methods converged on a rank of 3, replicating the result of Brunet et al [9], in which it was proposed that 3 factors yielded a more complete understanding of the esGolub data set with 200 features from 38 leukemia samples.

Figure 2. Application of the unit invariant knee (UIK) method on different algorithms: (A) "Brunet" and (B) "nsNMF." The optimal rank, which UIK represents, is 15 for the Brunet algorithm, whereas the UIK of the nsNMF algorithm reveals 14 as an optimum rank, similar to the "Lee" algorithm.

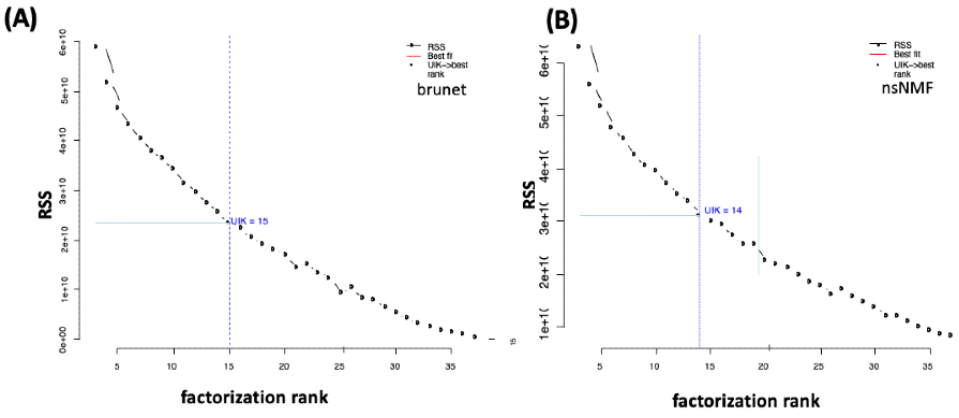
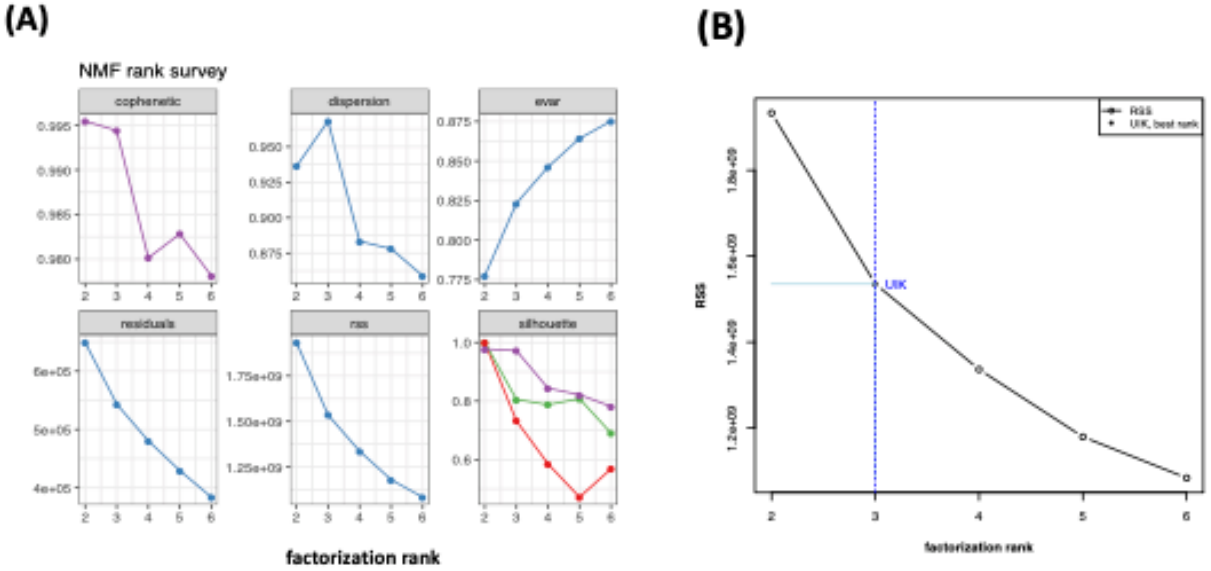


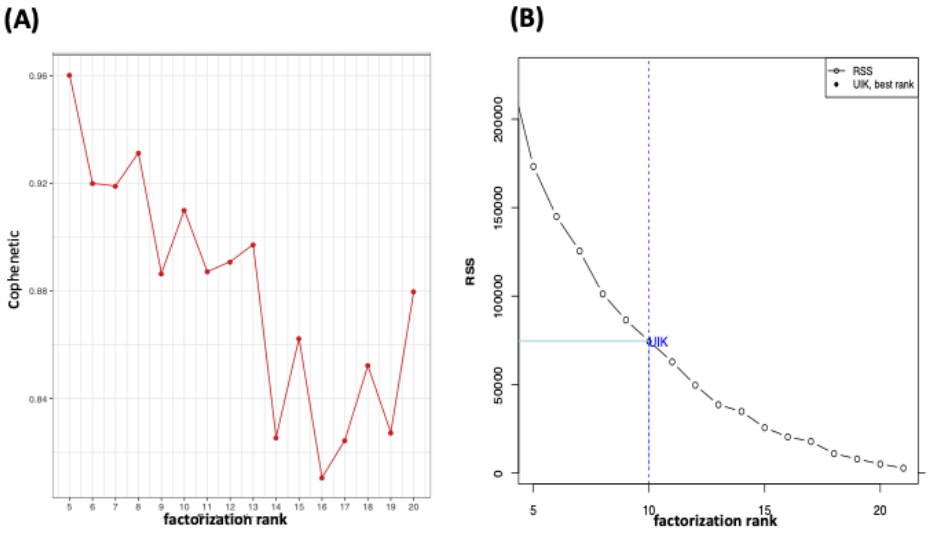
Figure 3. (A) Estimation of the optimal rank. Nonnegative matrix factorization (NMF) survey plot of quality measures obtained from factorization rank from 2 to 6 by running the target matrix esGolub [1:200] 10 times. (B) The function of factorization rank is selected as the emergence rank of the residual sum of squares (RSS) survey. For example, the rank range of 2 to 6 is between knee points detected by the R inflection package's uik()function at 3. Overall, the method of the UIK estimation was confirmed with former results.



Simulated Mutational Process Data

It is challenging to observe the rank factorization of the simulated data on the cophenetic coefficient curve (Figure 4A). Moreover, there is no clue in deciding rank factorization simply by observing the cophenetic correlation (Figure 4A) and the RSS (Figure 4B) plots. Nevertheless, the UIK method successfully validated the results of Alexandrov et al [46] and calculated 10 mutational signatures for the simulated data. From the perspective of Frigyesi et al [11], Figure 4B further demonstrates that the actual optimal value of r=10 with the UIK method evaluates the ability of each value of the rank to classify the samples into the same number of classes, which could be smaller than the cophenetic measure (Figure 4A). Despite a decline in the cophenetic correlation coefficient value for r=5, 8, 10, the clusters are stationary and reflected as robust by Brunet et al [9], which produces unmeaningful results that match the actual signatures. Alexandrov et al [46] considered that the biological significance of the 10th cluster, for r=10, is less clear with the cophenetic measure. The sharp decrease in the cophenetic correlation coefficient at r=13 indicates that substantially less stability is achieved using more than 10 clusters. Since this approach does not always provide a clear and consistent cutoff for the choice of r, Alexandrov et al [46] utilized the average silhouette width of the *N* clusters as a measure of reproducibility for the whole solution. Here, the method of UIK estimation with the former results of actual signatures according to Alexandrov et al [46] was validated (see Supplementary Data S3-S4 in Multimedia Appendix 1).

Figure 4. (A) It is complicated to locate the optimal rank with the cophenetic correlation coefficient approach. (B) However, the unit invariant knee (UIK) method can facilitate this decision more quickly and more accurately, which agrees with the number of signatures detected by Alexandrov et al [46]. RSS: residual sum of squares.



Discussion

Principal Results

The novel finding of this study is the ability to apply the UIK method in selecting optimal ranks based on the RSS curve of factorization ranks of the NMF technique. First, this study employed the Golub et al [38] data set and simulated mutational process data [46,47] utilizing the UIK method, which does not require averaging out the results from different runs of the nmf() function [50] or considering the variance between each run.

In the second module, the UIK precisely estimates simulated data with known dimensions. The UIK technique is free of a priori rank parameter input and does not require setting initial parameters that considerably affect the performance. Finally, this method was tested on gene expression data deconvolution, achieving optimal rank estimation.

The proposed uikNMF technique was tested on both experimental gene expression and simulated mutational processes data sets. Moreover, our recent study of utilization of the UIK technique on NMF revealed the genetic links of type 2 diabetes (T2D) that could lead to the development of Alzheimer disease (AD) [51]. The study extracted the most significant genes, or so-called "metagenes," using the elbow method in T2D data, which may be helpful for gaining insight into the mechanism of AD and the development of related therapeutics.

This study further shows that the UIK method provides a credible prediction for gene expression data and precisely estimates simulated data with known dimensions. The proposed UIK method based on the RSS curvature's first inflection point to estimate the optimal rank is theoretically superior or equivalent to existing implementation and software. All the undertaking is done with R programming and is freely available.

As future work, some software functionality ideas include adapting the UIK method on NMF rank estimation in a single function package to accommodate analyses of gene expression, mutational processes, and other biological data sets at the molecular level.

Limitations

The analysis has some limitations such that other NMF packages or software on gene expression research were not tested. This study demonstrates that the UIK method provides a credible prediction for gene expression data. However, it was simply assumed that the same algorithms of NMF are used, as far as the RSS and residual curves would be approximated the same way so that the UIK method would result in the same optimal ranks.

Comparison With Prior Work

One of the arguments related to the choice of rank is to remove noise and recover the signatures [52]. However, when it comes to NMF, the choice of noise is not obvious as the noisy version of the target matrix must be nonnegative as well, which suggests that injected noise may also introduce bias [53]. In addition, the selection of the noise distribution is yet another hyperparameter that is not obvious to select. To handle the noise issue, it is suggested to use gene expression data sets (ie, microarrays) with low-quality reads and genes with a very low number of reads removed before DEGs analysis. The DEGs would then be used as the target matrix for the uikNMF method, as previously demonstrated with T2D gene expression data [51].

Several methods have been developed to select the optimal rank factorization [50]. For example, Brunet et al [9] proposed grabbing the first value of r for which the cophenetic coefficient rate was declining, whereas Frigyesi et al [11] pondered the minimum value at which the decrease in the RSS is lower than the decay of the RSS simulated from random data. The aim of this study was to develop a method for deciding how and which approach performs better on an estimation of the latent factors on given different algorithms of NMF.

Conclusions

This study demonstrates that the elbow method provides a credible prediction for both gene expression data and for precisely estimating simulated mutational processes data with known dimensions. The suggested UIK method is faster than conventional methods with regard to usage of the consensus matrix as a benchmark for rank choice, while achieving considerably better computational adeptness without visual inspection on the curvatives. It is further argued that the suggested rank tuning method based on the elbow method with gene expression data is theoretically superior to the cophenetic measure. Lastly, the proposed method could be applied to other types of gene expression data sets to reveal the most significant genes (so-called "metagenes") in various diseases, including T2D and other metabolic diseases, and may further be helpful for understanding the underlying mechanism of AD and related neurological disorders.

Data Availability

The Golub gene expression [38] and simulated mutational processes [46] data sets are publicly available. The data and related R studio codes supporting the findings of the article are available in Multimedia Appendix 1.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Alexandrov et al [46] simulated mutational signatures data summary (Supplementary Data S1). Implementation of the comparison of Gaujoux estimates of the esGolub subdata set with the unit invariant knee (UIK) method (Supplementary Data S2). The rank survey plot of Alexandrov et al [46] simulated mutational signatures data (Supplementary Data S3). Application of the UIK method on Alexandrov et al [46] simulated mutational signatures data (Supplementary Data S4).
[PDF File (Adobe PDF File), 988 KB - bioinform_v4i1e43665_app1.pdf]

References

1. Laurberg H. Non-negative matrix factorization: theory and methods. PhD thesis. Institut for Elektroniske Systemer, Aalborg University Denmark. 2008. URL: https://vbn.aau.dk/ws/portalfiles/portal/316444854/HLA_thesis.pdf [accessed 2023-05-05]
2. Kameoka H, Ono N, Kashino K, Sagayama S. NMF: A new sparse representation for acoustic signals. 2009 Presented at: IEEE International Conference on Acoustics, Speech and Signal Processing; April 19-24, 2009; Taipei, Taiwan. [doi: 10.1109/icassp.2009.4960364]
3. Cantisani G, Essid S, Richard G. Neuro-steered music source separation with EEG-based auditory attention decoding and contrastive-NMF. 2021 Presented at: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 20211; June 6-11, 2021; Toronto, ON. [doi: 10.1109/icassp39728.2021.9413841]
4. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature 1999 Oct 21;401(6755):788-791. [doi: 10.1038/44565] [Medline: 10548103]
5. Ramanarayanan V, Katsamanis A, Narayanan S. Automatic data-driven learning of articulatory primitives from real-time mri data using convolutive nmf with sparseness constraints. 2011 Presented at: INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association; August 27-31, 2011; Florence, Italy. [doi: 10.21437/interspeech.2011-16]
6. Zhu L, Soldevila F, Moretti C, d'Arco A, Boniface A, Shao X, et al. Large field-of-view non-invasive imaging through scattering layers using fluctuating random illumination. Nat Commun 2022 Mar 18;13(1):1447. [doi: 10.1038/s41467-022-29166-y] [Medline: 35304460]
7. Zhang Y, Du N, Ge L, Jia K, Zhang A. A collective nmf method for detecting protein functional module from multiple data sources. 2012 Presented at: BCB '12: ACM Conference on Bioinformatics, Computational Biology and Biomedicine; October 8-10, 2012; Orlando, Florida. [doi: 10.1145/2382936.2383053]
8. Ye C, Toyoda K, Ohtsuki T. Blind source separation on non-contact heartbeat detection by non-negative matrix factorization algorithms. IEEE Trans Biomed Eng 2020 Feb;67(2):482-494. [doi: 10.1109/tbme.2019.2915762]
9. Brunet J, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. Proc Natl Acad Sci U S A 2004 Mar 23;101(12):4164-4169 [FREE Full text] [doi: 10.1073/pnas.0308531101] [Medline: 15016911]
10. Hutchins L, Murphy S, Singh P, Graber J. Position-dependent motif characterization using non-negative matrix factorization. Bioinformatics 2008 Dec 01;24(23):2684-2690 [FREE Full text] [doi: 10.1093/bioinformatics/btn526] [Medline: 18852176]
11. Frigyesi A, Höglund M. Non-negative matrix factorization for the analysis of complex gene expression data: identification of clinically relevant tumor subtypes. Cancer Inform 2008 May 29;6:275-292 [FREE Full text] [doi: 10.4137/cin.s606] [Medline: 19259414]
12. Pal NR, Aguan K, Sharma A, Amari S. Discovering biomarkers from gene expression data for predicting cancer subgroups using neural networks and relational fuzzy clustering. BMC Bioinformatics 2007 Jan 06;8(1):5 [FREE Full text] [doi: 10.1186/1471-2105-8-5] [Medline: 17207284]
13. Tsai Y, Lin C, Tseng GC, Chung I, Pal NR. Discovery of dominant and dormant genes from expression data using a novel generalization of SNR for multi-class problems. BMC Bioinformatics 2008 Oct 09;9(1):425 [FREE Full text] [doi: 10.1186/1471-2105-9-425] [Medline: 18842155]
14. Akçay S, Güven E, Afzal M, Kazmi I. Non-negative matrix factorization and differential expression analyses identify hub genes linked to progression and prognosis of glioblastoma multiforme. Gene 2022 May 25;824:146395. [doi: 10.1016/j.gene.2022.146395] [Medline: 35283227]
15. Bicciato S, Luchini A, Di Bello C. PCA disjoint models for multiclass cancer analysis using gene expression data. Bioinformatics 2003 Mar 22;19(5):571-578. [doi: 10.1093/bioinformatics/btg051] [Medline: 12651714]
16. Staffini A, Fujita K, Svensson AK, Chung U, Svensson T. Statistical methods for item reduction in a representative lifestyle questionnaire: pilot questionnaire study. Interact J Med Res 2022 Mar 18;11(1):e28692 [FREE Full text] [doi: 10.2196/28692] [Medline: 35302507]
17. Christopoulos D. Developing methods for identifying the inflection point of a convex/concave curve. arXiv. 2012. URL: https://arxiv.org/abs/1206.5478 [accessed 2023-05-05]
18. Inglehart R, Haerpfer C, Moreno A, Welzel C, Kizilova K, Diez-Medrano J. World Values Survey Round Five. Country-Pooled Datafile Version. WVS Database. 2005. URL: https://www.worldvaluessurvey.org/WVSDocumentationWV5.jsp [accessed 2023-05-05]
19. Christopoulos D. Introducing unit invariant knee (UIK) as an objective choice for elbow point in multivariate data analysis techniques. SSRN Journal. 2016. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3043076 [accessed 2023-05-05]

20. Cattell RB. The scree test for the number of factors. Multivariate Behav Res 1966 Apr 01;1(2):245-276. [doi: 10.1207/s15327906mbr0102_10] [Medline: 26828106]
21. Islam SA, Díaz-Gay M, Wu Y, Barnes M, Vangara R, Bergstrom EN, et al. Uncovering novel mutational signatures by extraction with SigProfilerExtractor. Cell Genom 2022 Nov 09;2(11):100179 [FREE Full text] [doi: 10.1016/j.xgen.2022.100179] [Medline: 36388765]
22. Tabak MA, Murray KL, Reed AM, Lombardi JA, Bay KJ. Automated classification of bat echolocation call recordings with artificial intelligence. Ecol Inform 2022 May;68:101526. [doi: 10.1016/j.ecoinf.2021.101526]
23. Saurette DD, Berg AA, Laamrani A, Heck RJ, Gillespie AW, Voroney P, et al. Effects of sample size and covariate resolution on field-scale predictive digital mapping of soil carbon. Geoderma 2022 Nov;425:116054. [doi: 10.1016/j.geoderma.2022.116054]
24. Maisog JM, DeMarco AT, Devarajan K, Young S, Fogel P, Luta G. Assessing methods for evaluating the number of components in non-negative matrix factorization. Mathematics 2021 Nov 02;9(22):2840 [FREE Full text] [doi: 10.3390/math9222840] [Medline: 35694180]
25. Muzzarelli L, Weis S, Eickhoff SB, Patil KR. Rank selection in non-negative matrix factorization: systematic comparison and a new MAD metric. 2019 Presented at: International Joint Conference on Neural Networks (IJCNN); July 14-19, 2019; Budapest, Hungary. [doi: 10.1109/ijcnn.2019.8852146]
26. Gaujoux R, Seoighe C. A flexible R package for nonnegative matrix factorization. BMC Bioinformatics 2010 Jul 02;11(1):367 [FREE Full text] [doi: 10.1186/1471-2105-11-367] [Medline: 20598126]
27. Marutho D, Handaka S, Wijaya E, Muljono. The determination of cluster number at k-mean using elbow method and purity evaluation on headline news. 2018 Presented at: International Seminar On Application For Technology of Information and Communication; September 21-22, 2018; Semarang, Indonesia. [doi: 10.1109/isemantic.2018.8549751]
28. Et-taleby A, Boussetta M, Benslimane M. Faults detection for photovoltaic field based on k-means, elbow, and average silhouette techniques through the segmentation of a thermal image. Int J Photoenergy 2020 Dec 4;2020:6617597. [doi: 10.1155/2020/6617597]
29. Liu Z, Tan V. Rank-one NMF-based initialization for NMF and relative error bounds under a geometric assumption. IEEE Transact Signal Process 2018;65(18):4717-4731. [doi: 10.1109/ita.2018.8503169]
30. Bandyopadhyay S, Thakur SS, Mandal JK. Product recommendation for e-commerce business by applying principal component analysis (PCA) and K-means clustering: benefit for the society. Innovations Syst Softw Eng 2020 Aug 25;17(1):45-52. [doi: 10.1007/s11334-020-00372-5]
31. Moltu C, Stefansen J, Svisdahl M, Veseth M. Negotiating the coresearcher mandate - service users' experiences of doing collaborative research on mental health. Disabil Rehabil 2012;34(19):1608-1616. [doi: 10.3109/09638288.2012.656792] [Medline: 22489612]
32. Vollmer Dahlke D, Fair K, Hong YA, Beaudoin CE, Pulczinski J, Ory MG. Apps seeking theories: results of a study on the use of health behavior change theories in cancer survivorship mobile apps. JMIR Mhealth Uhealth 2015 Mar 27;3(1):e31 [FREE Full text] [doi: 10.2196/mhealth.3861] [Medline: 25830810]
33. Revilla-Martín N, Budinski I, Puig-Montserrat X, Flaquer C, López-Baucells A. Monitoring cave-dwelling bats using remote passive acoustic detectors: a new approach for cave monitoring. Bioacoustics 2020 Sep 17;30(5):527-542. [doi: 10.1080/09524622.2020.1816492]
34. Ulfarsson MO, Solo V. Tuning parameter selection for nonnegative matrix factorization. 2013 Presented at: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing; May 26-31, 2013; Vancouver, BC. [doi: 10.1109/icassp.2013.6638936]
35. Glogoza M, Urbach J, Rosborough TK, Olet S, St Hill CA, Smith CS, et al. Tablet vs. station-based laptop ultrasound devices increases internal medicine resident point-of-care ultrasound performance: a prospective cohort study. Ultrasound J 2020 Apr 16;12(1):18 [FREE Full text] [doi: 10.1186/s13089-020-00165-8] [Medline: 32300979]
36. inflection-package: Finds the inflection point of a curve R package. RDRR. URL: https://rdrr.io/cran/inflection/ [accessed 2023-05-05]
37. Christopoulos DT. Reliable computations of knee point for a curve and introduction of a unit invariant estimation. ResearchGate. 2014. URL: https://www.researchgate.net/publication/268977798_Reliable_computations_of_knee_point_for_a_curve_and_introduction_of_a_unit_invariant_estimation [accessed 2023-05-05]
38. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999 Oct 15;286(5439):531-537. [doi: 10.1126/science.286.5439.531] [Medline: 10521349]
39. Golub ExpressionSet. NMF R Project. URL: https://nmf.r-forge.r-project.org/esGolub.html [accessed 2023-05-05]
40. Ilin A, Raiko T. Practical approaches to principal component analysis in the presence of missing values. J Machine Learn Res 2010;11:1957-2000.
41. Park PJ. Gene expression data and survival analysis. In: Shoemaker JS, Lin SM, editors. Methods of microarray data analysis. Boston, MA: Springer; 2005:21-34.

42. Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang C, Angelo M, et al. Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci U S A 2001 Dec 18;98(26):15149-15154 [FREE Full text] [doi: 10.1073/pnas.211566398] [Medline: 11742071]
43. Friedman N, Kaminski N. Statistical methods for analyzing gene expression data for cancer research. Ernst Schering Res Found Workshop 2002;109(38):109-131. [doi: 10.1007/978-3-662-04747-7_6] [Medline: 12060998]
44. Haferlach T, Kohlmann A, Bacher U, Schnittger S, Haferlach C, Kern W. Gene expression profiling for the diagnosis of acute leukaemia. Br J Cancer 2007 Feb 26;96(4):535-540 [FREE Full text] [doi: 10.1038/sj.bjc.6603495] [Medline: 17146476]
45. Hackstadt AJ, Hess AM. Filtering for increased power for microarray data analysis. BMC Bioinformatics 2009 Jan 08;10(1):11 [FREE Full text] [doi: 10.1186/1471-2105-10-11] [Medline: 19133141]
46. Alexandrov L, Nik-Zainal S, Wedge D, Campbell P, Stratton M. Deciphering signatures of mutational processes operative in human cancer. Cell Rep 2013 Jan 31;3(1):246-259 [FREE Full text] [doi: 10.1016/j.celrep.2012.12.008] [Medline: 23318258]
47. SigProfiler. MathWorks. URL: https://www.mathworks.com/matlabcentral/fileexchange/38724-sigprofiler [accessed 2023-05-05]
48. NMF: algorithms and framework for nonnegative matrix factorization (NMF). CRAN R project. URL: https://cran.r-project.org/web/packages/NMF/ [accessed 2023-05-05]
49. Bengtsson H, Jacobson A, Riedy J. R.matlab: Read and Write MAT Files and Call MATLAB from Within R. CRAN R project. 2018. URL: https://cran.r-project.org/web/packages/R.matlab/index.html [accessed 2023-05-05]
50. Gaujoux R. An introduction to NMF package Version 0. R Project for Statistical Computing. 2014. URL: http://nmf.r-forge.r-project.org/vignettes/NMF-vignette.pdf [accessed 2023-05-05]
51. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 1987 Nov;20:53-65. [doi: 10.1016/0377-0427(87)90125-7]
52. Afzal M, Alharbi KS, Alzarea SI, Alyamani NM, Kazmi I, Güven E. Revealing genetic links of type 2 diabetes that lead to the development of Alzheimer's disease. Heliyon 2023 Jan;9(1):e12202 [FREE Full text] [doi: 10.1016/j.heliyon.2022.e12202] [Medline: 36711310]
53. Lin X, Boutros PC. Optimization and expansion of non-negative matrix factorization. BMC Bioinformatics 2020 Jan 06;21(1):7 [FREE Full text] [doi: 10.1186/s12859-019-3312-5] [Medline: 31906867]

Abbreviations

AD: Alzheimer disease
ALL: acute lymphoblastic leukemia
DEG: differentially expressed gene
FA: factor analysis
MSE: mean squared error
NMF: nonnegative matrix factorization
PCA: principal component analysis
PRESS: predicted residual sum of squares
RSS: residual sum of squares
T2D: type 2 diabetes
UIK: unit invariant knee

Edited by E Uzun; submitted 19.10.22; peer-reviewed by S Özkan, M Banf, A Staffini; comments to author 19.12.22; revised version received 05.02.23; accepted 28.04.23; published 06.06.23.

Please cite as:
Guven E
Decision of the Optimal Rank of a Nonnegative Matrix Factorization Model for Gene Expression Data Sets Utilizing the Unit Invariant Knee Method: Development and Evaluation of the Elbow Method for Rank Selection
JMIR Bioinform Biotech 2023;4:e43665
URL: https://bioinform.jmir.org/2023/1/e43665
doi:10.2196/43665
PMID:

Original Paper

Genomic Insights Into the Evolution and Demographic History of the SARS-CoV-2 Omicron Variant: Population Genomics Approach

Kritika M Garg[1,2], PhD; Vinita Lamba[3,4], MSc; Balaji Chattopadhyay[1,3], PhD

[1]Department of Biology, Ashoka University, Sonipat, India

[2]Centre for Interdisciplinary Archaeological Research, Ashoka University, Sonipat, India

[3]Trivedi School of Biosciences, Ashoka University, Sonipat, India

[4]J William Fulbright College of Arts and Sciences, Department of Biological Sciences, University of Arkansas, Fayetteville, AR, United States

Corresponding Author:
Balaji Chattopadhyay, PhD
Trivedi School of Biosciences
Ashoka University
Rajiv Gandhi Education City
Sonipat, 131029
India
Phone: 91 8073119246
Email: balaji.chattopadhyay@ashoka.edu.in

Abstract

Background: A thorough understanding of the patterns of genetic subdivision in a pathogen can provide crucial information that is necessary to prevent disease spread. For SARS-CoV-2, the availability of millions of genomes makes this task analytically challenging, and traditional methods for understanding genetic subdivision often fail.

Objective: The aim of our study was to use population genomics methods to identify the subtle subdivisions and demographic history of the Omicron variant, in addition to those captured by the Pango lineage.

Methods: We used a combination of an evolutionary network approach and multivariate statistical protocols to understand the subdivision and spread of the Omicron variant. We identified subdivisions within the BA.1 and BA.2 lineages and further identified the mutations associated with each cluster. We further characterized the overall genomic diversity of the Omicron variant and assessed the selection pressure for each of the genetic clusters identified.

Results: We observed concordant results, using two different methods to understand genetic subdivision. The overall pattern of subdivision in the Omicron variant was in broad agreement with the Pango lineage definition. Further, 1 cluster of the BA.1 lineage and 3 clusters of the BA.2 lineage revealed statistically significant signatures of selection or demographic expansion (Tajima's $D<-2$), suggesting the role of microevolutionary processes in the spread of the virus.

Conclusions: We provide an easy framework for assessing the genetic structure and demographic history of SARS-CoV-2, which can be particularly useful for understanding the local history of the virus. We identified important mutations that are advantageous to some lineages of Omicron and aid in the transmission of the virus. This is crucial information for policy makers, as preventive measures can be designed to mitigate further spread based on a holistic understanding of the variability of the virus and the evolutionary processes aiding its spread.

(JMIR Bioinform Biotech 2023;4:e40673) doi:10.2196/40673

KEYWORDS

SARS-CoV-2; Omicron; evolutionary network; population subdivision; genome evolution; COVID-19; microevolution

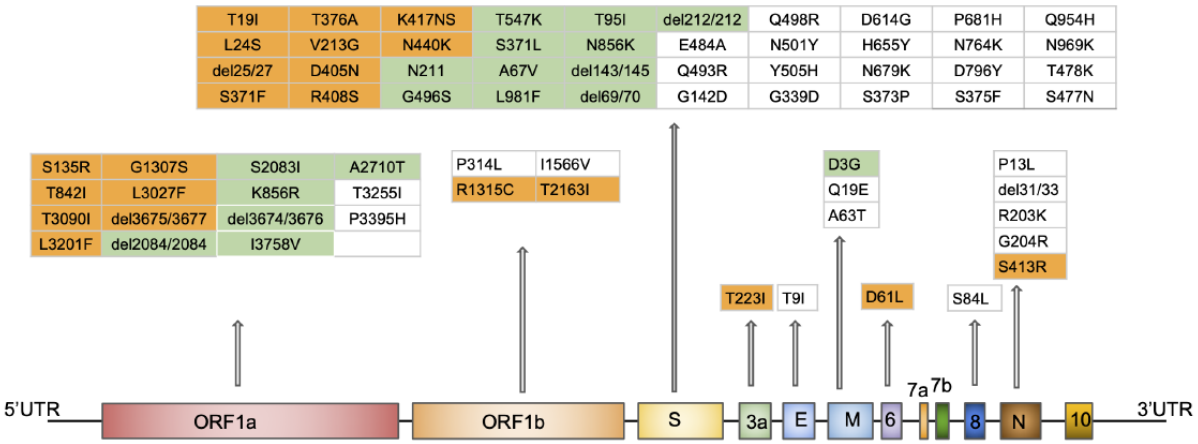
Introduction

The past 2 decades have witnessed multiple zoonotic coronavirus outbreaks, with the latest being the COVID-19 outbreak, which was caused by SARS-CoV-2. The virus emerged in Wuhan, China, and it quickly spread across the globe, resulting in more than 6.5 million deaths [1-3]. SARS-CoV-2 is a *Betacoronavirus* with a positive, single-stranded RNA genome. The genome is approximately 30 kilobases in length and encodes for 26 proteins (16 nonstructural, 4 structural, and 6 accessory proteins; Figure 1) [4,5].

Extensive genomic surveillance programs were established across the globe to monitor the evolution of the virus. This resulted in an exponential increase in the number of SARS-CoV-2 genomes that has in turn presented a unique set of challenges for data analysis [6-9]. With over 10 million genome sequences already available, new algorithms are being designed to tackle the deluge of data [6-9]. Most available analytical tools are designed to identify the overall evolutionary relationship between various lineages. However, obtaining a finer-level understanding of the diversity and subdivision within a lineage can provide important insights into pathogen evolution, particularly during ongoing pandemics [6]. Pango lineage classification is one such nomenclature method for identifying fine-scale, phylogenetically relevant clusters of SARS-CoV-2 based on the mutations in the spike protein [6].

In this study, we used population genomics methods to understand the subdivision of the Omicron lineage of SARS-CoV-2 as it spread across the globe, in an attempt to elucidate the evolutionary history of the variant. The Omicron variant was first identified within Botswana, Southern Africa, in November 2021, and within a short span of time, it emerged as the main variant driving SARS-CoV-2 infections across the globe, replacing the Delta variant [10,11]. The Omicron variant was also of immediate concern due to the large number of mutations observed in its spike protein (Figure 1). Among the 60 mutations that this variant accumulated when compared to the reference Wuhan sequence, the majority were concentrated in the spike protein (38 mutations in the BA.1 lineage and 31 mutations in the BA.2 lineage; Figure 1) [12,13]. Some of these mutations increased both the transmission ability and the antibody escape of the virus, allowing the Omicron variant to rapidly spread across the globe [11,13]. Given the high infection rate and rapid spread of the virus across the globe, alternative methods for inspecting fine-scale subdivision and transmission are necessary to understand the evolution of the virus and devise any strategy to reduce its spread. Thus, we investigated the subdivision and demographic history of the BA.1 and BA.2 lineages of Omicron, along with identifying mutations that are correlated with the spread of the virus.

Figure 1. The genome structure of SARS-CoV-2 with known mutations in the Omicron variant highlighted. Mutations unique to the BA.1 lineage are highlighted in green, and those unique to the BA.2 lineage are highlighted in orange. Mutations common to both lineages of Omicron are in plain black font. The list of mutations was obtained from Tzou et al [12]. ORF: open reading frame; UTR: untranslated region.



Methods

Data Matrix and Cleanup

We downloaded 20,067 SARS-CoV-2 genome sequences belonging to the Omicron lineage, which were available up to January 31, 2022, from the GISAID (Global Initiative on Sharing All Influenza Data) repository (Multimedia Appendices 1 and 2), retaining only high-coverage genomes (<1% undetermined nucleotide bases; <0.05% unique amino acid mutations) and genomes with a collection date for this study. Only sequences obtained from humans were used for all analyses. We retained 20,067 genomes, which were further filtered for quality by using Nextclade CLI (Nextstrain) [14]. Nextclade examines each query sequence for flaws that could suggest sequencing or assembly errors and assigns a score for each sequence. The quality score of a sequence is determined by the number of undetermined bases, ambiguous sites, private mutations, and stop codons. All sequences classified as good-quality sequences by Nextclade were selected for further analysis. We retained 14,002 good-quality sequences, of which most were from Denmark (n=11,272, 80.5%); the rest of the sequences were from 43 countries. We aligned the filtered SARS-CoV-2 genomes to the Wuhan reference genome (accession ID: MN908947.1) in Nextalign CLI [14], using default parameters. Further, we assigned the lineage for each sequence by using the pangolin web server (versions 3.1.20 and 4.0.6; accessed on March 3 and May 6, 2022, respectively) [6].

Genetic Subdivision Analysis

We used two different approaches to understand the population subdivision within our SARS-CoV-2 data set. For the first approach, we reconstructed an evolutionary network by using the program VENAS (Viral Genome Evolution Network Analysis System) [15]. VENAS can analyze thousands of genomes in a short span of time (a few minutes) and is a useful tool for tracking changes across a transmission chain. It identifies mutations across alignments and constructs a network based on hamming distances. In VENAS, we first estimated the effective parsimony-informative sites and minor allele frequency, using default settings, and retained 5253 genomes. These were then used to construct an evolutionary network, which was viewed in Cytoscape 3.9.1 (Cytoscape Consortium)

by using the prefuse force-directed method [16]. Finally, we analyzed the BA.1 (n=260) and BA.2 (n=4993) lineages separately to understand the fine-scale subdivision within each lineage.

For the second approach, we used the discriminant analysis of principal components (DAPC) method to understand the fine-scale subdivision patterns observed in each lineage (based on the Pango lineage definitions mentioned in the *Data Matrix and Cleanup* section). The DAPC is a useful method for detecting subdivision, as it maximizes between-group differences while minimizing within-group variability [17]. It is a relatively fast method for detecting complex subdivision patterns from genomic data. We used the filtered genomes obtained from VENAS and performed a DAPC for both lineages by using the R *adegenet* package (R Foundation for Statistical Computing) [17,18]. We first identified the optimal number of clusters within each data set, using the K-means algorithm, and then performed the DAPC. We further identified the unique mutations for each of the DAPC clusters and only considered mutations that were present in at least 70% of the sequences belonging to the cluster.

Genomic Diversity and Selection Analysis

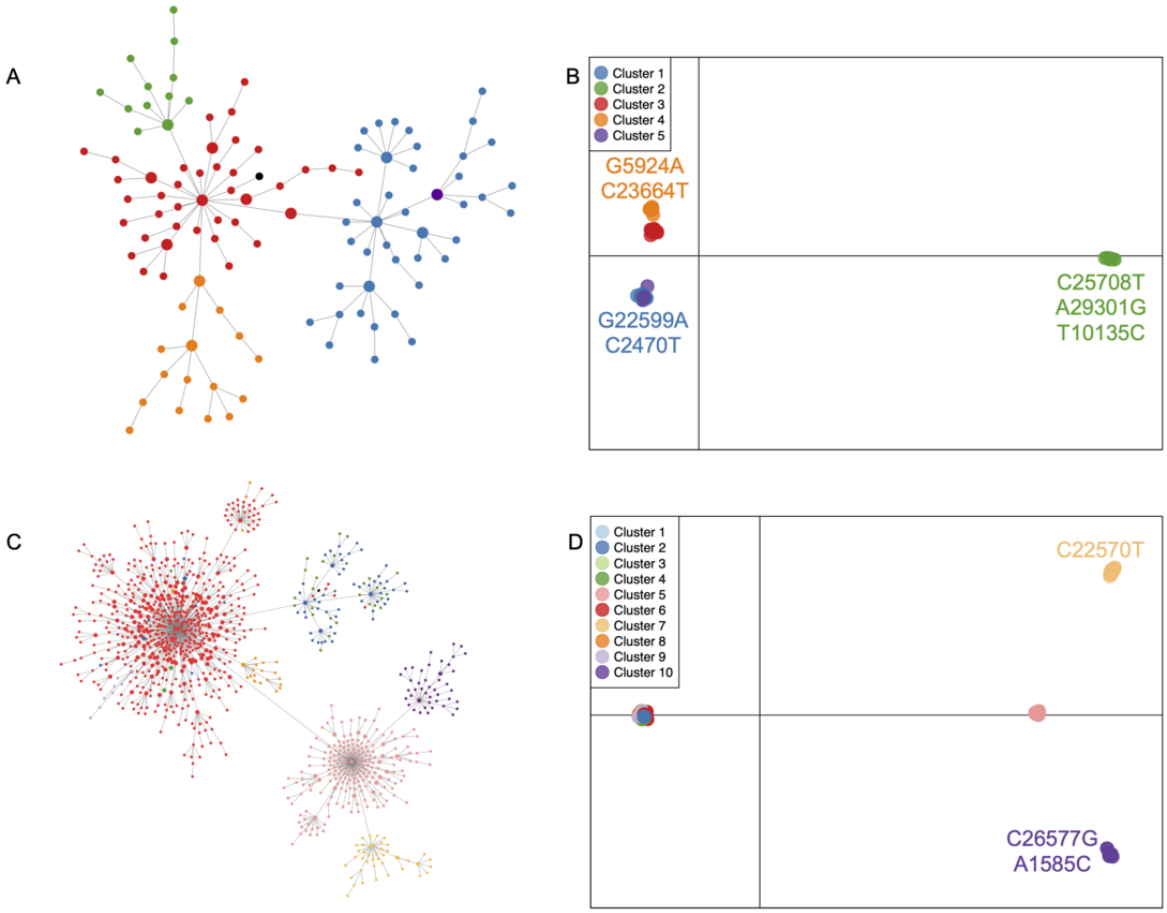
To estimate the level of genomic diversity within our data set, we characterized all substitutions in reference to the Wuhan genome by using VENAS. We considered all 14,003 good-quality sequences and identified the mutations for the 12 functional open reading frames (ORFs). We further estimated Tajima's D values for the spike protein sequences for all of the clusters identified in the DAPC, using MEGA (Molecular Evolutionary Genetics Analysis; version 10.2.6 [Pennsylvania State University]) [19]. The Tajima's D test is widely used to identify signatures of microevolutionary forces, such as population fluctuations and selections acting upon populations. We estimated the Tajima's D value for each genetic cluster identified by the DAPC to avoid confounding effects of population subdivision.

Results

Genetic Subdivision

We observed signatures of genetic subdivision in the BA.1 and BA.2 lineages of the Omicron variant. The patterns were broadly concordant between both approaches—the evolutionary network approach conducted in VENAS and the statistical approach using the DAPC. Although VENAS produced numerous nodes and groups (BA.1: n=111; BA.2: n=1046), these were nested within fewer, broader subdivisions retrieved by the DAPC (Figure 2). We identified 5 clusters for the BA.1 lineage and 10 clusters for the BA.2 lineage, using the DAPC. However, when we visualized our results, we observed that 2 clusters for the BA.1 lineage were clubbed together, and 4 clusters for the BA.2 lineage were clubbed together (Figure 2B and Figure 2D). Further, no sequences were assigned to clusters 1 and 3 for the BA.2 lineage. Thus, effectively, only 4 clusters for the BA.1 and BA.2 lineages were identified by the DAPC method. Private mutations were identified for 5 clusters (Figure 2B and Figure 2D).

Figure 2. The genetic subdivision observed in the Omicron lineage of SARS-CoV-2 based on haplotype networks (panel A: BA.1 lineage; panel C: BA.2 lineage). Panels B and D depict observed genetic subdivision based on the discriminant analysis of principal components (DAPC) for the BA.1 and BA.2 lineages, respectively. The Wuhan sequence is denoted by the black-colored node in panels A and C. Private mutations, if any, are depicted in the DAPC plot. A private mutation must be present in at least 70% of the sequences within the cluster.



Genomic Diversity

A detailed inspection of the pattern of substitution among the study genomes revealed that, as expected, the spike protein, ORF1a, and ORF1b harbored the maximum number of variations (Figure S1 in Multimedia Appendix 1). Gene coding for the envelope protein had the lowest rate of change. The most frequent mutations observed within our panel of genomes were C to T transition and G to T transversion (Figure S2 in Multimedia Appendix 1). Tajima's D values for the spike protein sequences were negative for all of the DAPC clusters, with significant values observed only for 4 clusters (Tajima's D<−2; Table 1).

Table 1. Tajima's D estimates for the various clusters that were identified by using the discriminant analysis of principal components.

Cluster ID	Number of sequences	Sampling location	Number of segregating sites	θ	Nucleotide diversity	Tajima's D
BA.1 lineage						
Cluster 1	41	Belgium, Germany, India, Japan, Mexico, Romania, South Africa, Switzerland, Taiwan, and United States of America	16	0.00293	0.00071	−2.425328[a]
Cluster 2	12	Denmark, Germany, India, Mexico, and United States of America	7	0.00182	0.00095	−1.866946
Cluster 3	40	Denmark, England, Germany, India, Japan, Slovenia, South Africa, Switzerland, Taiwan, Thailand, and United States of America	11	0.00203	0.00073	−1.948315
Cluster 4	16	Denmark, England, Germany, India, Romania, South Africa, Spain, and Switzerland	8	0.00189	0.00105	−1.598913
Cluster 5	1	South Africa	N/A[b]	N/A	N/A	N/A
BA.2 lineage						
Cluster 1	0	N/A	N/A	N/A	N/A	N/A
Cluster 2	61	Australia, Denmark, India, Singapore, and South Africa	8	0.001343	0.00028	−2.08083[a]
Cluster 3	0	N/A	N/A	N/A	N/A	N/A
Cluster 4	31	Denmark, India, Norway, and Singapore	2	0.000393	0.00010	−1.50558
Cluster 5	207	Denmark and Singapore	19	0.002527	0.00032	−2.31247[a]
Cluster 6	632	Denmark, Singapore, and South Africa	50	0.005591	0.00027	−2.59423[a]
Cluster 7	40	Denmark	3	0.000554	0.00019	−1.4309
Cluster 8	22	Denmark	1	0.000215	0.00007	−1.16240
Cluster 9	8	Denmark	1	0.000303	0.00020	−1.05482
Cluster 10	44	Denmark	3	0.000542	0.00027	−1.07839

[a]Tajima's D values of <−2 indicate significant demographic expansion or selection.

[b]N/A: not applicable.

Discussion

Study Overview

In this study, we investigated the effectiveness of population genomics methods to identify the fine-scale structure and demographic history of the Omicron lineage during the initial spread of the virus. Our study also highlights the utility of population genomics methods in handling large data sets and provides an analytical framework for future studies, which will help with understanding the genetic substructuring of the virus and identifying mutations that are potentially advantageous to the spread of the virus.

Fine-Scale Subdivision Within the Omicron Variant

Using a combination of population genetics methods, this study revealed cryptic, fine-scale substructuring within our data set. We observed a similar pattern of subdivision for each Omicron lineage, using both VENAS and the DAPC (Figure 2). Although both methods use different approaches, together they provide a robust understanding of the finer subdivision patterns within fast-evolving lineages. At the start of this study, Pango lineage definition 3.1.20 was available, which had divided the Omicron sequences into the BA.1, BA.1.1, and BA.2 lineages, and our population genomics–based clustering identified finer-level subdivision within these lineages. With the updated Pango lineage version 4.0.6, there is now a broad agreement in the lineage definitions between our methodology (DAPC and VENAS) and the Pango lineage.

We identified cluster-defining mutations that were later selected for Pango lineage definitions, such as the G22599A and G5924A mutations for BA.1.1 (clusters 1 and 5 in the DAPC) and BA.1.17 (cluster 4 in the DAPC), respectively (Figure 2B). The subdivision observed in our study, as well as some of the cluster-defining mutations (G5924A, G22599A, C2470T, and A29301G), also agrees with recent phylogenetic reconstructions (Figure 2B) [20,21].

We also recovered signatures of fine-scale subdivision within the updated Pango (version 4.0.6) definitions. For example, DAPC clusters 5, 7, and 10, which are all part of the BA.2.9 lineage from Denmark (Figure 2D), were segregated based on 3 unique mutations (Figure 2D). The mutation C22570T is unique to cluster 7, and the mutations C26577G and A1585C are unique to cluster 10 within our data set (Figure 2D). Thus,

our analytical regime could not only retrieve cluster-defining mutations in agreement with other methods but also identify finer subdivisions within existing Pango definitions and associated unique mutations.

Selection and Demographic History of the Omicron Variant

Tests for selection revealed that the evolution of the Omicron lineage could be attributed to microevolutionary processes, such as selection and demographic expansion. We used Tajima's D values to test for deviation of the identified clusters from neutrality. A negative Tajima's D value reflects a deficit of haplotypes in comparison to the number of segregating sites and is indicative of a recent selective sweep or a population expansion [22]. Significant negative Tajima's D values were observed for a subset of the DAPC clusters (BA.1 lineage: cluster 1; BA.2 lineage: clusters 2, 5, and 6; Table 1), suggesting that these clusters have undergone rapid expansion, experienced recent selective sweeps, or both. These attributes are indicators of greater transmissibility and thus make these clusters potential targets for surveillance and monitoring programs. For example, the spike protein mutation G22599A (S:R346K) is implicated in providing a transmission advantage and the antibody escape ability to the BA.1.1 variant. Although the population genomics framework adopted in our study identified this diagnostic mutation, which defines cluster 1 of the BA.1 lineage, the test for deviations from neutrality returned a significant negative value only for this cluster (Tajima's D=−2.425328), indicating the selective advantage, as well as signals of population expansion, of this cluster across the globe (Figure 2B) [23-25]. In addition to G22599A (S:R346K), we also identified the mutation C23664T (S:A701V), which, in conjunction with S:N501Y, provides a mild advantage to the virus by increasing the rate of infection [20]. However, the C23664T mutation is not unique to the Omicron lineage and is also observed in other SARS-CoV-2 variants of concern [20].

Interestingly, cluster 2 of the BA.1 lineage, which did not exhibit a signature of expansion or selection, also harbors 3 unique mutations (C25708T, A29301G, and T10135C), which have been independently identified as suppressor mutations associated with a reduction in the spread of the virus [26] (Figure 2A and Figure 2B; Table 1).

Future research efforts can use a similar analytical framework to swiftly identify mutations that are important for virus evolution, of which some might play an important role in facilitating the spread of a virus, while others may be detrimental to its transmission. We demonstrated that a combination of population genomics methods can be used to recover subtle variations within established lineage definitions and potentially aid in finding variants of concern. The identification of such target mutations is necessary from an epidemiological standpoint, as well as for vaccine development. This study provides an easy analytical framework that can be used by policy makers to identify variants of potential concern and understand the local demographic history and spread of a virus, thereby facilitating disease mitigation.

Conclusion

We provide an easy, computationally tractable framework for understanding the genetic subdivision and demographic history of SARS-CoV-2. Our framework can be quickly implemented to identify potentially important mutations that may be driving the spread of the virus. Such information can be very useful for deciphering the pattern of movement of variants and determining correlations with the local history of an outbreak.

Acknowledgments

BC acknowledges the startup funding from Trivedi School of Biosciences (TSB), Ashoka University, India. KMG acknowledges the support from the Department of Biotechnology-Ramalingaswami Fellowship (grant BT/HRD/35/02/2006). VL was supported by a TSB fellowship.

We gratefully acknowledge the authors from the originating laboratories, which were responsible for obtaining the specimens, and the submitting laboratories, where genetic sequence data were generated and shared via the GISAID (Global Initiative on Sharing All Influenza Data) Initiative, on which this research is based. A full acknowledgment table can be found in Multimedia Appendix 2.

Data Availability

The SARS-CoV-2 sequences analyzed during this study are available in the GISAID (Global Initiative on Sharing All Influenza Data) repository. Details of the sequences are available in Multimedia Appendix 2.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supporting information.
[DOCX File , 59 KB - bioinform_v4i1e40673_app1.docx]

Multimedia Appendix 2

Acknowledgment table.
[PDF File , 3198 KB - bioinform_v4i1e40673_app2.pdf]

References

1. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, et al. Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. N Engl J Med 2020 Mar 26;382(13):1199-1207 [FREE Full text] [doi: 10.1056/NEJMoa2001316] [Medline: 31995857]
2. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 2020 Mar;579(7798):270-273 [FREE Full text] [doi: 10.1038/s41586-020-2012-7] [Medline: 32015507]
3. World Health Organization. 2nd Global consultation on assessing the impact of SARS-CoV-2 variants of concern on public health interventions. World Health Organization. 2021 Jun 10. URL: https://www.who.int/publications/m/item/2nd-global-consultation-on-assessing-the-impact-of-sars-cov-2-variants-of-concern-on-public-health-interventions [accessed 2023-05-09]
4. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet 2020 Feb 22;395(10224):565-574 [FREE Full text] [doi: 10.1016/S0140-6736(20)30251-8] [Medline: 32007145]
5. Brant AC, Tian W, Majerciak V, Yang W, Zheng ZM. SARS-CoV-2: from its discovery to genome structure, transcription, and replication. Cell Biosci 2021 Jul 19;11(1):136 [FREE Full text] [doi: 10.1186/s13578-021-00643-z] [Medline: 34281608]
6. O'Toole Á, Scher E, Underwood A, Jackson B, Hill V, McCrone JT, et al. Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. Virus Evol 2021 Jul 30;7(2):veab064 [FREE Full text] [doi: 10.1093/ve/veab064] [Medline: 34527285]
7. McBroome J, Thornlow B, Hinrichs AS, Kramer A, De Maio N, Goldman N, et al. A daily-updated database and tools for comprehensive SARS-CoV-2 mutation-annotated trees. Mol Biol Evol 2021 Dec 09;38(12):5819-5824 [FREE Full text] [doi: 10.1093/molbev/msab264] [Medline: 34469548]
8. Chen C, Nadeau S, Yared M, Voinov P, Xie N, Roemer C, et al. CoV-Spectrum: analysis of globally shared SARS-CoV-2 data to identify and characterize new variants. Bioinformatics 2022 Mar 04;38(6):1735-1737 [FREE Full text] [doi: 10.1093/bioinformatics/btab856] [Medline: 34954792]
9. Sokhansanj BA, Rosen GL. Mapping data to deep understanding: Making the most of the deluge of SARS-CoV-2 genome sequences. mSystems 2022 Apr 26;7(2):e0003522 [FREE Full text] [doi: 10.1128/msystems.00035-22] [Medline: 35311562]
10. Paton RS, Overton CE, Ward T. The rapid replacement of the SARS-CoV-2 Delta variant by Omicron (B.1.1.529) in England. Sci Transl Med 2022 Jul 06;14(652):eabo5395 [FREE Full text] [doi: 10.1126/scitranslmed.abo5395] [Medline: 35503007]
11. Tian D, Sun Y, Xu H, Ye Q. The emergence and epidemic characteristics of the highly mutated SARS-CoV-2 Omicron variant. J Med Virol 2022 Jun;94(6):2376-2383 [FREE Full text] [doi: 10.1002/jmv.27643] [Medline: 35118687]
12. Tzou PL, Tao K, Pond SLK, Shafer RW. Coronavirus Resistance Database (CoV-RDB): SARS-CoV-2 susceptibility to monoclonal antibodies, convalescent plasma, and plasma from vaccinated persons. PLoS One 2022 Mar 09;17(3):e0261045 [FREE Full text] [doi: 10.1371/journal.pone.0261045] [Medline: 35263335]
13. Fan Y, Li X, Zhang L, Wan S, Zhang L, Zhou F. SARS-CoV-2 Omicron variant: recent progress and future perspectives. Signal Transduct Target Ther 2022 Apr 28;7(1):141 [FREE Full text] [doi: 10.1038/s41392-022-00997-x] [Medline: 35484110]
14. Aksamentov I, Roemer C, Hodcroft EB, Neher RA. Nextclade: clade assignment, mutation calling and quality control for viral genomes. J Open Source Softw 2021 Nov 30;6(67):3773 [FREE Full text] [doi: 10.21105/joss.03773]
15. Ling Y, Cao R, Qian J, Li J, Zhou H, Yuan L, et al. An interactive viral genome evolution network analysis system enabling rapid large-scale molecular tracing of SARS-CoV-2. Sci Bull (Beijing) 2022 Apr 15;67(7):665-669 [FREE Full text] [doi: 10.1016/j.scib.2022.01.001] [Medline: 35036033]
16. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003 Nov;13(11):2498-2504 [FREE Full text] [doi: 10.1101/gr.1239303] [Medline: 14597658]
17. Jombart T. adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 2008 Jun 01;24(11):1403-1405. [doi: 10.1093/bioinformatics/btn129] [Medline: 18397895]
18. R: The R Project for Statistical Computing. R Foundation for Statistical Computing. URL: https://www.R-project.org/ [accessed 2023-05-09]
19. Kumar S, Tamura K, Nei M. MEGA: Molecular Evolutionary Genetics Analysis software for microcomputers. Comput Appl Biosci 1994 Apr;10(2):189-191. [doi: 10.1093/bioinformatics/10.2.189] [Medline: 8019868]
20. Montaño RZ, Culasso ACA, Fernández F, Marquez N, Debat H, Salmerón M, et al. Evolution of SARS-CoV-2 during the first year of the COVID-19 pandemic in Northwestern Argentina. Virus Res 2022 Sep 28;323:198936 [FREE Full text] [doi: 10.1016/j.virusres.2022.198936] [Medline: 36181975]

21. Liu D, Cheng Y, Zhou H, Wang L, Fiel RH, Gruenstein Y, et al. Early introduction and community transmission of SARS-CoV-2 Omicron variant, New York, New York, USA. Emerg Infect Dis 2023 Feb;29(2):371-380 [FREE Full text] [doi: 10.3201/eid2902.220817] [Medline: 36692451]
22. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 1989 Nov;123(3):585-595 [FREE Full text] [doi: 10.1093/genetics/123.3.585] [Medline: 2513255]
23. Neher RA. Contributions of adaptation and purifying selection to SARS-CoV-2 evolution. Virus Evol 2022 Dec 10;8(2):veac113 [FREE Full text] [doi: 10.1093/ve/veac113]
24. Nutalai R, Zhou D, Tuekprakhon A, Ginn HM, Supasa P, Liu C, OPTIC consortium, ISARIC4C consortium, et al. Potent cross-reactive antibodies following Omicron breakthrough in vaccinees. Cell 2022 Jun 09;185(12):2116-2131.e18 [FREE Full text] [doi: 10.1016/j.cell.2022.05.014] [Medline: 35662412]
25. Zaman K, Shete AM, Mishra SK, Kumar A, Reddy MM, Sahay RR, et al. Omicron BA.2 lineage predominance in severe acute respiratory syndrome coronavirus 2 positive cases during the third wave in North India. Front Med (Lausanne) 2022 Nov 02;9:955930 [FREE Full text] [doi: 10.3389/fmed.2022.955930] [Medline: 36405589]
26. Yang HC, Wang JH, Yang CT, Lin YC, Hsieh HN, Chen PW, et al. Subtyping of major SARS-CoV-2 variants reveals different transmission dynamics based on 10 million genomes. PNAS Nexus 2022 Sep 01;1(4):pgac181 [FREE Full text] [doi: 10.1093/pnasnexus/pgac181] [Medline: 36714842]

Abbreviations

DAPC: discriminant analysis of principal components
GISAID: Global Initiative on Sharing All Influenza Data
MEGA: Molecular Evolutionary Genetics Analysis
ORF: open reading frame
VENAS: Viral Genome Evolution Network Analysis System

Edited by A Uzun; submitted 03.07.22; peer-reviewed by A Krishnan, A Alwin Prem Anand, S Sankar; comments to author 06.02.23; revised version received 07.03.23; accepted 05.05.23; published 12.06.23.

Please cite as:
Garg KM, Lamba V, Chattopadhyay B
Genomic Insights Into the Evolution and Demographic History of the SARS-CoV-2 Omicron Variant: Population Genomics Approach
JMIR Bioinform Biotech 2023;4:e40673
URL: https://bioinform.jmir.org/2023/1/e40673
doi:10.2196/40673
PMID:37456139

Original Paper

Secure Comparisons of Single Nucleotide Polymorphisms Using Secure Multiparty Computation: Method Development

Andrew Woods[1,2], BSc; Skyler T Kramer[2,3], BSc; Dong Xu[1,2,3], PhD; Wei Jiang[1], PhD

[1]Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, United States

[2]Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO, United States

[3]Institute for Data Science and Informatics, University of Missouri, Columbia, MO, United States

Corresponding Author:
Dong Xu, PhD
Department of Electrical Engineering and Computer Science
University of Missouri
227 Naka Hall
Columbia, MO, 65211-0001
United States
Phone: 1 5738822299
Email: xudong@missouri.edu

Abstract

Background: While genomic variations can provide valuable information for health care and ancestry, the privacy of individual genomic data must be protected. Thus, a secure environment is desirable for a human DNA database such that the total data are queryable but not directly accessible to involved parties (eg, data hosts and hospitals) and that the query results are learned only by the user or authorized party.

Objective: In this study, we provide efficient and secure computations on panels of single nucleotide polymorphisms (SNPs) from genomic sequences as computed under the following set operations: union, intersection, set difference, and symmetric difference.

Methods: Using these operations, we can compute similarity metrics, such as the Jaccard similarity, which could allow querying a DNA database to find the same person and genetic relatives securely. We analyzed various security paradigms and show metrics for the protocols under several security assumptions, such as semihonest, malicious with honest majority, and malicious with a malicious majority.

Results: We show that our methods can be used practically on realistically sized data. Specifically, we can compute the Jaccard similarity of two genomes when considering sets of SNPs, each with 400,000 SNPs, in 2.16 seconds with the assumption of a malicious adversary in an honest majority and 0.36 seconds under a semihonest model.

Conclusions: Our methods may help adopt trusted environments for hosting individual genomic data with end-to-end data security.

(JMIR Bioinform Biotech 2023;4:e44700) doi:10.2196/44700

KEYWORDS

secure multiparty computation; single nucleotide polymorphism; Variant Call Format; Jaccard similarity

Introduction

Background

With the dramatic decrease in sequencing costs and increase in consumer sequencing organizations, there is no shortage of genomics data. In fact, about 38 million people worldwide had taken a direct consumer genetics test from organizations like 23andMe, Ancestry, and Family Tree DNA by 2021 [1].

The genome is valuable for identifying health risks, predicting drug response, and revealing susceptibility to environmental factors. Genomic data are the foundation of personalized medicine. However, there are many privacy risks involved with access to genomic data. For example, health insurance companies can gain access to this data through genomic databanks via financial compensation, then deny coverage to a potential customer based on their genetic health risks. While the Genetic Information Nondiscrimination Act protects against

such discriminatory acts, detecting and enforcing such a law requires effort that could be mitigated if the company never gets the data in the first place. Traditional cryptographic methods may protect against data leakage, but they often prevent legitimate data queries for research and medical purposes. In today's ever-growing reliance on such data and collaboration, the need for secure computation on genomic data is rapidly growing. Thus, we propose methods in the realm of secure computation, which allows computation on private or sensitive data. We make use of secure computational methods to do secure similarity measures on genetic data without revealing anything about the data itself.

Genetic data is massive, so there are many different methods of expressing that data depending on the application the data is used for. We focus our attention on variants, which are often reported in the Variant Call Format (VCF) file [2] and can refer to substitutions, insertions, and deletions (indels); copy number variations; and others. These variants correlate with relationships between individuals and can be used to identify such relationships. The correlation carries over to properly chosen subsets of these variations, which we refer to as panels. Using these panels, we can perform operations on smaller sets with a fixed and reduced size to compute the similarity between two individuals more efficiently.

Related Work

There are many other works that discuss secure similarity comparisons of genetic data, typically in the context of approximating edit distance [3-6]. Aziz et al [3] used shingle set intersection as an alternative method of the similarity metric. However, edit distance is not the only method of comparison and lacks much information about the associated genetic material. For example, edit distance is positionally agnostic. Further, computing the edit distance takes O(mn) time for two strings of length m and n. Some other methods discussed performing set operations on variants [7], which aim to answer a different question than our approach. Our approach targets how similar two individuals are without identifying any genes or variants while they aim to identify the specific genes in the sets to help look for causal genes in diseases.

Some other works used variation sets to compare two genomes [6,8], but these works have some limitations. For example, the methods by Çetin et al [8] allow a data owner to outsource their data and securely query the data so that the server learns nothing about the VCF data. The specific operation asks if a small set of variants exists in some VCF data. The act of executing the query provides too much information about the stored VCF data and cannot be used for queries from outsiders. Zhu et al [6] focused on looking at small edit sets (VCFs) from shorter sequences and did not look at whole genome similarity, which is what our approach aims to look at. The method by Mahdi et al [4] aimed to securely compute the Hamming distance to search for the most similar sequences in a database using prefix tree queries. However, the method used a trusted party to encrypt the data and distributes decryption keys to researchers, but the expectation of the existence of a trusted party is not practical.

General Approach

In our scheme, we make use of secure multiparty computational techniques to allow secret sharing of filtered variation data. These techniques are built to allow a set of parties to compute on joint data without revealing anything about the input other than what can be derived from the output. Such methods typically use secret sharing schemes or homomorphic encryption schemes, which permit users to perform computations on the encrypted data without first decrypting it.

We consider two entities, each with an individual's genetic data to compare; both parties then agree on a set of important genetic variants through a public panel and ordering of those variants, and encode their genetic data into a binary vector based on the presence of the selected variants in their VCF files. The owners of the binary vectors then secretly share each element in the vector. Each sharing will consist of pieces, and each piece will be given to one of the computing servers. The servers then compute our protocol on the input shares to produce and send output shares to the user, who then reconstructs the shares to get the result.

In this problem domain, the participating parties (ie, individuals or organizations) can be classified into three categories: (1) database owners, (2) users, and (3) service providers (SPs).

- Database owners: These entities could be any health care organization, ancestry company, or genetic profiling company. They provide services for others to query their DNA data but do not want to expose them.
- Users: The users are parties who wish to make queries in the database. They may be individual patients, medical doctors looking for treatment for their patients by means of a similar patient query, other data owners, or researchers from universities.
- Service providers: The SPs can be any organization that offers computational infrastructures, such as Amazon (Amazon Web Service [AWS]), Microsoft (Azure), and Google (Google Cloud Platform [GCP]).

Note that the roles are not mutually exclusive, and a single party may take on multiple roles without compromising the system's security. For example, a database owner could play the role of one of the SPs and take part in the computation.

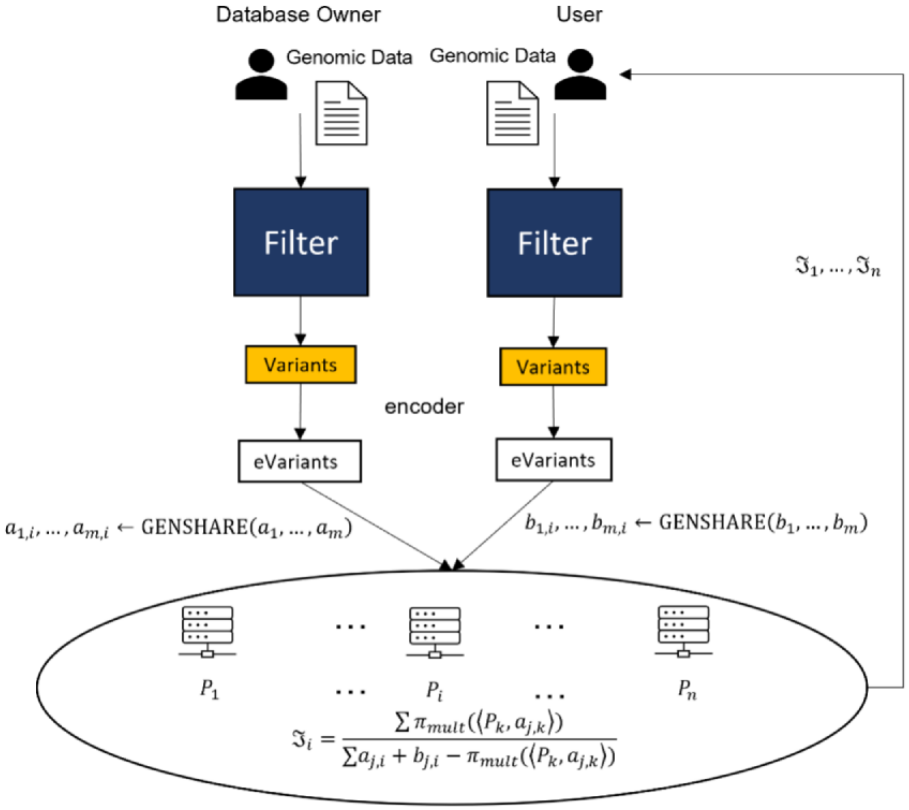
The goal is to provide secure methods to compute set operations on the two input sets of variants, as shown in Figure 1. We include methods to compute the union, intersection, set difference, symmetric difference, and Jaccard similarity on filtered sets of VCFs. We use these methods, for example, to compute the Jaccard similarity between two sets of variants. The protocol is broken up into three phases, the input phase, the computation phase, and the output phase.

- Computation phase: The SPs will perform the specified set operations on the sets of single nucleotide polymorphism (SNPs; or other types of variants) retrieving the sizes of the sets. These sizes can then be used to compute similarity metrics between the two input sequences. These similarity metrics can then be used according to the specific application, such as finding the top *k* matches.

- Output phase: The SPs receive the shares of the output. After this, the SPs can send their final shares to the user so that the user can reconstruct the result.

Figure 1 illustrates the general process when the parties are interested in the Jaccard similarity. The final values J_1, J_2, J_3 held by the servers are seemingly random and, thus, do not individually give any information of the actual Jaccard similarity J. However, the full set of values can be used to reconstruct J. Precisely, we have $J \leftarrow$ RECONSTRUCT ($\langle P_1,J_1\rangle,\langle P_2,J_2\rangle,\langle P_3,J_3\rangle$). The user can learn J by having each server P_i send their share J_i, then using the RECONSTRUCT function.

Figure 1. The database owner and the user filter their genomic data for variants and encode them into a binary string (eVariants) of the same length as the selected panel used for comparisons. Elements of the string are 1 if the implicated variant is present or 0 otherwise. They are then secretly shared with the service providers for secure computation. The service providers do not learn anything about the variants because they only receive secret shares of the encoded variants and none of the encoded variants themselves. Next, each player (Pi) computes the secret share (Ji) by using the secret shares received. Each server (Pi) then sends (Ji) to the user. The user then applies these shares to reconstruct the Jaccard similarity (J←RECONSTRUCT(Pi,Ji).



Contribution

We propose a method to securely compute the Jaccard similarity over two individuals' filtered set of genetic variants. We use the Multi-Party Secure, Privacy-Preserving, and Decentralized Zeus (MP-SPDZ) framework [9] to test the run time and communication costs of our approach. We also tested our approach on a few different popular secure multiparty computation paradigms considering different adversarial models. We then show that our approach provides useful information about the data when the filtered set is chosen properly. We make use of a highly informative but small SNP panel [10] with 4763 SNPs and VCFs from the Genome in a Bottle (GIAB) data set to show that there exists an SNP panel that can be used to identify familial relationships as an example application of our approach. These results are compared with the genomic comparison software BEDTools [11].

Methods

Preliminaries

In this section, we give an overview of the technical background that is needed for our protocols. We start by describing secure multiparty computation—the foundation that enables the execution of our protocols. We then introduce secret sharing and the two types used in the protocols behind the primitive operations in our protocols. We will then discuss the MP-SPDZ framework and how it helped us test the resource requirements behind our protocols.

Threat Model

Secure computation aims to provide guarantees on the privacy of data and correctness of computation. Any situation that potentially compromises these two guarantees is considered a threat. We will assume that the database owner and the user are both semihonest [12]. This means that the database owner and the user will follow the protocol as prescribed. However, we will consider both honest-but-curious and malicious adversarial models for the computing servers. The protocols we run from

the MP-SPDZ framework are based on those assumptions. The computational costs of the protocols are correlated with the assumptions of the behavior of the adversary. For example, if the adversary may do something not prescribed in the protocol, we must use extra measures to ensure that behavior does not cause any compromises to the security guarantees such as correctness or privacy.

Secure Multiparty Computation

Secure multiparty computation allows a set of *n* parties to compute a function over their private inputs without revealing anything about the input other than what can be derived from the output. There are many works providing methods for secure multiparty computation [13-20]. We adapted the methods provided by the MP-SPDZ framework [9] and its protocols. The protocols we used are Semihonest Oblivious Transfer (Semi-OT), MASCOT, Shamir, and Malicious Shamir (Mal-Shamir). Here, OT in Semi-OT and MASCOT both stand for oblivious transfer used to compute multiplication under additive sharing [19,21].

We give a more detailed description of the methods implemented in MP-SPDZ that are used to execute our protocols.

- Semihonest Shamir (Shamir): A semihonest protocol based on the Shamir sharing scheme [22]. This protocol requires at least three computing parties to be used. It is the semihonest equivalent of the honest-majority maliciously secure protocol Mal-Shamir. This protocol is referred to as "Shamir" in the MP-SPDZ framework.
- Mal-Shamir: A maliciously secure execution of the Shamir protocol. This protocol also requires at least three computing parties and can tolerate a minority of players deviating from the protocol and preserve correctness and privacy. This protocol is referred to as "Mal-Shamir" in the MP-SPDZ framework.
- Semi-OT: A semihonest equivalent of the MASCOT protocol, multiplication is based on the OT protocol. This protocol can be executed with as few as two computing parties. This protocol is referred to as "Semi" in the MP-SPDZ framework.
- MASCOT: A maliciously secure malicious majority protocol. This protocol makes use of oblivious transfer to compute multiplications and can also be executed with as few as two computing parties [19]. This protocol is referred to as "MASCOT" in the MP-SPDZ framework.

These protocols are specifications on how to perform addition and multiplication in a secure setting. We make use of these specifications so that our protocol can be executed securely. All these specifications rely on a notion called secret sharing.

Secret sharing forms the foundation of many secure multiparty computation protocols. A secret sharing is an encoding such that a single element (referred to here as a secret) is used to generate an array via a stochastic function (GENSHARE in 1). In a situation where shares should be turned back into the corresponding value, we use a deterministic function known as RECONSTRUCT.

Data

We tested our method on a small data set called GIAB as well as a simulated data set that we used to collect larger unspecified use panels that have no specified use only for measuring the scale of complexity growth as the panel size increases.

The simulated portion of the data was simulated according to the method described by Yue and Liti [23] with the GRCh38 Genome Reference Consortium Human Reference 38 from the University of California, Santa Cruz (UCSC) [24] as our reference genome. Specifically, simuG was used to generate a simulated SNP panel of 1 million SNPs and 2 simulated human VCF files with 5 million SNPs in each. The top "k" SNPs from the simulated SNP panel were pulled to test the impact of panel size on the protocol. Importantly, the simulated VCF files were simulated such that they only contained SNPs and no other variants.

In the VCF format, a single SNP element is a tuple containing the chromosome, position, reference allele, and alternate allele—labeled as "Chr," "Pos," "Ref," and "Alt" in the VCF file, respectively. The chromosome represents the chromosome on which the SNP is located, the position gives a direct location on the chromosome, the reference allele is the standard base to be compared within this location, and the alternate allele is the nonreference base found in this location for a specific genome. Thus, standard VCFs will only have entries for observed variants, not all possible variants. Other information is also provided in the file, but it is not important for our method.

We also tested our method on the GIAB data set [25]. This data set consists of two families, an Ashkenazim family and a Chinese family. Both families have a mother, a father, and a son. The Ashkenazim data set's IDs are HG002, HG003, and HG004 for the son, father, and mother, respectively. The IDs of the Chinese family are HG005, HG006, and HG007 for the son, father, and mother, respectively. To use these VCFs in our method, we made use of an SNP panel [10] with 4763 SNPs.

Secure Set Operations

In this section, we provide an overview of the secure operations that would be executed to compare two VCFs with SNP panels. Using the variant panel, we created an *m* bit string where each position of the bit string contains a "1" if the implicated variant is in the genome (as indicated by the VCF data) or a "0" otherwise. We can then embed the string into the field *F* to perform a computation that allows us to securely perform set operations based on the array. We are interested in the number of elements in a set from set operations. Though our protocols calculate the size of the sets produced by the set operations, they can be easily modified to produce the set itself by skipping the last step.

The protocols we describe in this section use subprotocols implemented by MP-SPDZ, and we denote these subprotocols with π. For example, π_{mult} multiplies two shared values together. These operations require communication between the parties; thus we use this symbol rather than a simple multiplication symbol. However, affine operations such as addition and multiplication by a constant such as 2 do not require

communication and are not written with a subprotocol symbol π.

We converted the two sets of genomic variants, *A*, *B* into two *m*-dimensional vectors , . Using these vectors, we can perform elementwise operations on the array to execute set operations and compute any similarity metric that relies on those set operations, such as Jaccard similarity. Here, we provide a formal description of the protocols.

Specifically, let the tuple of the genomic variants be (Chr, Pos, Ref, Alt). To generate the two *m*-dimensional vectors, we select a public panel of variants *S* to filter off and order to. The vector = (a_1, ..., a_n) is defined to be a_i = 1 if the *i*-th SNP ($Chri$, $Posi$, Ref_i, Alt_i) ∈ $S \cap A$; otherwise a_i = 0.

After computing step 1 in UNION, = 1 implies that at least = 1 or = 1. This means the sum of the elements of the vector is equal to the size of the union of the two sets *A* and *B*. Thus, in step 2, UNION computes the sharing of the size of the union.

After step 1 in INTERSECT, = 1 implies that both = 1 or = 1. Thus the sum of the elements in is equal to the size of the intersection of the two sets *A* and *B*. Thus on step 2, the protocol computes a sharing of the size of the intersection of the two sets.

After step 1 in DIFFERENCE, = 1 implies that = 1 and = 0. This corresponds to being the set difference, so computing the sum of in step 2 gives us a sharing of the size of the set difference $A \setminus B$.

After step 1 in SYMDIFFERENCE, = 1 implies that either = 1 or = 1, but not both. Thus, computing the sum in step 2 gives us the size of the symmetric set difference of *A* and *B*.

Jaccard similarity is the ratio of the size of the intersection to the union of two sets. Since $A \cap B \subseteq A \cup B$, we have $|A \cap B| \leq |A \cup B|$ and that $J \in [0, 1]$. Here, we use our previous protocols to first compute the sizes of the union and intersection, and we use the results to compute the Jaccard similarity.

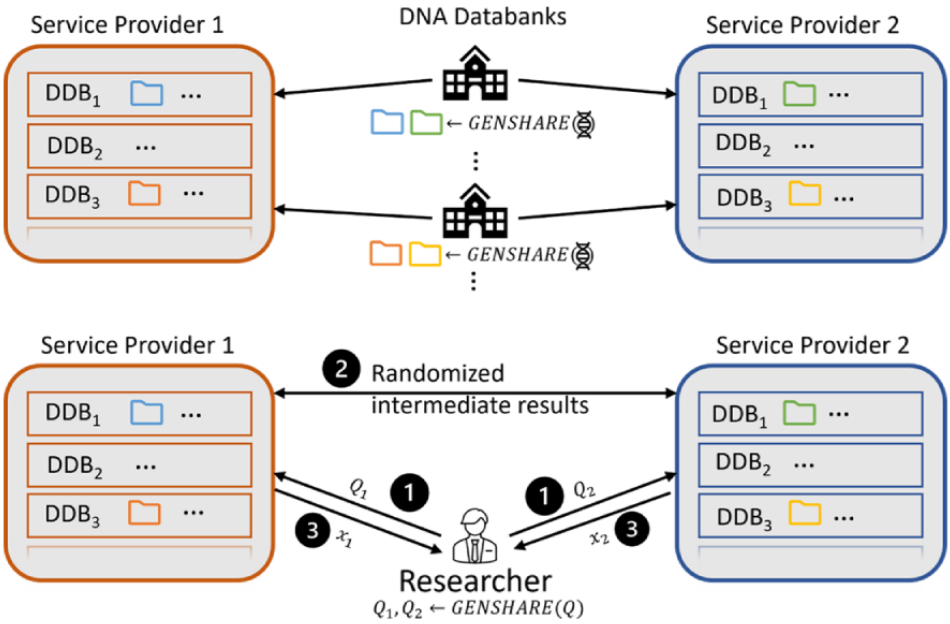
Network Setup

We had two different scenarios tested in our network setup, as shown in Figure 2. The first of which was a virtual network constructed for quick easy testing over a local area network. We then made use of CloudLab to host a real network. Using the virtual network, we ran our protocol on simulated data, and using the CloudLab network, we ran our methods on real-life data samples.

The virtual network was constructed using VirtualBox in the Ubuntu environment. We set up three virtual machines with access to the network and required the machines to talk on the network to communicate. The behavior of this network is consistent with that of a LAN. Using the three machines, we established a network and ran the protocols. Two machines provided input by secret sharing their data according to the protocol. The third machine helped to facilitate the Shamir and Mal-Shamir protocols. The third machine was not used during the Semi-OT and MASCOT executions.

For the CloudLab servers, we had three servers that could communicate with each other and provide computation. We used two of the servers as the input providers, and the third server helped to facilitate the realization of the Shamir and Mal-Shamir (three party) protocols. These servers better simulate a real-world computing environment.

Figure 2. Network setup. The institutions first share their data (top) with the two service providers. The researcher can then query the database (bottom) and get the information they require for their research.



MP-SPDZ Setup

We used MP-SPDZ [9] for the implementation of the secure protocols. The MP-SPDZ documentation provides a comprehensive guide in setting up the experiments. In our setup, we had three servers P_0, P_1, P_2. The servers P_0, P_1 were treated as users of the system, and they provided the input to the computation. This is because the two servers have their own privacy and security of interest, so there is less concern for collusion. If these two parties do not want to hide their information from each other, there is no reason to perform the computation. The last party P_2 is there to obtain the necessary three parties for Shamir and Mal-Shamir. This party in practice would be an individual who is selected by both of the parties who provide an input to the computation as this can help ensure that the extra party is not biased toward one of the input parties.

Results

We tested our protocols (ie, Shamir, Mal-Shamir, Semi-OT, and MASCOT) with simulated filtered genomic variants sets of up to 400,000 elements in length. Protocols were run over both a virtual network and a CloudLab network. The Shamir [22,26-28] and Mal-Shamir (maliciously secured by verification techniques by Chida et al [29]) protocols require third-party computation. The Semi-OT and MASCOT [19] protocols require at least two parties. In general, the honest majority is more efficient than the malicious majority protocols but at a minimum requires three parties to execute.

We measure the resources needed to compute an operation (set difference, symmetric difference, and Jaccard similarity) between two sets over the specified length of the array through time and network communication. The unit of time measured here is in seconds as measured by the framework, and the network communication is measured in megabytes. Figure 3 shows these results over an array of panel sizes.

We were able to calculate the Jaccard similarity over SNP panels of size 400,000 in 0.363 seconds with Shamir, 2.155 seconds with Mal-Shamir, 13.184 seconds with Semi-OT, and 113.397 seconds (about 2 minutes) with MASCOT. Shamir is the most efficient, whereas MASCOT takes the longest. The extra time MASCOT takes is due to extra security guarantees that it offers over the Shamir method. Figure 4 illustrates the growing complexity as the number of computing servers increases. Adding more servers has the benefit of improving the security of the system. When looking at Figure 4, one may notice that Shamir and Mal-Shamir do not have a statistic for two players. This is because these protocols do not support two players.

We also ran our method on the GIAB data set [25]. The comparison required converting the VCF and SNP panel files into a binary array to be used in the secure computation. This process took on average 96.430 seconds (about 1 and a half minutes) per file using our custom code. This process is a one-time need, and the resulting files can be used multiple times for other comparisons. We used our secure method in conjunction with the SNP panel provided by Murray et al [10]. Figure 5 compares the Jaccard similarity between filtered SNPs and the entire VCF using BEDTools [11], which indicates a high correlation. For convenience, we reiterate that HG002 is the child of HG003 and HG004, and that HG005 is the child of HG006 and HG007. Then we ran different secure computing methods on the network. The costs of the computation with the panel size of 4763 are presented in Table 1.

The comparison by Jaccard similarity took an average of 31.008 seconds between two files. This comparison done by BEDTools assumes that the VCF files are sorted but does not require any preprocessing beforehand.

Figure 3. Each figure shows the growth of complexity as the array length grows. While we chart Mal-Shamir, MASCOT, Semi-OT, and Shamir together in the same chart, they are not comparable in terms of resource use alone. The more expensive protocols tend to have stronger security guarantees. However, the security guarantees of Shamir and Mal-Shamir are frequent enough for standard usage and are practical to use on larger sets. Mal-Shamir: Malicious Shamir; Semi-OT: Semihonest Oblivious Transfer.

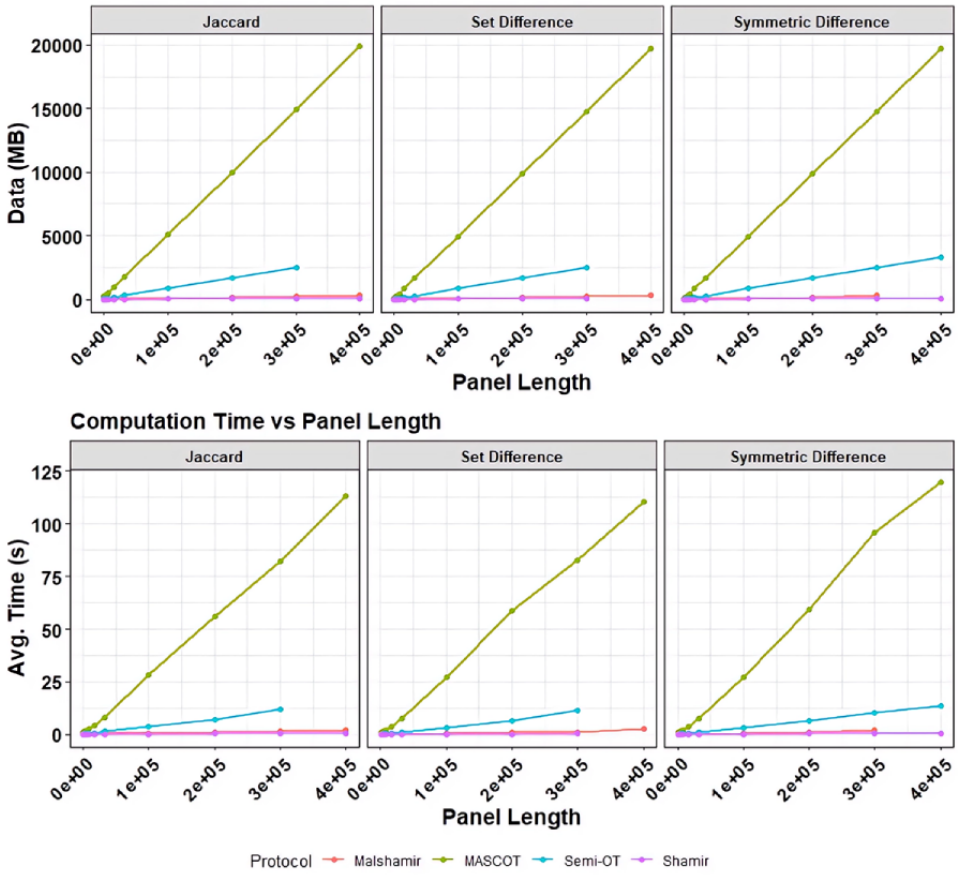


Figure 4. The communication complexity scales quadratically with the number of players. The benefit of increasing the number of computing servers is increased security.

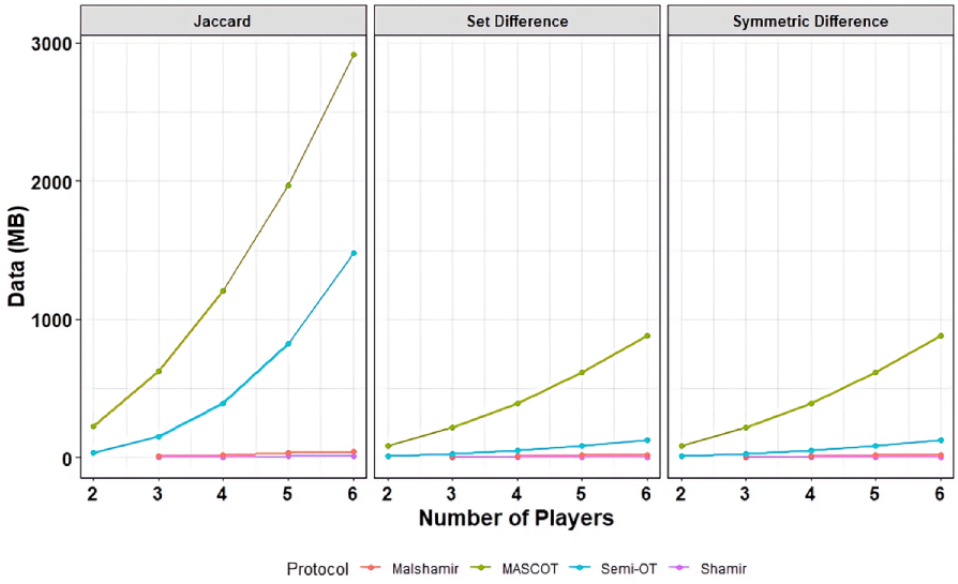


Figure 5. The Jaccard similarity is compared through filtered single nucleotide polymorphisms (Panel Jaccard) vs through the entire VCF using BEDTools (Whole VCF Jaccard). The correlation can be seen: HG002 matches HG003 and HG004 and HG005 matches HG006 and HG007. VCF: Variant Call Format.

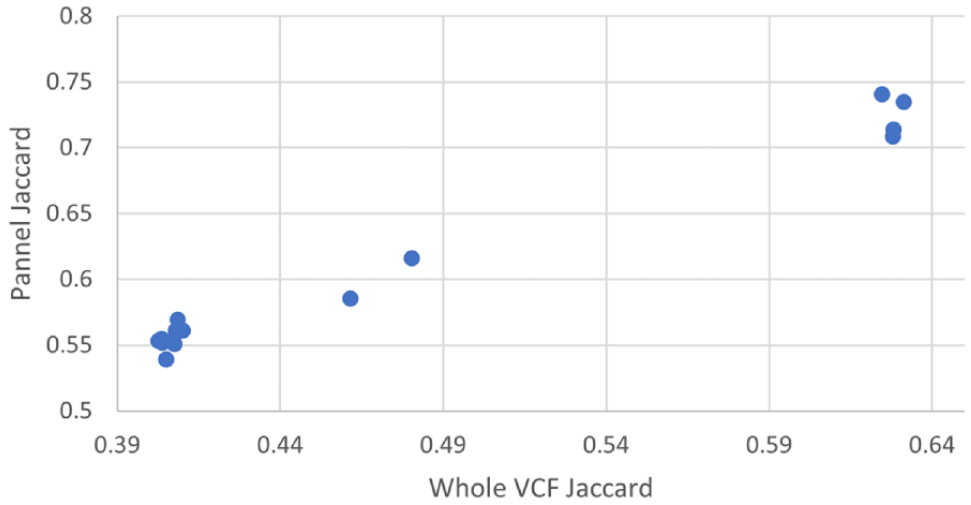


Table 1. This table shows the computation using the secure protocols from the filtered panel using 4763 single nucleotide polymorphisms after preprocessing.

Method	Time (s), average	Communication (MB)
Shamir	0.0350	1.671
Mal-Shamir[a]	0.1125	9.357
Semi-OT[b]	0.7085	40.570
MASCOT	5.4322	353.363

[a]Mal-Shamir: Malicious Shamir.

[b]Semi-OT: Semihonest Oblivious Transfer.

Discussion

There are several established protocols in a secure multiparty computation that allow for computation over a finite field. We make use of a framework called MP-SPDZ that securely implements these operations as described in many secure settings.

Our work provides an efficient secure method for computing similarity between two genomic sequences by considering predefined variant panels. Our study only considers the presence of a variant (ie, binary representation) and does not explicitly compute the set based on the actual allele (ie, nucleotide identity A, T, C, G) or combination of alleles (ie, heterozygous positions) represented at that location. Although this is sufficient for most practical applications, our methods can easily be extended to compute the set based on the explicit allele presented at the given location.

Our method can easily be extended to allow only results beyond a certain threshold. Such a modification can be done by performing an inequality check at the end of any of the protocols. The inequality check only needs to be performed once and adds constant time to the protocol when the number of parties is fixed.

We presented four protocols that can be used to execute the arithmetical operations of our protocols. Based on the results in the previous section Shamir and Mal-Shamir are faster but have different security guarantees from Semi-OT and MASCOT. Mal-Shamir provides realistic security guarantees while requiring similar computational and network resources to Shamir. Thus, we recommend using Mal-Shamir to execute our protocols. Shamir requires at least three parties, so we suggest executing the protocol using Semi-OT when only two parties can be used. Though MASCOT provides malicious security, the method is still impractical on realistic data sets if we expect results in real time. However, in some situations it may be reasonable to allow processing over a day, then even MASCOT will be practical.

Our development may pave the way for a practical protocol to share human variant data securely, which may help support large-scale variant applications for precision medicine.

Acknowledgments

This project was in part funded by the National Science Foundation (award 1946619) and the National Institutes of Health BD2K Training Grant T32HG009060 to STK.

Data Availability

Some data were simulated using the methods described in the Data subsection of the Methods; other data include the Genome in a Bottle data set [25].

Authors' Contributions

DX conceived the idea. AW provided the techniques/security, performed the computational experiments, and wrote the manuscript. STK provided the methods in biology. WJ provided supervision over the security of the methods. All authors edited the text.

Conflicts of Interest

None declared.

References

1. Guerrini C, Robinson J, Bloss C, Bash Brooks W, Fullerton S, Kirkpatrick B, et al. Family secrets: experiences and outcomes of participating in direct-to-consumer genetic relative-finder services. Am J Hum Genet 2022 Mar 03;109(3):486-497 [FREE Full text] [doi: 10.1016/j.ajhg.2022.01.013] [Medline: 35216680]
2. Danecek P, Auton A, Abecasis G, Albers C, Banks E, DePristo M, 1000 Genomes Project Analysis Group. The variant call format and VCFtools. Bioinformatics 2011 Aug 01;27(15):2156-2158 [FREE Full text] [doi: 10.1093/bioinformatics/btr330] [Medline: 21653522]
3. Aziz M, Alhadidi D, Mohammed N. Secure approximation of edit distance on genomic data. BMC Med Genomics 2017 Jul 26;10(Suppl 2):41 [FREE Full text] [doi: 10.1186/s12920-017-0279-9] [Medline: 28786362]
4. Mahdi M, Al Aziz MM, Alhadidi D, Mohammed N. Secure similar patients query on encrypted genomic data. IEEE J Biomed Health Inform 2019 Nov;23(6):2611-2618. [doi: 10.1109/JBHI.2018.2881086] [Medline: 30442622]
5. Tang H, Jiang X, Wang X, Wang S, Sofia H, Fox D, et al. Protecting genomic data analytics in the cloud: state of the art and opportunities. BMC Med Genomics 2016 Oct 13;9(1):63 [FREE Full text] [doi: 10.1186/s12920-016-0224-3] [Medline: 27733153]
6. Zhu D, Zhu H, Wang X, Lu R, Feng D. Efficient and privacy-preserving similar patients query scheme over outsourced genomic data. IEEE Trans Cloud Computing 2023 Apr 1;11(2):1286-1302. [doi: 10.1109/tcc.2021.3131287]
7. Jagadeesh KA, Wu DJ, Birgmeier JA, Boneh D, Bejerano G. Deriving genomic diagnoses without revealing patient genomes. Science 2017 Aug 18;357(6352):692-695. [doi: 10.1126/science.aam9710] [Medline: 28818945]
8. Çetin GS, Chen H, Laine K, Lauter K, Rindal P, Xia Y. Private queries on encrypted genomic data. BMC Med Genomics 2017 Jul 26;10(Suppl 2):45 [FREE Full text] [doi: 10.1186/s12920-017-0276-z] [Medline: 28786359]
9. Keller M. MP-SPDZ: a versatile framework for multi-party computation. In: Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security. 2020 Oct Presented at: CCS '20; November 9-13, 2020; Virtual p. 1575-1590. [doi: 10.1145/3372297.3417872]
10. Murray S, Oliphant A, Shen R, McBride C, Steeke RJ, Shannon SG, et al. A highly informative SNP linkage panel for human genetic studies. Nat Methods 2004 Nov;1(2):113-117. [doi: 10.1038/nmeth712] [Medline: 15782173]
11. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010 Mar 15;26(6):841-842 [FREE Full text] [doi: 10.1093/bioinformatics/btq033] [Medline: 20110278]
12. Canetti R. Universally composable security: a new paradigm for cryptographic protocols. In: Proceedings 42nd IEEE Symposium on Foundations of Computer Science. 2002 Presented at: 42nd IEEE Symposium on Foundations of Computer Science; October 8-11, 2001; Newport Beach, CA. [doi: 10.1109/SFCS.2001.959888]
13. Beerliová-Trubíniová Z, Hirt M. Perfectly-secure MPC with linear communication complexity. In: Canetti R, editor. Theory of Cryptography: Fifth Theory of Cryptography Conference, TCC 2008, New York, USA, March 19-21, 2008, Proceedings. Berlin, Heidelberg: Springer; 2008:213-230.
14. Ben-Sasson E, Fehr S, Ostrovsky R. Near-linear unconditionally-secure multiparty computation with a dishonest minority. In: Safavi-Naini R, Canetti R, editors. Advances in Cryptology -- CRYPTO 2012: 32nd Annual Cryptology Conference, Santa Barbara, CA, USA, August 19-23, 2012, Proceedings. Berlin: Springer; 2012:663-680.
15. Damgård I, Nielsen JB. Scalable and unconditionally secure multiparty computation. In: Advances in Cryptology - CRYPTO 2007: 27th Annual International Cryptology Conference, Santa Barbara, CA, USA, August 19-23, 2007, Proceedings. Berlin, Heidelberg: Springer; 2007:572-590.
16. Furukawa J, Lindell Y. Two-thirds honest-majority MPC for malicious adversaries at almost the cost of semi-honest. In: Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. 2019 Nov Presented at: CCS '19; November 11-15, 2019; London, United Kingdom p. 1557-1571. [doi: 10.1145/3319535.3339811]
17. Goyal V, Liu Y, Song Y. Communication-efficient unconditional MPC with guaranteed output delivery. In: Advances in Cryptology – CRYPTO 2019: 39th Annual International Cryptology Conference, Santa Barbara, CA, USA, August 18–22, 2019, Proceedings, Part II. Cham: Springer; 2019:85-114.
18. Goyal V, Song Y. Malicious security comes free in honest-majority MPC. Cryptology ePrint Archive Preprint posted online on February 10, 2020 [FREE Full text]

19. Keller M, Orsini E, Scholl P. MASCOT: faster malicious arithmetic secure computation with oblivious transfer. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. 2016 Oct Presented at: CCS '16; October 24-28, 2016; Vienna, Austria p. 830-842. [doi: 10.1145/2976749.2978357]
20. Lindell Y, Nof A. A framework for constructing fast MPC over arithmetic circuits with malicious adversaries and an honest-majority. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. 2017 Oct Presented at: CCS '17; October 30-November 3, 2017; Dallas, TX p. 259-276. [doi: 10.1145/3133956.3133999]
21. Gilboa N. Two party RSA key generation. In: Wiener M, editor. Advances in Cryptology - CRYPTO '99: 19th Annual International Cryptology Conference, Santa Barbara, California, USA, August 15-19, 1999 Proceedings. Berlin, Heidelberg: Springer; 1999:116-129.
22. Shamir A. How to share a secret. Commun ACM 1979 Nov 01;22(11):612-613. [doi: 10.1145/359168.359176]
23. Yue JG, Liti G. simuG: a general-purpose genome simulator. Bioinformatics 2019 Nov 01;35(21):4442-4444 [FREE Full text] [doi: 10.1093/bioinformatics/btz424] [Medline: 31116378]
24. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome Res 2002 Jun;12(6):996-1006 [FREE Full text] [doi: 10.1101/gr.229102] [Medline: 12045153]
25. Genome in a Bottle. National Institute of Standards and Technology. URL: https://www.nist.gov/programs-projects/genome-bottle [accessed 2023-06-20]
26. Ben-Or M, Goldwasser S, Wigderson A. Completeness theorems for non-cryptographic fault-tolerant distributed computation. In: Goldreich O, editor. Providing Sound Foundations for Cryptography: On the Work of Shafi Goldwasser and Silvio Micali. New York, NY: Association for Computing Machinery; Oct 2019:351-371.
27. Chaum D, Crépeau C, Damgard I. Multiparty unconditionally secure protocols. In: Proceedings of the Twentieth Annual ACM Symposium on Theory of Computing. 1988 Jan Presented at: STOC '88; May 2-4, 1988; Chicago, IL p. 11-19. [doi: 10.1145/62212.62214]
28. Cramer R, Damgård I, Maurer U. General secure multi-party computation from any linear secret-sharing scheme. In: Advances in Cryptology – EUROCRYPT 2000: International Conference on the Theory and Application of Cryptographic Techniques Bruges, Belgium, May 14-18, 2000 Proceedings. Berlin, Heidelberg: Springer; 2000:316-334.
29. Chida K, Hamada K, Ikarashi D, Kikuchi R, Genkin D, Lindell Y, et al. Fast large-scale honest-majority MPC for malicious adversaries. J Cryptology 2023 Apr 18;36(3):1. [doi: 10.1007/s00145-023-09453-7]

Abbreviations

AWS: Amazon Web Services
GCP: Google Cloud Platform
GIAB: Genome in a Bottle
Mal-Shamir: Malicious Shamir
MP-SPDZ: Multi-Party Secure, Privacy-Preserving, and Decentralized Zeus
Semi-OT: Semihonest Oblivious Transfer
SNP: single nucleotide polymorphism
SP: service provider
UCSC: University of California, Santa Cruz
VCF: Variant Call Format

Edited by T Leung, S Hacking; submitted 29.11.22; peer-reviewed by X Zhou, G Gürsoy, Y Lu; comments to author 25.01.23; revised version received 21.05.23; accepted 09.06.23; published 18.07.23.

Please cite as:
Woods A, Kramer ST, Xu D, Jiang W
Secure Comparisons of Single Nucleotide Polymorphisms Using Secure Multiparty Computation: Method Development
JMIR Bioinform Biotech 2023;4:e44700
URL: https://bioinform.jmir.org/2023/1/e44700
doi:10.2196/44700
PMID:

Original Paper

User and Usability Testing of a Web-Based Genetics Education Tool for Parkinson Disease: Mixed Methods Study

Noah Han[1,2], BS; Rachel A Paul[1,2], MS; Tanya Bardakjian[1,2,3], MS; Daniel Kargilis[1,4], BS; Angela R Bradbury[5], MD; Alice Chen-Plotkin[1,2], MD; Thomas F Tropea[1,2], MPH, MTR, DO

[1]Department of Neurology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States

[2]Department of Neurology, Pennsylvania Hospital, Philadelphia, PA, United States

[3]Sarepta Therapeutics, Cambridge, MA, United States

[4]Johns Hopkins University School of Medicine, Baltimore, MD, United States

[5]Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States

Corresponding Author:
Thomas F Tropea, MPH, MTR, DO
Department of Neurology
Perelman School of Medicine
University of Pennsylvania
330 South 9th Street
Philadelphia, PA, 19107
United States
Phone: 1 215 829 7731
Fax: 1 215 829 6606
Email: Thomas.Tropea@pennmedicine.upenn.edu

Abstract

Background: Genetic testing is essential to identify research participants for clinical trials enrolling people with Parkinson disease (PD) carrying a variant in the glucocerebrosidase (*GBA*) or leucine-rich repeat kinase 2 (*LRRK2*) genes. The limited availability of professionals trained in neurogenetics or genetic counseling is a major barrier to increased testing. Telehealth solutions to increase access to genetics education can help address issues around counselor availability and offer options to patients and family members.

Objective: As an alternative to pretest genetic counseling, we developed a web-based genetics education tool focused on *GBA* and *LRRK2* testing for PD called the Interactive Multimedia Approach to Genetic Counseling to Inform and Educate in Parkinson's Disease (IMAGINE-PD) and conducted user testing and usability testing. The objective was to conduct user and usability testing to obtain stakeholder feedback to improve IMAGINE-PD.

Methods: Genetic counselors and PD and neurogenetics subject matter experts developed content for IMAGINE-PD specifically focused on *GBA* and *LRRK2* genetic testing. Structured interviews were conducted with 11 movement disorder specialists and 13 patients with PD to evaluate the content of IMAGINE-PD in user testing and with 12 patients with PD to evaluate the usability of a high-fidelity prototype according to the US Department of Health and Human Services Research-Based Web Design & Usability Guidelines. Qualitative data analysis informed changes to create a final version of IMAGINE-PD.

Results: Qualitative data were reviewed by 3 evaluators. Themes were identified from feedback data of movement disorder specialists and patients with PD in user testing in 3 areas: content such as the topics covered, function such as website navigation, and appearance such as pictures and colors. Similarly, qualitative analysis of usability testing feedback identified additional themes in these 3 areas. Key points of feedback were determined by consensus among reviewers considering the importance of the comment and the frequency of similar comments. Refinements were made to IMAGINE-PD based on consensus recommendations by evaluators within each theme at both user testing and usability testing phases to create a final version of IMAGINE-PD.

Conclusions: User testing for content review and usability testing have informed refinements to IMAGINE-PD to develop this focused, genetics education tool for *GBA* and *LRRK2* testing. Comparison of this stakeholder-informed intervention to standard telegenetic counseling approaches is ongoing.

(JMIR Bioinform Biotech 2023;4:e45370) doi:10.2196/45370

KEYWORDS

Parkinson disease; genetic testing; teleneurology; patient education; neurology; genetic; usability; user testing; web-based; internet-based; web-based resource; mobile phone

Introduction

Parkinson disease (PD) is the second commonest neurodegenerative disease and the fastest-growing neurological disease worldwide [1,2]. Variants in leucine-rich repeat kinase 2 (*LRRK2*), glucocerebrosidase (*GBA*), *Parkin*, Parkinsonism-associated deglycase (*DJ-1*), VPS35 retromer complex component (*VPS35*), PTEN-induced kinase 1 (*PINK1*), and α-synuclein (*SNCA*) are identified in 10%-12% of PD cases [3-9]. Despite a genetic mutation frequency similar to some cancer syndromes where germline genetic testing is common [10,11], genetic testing is not standard in the evaluation and management of PD and is rarely conducted as part of clinical care [12]. However, knowing one's genetic status is already of key importance for research, as therapies targeting carriers of *GBA* and *LRRK2* variants are in clinical trials [13,14]. Additionally, patients with PD have expressed interest in learning their genetic information [15,16]. Research programs such as the multisite PDGENEration study sponsored by the Parkinson Foundation (ClinicalTrials.gov NCT04057794) and the University of Pennsylvania (UPenn) Molecular Integration in Neurological Diagnosis (MIND) Initiative [17] have been developed to increase genetic testing and counseling.

Recently proposed recommendations would expand clinical or research genetic testing to nearly all patients with PD [18]. However, the standard service delivery model for genetic testing includes pre- and posttest genetic counseling, which is limited by the availability of specialized genetic counselors or physicians with sufficient genetics training [19]. Indeed, there are only 125 genetic counselors specialized in neurogenetics offering in-person visits, and 82 genetic counselors offering telehealth visits listed on the National Society of Genetic Counselors' public directory for genetic counselors. We developed a web-based education tool focused on *GBA* and *LRRK2* genetic testing called the Interactive, Multimedia Approach to Genetic Counseling to Inform and Educate in Parkinson's Disease (IMAGINE-PD) to address this gap. IMAGINE-PD could be made available in research studies to increase genetics education prior to *GBA* and *LRRK2* testing to identify research-eligible patients with PD.

The goal of this work is to create a genetics education tool for *GBA* and *LRRK2* testing incorporating key stakeholder input. Structured interviews were conducted with movement disorder specialists (MDSs) and patients with PD to evaluate the content of IMAGINE-PD in user testing and with patients with PD to evaluate website usability according to the US Department of Health and Human Services (DHHS) Research-Based Web Design & Usability Guidelines [20]. Qualitative data analysis informed changes to create a final version of IMAGINE-PD.

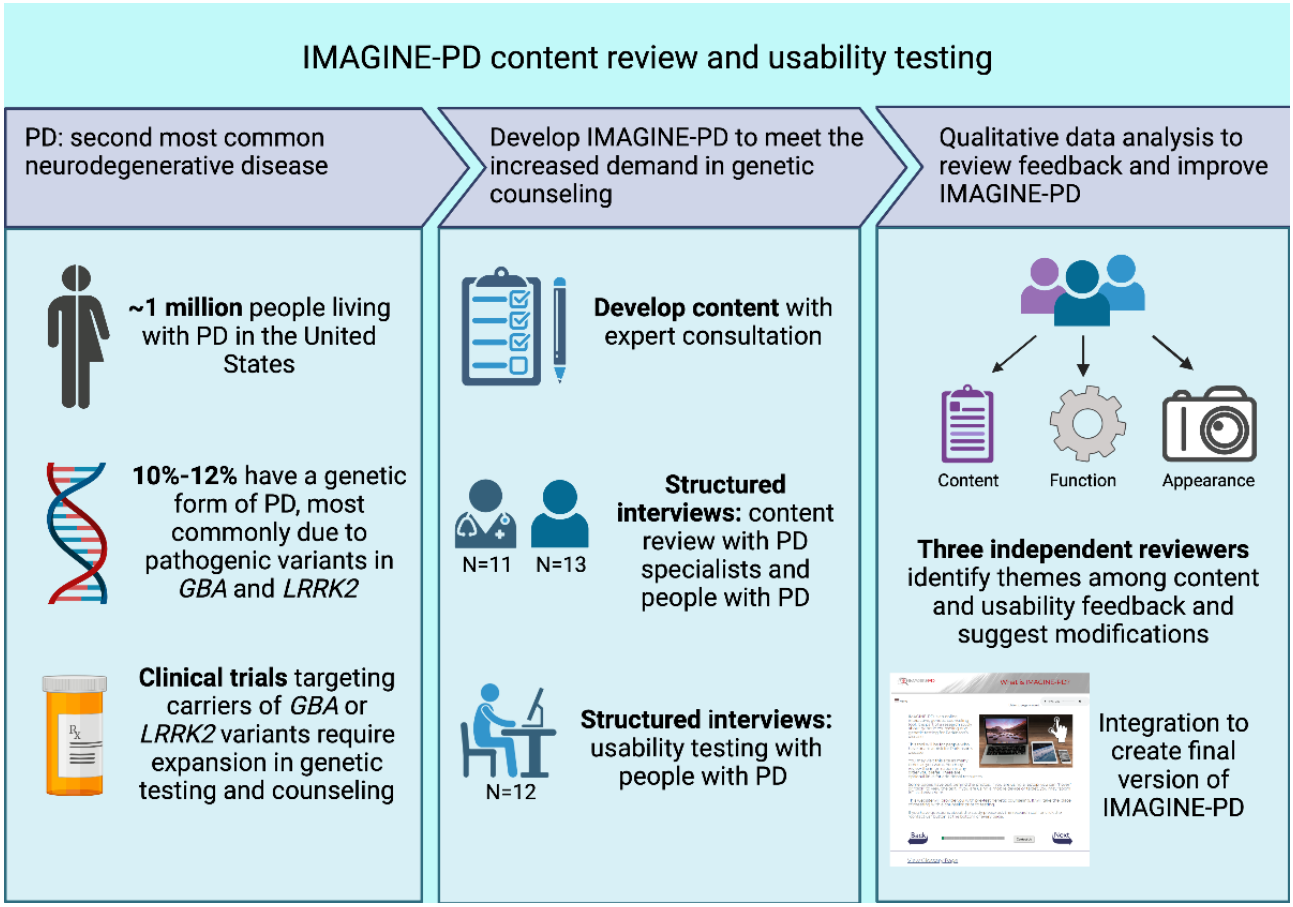
Methods

IMAGINE-PD Website Development

Website development and usability testing were conducted in accordance with the DHHS research-based web design and usability guidelines [20]. Core content of a genetic counseling tool for *GBA* and *LRRK2* variant genetic testing of PD was developed by the authors (TB, ACP, and TFT). *GBA* and *LRRK2* variant testing information was included to align with the MIND Initiative, which is a whole-clinic biobanking effort at UPenn that offers optional genetic counseling and clinical confirmation testing [17]. Content was then assigned to "primary" or essential, "secondary" or important but not essential, or "optional" categories based on their level of importance according to expert input from genetic counselors and movement disorder physicians (RAP, TB, ACP, and TFT). Five primary concepts included a review of PD, concepts of basic genetics, genetics of PD, disclosure of genetic results, and limitations and implications of genetic testing. For each primary concept, an audiovisual recording of a movement disorder physician or genetic counselor describing the concept was created. Secondary information included diagnosis, symptoms, and treatment of PD, an introduction to genetic counseling, a review of genetic testing (GT), types of genetic tests and results, risks and benefits of GT, a review of *GBA* including Gaucher disease and *LRRK2*, and a review of other genetic causes of PD that are not tested in the study (limitations). One or more slides were created to capture the secondary content important for genetic counseling including text, visuals, and an optional audio recording of the text presented. Links to outside reading were identified for optional material. All content was organized into a high-fidelity prototype website using the Wix website [21]. Once the preliminary set of material was created, we solicited feedback on content and website organization from genetic counselors and physician-scientists with expertise in neurogenetics. Changes were made to address points of feedback. A new website was created with assistance from The UPenn Center for Clinical Epidemiology and Biostatistics Clinical Research Computing Unit. This new website included the content from the prototype and was made compatible with common web browsers and operating systems to be accessible on typical connection speeds across a range of different resolutions and orientations to account for use on a computer, smartphone, or tablet computer. Number of visits and time spent on each page are captured. A computer graphics specialist with experience in medical artwork created animations [22]. With this prototype, user testing was first conducted, and changes were then incorporated into IMAGINE-PD after reviewing feedback from patients with PD. These changes included content revisions on both primary and secondary topics. Videos with movement disorders and neurogenetics specialists were professionally recorded after implementing content changes from user testing. Readability of each page was evaluated using Readable [23] to

ensure a Flesch-Kincaid readability index of below 9. Usability testing was conducted subsequently, and again feedback was reviewed, and refinements were made to the content, organization, and web interface. In the final version, the basic genetics video was broken into 2 videos for a total of 6 videos. Videos range from 51 to 92 seconds in length. A link was available below the video to read a transcript of the audio recording. The final version is hosted at UPenn [24]. A flowchart of the IMAGINE-PD web development and testing process can be viewed in Figure 1.

Figure 1. IMAGINE-PD website development flowchart. GBA: glucocerebrosidase; IMAGINE-PD: Interactive Multimedia Approach to Genetic Counseling to Inform and Educate in Parkinson's Disease; LRRK2: leucine-rich repeat kinase 2; PD: Parkinson disease.



Participants

In the user testing phase, 13 cognitively normal patients with PD who receive care at UPenn, referred by their physicians, were enrolled. MDSs at the UPenn Parkinson's Disease and Movement Disorders Center and the Philadelphia Veterans Affairs Parkinson's Disease Research, Education, and Clinical Center were asked to participate (excluding authors ACP and TFT) for a total of 11 neurologists or psychiatrists. In the usability testing phase, 12 cognitively normal patients with PD who receive care at UPenn, referred by their physicians, who did not participate in the user testing phase were enrolled. No patient or physician declined participation. All patients were previously enrolled in the MIND Initiative. In both phases, sample sizes exceeded the suggestions outlined in the DHHS usability testing guidelines [20]. Identified participants were approached in person, by phone, or by email. Informed consent was obtained, and study visits were performed in person by a single evaluator (NH).

User Testing

Structured interviews were conducted between a single evaluator with experience in conducting clinical research visits (NH) and a participant between October 31, 2019, and December 11, 2019. The evaluator neither had a prior relationship with any participant nor any assumptions or presuppositions about the outcomes of the interviews. First, participants were asked 11 questions about their general internet use, 4 questions about email and internet use for health information (adapted from Baker et al [25]), and 3 questions assessing background knowledge in genetics, PD, and genetic testing. During the interview, the evaluator asked questions of each participant, and data were entered directly into Research Electronic Data Capture (REDCap; Vanderbilt University) [26,27]. The evaluator navigated to each web page in the prototype allowing the participant to view and listen to all material on that page with unlimited time. Prompts were given to ensure all participants interacted with all aspects of each page. Feedback was solicited using 4 open-ended questions, a 1-10 rating of usefulness, and a 1-4 rating of clarity of presentation. After all web pages were reviewed, 7 additional questions were asked pertaining to the entire series of web pages and to solicit overall comments. Questionnaires are available in Multimedia Appendix 1.

Usability Testing

The second iteration of IMAGINE-PD was created after the results of user testing were reviewed, and changes were implemented. For usability testing, structured videoconference interviews were conducted between evaluator (NH) and participant between August 18, 2020, and September 14, 2020. Again, the evaluator neither had a prior relationship with any participant nor any assumptions or presuppositions about the outcomes of the interviews. First, participants were asked 8 questions about their internet use. The participant navigated to each web page with unlimited time to view all content. Participants were asked to use the "share-screen" function so the evaluator could observe their navigation of the website. Prompts were given to ensure all participants interacted with all aspects of each page. Feedback was solicited using 8 open-ended questions asked for each page. Questionnaires are available in Multimedia Appendix 1.

Statistical Analysis

Descriptive statistics are presented for all demographic data and scales. A qualitative content analysis plan was developed in advance of data review in consultation with Judy Shea, PhD. For both user testing review and usability testing phases, all data were collected into a spreadsheet excluding identifying information. Three evaluators (NH, RAP, and TFT) were given the following rules: (1) review the data from each web page for PD providers and patients with PD separately and identify 1 or more key themes per page, if a theme is apparent, and (2) if more than 1 theme is apparent, order themes for each page in order of importance or frequency. Data were independently reviewed by each evaluator to improve the trustworthiness of the analysis. Subsequently, the 3 evaluators met to establish a consensus on the key themes. Suggestions for changes to the website were made after all themes were reviewed and discussed among the evaluators.

Ethics Approval

Institutional review board (IRB) approval (UPenn IRB 834311) was obtained before initiating the study, and informed consent was obtained from all participants prior to any study activities. This research adheres to the principles set out in the Declaration of Helsinki.

Results

Website Development

The final version includes 27 web pages, including 3 video-only pages and 3 video pages with animations. Audiovisual pages also include accessible text for participants with hearing impairment. Each nonvideo page included audio recording of the text for auditory learners. The website is accessible by and formatted for computer browsers, mobile devices, or tablets. The length of time to view all videos is 7 minutes 42 seconds, and the estimated time to review all website content is 25-40 minutes. The Flesch-Kincaid reading level for all pages was a median of 8.3 (IQR 7.375-8.25) indicating all pages on the website maintained approximately an eighth-grade leading level. Screenshots of each web page are available in Multimedia Appendix 2.

User Testing

Participant demographics are summarized in Table 1. Email and internet use for health information is described in Figure 2A and Table S1 in Multimedia Appendix 1. PD, genetics, and genetic testing knowledge are reported in Figure 2B.

Table 1. Cohort description.

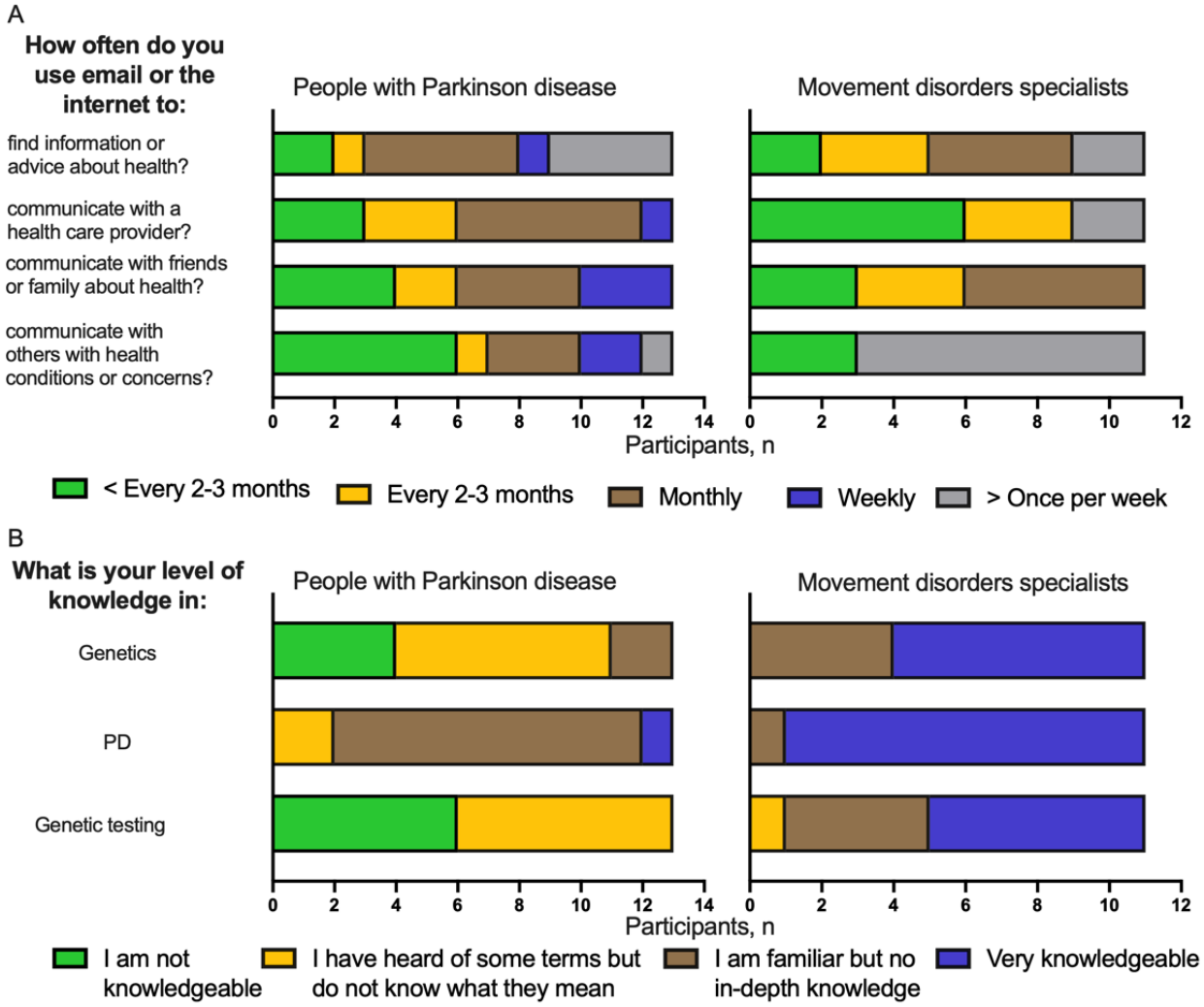
	Content review		Usability
	PD[a] (N=13)	Physicians (N=11)	PD (N=12)
Age at test (years), median (IQR)	64 (63-69)	N/A[b]	63 (59-72)
Sex, n (%)			
Female	3 (23)	4 (36)	5 (42)
Male	10 (77)	7 (64)	7 (58)
Disease duration (years), median (IQR)	11 (7-13)	N/A	8 (5-13)
Education (years), median (IQR)	17 (15-18)	N/A	16 (16-17)[c]

[a]PD: patients with Parkinson disease.

[b]N/A: not applicable.

[c]Two values excluded due to missing data.

Figure 2. (A) Patients with PD and MDS self-reported use of the internet and email for health information. (B) PD patients and MDS self-reported knowledge in genetics, PD, and genetic testing (GT). MDS: movement disorder specialist; PD: Parkinson disease.



For each web page, participants were asked to rate the usefulness (1 being the least useful and 10 being the most useful). The average usefulness across all pages for patients with PD was 8.49 (SD 0.65) and for MDS was 8.85 (SD 0.64). The results of usefulness for each page are found in Table S2 in Multimedia Appendix 1.

In user testing, content feedback pertained to the volume and level of detail of information (too much information), the complexity of the content (wording is too technical), and the focus of the content (make it more related to PD). Specifically, the amount of information was reported to be too much by patients on 13 occasions by 3 different physicians and 4 different patients pertaining to 11 different pages. Representative comments included “Gaucher's disease may be a little confusing. May not need to go into so much detail” and “alpha synuclein was confusing. Maybe too much content here.” On 6 different occasions, 6 different patients reported the topic, and the language used to be too technical including words like “bradykinesia” and “brain imaging.” Content was mentioned on 3 occasions by 2 different physicians and 1 patient pertaining to 3 different pages. Representative comments included “focus more on PD related genetic testing,” and “mention these are two genes we are looking at and this is why it is related to PD.” To address these points, we made changes to focus and simplify the content and to use plain language at a Flesch-Kincaid reading level 9 or below.

Three domains of feedback were identified on user testing analysis: content, function, and appearance. These domains were consistent between patient and physician participants, and suggestions for change were made based on the combined review of both sets of responses. Changes were made on 20 pages in response to content feedback, 4 pages in response to function feedback, and 4 pages in response to appearance feedback. A summary of the user testing qualitative analysis and recommendations for change are shown in Tables 2-4.

Table 2. Summary of content changes made in response to user testing.

Page	Content feedback	Changes made
1. Introduction	Want more information on this page	Short summary was added to introduction
3. What is IMAGINE-PD[a]	Not sure how the website is relevant to PD[b]	Clarified study relevance to PD
4. Summary	Wording is too technical	Simplified vocabulary
5. What is PD	Wording is too technical	Simplified vocabulary
6. How is PD diagnosed	Want more information on nonmotor symptoms	Information on nonmotor symptoms included
7. What causes PD	Too much information on page	Information was cut down and simplified
8. How is PD treated	Want less information on drugs, more information on other treatments	Drug section shortened and added other PD treatments
9. What is genetics	Too much information on page	Information was shortened and simplified
10. Basics of genetics	Too much information on page and too technical	Information was shortened and simplified
11. What is GT[c]	Wording is too technical and not sure how information is related to PD	Simplified vocabulary, clarified relevance
12.5. Reasons for GT	Page should be included as a separate page	Separate page was made for information
13. Benefits and risks	Make it more related to PD and more information on benefits	Clarified PD relevance
14. GT process	Too much detail	Simplified a shortened page
15. How genetics affects PD	Too much information	Shortened page
16. *GBA*[d] and PD	Information too technical and want to know why this gene is being studied	Simplified vocabulary and clarified the relevance of *GBA*
17. *LRRK2*[e] and PD	Information too technical and want to know why this gene is being studied	Simplified vocabulary and clarified the relevance of *LRRK2*
18. Other rare genes for PD	Language seems too technical, why only *GBA* and *LRRK2* being tested	Simplified language, provides more resources for additional information on other genes
19. How do I get results	Language seems too technical, want more information	Clarified next steps and reviewed previously introduced topics to clarify information
20. VUS[f] and unexpected results	Wording too technical	Simplified language
21. Implications and limitations	Confused about previously introduced topics	Added information to clarify

[a]IMAGINE-PD: Interactive Multimedia Approach to Genetic Counseling to Inform and Educate in Parkinson's Disease.
[b]PD: Parkinson disease.
[c]GT: genetic testing.
[d]GBA: glucocerebrosidase.
[e]LRRK2: leucine-rich repeat kinase 2.
[f]VUS: variant of uncertain significance.

Table 3. Summary of function changes made in response to user testing.

Page	Function feedback	Changes made
2. Instructions	Did not know that there was more information at the bottom and needed to scroll	Shortened page to minimize scrolling
4. Summary	Picture was clickable but did not lead to any new page	Fixed error so picture was no longer clickable
5. What is PD[a]	Already knew information, wanted a skip function	Shortened page, so not as much time would be spent on it
6. How is PD diagnosed	In page slides were confusing	Made in page slides more obvious

[a]PD: Parkinson disease.

Table 4. Summary of appearance changes made in response to user testing.

Page	Appearance feedback	Changes made
1. Introduction	Make title clearer and more obvious	Title was enlarged and bolded
3. What is IMAGINE-PD[a]	Use bullet points to make page easier to read	Used bullet points
4. Summary	Picture was not relevant	Used a different picture
12. When to consider GT[b]	Did not like picture	Used a different picture

[a]IMAGINE-PD: Interactive Multimedia Approach to Genetic Counseling to Inform and Educate in Parkinson's Disease.
[b]GT: genetic testing.

At the conclusion, participants were asked to provide summary feedback. A summary of responses is found in Table S3 in Multimedia Appendix 1. Comments regarding the order of the presented information, the web-based tools such as the buttons, links, and menus, and overall comments are provided in Table S4 in Multimedia Appendix 1.

Usability Testing

Participant demographics are summarized in Table 1. All participants reported using the internet or email within the past 12 months. Overall, 1 (8.3%) participant reported dial-up network use, 9 (75%) participants reported broadband network use, 7 (58%) participants reported using smartphones, and 8 (67%) participants reported accessing the internet via a Wi-Fi network. All participants reported using the internet to communicate with a health care provider, and 9 (75%) participants reported using the internet to search for health or medical information.

In usability testing, technical language was reported on 10 occasions, while the volume of information was reported on 6 occasions. Representative comments included "technical and a little difficult to understand" and "it was a little too much info in one video." Based on usability testing feedback, further refinements were made as outlined in Tables 5-7. In addition to simplifying language, we included a glossary page, accessible via a link on every page, that defined key terms organized by concept. We also included a "Contact Us" page allowing participants to request additional information from a genetic counselor. Key points of feedback about website function included requests for a progress bar and mobile compatibility, which were both addressed for the final version. Improved audio quality was also suggested. Additionally, feedback on the difficulty of navigation through the website was mentioned by patients with PD and was addressed by creating clear instructions for navigation at the beginning of the website and large labels for website progression. Feedback about appearance included font color and size as well as picture choice. Final pictures were selected to ensure the representation of individuals of diverse race, ethnicity, age, and sex.

Table 5. Summary of content changes made in response to usability testing.

Page	Content feedback	Changes made
1. Introduction	Want more information on this page	Short summary was added to introduction
2. What is IMAGINE-PD[a]	Too much information	Cut down on information
6. Genetics introduction	Too much information all at once	Added animations and broke page to 2 pages
7. VUS[b]	Wording is too technical	Removed to focus on the relevant testing and simplify material
12. Implications and limitations to GT[c]	Accent was hard to follow at first	Added page with script of video
14. *GBA*[d] page 1	Confused about *GBA* in Ashkenazi Jews, information too technical	Information was cut down and simplified
15. *GBA* page 2	Information too technical	Simplified information
16. *GBA* page 3	Want more relevance to PD[e]	Clarified the relevance of *GBA* to PD
17. *LRRK2*[f] page 1	Information too technical	Simplified information
18. *LRRK2* page 2	Information too technical	Simplified information

[a]IMAGINE-PD: Interactive Multimedia Approach to Genetic Counseling to Inform and Educate in Parkinson's Disease.
[b]VUS: variant of uncertain significance.
[c]GT: genetic testing.
[d]GBA: glucocerebrosidase.
[e]PD: Parkinson disease.
[f]LRRK2: leucine-rich repeat kinase 2.

Table 6. Summary of function changes made in response to usability testing.

Page	Function feedback	Changes made
1. Title page	Did not know what to do	Clarified next steps, enlarged arrow to move to next page
2. What is IMAGINE-PD[a]	Hovering function was not compatible with mobile device	Made function easier to use with mobile device
3. What is PD[b]	Want a progress bar to see the length of the website	Added a progress bar to the website
4. PD diagnosis or treatment	In page slides were confusing	Removed in page slides
8. What is GT[c]	Page incompatible with mobile	Improved mobile compatibility
9. When to consider GT	Hard to navigate between external resources page	Provide instructions on navigation

[a]IMAGINE-PD: Interactive Multimedia Approach to Genetic Counseling to Inform and Educate in Parkinson's Disease.

[b]PD: Parkinson disease.

[c]GT: genetic testing.

Table 7. Summary of appearance changes made in response to usability testing.

Page	Appearance feedback	Changes made
1. Title page	Make title acronym clearer and change font color	Title acronym was made more obvious and changed font color
4. PD[a] diagnosis or treatment	Some of the text was too small	Increased text font
7. VUS[b]	Graphic was confusing	Clarified graphic
9. When to consider GT[c]	Want more diverse pictures	Included more diverse pictures
10. Benefits and risks	Colors were too vibrant and hurt eyes	Toned down color scheme
15. *GBA*[d] page 2	Did not like picture	Changed the picture
20. Conclusion	Did not like red font	Changed font color

[a]PD: Parkinson disease.

[b]VUS: variant of uncertain significance.

[c]GT: genetic testing.

[d]GBA: glucocerebrosidase.

The same 3 themes were again identified on usability testing qualitative analysis: content, function, and appearance. Changes were made on 10 pages in response to content feedback, 6 pages in response to function feedback, and 7 pages in response to appearance feedback. A summary of the usability qualitative analysis and recommendations for change are shown in Tables 5-7.

Discussion

Principal Results

In this study, we evaluated a web-based genetics education tool for PD using evidence-based research methods to refine the content and conduct usability testing. First, content was developed based on expert opinion, and a high-fidelity prototype was created. Next, user testing was conducted through structured interviews with MDS and patients with PD to evaluate website content. Subsequently, usability testing was conducted via structured interviews with patients with PD. Using qualitative data analysis in both phases, we identified 3 domains of feedback (content, function, and appearance) and addressed feedback by incorporating changes to IMAGINE-PD to create a final version.

Comparison With Prior Work

Some studies in PD and Alzheimer disease have used alternative media forms as pretest education tools. For instance, the PDGENEration study, which offers genetic testing and counseling for PD, provides a pretest education tool that is a prerecorded video covering essential topics in PD genetics and genetic testing that was created by experts in neurogenetics and genetic counseling for PD (ClinicalTrials.gov NCT04057794). Additionally, the Alzheimer's Prevention Initiative, which conducts apolipoprotein E testing in cognitively unimpaired people 60 years or older, uses a self-directed learning technique providing a brochure and a video covering content typically addressed in a pretest counseling session coupled with multiple-choice questions to reinforce learning [28]. To our knowledge, neither approach has undergone usability testing to incorporate the input of the end user as we demonstrate in this study.

Beyond neurogenetics, these results can also be viewed in the context of alternative genetic education tools, where more robust

efforts to develop alternate education and disclosure methods are underway. A novel, web-based genetics education for a polygenic risk score of alcohol use disorders was evaluated in a randomized clinical trial of 325 college students. The tool was shown to improve user knowledge compared to general alcohol-use education alone [29]. In another study, a tool named Decision-Aid and E-Counselling for Inherited Disorder Evaluation [30] was developed to educate about genome-wide sequencing. Decision-Aid and E-Counselling for Inherited Disorder Evaluation was compared to pretest genetic counseling with a counselor; the genetics education methods were equivalent in conveying knowledge and were highly satisfactory to participants [31]. In the ongoing Communication and Education in Tumor Profiling and the Returning Genetic Research Panel Results for Breast Cancer Susceptibility studies, a web-based educational tool for pretest education was developed. These served as a guide for user testing and usability testing of IMAGINE-PD [32,33]. Furthermore, in the Study of an eHealth Delivery Alternative for Cancer Genetic Testing for Hereditary Predisposition in Metastatic Cancer Patients, web-based alternatives to traditional provider–mediated counseling and results disclosure are being evaluated. The outcomes of these studies will be informative for understanding the use of web-based genetic education tools even beyond inherited cancer syndromes. However, differences in the target populations, genetic testing performed, and the implications of the genetic test results between IMAGINE-PD and the Communication and Education in Tumor Profiling, Returning Genetic Research Panel Results for Breast Cancer Susceptibility, and Study of an eHealth Delivery Alternative for Cancer Genetic Testing for Hereditary Predisposition in Metastatic Cancer Patients studies necessitated the rigorous user testing and usability testing that we report here.

This study has several strengths that should be noted. First, the content for IMAGINE-PD was developed by experts in genetic counseling in PD, neurogenetics experts, and movement disorder physicians. Second, user testing included referring providers for neurogenetic services for patients with PD (movement disorder physicians) and the intended end user (patients with PD). Third, this evaluation followed the DHHS guidelines for user testing and usability testing, nearly doubling the recommended sample size in each phase for this type of research.

Limitations

Some limitations of this study should be acknowledged. First, the content of IMAGINE-PD is focused on targeted variant testing in *GBA* and *LRRK2*, limiting its scope and generalizability to other PD genetic testing. This was intentional to match the research-based *GBA* and *LRRK2* screening being performed in the MIND Initiative at UPenn [17]. IMAGINE-PD may serve as the genetics education tool for MIND participants. Additionally, the MIND Initiative conducts the same genetic test for all involved study participants. As a result, the pretest education in the IMAGINE-PD tool deviates from a typical pretest genetic counseling session that would involve obtaining a family history and making decisions about test choices (family variant testing, multigene panel testing, exome, or genome sequencing). Instead, this is a scalable approach to screen everyone in the UPenn PD clinic for variants within the 2 most common genes associated with PD and identify potentially eligible participants for clinical trials enrolling carriers of variants in *GBA* or *LRRK2*. During a separate disclosure visit for *GBA* or *LRRK2*, a personal and family history could be reviewed in detail, and additional testing could be pursued afterward if indicated. Additionally, specialized content could be developed and added to subsequent versions of IMAGINE-PD to accommodate other specific types of diagnostic genetic testing. Second, all participants (physicians and patients) had a high level of education and experience with technology and computer use and interest in using a web-based pretest education tool, which probably does not capture the breadth of patients with PD and may overestimate the user experience. Ongoing evaluation of this tool will be necessary to determine which patients will be able to successfully use a web-based educational platform and who would be better served by other methods such as in-person or live telemedicine counseling with a genetic counselor. Although patients with PD were not involved in content creation, the content was derived from principals based on genetic counseling expertise. Patient feedback on content was solicited during both user and usability testing, where participants were given free answer choices to provide input on any additional topics they would want to have included. In the future, supplementary content could be developed to address deficits in patient or family-member comprehension, low literacy or education level, identifying potential psychosocial concerns to prompt additional counseling or comfort with technology. Making this tool accessible while in clinic on a smartphone, tablet, or computer screen may help to address limitations in access to technology. Third, participants elected to be enrolled in a genetic biobanking study and therefore may not represent the community with PD more broadly.

Conclusions

In summary, we present our findings from user testing and usability testing for the development of IMAGINE-PD, a web-based pretest genetic education tool. We describe a phased review and iterative process of refining the content, appearance, and functionality based on expert review as well as physician and patient feedback according to DHHS guidelines. The final version [24], which can be made available by request to the authors, will undergo further evaluation to compare it to standard telegenetic counseling with a genetic counselor (ClinicalTrials.gov NCT04527146) measuring satisfaction, impact, and comprehension. As a web-based learning tool accessible by internet, IMAGINE-PD has the potential to improve access to neurogenetic services for patients with PD interested in learning about their eligibility for *LRRK2*- or *GBA*-directed clinical trials.

Acknowledgments

The authors would like to acknowledge patients for their generous participation in this study, the clinical research associates at the UPenn, and the expert reviewers of the preliminary material. The authors would also like to thank Dr Judy Shea, at the UPenn who provided guidance on qualitative data analysis, Susan Paolin who created the animations, and Stephen Durborow who created the website. Support and funding were provided by the Penn Center for Precision Medicine, the Department of Neurology at the UPenn, The Parkinson Council, and the Parker Family Chair (ACP).

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author upon reasonable request. All deidentified data can be made available upon request of the authors.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplemental tables and supplemental questionnaires for user and usability testing of IMAGINE-PD.
[PDF File (Adobe PDF File), 1357 KB - bioinform_v4i1e45370_app1.pdf]

Multimedia Appendix 2
Screenshots of all web pages in IMAGINE-PD. IMAGINE-PD: Interactive Multimedia Approach to Genetic Counseling to Inform and Educate in Parkinson’s Disease.
[PDF File (Adobe PDF File), 5807 KB - bioinform_v4i1e45370_app2.pdf]

References

1. Marras C, Beck JC, Bower JH, Roberts E, Ritz B, Ross GW, Parkinson’s Foundation P4 Group. Prevalence of Parkinson's disease across North America. NPJ Parkinsons Dis 2018;4:21 [FREE Full text] [doi: 10.1038/s41531-018-0058-0] [Medline: 30003140]
2. GBD 2016 Parkinson's Disease Collaborators. Global, regional, and national burden of Parkinson's disease, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet Neurol 2018 Nov;17(11):939-953 [FREE Full text] [doi: 10.1016/S1474-4422(18)30295-3] [Medline: 30287051]
3. Kitada T, Asakawa S, Hattori N, Matsumine H, Yamamura Y, Minoshima S, et al. Mutations in the parkin gene cause autosomal recessive juvenile parkinsonism. Nature 1998 Apr 09;392(6676):605-608. [doi: 10.1038/33416] [Medline: 9560156]
4. Valente EM, Bentivoglio AR, Dixon PH, Ferraris A, Ialongo T, Frontali M, et al. Localization of a novel locus for autosomal recessive early-onset parkinsonism, PARK6, on human chromosome 1p35-p36. Am J Hum Genet 2001 Apr;68(4):895-900 [FREE Full text] [doi: 10.1086/319522] [Medline: 11254447]
5. Bonifati V, Rizzu P, Squitieri F, Krieger E, Vanacore N, van Swieten JC, et al. DJ-1(PARK7), a novel gene for autosomal recessive, early onset parkinsonism. Neurol Sci 2003 Oct;24(3):159-160. [doi: 10.1007/s10072-003-0108-0] [Medline: 14598065]
6. Polymeropoulos MH, Lavedan C, Leroy E, Ide SE, Dehejia A, Dutra A, et al. Mutation in the alpha-synuclein gene identified in families with Parkinson's disease. Science 1997 Jun 27;276(5321):2045-2047. [doi: 10.1126/science.276.5321.2045] [Medline: 9197268]
7. Lesage S, Brice A. Parkinson's disease: from monogenic forms to genetic susceptibility factors. Hum Mol Genet 2009 Apr 15;18(R1):R48-R59 [FREE Full text] [doi: 10.1093/hmg/ddp012] [Medline: 19297401]
8. Aharon-Peretz J, Rosenbaum H, Gershoni-Baruch R. Mutations in the glucocerebrosidase gene and Parkinson's disease in Ashkenazi Jews. N Engl J Med 2004 Nov 04;351(19):1972-1977 [FREE Full text] [doi: 10.1056/NEJMoa033277] [Medline: 15525722]
9. Sidransky E, Lopez G. The link between the GBA gene and parkinsonism. Lancet Neurol 2012 Nov;11(11):986-998 [FREE Full text] [doi: 10.1016/S1474-4422(12)70190-4] [Medline: 23079555]
10. Norquist BM, Harrell MI, Brady MF, Walsh T, Lee MK, Gulsuner S, et al. Inherited mutations in women with ovarian carcinoma. JAMA Oncol 2016 Apr;2(4):482-490 [FREE Full text] [doi: 10.1001/jamaoncol.2015.5495] [Medline: 26720728]
11. Ma H, Brosens LAA, Offerhaus GJA, Giardiello FM, de Leng WWJ, Montgomery EA. Pathology and genetics of hereditary colorectal cancer. Pathology 2018 Jan;50(1):49-59. [doi: 10.1016/j.pathol.2017.09.004] [Medline: 29169633]

12. Alcalay RN, Kehoe C, Shorr E, Battista R, Hall A, Simuni T, et al. Genetic testing for Parkinson disease: current practice, knowledge, and attitudes among US and Canadian movement disorders specialists. Genet Med 2020 Mar;22(3):574-580 [FREE Full text] [doi: 10.1038/s41436-019-0684-x] [Medline: 31680121]
13. Mullin S, Smith L, Lee K, D'Souza G, Woodgate P, Elflein J, et al. Ambroxol for the treatment of patients with Parkinson disease with and without glucocerebrosidase gene mutations: a nonrandomized, noncontrolled trial. JAMA Neurol 2020 Apr 01;77(4):427-434 [FREE Full text] [doi: 10.1001/jamaneurol.2019.4611] [Medline: 31930374]
14. Our pipeline. Denali Therapeutics. URL: https://www.denalitherapeutics.com/pipeline [accessed 2023-07-14]
15. Falcone DC, Wood EM, Xie SX, Siderowf A, Van Deerlin VM. Genetic testing and Parkinson disease: assessment of patient knowledge, attitudes, and interest. J Genet Couns 2011 Aug;20(4):384-395 [FREE Full text] [doi: 10.1007/s10897-011-9362-0] [Medline: 21476119]
16. Gupte M, Alcalay RN, Mejia-Santana H, Raymond D, Saunders-Pullman R, Roos E, et al. Interest in genetic testing in Ashkenazi Jewish Parkinson's disease patients and their unaffected relatives. J Genet Couns 2015 Apr;24(2):238-246 [FREE Full text] [doi: 10.1007/s10897-014-9756-x] [Medline: 25127731]
17. Tropea TF, Amari N, Han N, Rick J, Suh E, Akhtar RS, et al. Whole clinic research enrollment in Parkinson's disease: the molecular integration in neurological diagnosis (MIND) study. J Parkinsons Dis 2021;11(2):757-765 [FREE Full text] [doi: 10.3233/JPD-202406] [Medline: 33492247]
18. Cook L, Schulze J, Kopil C, Hastings T, Naito A, Wojcieszek J, et al. Genetic testing for Parkinson disease: are we ready? Neurol Clin Pract 2021 Feb;11(1):69-77 [FREE Full text] [doi: 10.1212/CPJ.0000000000000831] [Medline: 33968475]
19. Rubanovich CK, Cheung C, Mandel J, Bloss CS. Physician preparedness for big genomic data: a review of genomic medicine education initiatives in the United States. Hum Mol Genet 2018 Aug 01;27(R2):R250-R258 [FREE Full text] [doi: 10.1093/hmg/ddy170] [Medline: 29750248]
20. Research-based web design & usability guidelines. United States General Services Administration. URL: https://www.usability.gov/sites/default/files/documents/guidelines_book.pdf [accessed 2023-07-15]
21. Wix. URL: https://www.wix.com/ [accessed 2023-07-14]
22. SP Motion Design. URL: https://www.spmotiondesign.com/ [accessed 2023-07-14]
23. Readable. URL: https://readable.com/ [accessed 2023-07-21]
24. IMAGINE-PD. Penn Medicine. URL: http://www.pennmedicine.org/imaginepd [accessed 2023-07-14]
25. Baker L, Wagner TH, Singer S, Bundorf MK. Use of the Internet and e-mail for health care information: results from a national survey. JAMA 2003 May 14;289(18):2400-2406 [FREE Full text] [doi: 10.1001/jama.289.18.2400] [Medline: 12746364]
26. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381 [FREE Full text] [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
27. Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O'Neal L, REDCap Consortium. The REDCap consortium: building an international community of software platform partners. J Biomed Inform 2019 Jul;95:103208 [FREE Full text] [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
28. Langlois CM, Bradbury A, Wood EM, Roberts JS, Kim SYH, Riviere ME, et al. Alzheimer's Prevention Initiative Generation Program: development of an genetic counseling and disclosure process in the context of clinical trials. Alzheimers Dement (N Y) 2019;5:705-716 [FREE Full text] [doi: 10.1016/j.trci.2019.09.013] [Medline: 31921963]
29. Driver MN, Kuo SIC, Petronio L, Brockman D, Dron JS, Austin J, et al. Evaluating the impact of a new educational tool on understanding of polygenic risk scores for alcohol use disorder. Front Psychiatry 2022;13:1025483 [FREE Full text] [doi: 10.3389/fpsyt.2022.1025483] [Medline: 36506445]
30. Birch P, Adam S, Bansback N, Coe RR, Hicklin J, Lehman A, et al. DECIDE: a decision support tool to facilitate parents' choices regarding genome-wide sequencing. J Genet Couns 2016;25(6):1298-1308. [doi: 10.1007/s10897-016-9971-8] [Medline: 27211035]
31. Adam S, Birch PH, Coe RR, Bansback N, Jones AL, Connolly MB, et al. Assessing an interactive online tool to support parents' genomic testing decisions. J Genet Couns 2018 Jul 23;28(1):10-17 [FREE Full text] [doi: 10.1007/s10897-018-0281-1] [Medline: 30033481]
32. Bradbury AR, Lee JW, Gaieski JB, Li S, Gareen IF, Flaherty KT, et al. A randomized study of genetic education versus usual care in tumor profiling for advanced cancer in the ECOG-ACRIN Cancer Research Group (EAQ152). Cancer 2022 Apr 01;128(7):1381-1391 [FREE Full text] [doi: 10.1002/cncr.34063] [Medline: 34890045]
33. Gaieski JB, Patrick-Miller L, Egleston BL, Maxwell KN, Walser S, DiGiovanni L, et al. Research participants' experiences with return of genetic research results and preferences for web-based alternatives. Mol Genet Genomic Med 2019 Sep;7(9):e898 [FREE Full text] [doi: 10.1002/mgg3.898] [Medline: 31376244]

Abbreviations

DHHS: US Department of Health and Human Services
DJ-1: Parkinsonism-associated deglycase

GBA: glucocerebrosidase
GT: genetic testing
IMAGINE-PD: Interactive Multimedia Approach to Genetic Counseling to Inform and Educate in Parkinson's Disease
IRB: institutional review board
LRRK2: leucine-rich repeat kinase 2
MDS: movement disorder specialist
MIND: Molecular Integration in Neurological Diagnosis
PD: Parkinson disease
Pink1: PTEN-induced kinase 1
REDCap: Research Electronic Data Capture
SNCA: α-synuclein
UPenn: University of Pennsylvania
VPS35: VPS35 retromer complex component

Edited by E Uzun; submitted 28.12.22; peer-reviewed by M Nance, R Schneider, K Kaphingst; comments to author 16.02.23; revised version received 16.03.23; accepted 06.07.23; published 30.08.23.

Please cite as:
Han N, Paul RA, Bardakjian T, Kargilis D, Bradbury AR, Chen-Plotkin A, Tropea TF
User and Usability Testing of a Web-Based Genetics Education Tool for Parkinson Disease: Mixed Methods Study
JMIR Bioinform Biotech 2023;4:e45370
URL: https://bioinform.jmir.org/2023/1/e45370
doi:10.2196/45370
PMID:

Editorial

Introducing JMIR Bioinformatics and Biotechnology: A Platform for Interdisciplinary Collaboration and Cutting-Edge Research

Ece Dilber Gamsiz Uzun[1,2], MSci, PhD

[1]Department of Pathology and Laboratory Medicine, Rhode Island Hospital, Providence, RI, United States

[2]Department of Pathology and Laboratory Medicine, Brown University, Providence, RI, United States

Corresponding Author:
Ece Dilber Gamsiz Uzun, MSci, PhD
Department of Pathology and Laboratory Medicine
Rhode Island Hospital
593 Eddy Street
Providence, RI, 02903
United States
Email: dilber_gamsiz@brown.edu

Abstract

JMIR Bioinformatics and Biotechnology supports interdisciplinary research and welcomes contributions that push the boundaries of bioinformatics, genomics, artificial intelligence, and pathology informatics.

(JMIR Bioinform Biotech 2023;4:e48631) doi:10.2196/48631

KEYWORDS

bioinformatics; biotechnology; artificial intelligence; genomic; informatics; interdisciplinary research

Introduction

Bioinformatics is a rapidly evolving field that has transformed the way we study and understand biological systems. With the advent of large-scale genomic data and advances in computational tools and algorithms, we are now able to discover hidden patterns in biological data. One of the key areas where bioinformatics is making a significant impact is in the detection and interpretation of genomic variations. Genomic variations can have important implications for disease susceptibility, drug response, and other biological processes. With the help of advanced algorithms and tools, researchers can now detect genomic variations with high accuracy and precision. This has opened new avenues for drug discovery and precision medicine [1].

The rapid advances in artificial intelligence (AI) applications in the fields of genomics and pathology informatics have led the development of AI-based models for disease risk prediction and drug discovery. AI algorithms can be trained to analyze large-scale genomic data and identify hidden patterns. This has led to significant improvements in disease diagnosis, prognosis, and treatment. AI-based tools can now identify genetic markers that are associated with specific diseases, such as cancer, and help clinicians select the most effective treatment options [2-4].

Network biology is another area where bioinformatics is making significant advances. By analyzing large-scale genomic data sets, researchers can identify key pathways, protein-protein interactions, and networks that are involved in disease pathogenesis. This information can aid in the development of new drugs and therapies [5,6]. Genomic data visualization has led to new and innovative ways for researchers to gain insights into the structure and function of biological systems. This has important implications for understanding disease mechanisms, as well as for developing new diagnostic and therapeutic tools [7]. *JMIR Bioinformatics and Biotechnology* will support the development of new bioinformatics analysis tools, novel algorithms, advanced AI-based predictive models, and network biology studies. We would like to foster not only basic science and algorithm development but also translational research studies in the focus areas described above. We will also explore the use of new technologies for drug discovery and therapy development in cancer and other complex disorders.

The Scope of JMIR Bioinformatics and Biotechnology

JMIR Bioinformatics and Biotechnology aims to publish cutting-edge research in the fields of bioinformatics, genomics, and pathology informatics. The scope of the journal includes the development and application of genomic variation detection algorithms and tools including single-cell sequencing and spatial transcriptomics; AI-based predictive models; pathology informatics, including image analysis; mathematical modeling

in biological systems, including drug delivery and discovery; genomic data visualization; network biology; and cancer genomic data analysis. *JMIR Bioinformatics and Biotechnology* will be a platform for interdisciplinary collaborations between bioinformaticians, biologists, computer scientists, mathematicians, and clinicians to address the challenges of integrating large-scale genomic data with clinical and pathological information. The journal welcomes original research articles, review articles, and perspectives, as well as submissions on methodological advances and computational tools in these areas.

Although we welcome translational research studies, *JMIR Bioinformatics and Biotechnology* will not consider manuscripts describing medical informatics–related projects without a focus on bioinformatics or genomics. We aim to focus on the bioinformatics applications within the medical informatics field. We welcome you to *JMIR Bioinformatics and Biotechnology* and hope that you will consider contributing to the fast-moving bioinformatics field!

Conflicts of Interest

EDGU is the Editor-in-Chief of *JMIR Bioinformatics and Biotechnology*.

References

1. Wooller S, Benstead-Hume G, Chen X, Ali Y, Pearl FMG. Bioinformatics in translational drug discovery. Biosci Rep 2017 Aug 31;37(4) [FREE Full text] [doi: 10.1042/BSR20160180] [Medline: 28487472]
2. Shimizu H, Nakayama KI. Artificial intelligence in oncology. Cancer Sci 2020 May 21;111(5):1452-1460 [FREE Full text] [doi: 10.1111/cas.14377] [Medline: 32133724]
3. Tran KA, Kondrashova O, Bradley A, Williams ED, Pearson JV, Waddell N. Deep learning in cancer diagnosis, prognosis and treatment selection. Genome Med 2021 Sep 27;13(1):152 [FREE Full text] [doi: 10.1186/s13073-021-00968-x] [Medline: 34579788]
4. Alam MR, Abdul-Ghafar J, Yim K, Thakur N, Lee SH, Jang H, et al. Recent applications of artificial intelligence from histopathologic image-based prediction of microsatellite instability in solid cancers: a systematic review. Cancers (Basel) 2022 May 24;14(11):2590 [FREE Full text] [doi: 10.3390/cancers14112590] [Medline: 35681570]
5. Menche J, Sharma A, Kitsak M, Ghiassian SD, Vidal M, Loscalzo J, et al. Disease networks. Uncovering disease-disease relationships through the incomplete interactome. Science 2015 Feb 20;347(6224):1257601 [FREE Full text] [doi: 10.1126/science.1257601] [Medline: 25700523]
6. Barabási AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet 2011 Jan 17;12(1):56-68 [FREE Full text] [doi: 10.1038/nrg2918] [Medline: 21164525]
7. Dunn W, Burgun A, Krebs M, Rance B. Exploring and visualizing multidimensional data in translational research platforms. Brief Bioinform 2017 Nov 01;18(6):1044-1056 [FREE Full text] [doi: 10.1093/bib/bbw080] [Medline: 27585944]

Abbreviations

AI: artificial intelligence

Edited by T Leung; submitted 01.05.23; this is a non–peer-reviewed article;accepted 04.05.23; published 12.06.23.

Please cite as:
Gamsiz Uzun ED
Introducing JMIR Bioinformatics and Biotechnology: A Platform for Interdisciplinary Collaboration and Cutting-Edge Research
JMIR Bioinform Biotech 2023;4:e48631
URL: https://bioinform.jmir.org/2023/1/e48631
doi:10.2196/48631
PMID:

Original Paper

The Differentially Expressed Genes Responsible for the Development of T Helper 9 Cells From T Helper 2 Cells in Various Disease States: Immuno-Interactomics Study

Manoj Khokhar[1], MSc; Purvi Purohit[1], MSc, PhD; Ashita Gadwal[1], MSc; Sojit Tomo[1], MD, DNB; Nitin Kumar Bajpai[2], MD, DM; Ravindra Shukla[3], MD, DM

[1]Department of Biochemistry, All India Institute of Medical Sciences Jodhpur, Jodhpur, India

[2]Department of Nephrology, All India Institute of Medical Sciences Jodhpur, Jodhpur, India

[3]Department of Endocrinology and Metabolism, All India Institute of Medical Sciences Jodhpur, Jodhpur, India

Corresponding Author:
Purvi Purohit, MSc, PhD
Department of Biochemistry
All India Institute of Medical Sciences Jodhpur
Basni Industrial Area Phase-2
Jodhpur, 342005
India
Phone: 91 09928388223
Email: dr.purvipurohit@gmail.com

Abstract

Background: T helper (Th) 9 cells are a novel subset of Th cells that develop independently from Th2 cells and are characterized by the secretion of interleukin (IL)-9. Studies have suggested the involvement of Th9 cells in variable diseases such as allergic and pulmonary diseases (eg, asthma, chronic obstructive airway disease, chronic rhinosinusitis, nasal polyps, and pulmonary hypoplasia), metabolic diseases (eg, acute leukemia, myelocytic leukemia, breast cancer, lung cancer, melanoma, pancreatic cancer), neuropsychiatric disorders (eg, Alzheimer disease), autoimmune diseases (eg, Graves disease, Crohn disease, colitis, psoriasis, systemic lupus erythematosus, systemic scleroderma, rheumatoid arthritis, multiple sclerosis, inflammatory bowel disease, atopic dermatitis, eczema), and infectious diseases (eg, tuberculosis, hepatitis). However, there is a dearth of information on its involvement in other metabolic, neuropsychiatric, and infectious diseases.

Objective: This study aims to identify significant differentially altered genes in the conversion of Th2 to Th9 cells, and their regulating microRNAs (miRs) from publicly available Gene Expression Omnibus data sets of the mouse model using in silico analysis to unravel various pathogenic pathways involved in disease processes.

Methods: Using differentially expressed genes (DEGs) identified from 2 publicly available data sets (GSE99166 and GSE123501) we performed functional enrichment and network analyses to identify pathways, protein-protein interactions, miR-messenger RNA associations, and disease-gene associations related to significant differentially altered genes implicated in the conversion of Th2 to Th9 cells.

Results: We extracted 260 common downregulated, 236 common upregulated, and 634 common DEGs from the expression profiles of data sets GSE99166 and GSE123501. Codifferentially expressed ILs, cytokines, receptors, and transcription factors (TFs) were enriched in 7 crucial Kyoto Encyclopedia of Genes and Genomes pathways and Gene Ontology. We constructed the protein-protein interaction network and predicted the top regulatory miRs involved in the Th2 to Th9 differentiation pathways. We also identified various metabolic, allergic and pulmonary, neuropsychiatric, autoimmune, and infectious diseases as well as carcinomas where the differentiation of Th2 to Th9 may play a crucial role.

Conclusions: This study identified hitherto unexplored possible associations between Th9 and disease states. Some important ILs, including *CCL1* (chemokine [C-C motif] ligand 1), *CCL20* (chemokine [C-C motif] ligand 20), *IL-13*, *IL-4*, *IL-12A*, and *IL-9*; receptors, including *IL-12RB1*, *IL-4RA* (interleukin 9 receptor alpha), *CD53* (cluster of differentiation 53), *CD6* (cluster of differentiation 6), *CD5* (cluster of differentiation 5), *CD83* (cluster of differentiation 83), *CD197* (cluster of differentiation 197), *IL-1RL1* (interleukin 1 receptor-like 1), *CD101* (cluster of differentiation 101), *CD96* (cluster of differentiation 96), *CD72* (cluster of differentiation 72), *CD7* (cluster of differentiation 7), *CD152* (cytotoxic T lymphocyte–associated protein 4), *CD38* (cluster of differentiation 38), *CX3CR1* (chemokine [C-X3-C motif] receptor 1), *CTLA2A* (cytotoxic T lymphocyte–associated protein 2

alpha), *CTLA28*, and *CD196* (cluster of differentiation 196); and TFs, including *FOXP3* (forkhead box P3), *IRF8* (interferon regulatory factor 8), *FOXP2* (forkhead box P2), *RORA* (RAR-related orphan receptor alpha), *AHR* (aryl-hydrocarbon receptor), *MAF* (avian musculoaponeurotic fibrosarcoma oncogene homolog), *SMAD6* (SMAD family member 6), *JUN* (Jun proto-oncogene), *JAK2* (Janus kinase 2), *EP300* (E1A binding protein p300), *ATF6* (activating transcription factor 6), *BTAF1* (B-TFIID TATA-box binding protein associated factor 1), *BAFT* (basic leucine zipper transcription factor), *NOTCH1* (neurogenic locus notch homolog protein 1), *GATA3* (GATA binding protein 3), *SATB1* (special AT-rich sequence binding protein 1), *BMP7* (bone morphogenetic protein 7), and *PPARG* (peroxisome proliferator–activated receptor gamma, were able to identify significant differentially altered genes in the conversion of Th2 to Th9 cells. We identified some common miRs that could target the DEGs. The scarcity of studies on the role of Th9 in metabolic diseases highlights the lacunae in this field. Our study provides the rationale for exploring the role of Th9 in various metabolic disorders such as diabetes mellitus, diabetic nephropathy, hypertensive disease, ischemic stroke, steatohepatitis, liver fibrosis, obesity, adenocarcinoma, glioblastoma and glioma, malignant neoplasm of stomach, melanoma, neuroblastoma, osteosarcoma, pancreatic carcinoma, prostate carcinoma, and stomach carcinoma.

(JMIR Bioinform Biotech 2023;4:e42421) doi:10.2196/42421

KEYWORDS

Th9 cells; Th2 cells; autoimmune diseases; DEGs; interleukins

Introduction

$CD4^+$ T helper (Th) cells have been classified into different subsets based on the cytokine profile that each subset secretes and their distinct role in regulating immunity and inflammation. Previous studies have shown that immune cells play a role in various metabolic [1-3] and infectious [3-7] diseases. Th9 cells are a subset of $CD4^+$ Th cells that develop from naïve T cells and release interleukin (IL)-9. The generation of Th9 cells from naïve Th0 cells requires a Th2 state as an intermediate. While both Th2 and Th9 cells express *PU.1* (spleen focus forming virus [SFFV] proviral integration oncogenes), *IRF4* (interferon regulatory factor 4), and *GATA3* (GATA binding protein 3), the latter have upregulated expression of *IRF4* and suppressed *PU.1*. The Th2 cells, generated during Th0 cell differentiation, further evolve into Th9 cells in the presence of activated *Smad3/Smad4* and *IRF4* pathways. The prolonged transforming growth factor beta (*TGF*β) stimulation transforms the Th2 cells into Th9 cells and alters the cytokine secretion pattern from an *IL-4*–dominant phenotype to an *IL-9*–dominant one [8]. Th9 cells produce *IL-9*, which is crucial in regulating autoimmune and allergic reactions [9]. Various other cytokines also affect the development of Th9 cells and *IL-9* production. *IL-23* inhibits *IL-9* production, whereas *IL-1* and *IL-33* stimulate the production of *IL-9* in T cells [10,11]. Similarly, *IL-25* stimulates the release of *IL-9* from T cells [12]. In addition, costimulatory receptors, such as *OX40*, have been found to be a stimulant for the development of Th9 cells [13]. Thus, the development of Th9 cells is a result of integrating multiple positive and negative signals in the form of cytokines and costimulation from surface receptors.

Th9 cells can manifest differently in various diseases. Th9 cells have been demonstrated to incite allergic airway disease [14]. Th9 cells have also been implicated in tumor immunity [10]. Interestingly, the evolution of Th2 to Th9 cells does influence the pathophysiology of multiple diseases. The nitric oxide–mediated airway inflammation has been attributed to the inducing effect of nitric oxide on the development of Th9 cells [15]. The tricarboxylic acid cycle metabolite succinate stimulates Th9 cell differentiation and leads to Th9 cell–mediated tumor regression. Similarly, Th9 differentiation resulting from *IL-35* stimulation accentuates the inflammatory process and leads to an immunoglobulin (Ig) class switch toward IgG4 in IgG4-related diseases [16].

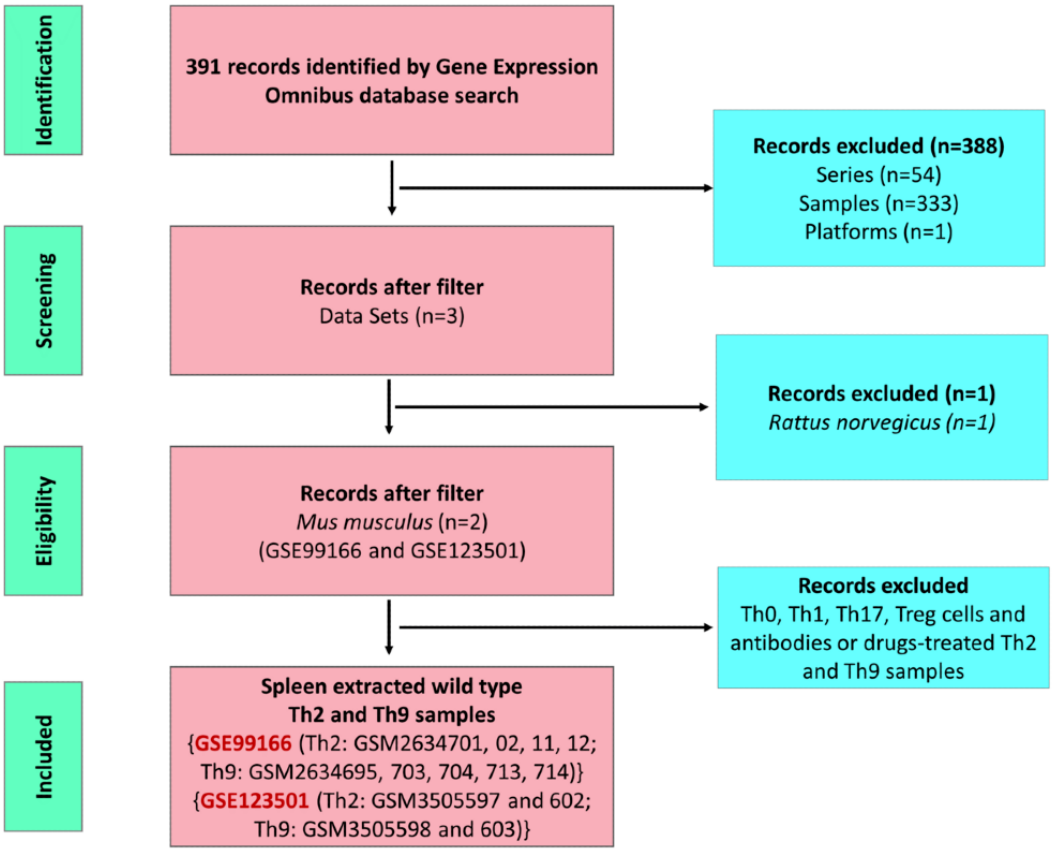
Unfortunately, the experimental approach to Th9 cells has been riddled with difficulty, because a selective deficiency model for Th9 lineage has not yet been defined. In addition, factors needed to develop Th9 cells such as *IL-4* and *IRF4* are required to develop other Th subsets [17]. Our study aimed to compare the transcriptome of Th2 and Th9 cells to identify the pattern of changes in the expression of various genes when the Th2 cells get differentiated into Th9 cells. We also aimed to assess these genes, which are markedly altered in the transition of Th2 to Th9 cells, in various other diseases to enlist the possible diseases in which Th9 cells may play a crucial role.

Methods

Expression Profiling: Gene Expression Omnibus Assay to Data Mining for Th2 to Th9 Cells Differentiation

We performed a search in the Gene Expression Omnibus (GEO) database using several keywords, including “Healthy Control,” “Wild Type,” “Mice,” “*Mus musculus*,” “Th9,” “Th2,” and “Expression profiling by array” from January 1, 2012, to December 17, 2020, and selected 2 gene series expressions (GSEs) data for further study: GSE99166 and GSE123501. GSE99166 contained 4 samples of Th2 wild-type cells (GSM2634701, GSM2634702, GSM2634711, and GSM2634712) and 5 samples of Th9 wild-type cells (GSM2634695, GSM2634703, GSM2634704, GSM2634713, and GSM2634714) from the spleen. GSE123501 contained 2 samples of Th2 wild-type cells (GSM3505597 and GSM3505602) and another 2 samples of Th9 wild-type cells (GSM350598 and GSM3505603) from the spleen (Figure 1).

Figure 1. Flow diagram illustrating the data collection process and the number of data sets considered for inclusion. Th: T helper; Treg: T regulatory cell.

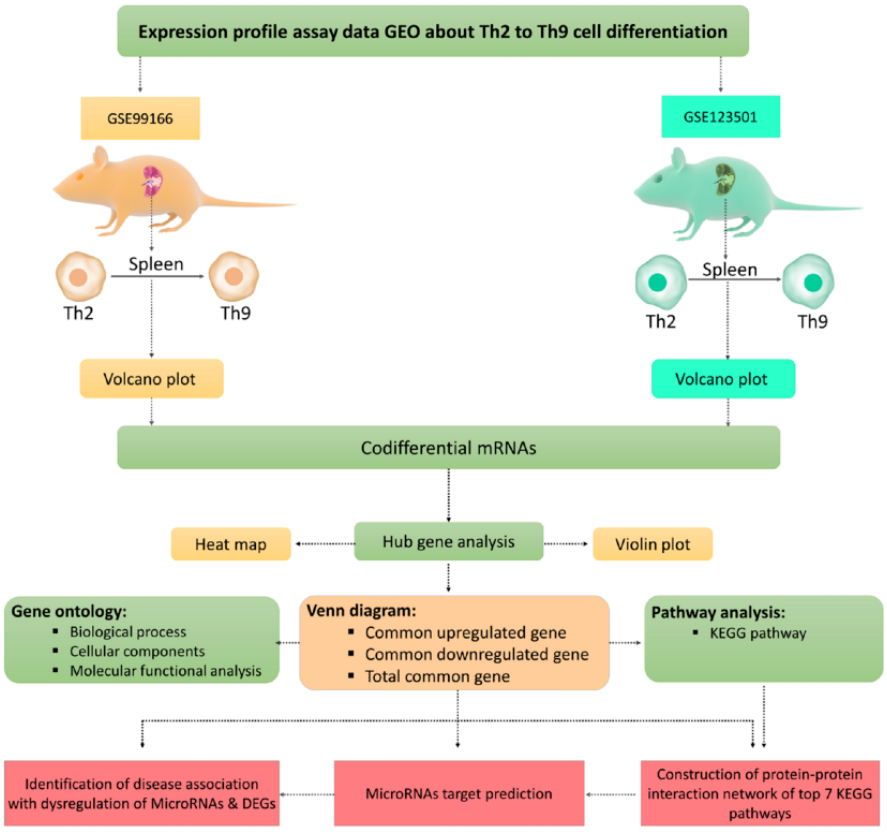


Assortment and Identification of Codifferentially Expressed Messenger RNAs From the Spleen (2 Different) Data Sets

The differentially expressed genes (DEGs) were obtained from the 13 samples of 2 different data sets (GSE99166 and GSE123501) using the GREIN (GEO RNA-seq Experiments Interactive Navigator) platform (BD2K-LINCS Data Coordination and Integration Center). This interactive online web tool analyses GEO RNA-seq data [18]. The DEGs extracted from the data sets comprised genes from Th2 and Th9 cells. As we wanted to assess the alteration of genes during the conversion of Th2 to Th9 cells, the analysis was performed with DEGs of Th2 cells as the standard to which DEGs of Th9 cells were compared. The workflow for the data processing and analysis is portrayed in Figure 2.

The DEGs were considered upregulated when the expression of genes in Th9 cells was higher than that in Th2 cells. The cutoff for the selection was kept at $P<.05$, and overlapping DEGs between 2 data sets (GSE99166 and GSE123501) on comparison of Th2 and Th9 cells were identified by the Venn diagram tool [19,20]. In addition, the common upregulated, downregulated, and oppositely regulated DEGs of these 2 data sets (GSE99166 and GSE123501) were identified. The fold change expression distribution was visualized by a heat map and violin plot using the Linear Models for the Microarray Data (limma) Package of R (R Foundation for Statistical Computing) and Orange Data Mining (University of Ljubljana) [21,22].

Figure 2. Flowchart of the data processing and analysis for Th2 to Th9 cells differentiation. DEG: differentially expressed gene; KEGG: Kyoto Encyclopedia of Genes and Genomes; mRNA: messenger RNA; Th: T helper.



Functional Enrichment of Gene Ontology for Common, Regulated DEGs

The codifferential genes were divided into 3 parts, namely, (1) common upregulated, (2) common downregulated, and (3) common, oppositely regulated. The top ranked ontological features of all DEGs were analyzed with STRING. The Gene Ontology (GO) terms included the following 3 categories: biological processes, cellular components, and molecular functions. The significant GO terms regulating genes are presented in a radar graph with a negative log10 (false discovery rate). We defined $P<.05$ as a significant value.

Kyoto Encyclopedia of Genes and Genomes Pathway Analysis of Top Ranked Significant, Common, Regulated DEGs

We searched the functionally significant Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways for top ranked significantly altered DEGs using the STRING and WikiPathways databases. We identified important genes participating in each pathway, and selected the top 7 pathways based on negative log10 (false discovery rate) and P values (<.05) that were important for further study.

Genes Assortment and Construction of a Protein-Protein Interaction Network of the Top Enriched Pathways

We downloaded the complete gene list of the top ranked 7 individual pathways with an interaction network from the KEGG database. We revisualized and constructed the pathway with the help of Cytoscape (Cytoscape Team/Institute for Systems Biology; an open-source software platform for visualizing complex networks and integrating these with any type of attribute data) [23] and marked the DEGs that play a significant role in the differentiation of Th2 to Th9 cells.

Identification of Top Regulatory MicroRNAs Involved in the Th2 to Th9 Differentiation Pathways

The top 10 microRNAs (miRs) that targeted the hub genes were predicted by the well-established miR target prediction database miRNet version 22.0 [24], with special emphasis on the selected organism. Default values for the degree of interaction and betweenness were selected. Common miRs and their targeted messenger RNAs (mRNAs) of all groups were sorted by the Venn diagram.

Construction of a Gene-Disease–Based Genomic Pathway Interaction Network

The DEGs that were identified to play a significant role in Th2 to Th9 differentiation were further analyzed for their involvement in various pathways pertaining to specific diseases using DisGeNET (IBI Group) [25], a discovery platform that describes genes, transcription factors (TFs), chemokines, and IL in association with various specific diseases.

Ethical Considerations

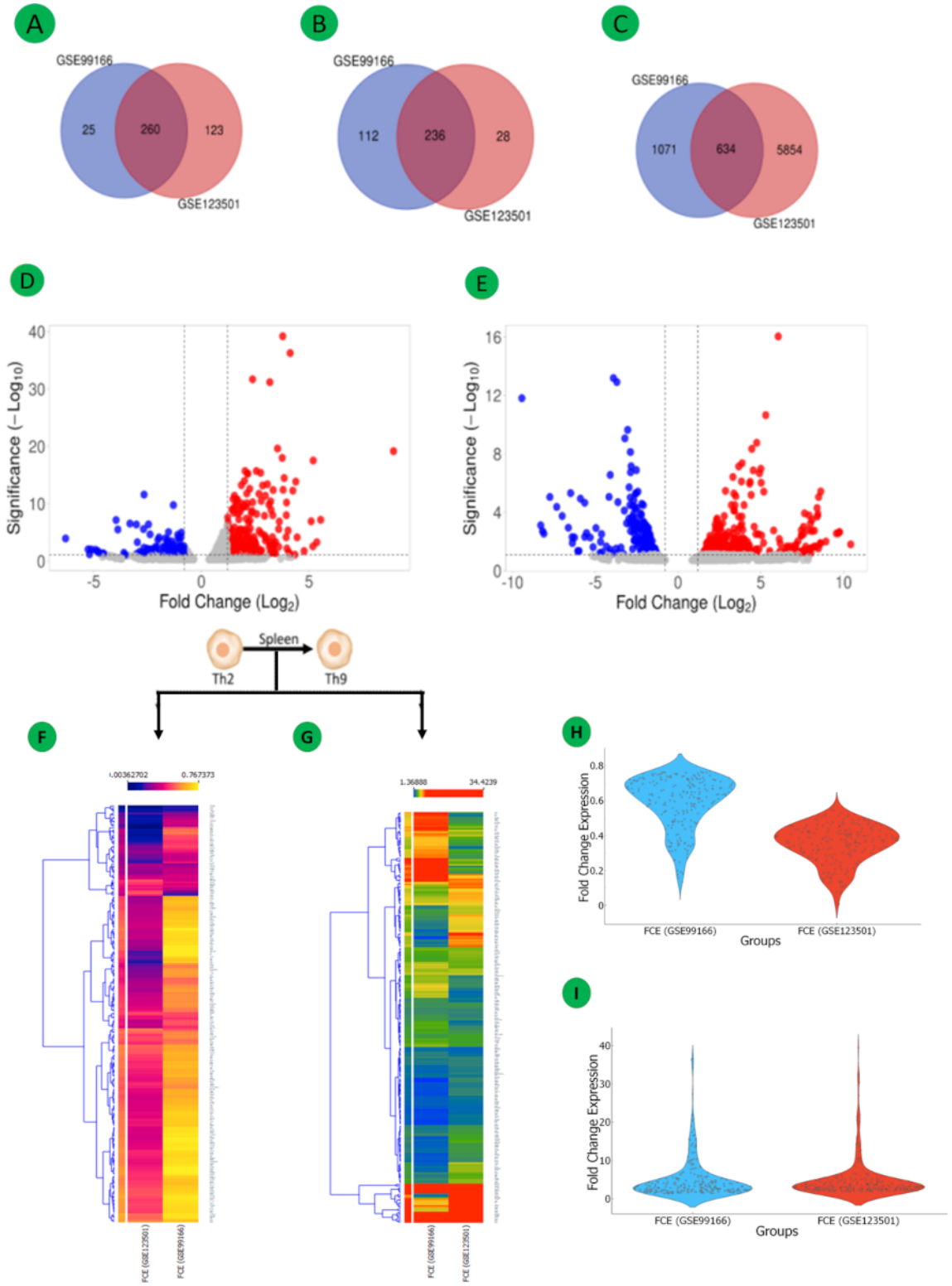
The study was approved by the Institutional Ethics Committee of All India Institute of Medical Sciences (AIIMS) Jodhpur (certificate reference number AIIMS/IEC/2019-20/792).

Results

Assortment of Significant DEGs in the Differentiation of Th2 to Th9 Cells

The *Mus musculus* (C57BL/6) mRNA expression profiles of GSE99167 and GSE123501, which were selected for this study, included the expression profiles of Th2 and Th9 cells obtained from the spleen. We extracted and compared mice spleen samples from 2 different studies to identify genes that are involved in the differentiation of Th2 to Th9 cells. In both groups, 254 common mRNAs were identified, and 634 common DEGs were identified, of which 236 were downregulated and 260 were upregulated. We performed a quality assessment of the selected samples for our expression profiles (Figures 3A-3I; see Tables S1 and S2 in Multimedia Appendix 1, and Multimedia Appendix 2 for larger version of figures).

Figure 3. Differential mRNA expression of the 2 data sets (GSE99166 and GSE123501) for Th2 to Th9 differentiation: Venn diagrams (A-C) of the 3 data sets' total common, downregulated, and upregulated DEGs; volcano plot for the data sets GSE99166 (D) and GSE123501 (E); and heat maps of common downregulated (F) and common upregulated (G) DEGs extracted from the data sets consisting of genes from Th2 and Th9 cells. FCE levels are displayed in ascending order from blue to yellow. Violin plots (H and I) showing FCE distribution of both data sets for common upregulated and common downregulated DEGs. DEG: differentially expressed gene; FCE: fold change expression; mRNA: messenger RNA; Th: T helper.

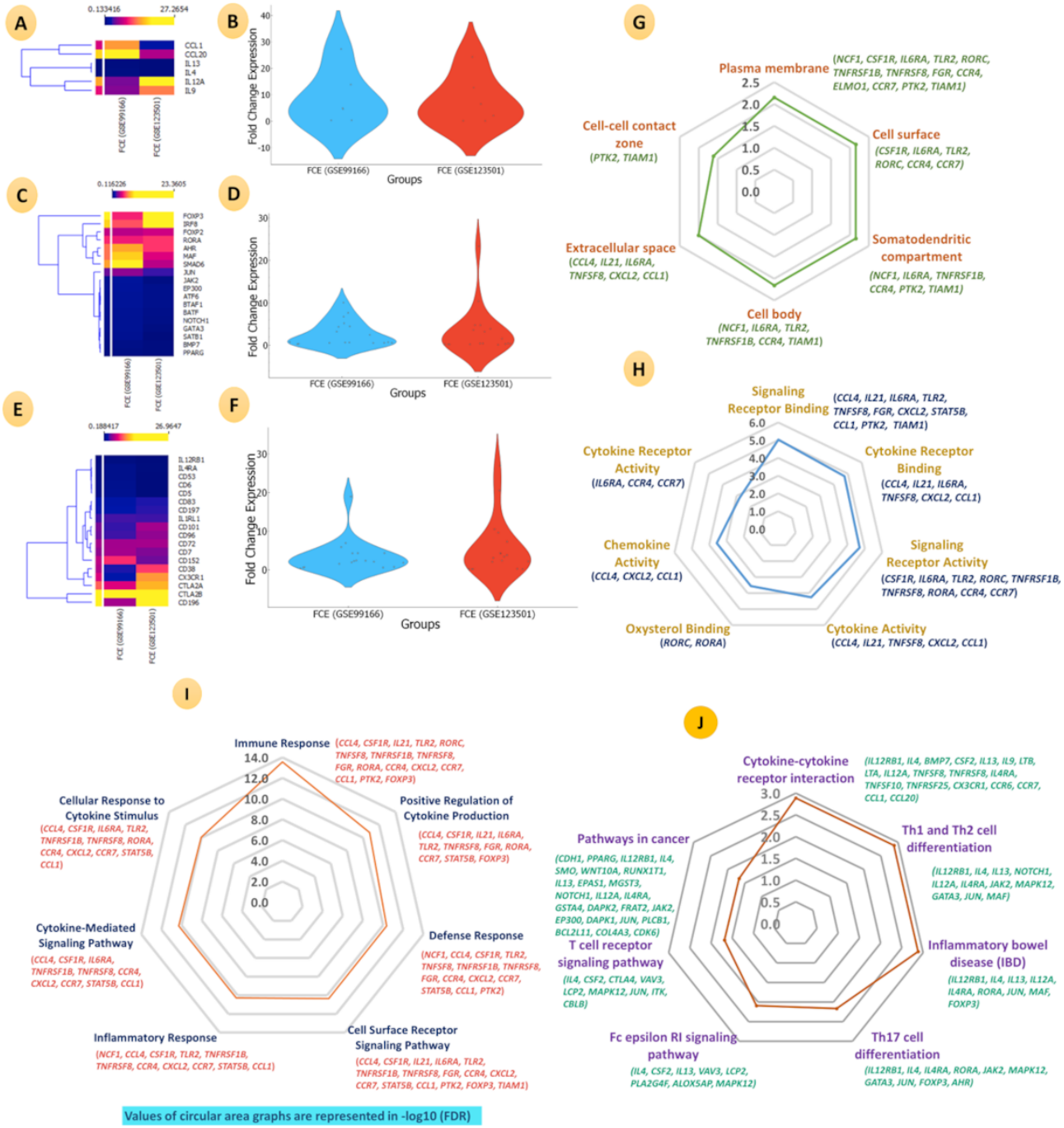


Identification and Assortment of Codifferentially Expressed ILs, Cytokines, Receptors, and TFs

Our analysis identified genes encoding various ILs and receptors whose differential expression may determine the differentiation of Th2 to Th9 cells. Some important ILs identified were *CCL1* (chemokine [C-C motif] ligand 1), *CCL20* (chemokine [C-C motif] ligand 20), *IL-13*, *IL-4*, *IL-12A*, and *IL-9*. The important receptors identified in our analysis were *IL-12RB1*, *IL-4RA* (interleukin 4 receptor alpha), *CD53* (cluster of differentiation 53), *CD6* (cluster of differentiation 6), *CD5* (cluster of differentiation 5), *CD83* (cluster of differentiation 83), *CD197*

(cluster of differentiation 197), *IL-1RL1* (interleukin 1 receptor-like 1), *CD101* (cluster of differentiation 101), *CD96* (cluster of differentiation 96), *CD72* (cluster of differentiation 72), *CD7* (cluster of differentiation 7), *CD152* (cytotoxic T lymphocyte–associated protein 4), *CD38* (cluster of differentiation 38), *CX3CR1* (chemokine [C-X3-C motif] receptor 1), *CTLA2A* (cytotoxic T lymphocyte–associated protein 2 alpha), *CTLA28*, and *CD196* (cluster of differentiation 196). In addition, the differential expression of various TFs such as *FOXP3* (forkhead box P3), *IRF8* (interferon regulatory factor 8), *FOXP2* (forkhead box P2), *RORA* (RAR-related orphan receptor alpha), *AHR* (aryl-hydrocarbon receptor), *MAF* (avian musculoaponeurotic fibrosarcoma oncogene homolog), *SMAD6* (SMAD family member 6), *JUN* (Jun proto-oncogene), *JAK2* (Janus kinase 2), *EP300* (E1A binding protein p300), *ATF6* (activating transcription factor 6), *BTAF1* (B-TFIID TATA-box binding protein associated factor 1), *BAFT* (basic leucine zipper transcription factor), *NOTCH1* (neurogenic locus notch homolog protein 1), *GATA3*, *SATB1* (special AT-rich sequence binding protein 1), *BMP7* (bone morphogenetic protein 7), and *PPARG* (peroxisome proliferator–activated receptor gamma) may influence the differentiation of Th2 to Th9 cells. The expression of the aforementioned immune regulators is represented by a heat map and Venn diagram in Figures 4A-4F (also see Tables S3-S5 in Multimedia Appendix 1, and Multimedia Appendix 2 for larger version of figures).

Figure 4. Differential mRNA expression of the 2 data sets (GSE99166 and GSE123501) for Th2 to Th9 differentiation: (A, B) interleukin, cytokines/chemokines; (C, D) immune regulating transcription factor; (E, F) immune receptor heat maps and violin plot of all, downregulated, and upregulated DEGs; (G, J) enrichment analysis of common DEGs; (G) Gene Ontology cellular component terms; (H) Gene Ontology molecular function terms; (I) Gene Ontology biological process terms; and (J) the KEGG pathway. The radar plot or circular area graph values are represented in the -log10 (FDR). DEG: differentially expressed gene; FCE: fold change expression; FDR: false discovery rate; KEGG: Kyoto Encyclopedia of Genes and Genomes; mRNA: messenger RNA; Th: T helper.



Functional Enrichment and KEGG Pathway Analysis of DEGs Involved in the Transition of Th2 to Th9 Cells

A GO analysis of DEGs classified them into 3 functional classes (Figures 4G-4I; see Multimedia Appendix 2 for larger versions of figures): cellular component, biological process, and molecular function.

The enrichments for the 3 DEG classes with significantly altered expression are shown in Tables S6-S8 in Multimedia Appendix 1. In the KEGG pathway enrichment analysis, the identified genes were enriched in various KEGG pathways such as cytokines-cytokines interaction, Th1 and Th2 cell differentiation, inflammatory bowel disease (IBD), Th17 cell differentiation, the Fc epsilon RI signaling pathway, the T-cell receptor signaling pathway, and pathways in cancer (Figure 4J and Tables S9 and S10 in Multimedia Appendix 1; see Multimedia Appendix 2 for larger version of figures).

Construction of the Protein-Protein Interaction Network of DEGs Involved in the Transition of Th2 to Th9 Cells

We downloaded the complete protein-protein interaction (PPI) network of the identified KEGG pathways from the KEGG database. The Cytoscape software was used for the construction of the network. The significantly altered DEGs of cytokines, chemokines, receptors, and TFs were highlighted in the respective networks. Our analysis of the KEGG pathway enrichment and PPI network demonstrated that the genes that had a significantly altered expression in Th9 cells when compared with Th2 cells also played a significant role in other immune regulating pathways. These affected pathways were mainly involved in cytokines-cytokines interaction, Th1 and Th2 differentiation, CTLA4 (cytotoxic T lymphocyte–associated protein 4) regulation, T-cell receptor signaling, Fc epsilon signaling, Th17 cell differentiation, IBD, and cancer. The concurrent presence of these genes in the aforementioned pathways highlights the significance of the differentiation of Th2 to Th9 in diseases where these pathways are affected. The role of the identified DEGs in these pathways and their interaction with other genes has been depicted in Figures 5-7. See Multimedia Appendix 2 for larger images.

Figure 5. Illustration of the Th2 to Th9 differentiation, mainly the 7 pathways involved in this regulatory mechanism: (A) heat maps expression (the upper section of the heat map shows FCE values represented by varying color densities); (B) PPIs networks (top significant DEGs of the network are illustrated in cyan); (C) common posttranscriptional regulatory microRNA pathways—(1) cytokine-cytokine receptor interaction, (2) Th1 and Th2 cell differentiation, and (3) inflammatory bowel disease. FCE: fold change expression; PPI: protein-protein interaction; Th: T helper.

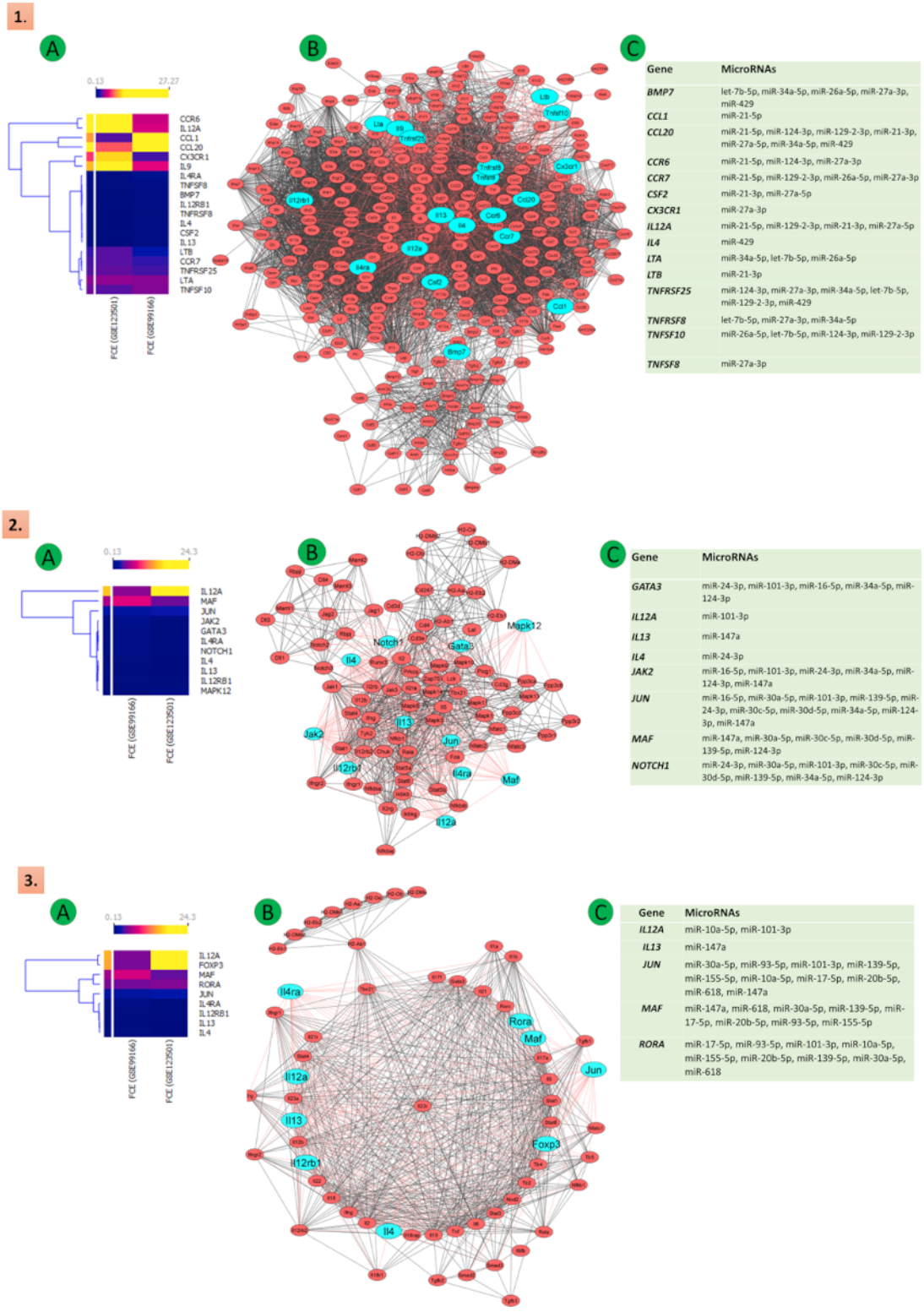
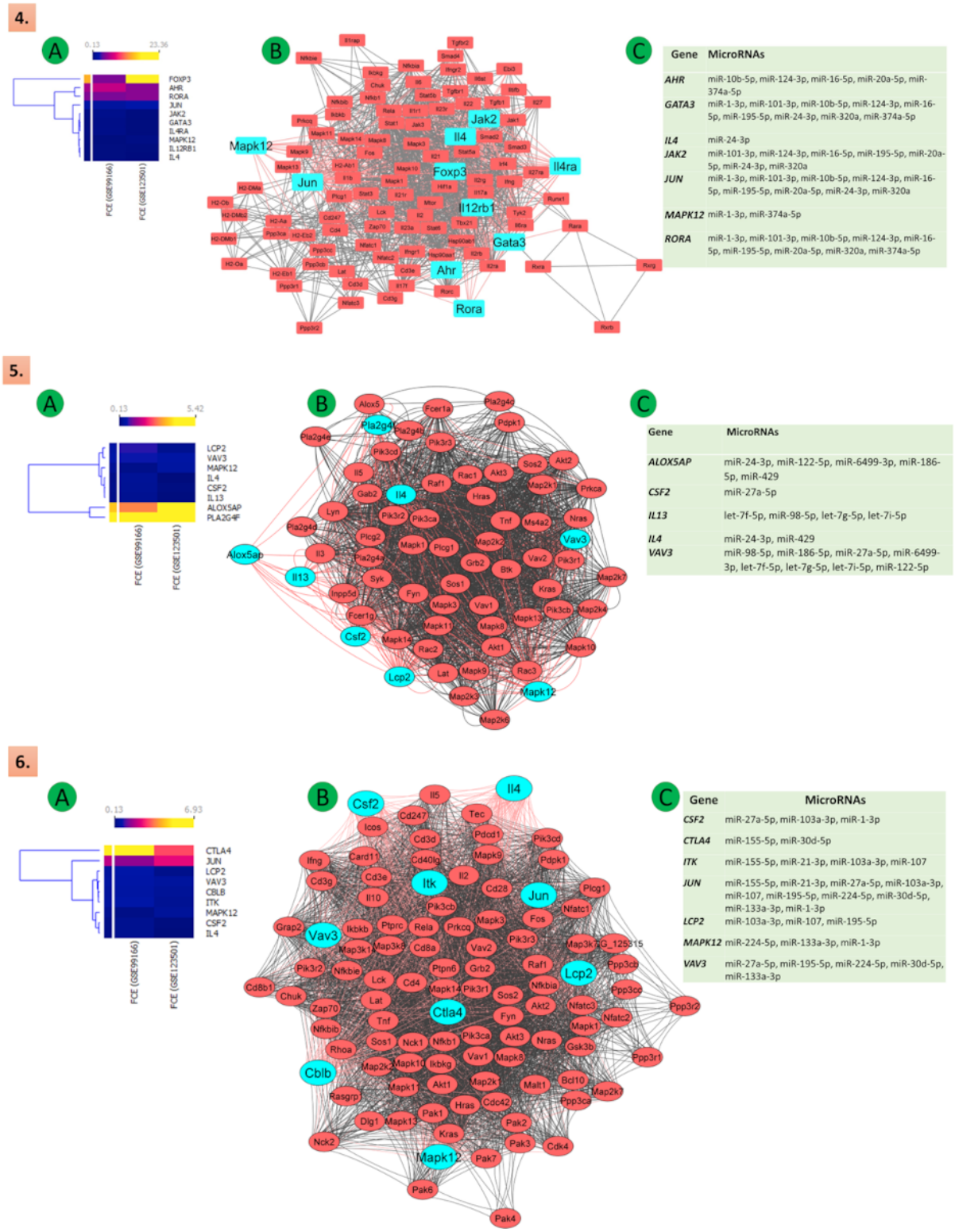


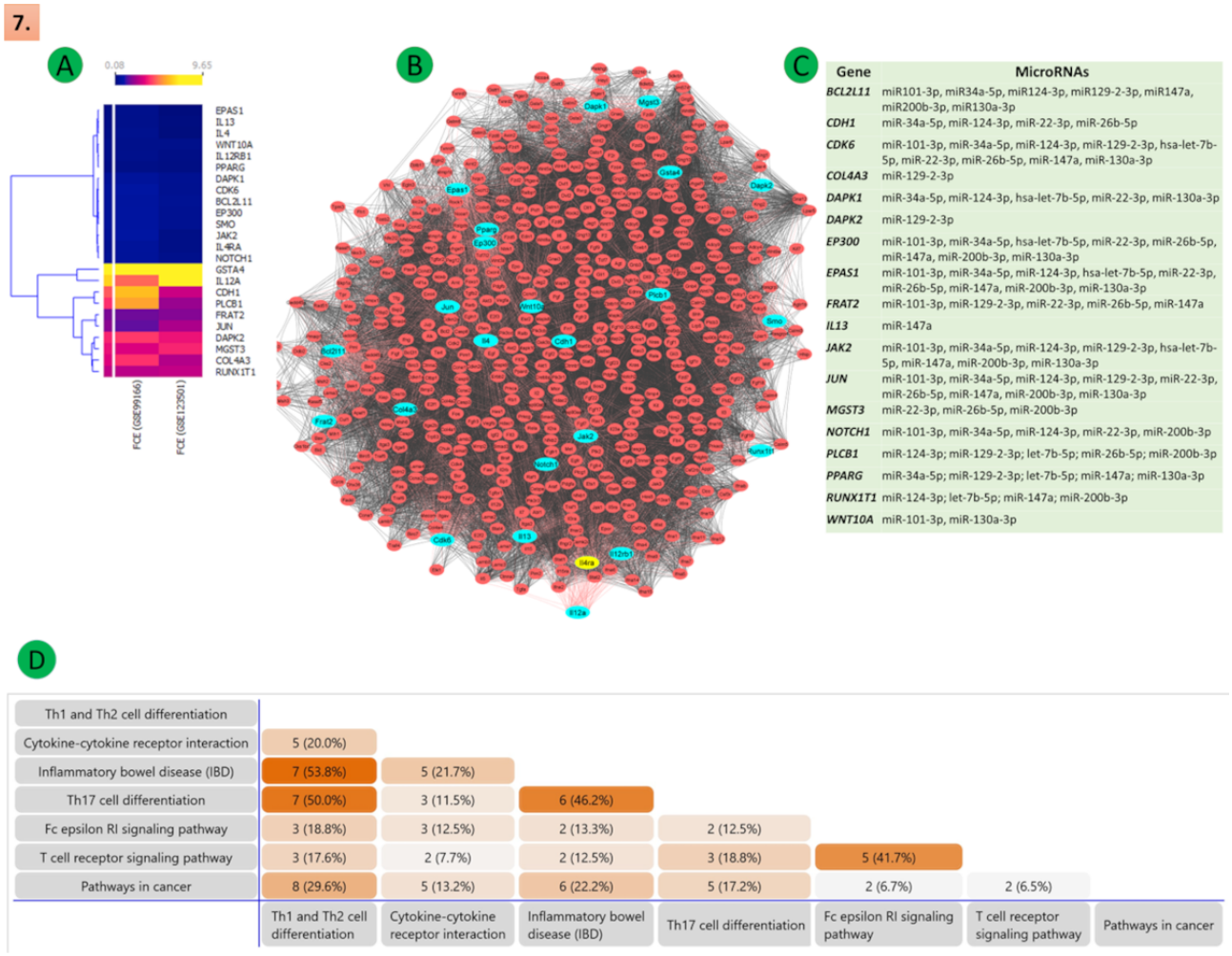
Figure 6. Illustration of the Th2 to Th9 differentiation, mainly the 7 pathways involved in this regulatory mechanism: (A) heat maps expression (the upper section of the heat map shows FCE values represented by varying color densities); (B) PPIs networks (top significant DEGs of the network are illustrated in cyan); (C) common posttranscriptional regulatory microRNA pathways—(4) Th17 cell differentiation, (5) Fc epsilon and RI pathway, (6) T-cell receptor signaling pathway. DEG: differentially expressed gene; FCE: fold change expression; PPI: protein-protein interaction; Th: T helper.



Assessment of Gene Similarity in Pathways Identified in the KEGG Pathway Enrichment Analysis

We performed a gene similarity analysis to find similar genes in all the 7 KEGG pathways identified with the help of the Venn diagram and calculate the percentage of similarity among the genes that were altered. We observed that 7/13 (54%) genes were similar between the "Th1 and Th2 cell differentiation" and "IBD" pathways, whereas 7/14 (50%) genes were similar between the "Th1 and Th2 cell differentiation" and "Th17 cell differentiation" pathways (Figure 7A; see Multimedia Appendix 2 for larger images).

Figure 7. Illustration of the Th2 to Th9 differentiation, mainly the 7 pathways involved in this regulatory mechanism: (A) heat maps expression (the upper part of the heat map shows FCE values represented by varying color densities); (B) PPIs networks (top significant DEGs of the network are presented in cyan); (C) common posttranscriptional regulatory microRNA pathways; (D) gene or DEGs similarity in regulatory pathways. DEG: differentially expressed gene; FCE: fold change expression; PPI: protein-protein interaction; Th: T helper.



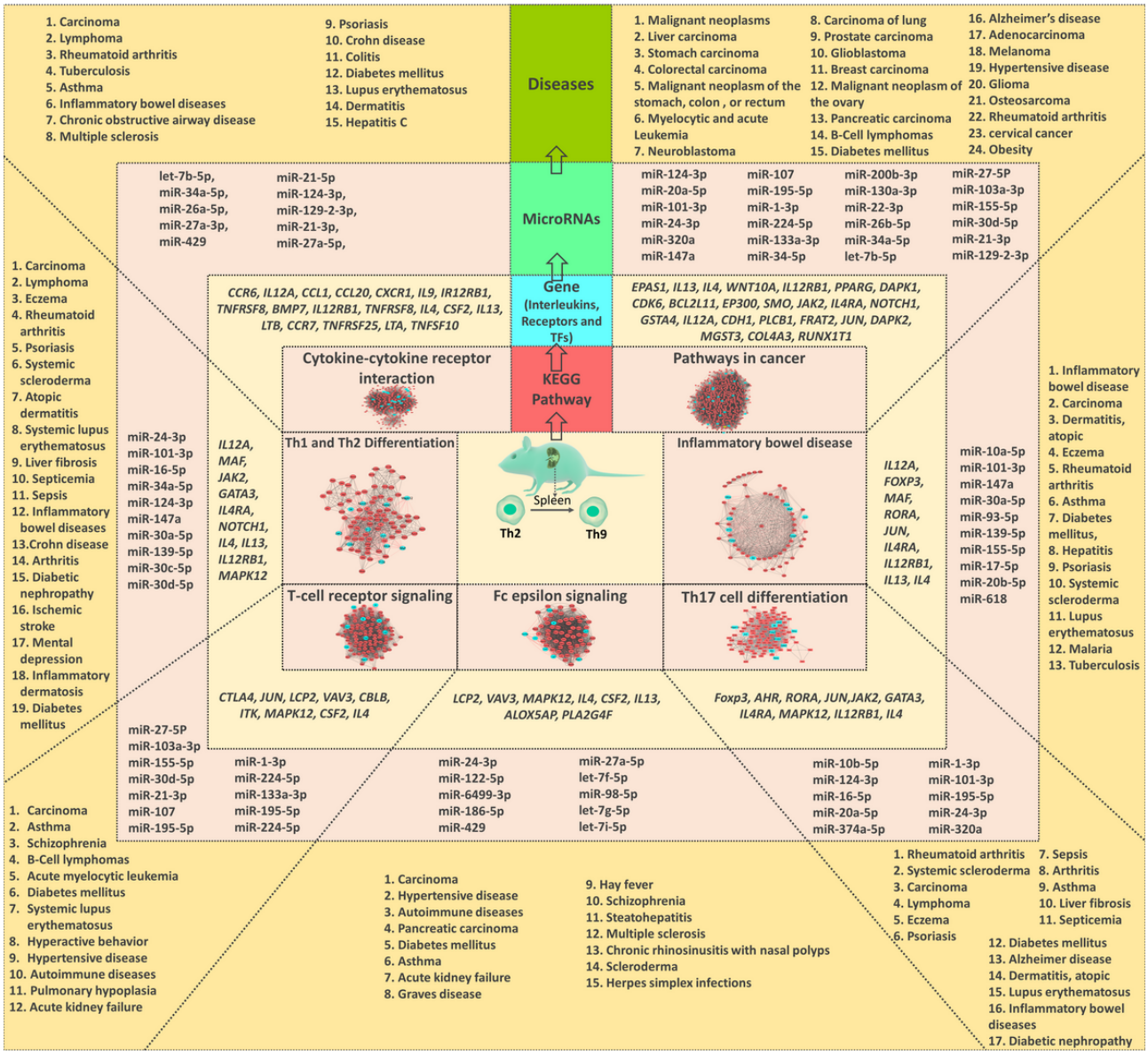
Prediction of miRs That Target the DEGs Involved in the Transition of Th2 to Th9 cells

To explore the posttranscriptional regulation of the identified DEGs, we predicted the miRs that could target the identified DEGs. We identified the following 53 common miRs that could target the DEGs listed in our analysis: let-7b-5p, let-7f-5p, let-7g-5p, let-7i-5p, miR-1-3p, miR-101-3p, miR-103a-3p, miR-107, miR-10a-5p, miR-10b-5p, miR-122-5p, miR-124-3p, miR-129-3p, miR-130a-3p, miR-133a-3p, miR-139-5p, miR-147a, miR-155-5p, miR-16-5p, miR-17-5p, miR-186-5p, miR-195-5p, miR-200b-3p, miR-20a-3p, miR-20a-5p, miR-20b-5p, miR-21-3p, miR-21-5p, miR-22-3p, miR-224-5p, miR-24-3p, miR-26a-5p, miR-26b-5p, miR-27-5p, miR-27a-3p, miR-302a, miR-30a-5p, miR-30c-5p, miR-30d-5p, miR-320a, miR-34-5p, miR-374-5p, miR-426, miR-429, miR-618, miR-6499-3p, miR-93-5p, miR-98-5p, miR-103a-3p, miR-139-5p, miR-147a, miR-195-5p, and miR-27a-5p.

Identification of Diseases Associated With Dysregulation of the Identified miRs and DEGs

Subsequent to the identification of pathways affected as a result of the alteration of DEGs found in our analysis, we further searched for possible diseases whose pathogenesis is affected by alterations in these pathways. We listed the diseases where alterations in the aforementioned 7 pathways have already been documented, and these were as follows: metabolic diseases (eg, diabetes mellitus, diabetic nephropathy, hyperactive behavior, hypertensive disease, ischemic stroke, steatohepatitis, liver fibrosis, obesity), allergic and pulmonary diseases (eg, asthma, chronic obstructive airway disease, chronic rhinosinusitis, nasal polyps, pulmonary hypoplasia, hay fever), carcinomas (eg, acute leukemia and myelocytic leukemia, B-cell lymphomas, lymphoma, adenocarcinoma, breast carcinoma, carcinoma of the lung, cervical cancer, colorectal carcinoma, glioblastoma and glioma, liver carcinoma, malignant neoplasm of the stomach, melanoma, neuroblastoma, osteosarcoma, pancreatic carcinoma, prostate carcinoma, stomach carcinoma), neuropsychiatric disorders (eg, mental depression, schizophrenia, Alzheimer disease), autoimmune diseases (eg, Graves disease, Crohn disease, colitis, psoriasis, systemic lupus erythematosus [SLE], systemic scleroderma, rheumatoid arthritis, multiple sclerosis [MS], IBD, atopic dermatitis, eczema), and infectious diseases (eg, sepsis, septicemia, tuberculosis, hepatitis, herpes simplex infections, malaria; Figure 8).

Figure 8. The diagram illustrates the molecular regulatory process of Th2 to Th9 differentiation in mouse spleen. This regulatory mechanism is regulated by 7 major pathways controlled by some specific immune regulatory transcription factors, receptors, and cytokines/chemokines. The standard and specific microRNAs play a crucial role in the posttranscriptional regulatory mechanism. These 7 pathways regulate DEGs misexpression involved in some critical diseases. DEG: differentially expressed gene; KEGG: Kyoto Encyclopedia of Genes and Genomes; Th: T helper.



Discussion

Principal Findings

In this study, we compared 2 different data sets (GSE99166 and GSE123501) that have compared the mRNA expression in Th2 and Th9 cells. We identified common DEGs that have significantly altered expression between Th2 and Th9 cells from these 2 data sets. Sequential assessment of the DEGs and miRs that had significantly altered expression between Th2 and Th9 cells allows to identify disease states that affect the differentiation process. Although this analysis does not answer whether differentiation of Th2 to Th9 is the cause or the effect of the disease state, it does unravel the possibility of hitherto unknown associations between various diseases and the process of differentiation of Th2 to Th9 cells. Our analysis indicates that differentiation of Th2 to Th9 may play a crucial role via the alteration of DEGs (Table 1) and miRs (Table 2) in various metabolic diseases, allergic and pulmonary diseases, carcinomas, neuropsychiatric disorders, autoimmune diseases, and infectious diseases. In concordance with the existing literature, it was revealed that Th9 cells might play a major role in erythematosus, MS, IBDs, and psoriasis. The role of Th9 cells in autoimmune disease has already been explored in multiple studies [26], including in Graves disease [27], Crohn disease [28-30], psoriasis [31], SLE [32-35], systemic scleroderma [36], rheumatoid arthritis [37-40], MS [26,36,41,42], IBD [26,29,30,43], and atopic dermatitis/eczema [44], which have demonstrated an increased level of differentiation of Th2 to Th9 cells. Th9 cells and *IL-9* have been observed in peripheral blood mononuclear cells and synovial fluid from patients with rheumatoid arthritis. Toll-like receptor 2 (*TLR2*) stimulates naïve $CD4^+$ T cells for *IL-9* secretion and Th9 differentiation by increasing the expression of TFs *BATF* and *PU.1*. *TLR2* activation results in increased expression of *IL-33* and its receptor ST2, augmenting *IL-9* gene expression and Th9 cell

development [45]. Similarly, in patients with SLE, Th9 cell differentiation is suppressed by repression of *IRF4* expression [46]. Although the role of Th9 has been explored in experimental models of MS and IBD, there is insufficient evidence regarding its role in humans. Th9 cells have been shown to play a pathogenic role in experimental autoimmune encephalomyelitis, an animal model of MS [47]. However, only limited studies have assessed Th9 cells in human patients with MS. The skin toxicity of Th9 cells makes them a crucial link in the pathophysiology of multiple skin diseases [48]. Our study highlights the possibility of Th9 playing a crucial role in the pathophysiology of various autoimmune skin diseases such as eczema, atopic dermatitis, psoriasis, and dermatitis. A predominant expression of *IL-9* from Th9 cells was observed to be a characteristic immunologic signature in psoriatic arthritis [49]. Similarly, *IL-9* and *PU.1* gene expressions in atopic dermatitis were higher and associated with disease severity [50]. In addition, the Th9 cell percentage in patients with atopic dermatitis correlated with serum IgE levels, highlighting the link between allergy and the development of Th9 cells [51]. Our in silico analysis further reiterated the involvement of Th9 in various autoimmune pathways. The involvement of *IL-9* and Th9 cells in allergic response can also be seen in other diseases. One such allergic disease in which Th9 cells have been recently explored is asthma. Patients with allergic asthma have increased peripheral blood Th9 cells and elevated levels of serum *IL-9* [51]. *SGK1* (serum/glucocorticoid regulated kinase 1) has been shown to enhance the differentiation of Th9 by modulating the nuclear factor kappa B (*NF-κB*) signaling pathway in patients with asthma [52]. The activation of *MAPK* (mitogen-activated protein kinase) has also been attributed to the activation of Th9 cells in mice models of asthma [53]. Interestingly, *IL-9* and *IL-13* have been elevated in patients with chronic obstructive airway disease compared with asthma [54]. However, so far, the Th9 cells have not been explored for their significance in the pathophysiology of chronic obstructive pulmonary disease. Interestingly, apart from asthma, our in silico analysis highlighted chronic obstructive airway disease, tuberculosis, and chronic rhinosinusitis with nasal polyps as major airway diseases in which Th9 cells may play a crucial role. Our findings are in sync with the study of Ye et al [55], which demonstrated tuberculous pleural effusion to be chemotactic for Th9 cells, while pleural mesothelial cells in tuberculosis stimulated the Th9 cell differentiation. This in silico analysis also highlights the possible role of Th9 in neuropsychiatric diseases. Very few studies have explored the role of Th9 in neuropsychiatric disorders. Saresella et al [56] have demonstrated an increase in the activity of Th9 lymphocytes, while postthymic maturation pathways showed an accumulation of differentiated effector T lymphocytes ($CD4^+$). In Alzheimer disease, schizophrenia, and multiple-episode schizophrenia, although *IL-9* has been elevated, limited studies have been performed to assess the role of Th9 cells in the pathophysiology of the diseases [56,57]. In addition to the aforementioned diseases, this study identified malignancies as one of the disease states that could be affected by the development of Th9 cells. The role of Th9 cells in modulating immunity in cancer has been widely explored. Th9 cells contribute to antitumor immunity by enhancing the recruitment and activation of mast cells, natural killer cells, CD8 T cells, and dendritic cells in the tumor microenvironment. The antitumor effect of Th9 cells has been documented in various animal studies. Lu et al [58] have demonstrated the protective effects of *IL-9* and Th9 on tumor development. The tumor-specific Th9 cells promoted the activation of $CD8^+$ cytotoxic T lymphocytes by recruiting dendritic cells into tumor tissues and subsequently presenting tumor antigens in tumor-draining LNs. Th9 cells in tumor tissues mount an inflammatory response via CTL in a *CCL20/CCR6* (chemokine [C-C motif] receptor 6)-dependent manner [59,60]. Wang et al [61] also demonstrated that Th9-enriched $CD4^+$ T cells significantly increased the expansion of activated $CD8^+$ T cells in a manner that was dependent on the expression of *IL-9R* (interleukin 9 receptor). Th9 thus seems to enhance antitumor immune response through T-cell cytotoxicity and play a crucial role in controlling the progression of cancer [62]. Apart from Th9 cells, the cytokine *IL-9* has also been widely explored in cancers. Expression of *IL-9* in the serum and circulating $CD4^+$ T cells was significantly upregulated in patients with breast cancer compared with healthy controls [63]. Purwar et al [10] demonstrated that *IL-9* depletion in *RORγt*-deficient mice promoted melanoma growth. Zheng et al [64] demonstrated that Th9 cells produce *IL-9* to induce glioma cell apoptosis and inhibit tumor growth. Interestingly, tumor-specific Th9 cells displayed a unique *PU.1-TRAF6-NF-kB* activation–driven hyperproliferative feature, suggesting a persistence mechanism rather than an antiapoptotic strategy. This equips tumor-specific Th9 cells to become a more effective $CD4^+$ T-cell subset for adoptive cancer therapy [65]. Although Th9 cells play an important role in tumor suppression, they have not been studied in various cancer subtypes. Our analysis suggests a possible role for Th9 in different cancer types such as malignant neoplasm of the stomach, melanoma, neuroblastoma, osteosarcoma, pancreatic carcinoma, and prostate carcinoma. Finally, our study also highlights the possible role of Th9 in different metabolic diseases. Interestingly, to our knowledge, no study has yet explored the role of Th9 in metabolic diseases such as diabetes and obesity. We want to highlight these lacunae to open up newer research attempts that would explore the role of Th9 in metabolic diseases. The insights into the role of Th9 in metabolic diseases would better help delineate the role of immunological dysregulation in developing metabolic diseases.

Table 1. Role of various differentially expressed genes in the differentiation of T helper 9 cells and production of *IL-9*[a].

Cytokine or ligand	Receptor	Transcription factors	Effect on T helper 9 cell differentiation	References
IL-6	*IL-6R* and *gp130*	*STAT1*[b] and *STAT3*	Both increases and decreases	[66,67]
IL-10	*IL-10R1*[c] and *IL-10R2*	*STAT1* and *STAT3*	Both increases and decreases	[68-70]
IL-23	*IL-23R* and *IL-12RB1*	*STAT3*	Decreases	[71,72]
IL-27	*IL-27R* and *gp130*	*STAT1*	Decreases	[73]
IL-27	*IL-27R*	*IFN-γ*[d]	Decreases	[74]
IL-1α	*IL-1R1* and *IL-1RACP*	*NF-κB*[e], *MYD88*[f], and *IRAK*[g]	Increases	[75,76]
IL-1β	*IL-1R1* and *IL-1RACP*	*MYD88*, *IRAK*, *NF-κB*, *STAT1*, *IL-9*, and *IRF1*[h]	Increases	[77-80]
IL-2	*IL-2Rα*, *IL-2Rβ*, and *γc*	*STAT5*, *IL-9*, *BCL-6*[i], *IRF4*, and *GATA3*[j]	Increases	[75,81,82]
IL-4	*IL-4Rα* and *γ-chain*	*STAT6*, *FOXP3*[k], *IL-9*	Increases	[83-85]
IL-21	*IL-21R* and *γ-chain*	*IL-1β*, *BCL-6*, *STAT1*, and *STAT3*	Increases	[77,82]
IL-25	*IL-17RB*	*ACT1*[l] and *TRAF6?*[m]	Increases	[86]
IL-33	*IL-1RL1* and *IL-1RACP*	Unknown	Increases	[87]
IFNα and *IFNβ*	*IFNAR1*[n] and *IFNAR2*	*STAT1*	Increases	[69]
TGFβ[o]	*TGFβR2*	*SMAD*[p], *IL-9*, *PU.1*[q], *FOXP3*	Increases	[85,88,89]
TSLP[r]	*TSLPR*[s] and *IL-7Rα*	*STAT5*, *IL-9*	Increases	[81]
Activin A	*ACTRII*[t] and *ALK4*[u]	*SMAD*, *TGFβ*	Increases	[90]
CGRP[v]	N/A[w]	*PKA*[x], *NFATC2*[y], *GATA3*, and *PU.1*	Increases	[91]
Nitric oxide	N/A	*p53*[z], *IL-2*, *STAT5*, *IL-4Rα*, *TGFβR2*	Increases	[15]
TL1A[aa]	*DR3*[bb]	*IL-2*, *STAT5*	Increases	[92]
Notch	Jagged	*NICD1*[cc]	Increases	[67]
IFNγ	*IFNGR1*[dd] and *IFNGR2*[ee]	*STAT1*	Decreases	[73]
PDL2[ff]	*PD1*[gg]	*SHP2*[hh]	Decreases	[93]

[a]IL: interleukin.
[b]*STAT*: signal transducer and activator of transcription.
[c]*ILxR*: interleukin receptor (where x corresponds to the interleukin number).
[d]IFN: interferon.
[e]*NF-κB*: nuclear factor kappa B.
[f]*MYD88*: myeloid differentiation primary response gene 88.
[g]*IRAK*: interleukin-1 receptor-associated kinase 1.
[h]*IRF*: interferon regulatory factor.
[i]BCL-6: B-cell leukemia/lymphoma 6.
[j]*GATA3*: GATA binding protein 3.
[k]*FOXP3*: forkhead box P3.
[l]*ACT1*: actin-related gene 1.
[m]*TRAF6*: TNF receptor–associated factor 6.
[n]*IFNAR*: interferon (alpha and beta) receptor.
[o]*TGF*: transforming growth factor.
[p]*SMAD*: SMAD family member.
[q]*PU.1*: spleen focus forming virus (SFFV) proviral integration oncogene.
[r]*TSLP*: thymic stromal lymphopoietin.
[s]*TSLPR*: thymic stromal lymphopoietin receptor.

[t]*ACTRII*: activin receptor type 2.
[u]*ALK4*: activin A receptor, type 1B.
[v]*CGRP*: calcitonin/calcitonin-related polypeptide.
[w]N/A: not applicable.
[x]*PKA*: protein kinase A.
[y]*NFATC2*: nuclear factor of activated T cells, cytoplasmic, calcineurin dependent 2.
[z]*p53*: transformation-related protein 53.
[aa]*TL1A*: tumor necrosis factor (ligand) superfamily, member 15.
[bb]*DR3*: death-domain receptor 3 (tumor necrosis factor receptor superfamily).
[cc]*NICD1*: notch1 intracellular domain 1.
[dd]*IFNGR1*: interferon gamma receptor 1.
[ee]*IFNGR2*: interferon gamma receptor 2.
[ff]*PDL2*: programmed cell death 1 ligand 2.
[gg]*PD1*: programmed cell death protein 1.
[hh]*SHP2*: protein tyrosine phosphatase, nonreceptor type 11.

Table 2. Role of various microRNAs and target genes in the differentiation of Th9[a] cells in various disease conditions.

MicroRNA	Study model	Type of disease	Level of microRNA	Molecular target gene	Differentiation of Th9	Reference
miR-145	Mouse	Liver cancer	Upregulated	Reducing the expression of *HIF-1α*[b]	Increased	[94]
miR-155	Mouse	Wound	Upregulated	Increased *c-MAF1*[c], *SOCS1*[d], *CXCL1*[e], *CXCL2*[f], *IL-9R*[g]/*IL-9*[h], *IL-17R*[i]/*IL-17A*	Increased	[95]
miR-155	Human and mouse	Acute graft-versus-host disease	Upregulated	*TNF-α*[j]	Increased	[96]
miR-15b/miR-16	Mouse	N/A[k]	Upregulated	Decreased *HIF-2α* expression	Decreased the *IL-9* level in overexpressed Th9 cells	[97]
miR-493-5p	Both human and mouse	Asthma	Downregulated	Decreased *FOXO1*[l] expression	Decreased	[98]
miR-143 and miR-145	Mouse	N/A	Upregulated	*NFATC1*[m] downregulation	Decreased	[99]
miR-155	Human	Methicillin-resistant *Staphylococcus aureus* pneumonia	Upregulated	Decreased *SIRT1*[n]	Increased Th9/*IL-9*	[100]
miR-148a-3p	Mouse	Allergic rhinitis	Upregulated	Increased *IRF4*[o]	Increased	[101]

[a]Th9: T helper 9.
[b]*HIF*: hypoxia-inducible factor.
[c]*MAF*: avian musculoaponeurotic fibrosarcoma oncogene homolog.
[d]*SOCS1*: suppressor of cytokine signaling 1.
[e]*CXCL1*: chemokine (C-X-C motif) ligand 1.
[f]*CXCL2*: chemokine (C-X-C motif) ligand 2.
[g]*IL-9R*: interleukin 9 receptor.
[h]*IL*: interleukin.
[i]*IL-17R*: interleukin 17 receptor.
[j]*TNF*: tumor necrosis factor.
[k]N/A: not applicable.
[l]*FOXO1*: forkhead box O1.
[m]*NFATC1*: nuclear factor of activated T cells, cytoplasmic, calcineurin dependent 1.
[n]*SIRT1*: sirtuin 1.
[o]*IRF*: interferon regulatory factor.

Limitations

The main limitation of the study is that the analysis is based on an in silico method where only a few specific wild-type samples from data sets of previous studies were included; therefore, further validation of the identified genes and miRNAs is required in various animal models and human diseases. The data sets were compiled using different arrays on the Affymetrix platform, which may account for some of the variability in the results. However, the functional enrichment for the mRNAs highlighted some significant pathways related to immune regulation and its derangements.

Conclusions

This study identified common DEGs of ILs, receptors, and TFs that have significantly altered expression between Th2 and Th9 cells. The KEGG pathway enrichment analysis identified cytokines-cytokines interaction, Th1 and Th2 differentiation, T-cell receptor signaling regulation via CTLA4, Fc epsilon signaling, and Th17 cell differentiation as the significant pathways affected by the identified DEGs. Our study identified hitherto unexplored possible associations between Th9 and disease states. The interactome analysis also identified pathways that are involved in various metabolic diseases, allergic and pulmonary diseases, carcinomas, neuropsychiatric disorders, autoimmune diseases, and infectious diseases, where differentiation of Th2 to Th9 may play a crucial role. The scarcity of studies on the role of Th9 in metabolic diseases highlights the lacunae in this field. Thus, our study provides the rationale for exploring the role of Th9 in various metabolic disorders.

Acknowledgments

The authors are grateful to the All India Institute of Medical Sciences Jodhpur for providing the research facility to perform this in silico experiment. MK is supported by a senior research fellowship of The University Grants Commission of India (number NOV2017-361200).

Data Availability

Publicly available GEO data sets were used for the analysis in this study. These data sets can be accessed online [102,103].

Conflicts of Interest

None declared.

Multimedia Appendix 1
Fold change expression of the significant differentially expressed genes analyzed.
[DOCX File , 64 KB - bioinform_v4i1e42421_app1.docx]

Multimedia Appendix 2
Higher resolution images for Figures 3-7.
[DOCX File , 6052 KB - bioinform_v4i1e42421_app2.docx]

References

1. Khokhar M, Roy D, Tomo S, Gadwal A, Sharma P, Purohit P. Novel Molecular Networks and Regulatory MicroRNAs in Type 2 Diabetes Mellitus: Multiomics Integration and Interactomics Study. JMIR Bioinform Biotech 2022 Feb 23;3(1):e32437. [doi: 10.2196/32437]
2. Gadwal A, Purohit P, Khokhar M. Vishnoi DrJR, Pareek DrP, Choudhary DrR, Elhence DrP, Mithu Banerjee1 DrM, Sharma DrP. Identification of potential key genes and their regulatory microRNAs and transcription factors in lymph node and skin metastasis in breast cancer using in silico analysis. (Preprint). JMIR Bioinformatics and Biotechnology; 2022 Dec 2022 Dec 27:1-22. [doi: 10.2196/preprints.45357]
3. Khokhar M, Roy D, Bajpai NK, Bohra GK, Yadav D, Sharma P, et al. Metformin mediates MicroRNA-21 regulated circulating matrix metalloproteinase-9 in diabetic nephropathy: an in-silico and clinical study. Arch Physiol Biochem 2021 Jun 04:1-11. [doi: 10.1080/13813455.2021.1922457] [Medline: 34087084]
4. Khokhar M, Tomo S, Purohit P. MicroRNAs based regulation of cytokine regulating immune expressed genes and their transcription factors in COVID-19. Meta Gene 2022 Feb;31:100990 [FREE Full text] [doi: 10.1016/j.mgene.2021.100990] [Medline: 34722158]
5. Khokhar M, Purohit P, Roy D, Tomo S, Gadwal A, Modi A, et al. Acute kidney injury in COVID 19 - an update on pathophysiology and management modalities. Arch Physiol Biochem 2020 Dec 15:1-14. [doi: 10.1080/13813455.2020.1856141] [Medline: 33320717]
6. Iwalokun BA, Olalekan A, Adenipekun E, Ojo O, Iwalokun SO, Mutiu B, et al. Improving the Understanding of the Immunopathogenesis of Lymphopenia as a Correlate of SARS-CoV-2 Infection Risk and Disease Progression in African

Patients: Protocol for a Cross-sectional Study. JMIR Res Protoc 2021 Mar 04;10(3):e21242 [FREE Full text] [doi: 10.2196/21242] [Medline: 33621190]
7. Ghafouri F, Ahangari Cohan R, Samimi H, Hosseini Rad S, Naderi M, Noorbakhsh F, et al. Development of a Multiepitope Vaccine Against SARS-CoV-2: Immunoinformatics Study. JMIR Bioinform Biotech 2022 Jul 19;3(1):e36100 [FREE Full text] [doi: 10.2196/36100] [Medline: 35891920]
8. Abdelaziz M, Wang H, Cheng J, Xu H. Th2 cells as an intermediate for the differentiation of naïve T cells into Th9 cells, associated with the Smad3/Smad4 and IRF4 pathway. Exp Ther Med 2020 Mar 03;19(3):1947-1954 [FREE Full text] [doi: 10.3892/etm.2020.8420] [Medline: 32104253]
9. Noelle RJ, Nowak EC. Cellular sources and immune functions of interleukin-9. Nat Rev Immunol 2010 Oct 17;10(10):683-687 [FREE Full text] [doi: 10.1038/nri2848] [Medline: 20847745]
10. Purwar R, Schlapbach C, Xiao S, Kang HS, Elyaman W, Jiang X, et al. Robust tumor immunity to melanoma mediated by interleukin-9-producing T cells. Nat Med 2012 Aug 8;18(8):1248-1253 [FREE Full text] [doi: 10.1038/nm.2856] [Medline: 22772464]
11. Guo L, Wei G, Zhu J, Liao W, Leonard WJ, Zhao K, et al. IL-1 family members and STAT activators induce cytokine production by Th2, Th17, and Th1 cells. Proc Natl Acad Sci U S A 2009 Aug 11;106(32):13463-13468 [FREE Full text] [doi: 10.1073/pnas.0906988106] [Medline: 19666510]
12. Angkasekwinai P, Chang SH, Thapa M, Watarai H, Dong C. Regulation of IL-9 expression by IL-25 signaling. Nat Immunol 2010 Mar 14;11(3):250-256 [FREE Full text] [doi: 10.1038/ni.1846] [Medline: 20154671]
13. Xiao X, Balasubramanian S, Liu W, Chu X, Wang H, Taparowsky EJ, et al. OX40 signaling favors the induction of T(H)9 cells and airway inflammation. Nat Immunol 2012 Oct 29;13(10):981-990 [FREE Full text] [doi: 10.1038/ni.2390] [Medline: 22842344]
14. Staudt V, Bothur E, Klein M, Lingnau K, Reuter S, Grebe N, et al. Interferon-regulatory factor 4 is essential for the developmental program of T helper 9 cells. Immunity 2010 Aug 27;33(2):192-202 [FREE Full text] [doi: 10.1016/j.immuni.2010.07.014] [Medline: 20674401]
15. Niedbala W, Besnard A, Nascimento DC, Donate PB, Sonego F, Yip E, et al. Nitric oxide enhances Th9 cell differentiation and airway inflammation. Nat Commun 2014 Aug 07;5(1):4575 [FREE Full text] [doi: 10.1038/ncomms5575] [Medline: 25099390]
16. Zhang J, Lian M, Li B, Gao L, Tanaka T, You Z, et al. Interleukin-35 Promotes Th9 Cell Differentiation in IgG4-Related Disorders: Experimental Data and Review of the Literature. Clin Rev Allergy Immunol 2021 Feb 25;60(1):132-145. [doi: 10.1007/s12016-020-08803-8] [Medline: 32712804]
17. Kaplan MH. Th9 cells: differentiation and disease. Immunol Rev 2013 Mar 13;252(1):104-115 [FREE Full text] [doi: 10.1111/imr.12028] [Medline: 23405898]
18. Mahi NA, Najafabadi MF, Pilarczyk M, Kouril M, Medvedovic M. GREIN: An Interactive Web Platform for Re-analyzing GEO RNA-seq Data. Sci Rep 2019 May 20;9(1):7580 [FREE Full text] [doi: 10.1038/s41598-019-43935-8] [Medline: 31110304]
19. Pathan M, Keerthikumar S, Chisanga D, Alessandro R, Ang C, Askenase P, et al. A novel community driven software for functional enrichment analysis of extracellular vesicles data. J Extracell Vesicles 2017 Dec;6(1):1321455 [FREE Full text] [doi: 10.1080/20013078.2017.1321455] [Medline: 28717418]
20. Pathan M, Keerthikumar S, Ang C, Gangoda L, Quek CY, Williamson NA, et al. FunRich: An open access standalone functional enrichment and interaction network analysis tool. Proteomics 2015 Aug 17;15(15):2597-2601. [doi: 10.1002/pmic.201400515] [Medline: 25921073]
21. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015 Apr 20;43(7):e47 [FREE Full text] [doi: 10.1093/nar/gkv007] [Medline: 25605792]
22. Orange Data Mining. URL: https://orangedatamining.com/citation/ [accessed 2021-04-13]
23. Otasek D, Morris JH, Bouças J, Pico AR, Demchak B. Cytoscape Automation: empowering workflow-based network analysis. Genome Biol 2019 Sep 02;20(1):185 [FREE Full text] [doi: 10.1186/s13059-019-1758-4] [Medline: 31477170]
24. Chang L, Zhou G, Soufan O, Xia J. miRNet 2.0: network-based visual analytics for miRNA functional analysis and systems biology. Nucleic Acids Res 2020 Jul 02;48(W1):W244-W251 [FREE Full text] [doi: 10.1093/nar/gkaa467] [Medline: 32484539]
25. Piñero J, Saüch J, Sanz F, Furlong LI. The DisGeNET cytoscape app: Exploring and visualizing disease genomics data. Comput Struct Biotechnol J 2021;19:2960-2967 [FREE Full text] [doi: 10.1016/j.csbj.2021.05.015] [Medline: 34136095]
26. Deng Y, Wang Z, Chang C, Lu L, Lau CS, Lu Q. Th9 cells and IL-9 in autoimmune disorders: Pathogenesis and therapeutic potentials. Hum Immunol 2017 Feb;78(2):120-128. [doi: 10.1016/j.humimm.2016.12.010] [Medline: 28040536]
27. Janyga S, Marek B, Kajdaniuk D, Ogrodowczyk-Bobik M, Urbanek A, Bułdak. CD4+ cells in autoimmune thyroid disease. Endokrynol Pol 2021;72(5):572-583 [FREE Full text] [doi: 10.5603/EP.a2021.0076] [Medline: 34647609]
28. Giuffrida P, Corazza GR, Di Sabatino A. Old and New Lymphocyte Players in Inflammatory Bowel Disease. Dig Dis Sci 2018 Feb 23;63(2):277-288. [doi: 10.1007/s10620-017-4892-4] [Medline: 29275447]

29. Weigmann B, Neurath MF. Th9 cells in inflammatory bowel diseases. Semin Immunopathol 2017 Jan 11;39(1):89-95. [doi: 10.1007/s00281-016-0603-z] [Medline: 27837255]
30. Fonseca-Camarillo G, Yamamoto-Furusho JK. Immunoregulatory Pathways Involved in Inflammatory Bowel Disease. Inflammatory Bowel Diseases 2015;21(9):2188-2193. [doi: 10.1097/mib.0000000000000477]
31. Solberg S, Aarebrot A, Sarkar I, Petrovic A, Sandvik L, Bergum B, et al. Mass cytometry analysis of blood immune cells from psoriasis patients on biological therapy. Eur J Immunol 2021 Mar;51(3):694-702 [FREE Full text] [doi: 10.1002/eji.202048857] [Medline: 33226128]
32. Yap D, Lai K. Pathogenesis of renal disease in systemic lupus erythematosus--the role of autoantibodies and lymphocytes subset abnormalities. Int J Mol Sci 2015 Apr 09;16(4):7917-7931 [FREE Full text] [doi: 10.3390/ijms16047917] [Medline: 25860947]
33. Ciccia F, Guggino G, Ferrante A, Cipriani P, Giacomelli R, Triolo G. Interleukin-9 and T helper type 9 cells in rheumatic diseases. Clin Exp Immunol 2016 Aug;185(2):125-132 [FREE Full text] [doi: 10.1111/cei.12807] [Medline: 27159882]
34. Liu S, Ye D, Lou J, Fan Z, Ye D. No evidence for a genetic association of IRF4 with systemic lupus erythematosus in a Chinese population. Z Rheumatol 2014 Aug 1;73(6):565-570. [doi: 10.1007/s00393-013-1279-6] [Medline: 24292686]
35. Medrano-Campillo P, Sarmiento-Soto H, Álvarez-Sánchez N, Álvarez-Ríos AI, Guerrero JM, Rodríguez-Prieto I, et al. Evaluation of the immunomodulatory effect of melatonin on the T-cell response in peripheral blood from systemic lupus erythematosus patients. J Pineal Res 2015 Mar 04;58(2):219-226. [doi: 10.1111/jpi.12208] [Medline: 25612066]
36. Guggino G, Lo Pizzo M, Di Liberto D, Rizzo A, Cipriani P, Ruscitti P, et al. Interleukin-9 over-expression and T helper 9 polarization in systemic sclerosis patients. Clin Exp Immunol 2017 Nov;190(2):208-216 [FREE Full text] [doi: 10.1111/cei.13009] [Medline: 28681919]
37. Vyas SP, Srivastava RN, Goswami R. Calcitriol attenuates TLR2/IL-33 signaling pathway to repress Th9 cell differentiation and potentially limits the pathophysiology of rheumatoid arthritis. Mol Cell Biochem 2021 Jan 23;476(1):369-384. [doi: 10.1007/s11010-020-03914-4] [Medline: 32965596]
38. Talotta R, Berzi A, Doria A, Batticciotto A, Ditto M, Atzeni F, et al. The Immunogenicity of Branded and Biosimilar Infliximab in Rheumatoid Arthritis According to Th9-Related Responses. Int J Mol Sci 2017 Oct 12;18(10):2127 [FREE Full text] [doi: 10.3390/ijms18102127] [Medline: 29023386]
39. Ciccia F, Guggino G, Rizzo A, Manzo A, Vitolo B, La Manna MP, et al. Potential involvement of IL-9 and Th9 cells in the pathogenesis of rheumatoid arthritis. Rheumatology (Oxford) 2015 Dec 15;54(12):2264-2272. [doi: 10.1093/rheumatology/kev252] [Medline: 26178600]
40. Vyas SP, Goswami R. A Decade of Th9 Cells: Role of Th9 Cells in Inflammatory Bowel Disease. Front Immunol 2018 May 24;9:1139 [FREE Full text] [doi: 10.3389/fimmu.2018.01139] [Medline: 29881387]
41. Trad S, Granel B, Parizot C, Dorgham K, Hanslik T, Marie I, et al. [Cytokines and T cell differentiation in systemic sclerosis]. Rev Med Interne 2011 Aug;32(8):472-485. [doi: 10.1016/j.revmed.2010.07.015] [Medline: 20850209]
42. Liu M, Wu W, Sun X, Yang J, Xu J, Fu W, et al. New insights into CD4(+) T cell abnormalities in systemic sclerosis. Cytokine Growth Factor Rev 2016 Apr;28:31-36. [doi: 10.1016/j.cytogfr.2015.12.002] [Medline: 26724976]
43. Hisamatsu T, Erben U, Kühl AA. The Role of T-Cell Subsets in Chronic Inflammation in Celiac Disease and Inflammatory Bowel Disease Patients: More Common Mechanisms or More Differences? Inflamm Intest Dis 2016 Jul 9;1(2):52-62 [FREE Full text] [doi: 10.1159/000445133] [Medline: 29922658]
44. Auriemma M, Vianale G, Amerio P, Reale M. Cytokines and T cells in atopic dermatitis. Eur Cytokine Netw 2013 Mar;24(1):37-44 [FREE Full text] [doi: 10.1684/ecn.2013.0333] [Medline: 23608610]
45. Karim AF, Reba SM, Li Q, Boom WH, Rojas RE. Toll like Receptor 2 engagement on CD4 T cells promotes TH9 differentiation and function. Eur J Immunol 2017 Sep 18;47(9):1513-1524 [FREE Full text] [doi: 10.1002/eji.201646846] [Medline: 28665005]
46. Sheng Y, Zhang J, Li K, Wang H, Wang W, Wen L, et al. Bach2 overexpression represses Th9 cell differentiation by suppressing IRF4 expression in systemic lupus erythematosus. FEBS Open Bio 2021 Feb 22;11(2):395-403 [FREE Full text] [doi: 10.1002/2211-5463.13050] [Medline: 33249782]
47. Al-Mazroua HA, Nadeem A, Ansari MA, Attia SM, Bakheet SA, Albekairi TH, et al. CCR1 antagonist ameliorates experimental autoimmune encephalomyelitis by inhibition of Th9/Th22-related markers in the brain and periphery. Mol Immunol 2022 Apr;144:127-137. [doi: 10.1016/j.molimm.2022.02.017] [Medline: 35219910]
48. Schlapbach C, Gehad A, Yang C, Watanabe R, Guenova E, Teague JE, et al. Human TH9 cells are skin-tropic and have autocrine and paracrine proinflammatory capacity. Sci Transl Med 2014 Jan 15;6(219):219ra8 [FREE Full text] [doi: 10.1126/scitranslmed.3007828] [Medline: 24431112]
49. Mauro D, Simone D, Bucci L, Ciccia F. Novel immune cell phenotypes in spondyloarthritis pathogenesis. Semin Immunopathol 2021 Apr 10;43(2):265-277 [FREE Full text] [doi: 10.1007/s00281-021-00837-0] [Medline: 33569634]
50. Hamza AM, Omar SS, Abo El-Wafa RAH, Elatrash MJ. Expression levels of transcription factor PU.1 and interleukin-9 in atopic dermatitis and their relation to disease severity and eruption types. Int J Dermatol 2017 May 22;56(5):534-539. [doi: 10.1111/ijd.13579] [Medline: 28229452]
51. Ma L, Xue H, Guan X, Shu C, Zhang J, Yu J. Possible pathogenic role of T helper type 9 cells and interleukin (IL)-9 in atopic dermatitis. Clin Exp Immunol 2014 Jan;175(1):25-31 [FREE Full text] [doi: 10.1111/cei.12198] [Medline: 24032555]

52. Wu X, Jiang W, Wang X, Zhang C, Cai J, Yu S, et al. SGK1 enhances Th9 cell differentiation and airway inflammation through NF-κB signaling pathway in asthma. Cell Tissue Res 2020 Dec 28;382(3):563-574. [doi: 10.1007/s00441-020-03252-3] [Medline: 32725426]
53. Huang M, Wei Y, Dong J. Epimedin C modulates the balance between Th9 cells and Treg cells through negative regulation of noncanonical NF-κB pathway and MAPKs activation to inhibit airway inflammation in the ovalbumin-induced murine asthma model. Pulm Pharmacol Ther 2020 Dec;65:102005 [FREE Full text] [doi: 10.1016/j.pupt.2021.102005] [Medline: 33636365]
54. Bai Y, Zhou Q, Fang Q, Song L, Chen K. Inflammatory Cytokines and T-Lymphocyte Subsets in Serum and Sputum in Patients with Bronchial Asthma and Chronic Obstructive Pulmonary Disease. Med Sci Monit 2019 Mar 25;25:2206-2210. [doi: 10.12659/msm.913703]
55. Ye Z, Yuan M, Zhou Q, Du R, Yang W, Xiong X, et al. Differentiation and recruitment of Th9 cells stimulated by pleural mesothelial cells in human Mycobacterium tuberculosis infection. PLoS One 2012 Feb 20;7(2):e31710 [FREE Full text] [doi: 10.1371/journal.pone.0031710] [Medline: 22363712]
56. Saresella M, Calabrese E, Marventano I, Piancone F, Gatti A, Alberoni M, et al. Increased activity of Th-17 and Th-9 lymphocytes and a skewing of the post-thymic differentiation pathway are seen in Alzheimer's disease. Brain Behav Immun 2011 Mar;25(3):539-547. [doi: 10.1016/j.bbi.2010.12.004] [Medline: 21167930]
57. Frydecka D, Krzystek-Korpacka M, Lubeiro A, Stramecki F, Stańczykiewicz B, Beszłej JA, et al. Profiling inflammatory signatures of schizophrenia: A cross-sectional and meta-analysis study. Brain Behav Immun 2018 Jul;71:28-36. [doi: 10.1016/j.bbi.2018.05.002] [Medline: 29730395]
58. Lu Y, Hong S, Li H, Park J, Hong B, Wang L, et al. Th9 cells promote antitumor immune responses in vivo. J. Clin. Invest 2012 Oct 15;122(11):4160-4171. [doi: 10.1172/jci65459]
59. Zhou Y, Sonobe Y, Akahori T, Jin S, Kawanokuchi J, Noda M, et al. IL-9 promotes Th17 cell migration into the central nervous system via CC chemokine ligand-20 produced by astrocytes. J Immunol 2011 Apr 01;186(7):4415-4421. [doi: 10.4049/jimmunol.1003307] [Medline: 21346235]
60. Yamasaki A, Saleh A, Koussih L, Muro S, Halayko AJ, Gounni AS. IL-9 induces CCL11 expression via STAT3 signalling in human airway smooth muscle cells. PLoS One 2010 Feb 12;5(2):e9178 [FREE Full text] [doi: 10.1371/journal.pone.0009178] [Medline: 20169197]
61. Wang C, Lu Y, Chen L, Gao T, Yang Q, Zhu C, et al. Th9 cells are subjected to PD-1/PD-L1-mediated inhibition and are capable of promoting CD8 T cell expansion through IL-9R in colorectal cancer. Int Immunopharmacol 2020 Jan;78:106019. [doi: 10.1016/j.intimp.2019.106019] [Medline: 31776089]
62. Chauhan SR, Singhal PG, Sharma U, Bandil K, Chakraborty K, Bharadwaj M. Corrigendum to “Th9 cytokines curb cervical cancer progression and immune evasion” [80 (2019) 1020–1025]. Human Immunology 2020 Feb;81(2-3):125. [doi: 10.1016/j.humimm.2019.12.007]
63. You F, Zhang J, Cui T, Zhu R, Lv C, Tang H, et al. Th9 cells promote antitumor immunity via IL-9 and IL-21 and demonstrate atypical cytokine expression in breast cancer. International Immunopharmacology 2017 Nov;52:163-167. [doi: 10.1016/j.intimp.2017.08.031]
64. Zheng H, Yang B, Xu D, Wang W, Tan J, Sun L, et al. Induction of specific T helper-9 cells to inhibit glioma cell growth. Oncotarget 2017 Jan 17;8(3):4864-4874 [FREE Full text] [doi: 10.18632/oncotarget.13981] [Medline: 28002799]
65. Lu Y, Wang Q, Xue G, Bi E, Ma X, Wang A, et al. Th9 Cells Represent a Unique Subset of CD4 T Cells Endowed with the Ability to Eradicate Advanced Tumors. Cancer Cell 2018 Jun 11;33(6):1048-1060.e7 [FREE Full text] [doi: 10.1016/j.ccell.2018.05.004] [Medline: 29894691]
66. Veldhoen M, Uyttenhove C, van Snick J, Helmby H, Westendorf A, Buer J, et al. Transforming growth factor-beta 'reprograms' the differentiation of T helper 2 cells and promotes an interleukin 9-producing subset. Nat Immunol 2008 Dec 19;9(12):1341-1346. [doi: 10.1038/ni.1659] [Medline: 18931678]
67. Elyaman W, Bassil R, Bradshaw E, Orent W, Lahoud Y, Zhu B, et al. Notch receptors and Smad3 signaling cooperate in the induction of interleukin-9-producing T cells. Immunity 2012 Apr 20;36(4):623-634 [FREE Full text] [doi: 10.1016/j.immuni.2012.01.020] [Medline: 22503540]
68. Chang H, Sehra S, Goswami R, Yao W, Yu Q, Stritesky GL, et al. The transcription factor PU.1 is required for the development of IL-9-producing T cells and allergic inflammation. Nat Immunol 2010 Jun 2;11(6):527-534 [FREE Full text] [doi: 10.1038/ni.1867] [Medline: 20431622]
69. Wong MT, Ye JJ, Alonso MN, Landrigan A, Cheung RK, Engleman E, et al. Regulation of human Th9 differentiation by type I interferons and IL-21. Immunol Cell Biol 2010 Aug 27;88(6):624-631 [FREE Full text] [doi: 10.1038/icb.2010.53] [Medline: 20421880]
70. Ramming A, Druzd D, Leipe J, Schulze-Koops H, Skapenko A. Maturation-related histone modifications in the PU.1 promoter regulate Th9-cell development. Blood 2012 May 17;119(20):4665-4674 [FREE Full text] [doi: 10.1182/blood-2011-11-392589] [Medline: 22446486]
71. Jäger A, Dardalhon V, Sobel R, Bettelli E, Kuchroo V. Th1, Th17, and Th9 effector cells induce experimental autoimmune encephalomyelitis with different pathological phenotypes. J Immunol 2009 Dec 01;183(11):7169-7177 [FREE Full text] [doi: 10.4049/jimmunol.0901906] [Medline: 19890056]

72. Elyaman W, Bradshaw EM, Uyttenhove C, Dardalhon V, Awasthi A, Imitola J, et al. IL-9 induces differentiation of TH17 cells and enhances function of FoxP3+ natural regulatory T cells. Proc Natl Acad Sci U S A 2009 Aug 04;106(31):12885-12890 [FREE Full text] [doi: 10.1073/pnas.0812530106] [Medline: 19433802]
73. Murugaiyan G, Beynon V, Pires Da Cunha A, Joller N, Weiner H. IFN-γ limits Th9-mediated autoimmune inflammation through dendritic cell modulation of IL-27. J Immunol 2012 Dec 01;189(11):5277-5283 [FREE Full text] [doi: 10.4049/jimmunol.1200808] [Medline: 23125412]
74. Xiong P, Liu T, Huang H, Yuan Y, Zhang W, Fu L, et al. IL-27 overexpression alleviates inflammatory response in allergic asthma by inhibiting Th9 differentiation and regulating Th1/Th2 balance. Immunopharmacol Immunotoxicol 2022 Oct 13;44(5):712-718. [doi: 10.1080/08923973.2022.2077755] [Medline: 35695698]
75. Schmitt E, Germann T, Goedert S, Hoehn P, Huels C, Koelsch S, et al. IL-9 production of naive CD4+ T cells depends on IL-2, is synergistically enhanced by a combination of TGF-beta and IL-4, and is inhibited by IFN-gamma. J Immunol 1994 Nov 01;153(9):3989-3996. [Medline: 7930607]
76. Uyttenhove C, Brombacher F, Van Snick J. TGF-β interactions with IL-1 family members trigger IL-4-independent IL-9 production by mouse CD4(+) T cells. Eur J Immunol 2010 Aug 10;40(8):2230-2235 [FREE Full text] [doi: 10.1002/eji.200940281] [Medline: 20540113]
77. Végran F, Berger H, Boidot R, Mignot G, Bruchard M, Dosset M, et al. The transcription factor IRF1 dictates the IL-21-dependent anticancer functions of TH9 cells. Nat Immunol 2014 Aug 29;15(8):758-766. [doi: 10.1038/ni.2925] [Medline: 24973819]
78. Horka H, Staudt V, Klein M, Taube C, Reuter S, Dehzad N, et al. The tick salivary protein sialostatin L inhibits the Th9-derived production of the asthma-promoting cytokine IL-9 and is effective in the prevention of experimental asthma. J Immunol 2012 Mar 15;188(6):2669-2676 [FREE Full text] [doi: 10.4049/jimmunol.1100529] [Medline: 22327077]
79. Anuradha R, George P, Hanna L, Chandrasekaran V, Kumaran P, Nutman T, et al. IL-4-, TGF-β-, and IL-1-dependent expansion of parasite antigen-specific Th9 cells is associated with clinical pathology in human lymphatic filariasis. J Immunol 2013 Sep 01;191(5):2466-2473 [FREE Full text] [doi: 10.4049/jimmunol.1300911] [Medline: 23913964]
80. Beriou G, Bradshaw E, Lozano E, Costantino C, Hastings W, Orban T, et al. TGF-beta induces IL-9 production from human Th17 cells. J Immunol 2010 Jul 01;185(1):46-54 [FREE Full text] [doi: 10.4049/jimmunol.1000356] [Medline: 20498357]
81. Yao W, Zhang Y, Jabeen R, Nguyen E, Wilkes D, Tepper R, et al. Interleukin-9 is required for allergic airway inflammation mediated by the cytokine TSLP. Immunity 2013 Feb 21;38(2):360-372 [FREE Full text] [doi: 10.1016/j.immuni.2013.01.007] [Medline: 23376058]
82. Liao W, Spolski R, Li P, Du N, West EE, Ren M, et al. Opposing actions of IL-2 and IL-21 on Th9 differentiation correlate with their differential regulation of BCL6 expression. Proc Natl Acad Sci U S A 2014 Mar 04;111(9):3508-3513 [FREE Full text] [doi: 10.1073/pnas.1301138111] [Medline: 24550509]
83. Dardalhon V, Awasthi A, Kwon H, Galileos G, Gao W, Sobel RA, et al. IL-4 inhibits TGF-beta-induced Foxp3+ T cells and, together with TGF-beta, generates IL-9+ IL-10+ Foxp3(-) effector T cells. Nat Immunol 2008 Dec 09;9(12):1347-1355 [FREE Full text] [doi: 10.1038/ni.1677] [Medline: 18997793]
84. Jabeen R, Goswami R, Awe O, Kulkarni A, Nguyen ET, Attenasio A, et al. Th9 cell development requires a BATF-regulated transcriptional network. J Clin Invest 2013 Nov;123(11):4641-4653 [FREE Full text] [doi: 10.1172/JCI69489] [Medline: 24216482]
85. Goswami R, Jabeen R, Yagi R, Pham D, Zhu J, Goenka S, et al. STAT6-dependent regulation of Th9 development. J Immunol 2012 Feb 01;188(3):968-975 [FREE Full text] [doi: 10.4049/jimmunol.1102840] [Medline: 22180613]
86. Vink A, Renauld J, Warnier G, Van Snick J. Interleukin-9 stimulates in vitro growth of mouse thymic lymphomas. Eur J Immunol 1993 May;23(5):1134-1138. [doi: 10.1002/eji.1830230523] [Medline: 8477807]
87. Blom L, Poulsen BC, Jensen BM, Hansen A, Poulsen LK. IL-33 induces IL-9 production in human CD4+ T cells and basophils. PLoS One 2011 Jul 6;6(7):e21695 [FREE Full text] [doi: 10.1371/journal.pone.0021695] [Medline: 21765905]
88. Tamiya T, Ichiyama K, Kotani H, Fukaya T, Sekiya T, Shichita T, et al. Smad2/3 and IRF4 play a cooperative role in IL-9-producing T cell induction. J Immunol 2013 Sep 01;191(5):2360-2371. [doi: 10.4049/jimmunol.1301276] [Medline: 23913959]
89. Wang A, Pan D, Lee Y, Martinez G, Feng X, Dong C. Cutting edge: Smad2 and Smad4 regulate TGF-β-mediated Il9 gene expression via EZH2 displacement. J Immunol 2013 Nov 15;191(10):4908-4912 [FREE Full text] [doi: 10.4049/jimmunol.1300433] [Medline: 24108699]
90. Jones CP, Gregory LG, Causton B, Campbell GA, Lloyd CM. Activin A and TGF-β promote T(H)9 cell-mediated pulmonary allergic pathology. J Allergy Clin Immunol 2012 Apr;129(4):1000-10.e3 [FREE Full text] [doi: 10.1016/j.jaci.2011.12.965] [Medline: 22277204]
91. Houssiau FA, Schandené L, Stevens M, Cambiaso C, Goldman M, van Snick J, et al. A cascade of cytokines is responsible for IL-9 expression in human T cells. Involvement of IL-2, IL-4, and IL-10. J Immunol 1995 Mar 15;154(6):2624-2630. [Medline: 7876537]
92. Richard A, Tan C, Hawley E, Gomez-Rodriguez J, Goswami R, Yang X, et al. The TNF-family ligand TL1A and its receptor DR3 promote T cell-mediated allergic immunopathology by enhancing differentiation and pathogenicity of IL-9-producing T cells. J Immunol 2015 Apr 15;194(8):3567-3582 [FREE Full text] [doi: 10.4049/jimmunol.1401220] [Medline: 25786692]

93. Kerzerho J, Maazi H, Speak AO, Szely N, Lombardi V, Khoo B, et al. Programmed cell death ligand 2 regulates TH9 differentiation and induction of chronic airway hyperreactivity. J Allergy Clin Immunol 2013 Apr;131(4):1048-57, 1057.e1 [FREE Full text] [doi: 10.1016/j.jaci.2012.09.027] [Medline: 23174661]
94. Huang Y, Jiang H, Shi Q, Qiu X, Wei X, Zhang X, et al. miR-145 Inhibits Th9 Cell Differentiation by Suppressing Activation of the PI3K/Akt/mTOR/p70S6K/HIF-1α Pathway in Malignant Ascites from Liver Cancer. Onco Targets Ther 2020;13:3789-3800 [FREE Full text] [doi: 10.2147/OTT.S245346] [Medline: 32440147]
95. Wang C, Zhu H, Zhu Y. Knockout of MicroRNA-155 Ameliorates the Th17/Th9 Immune Response and Promotes Wound Healing. Curr Med Sci 2019 Dec 16;39(6):954-964. [doi: 10.1007/s11596-019-2128-x] [Medline: 31845227]
96. Zhang R, Wang X, Hong M, Luo T, Zhao M, Shen H, et al. Endothelial microparticles delivering microRNA-155 into T lymphocytes are involved in the initiation of acute graft-versus-host disease following allogeneic hematopoietic stem cell transplantation. Oncotarget 2017 Apr 04;8(14):23360-23375 [FREE Full text] [doi: 10.18632/oncotarget.15579] [Medline: 28423578]
97. Singh Y, Garden OA, Lang F, Cobb BS. MicroRNAs regulate T-cell production of interleukin-9 and identify hypoxia-inducible factor-2α as an important regulator of T helper 9 and regulatory T-cell differentiation. Immunology 2016 Sep 11;149(1):74-86 [FREE Full text] [doi: 10.1111/imm.12631] [Medline: 27278750]
98. Rao X, Dong H, Zhang W, Sun H, Gu W, Zhang X, et al. MiR-493-5p inhibits Th9 cell differentiation in allergic asthma by targeting FOXO1. Respir Res 2022 Oct 17;23(1):286 [FREE Full text] [doi: 10.1186/s12931-022-02207-2] [Medline: 36253857]
99. Qiu X, Shi Q, Huang Y, Jiang H, Qin S. miR-143/145 inhibits Th9 cell differentiation by targeting NFATc1. Mol Immunol 2021 Apr;132:184-191 [FREE Full text] [doi: 10.1016/j.molimm.2021.01.001] [Medline: 33446394]
100. Tian K, Xu W. MiR-155 regulates Th9 differentiation in children with methicillin-resistant Staphylococcus aureus pneumonia by targeting SIRT1. Hum Immunol 2021 Oct;82(10):775-781. [doi: 10.1016/j.humimm.2021.07.002] [Medline: 34294459]
101. Li L, Deng J, Huang T, Liu K, Jiang X, Chen X, et al. IRF4 transcriptionally activate HOTAIRM1, which in turn regulates IRF4 expression, thereby affecting Th9 cell differentiation and involved in allergic rhinitis. Gene 2022 Mar 01;813:146118. [doi: 10.1016/j.gene.2021.146118] [Medline: 34929342]
102. Schwartz DS, Meylan F, Shih HY, Mikami Y, Petermann FP, Sun HW, et al. GSE123501: Retinoic acid receptor alpha represses a Th9 transcriptional and epigenomic program to reduce allergic pathology. NCBI. 2019. URL: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE123501 [accessed 2023-02-16]
103. GSE99167: CD4+ T cells. NCBI. 2017. URL: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE99167 [accessed 2023-02-16]

Abbreviations

***ACT1*:** actin-related gene 1
***ACTRII*:** activin receptor type 2
***AHR*:** aryl-hydrocarbon receptor
***ALK4*:** activin A receptor, type 1B
***ATF6*:** activating transcription factor 6
***BAFT*:** basic leucine zipper transcription factor
***BCL6*:** B-cell leukemia/lymphoma 6
***BMP7*:** bone morphogenetic protein 7
***BTAF1*:** B-TFIID TATA-box binding protein associated factor 1
***CCL1*:** chemokine (C-C motif) ligand 1
***CCL20*:** chemokine (C-C motif) ligand 20
***CCR6*:** chemokine (C-C motif) receptor 6
CD: cluster of differentiation
***CGRP*:** calcitonin/calcitonin-related polypeptide, alpha
CTLA: cytotoxic T lymphocyte–associated protein
***CX3CR1*:** chemokine (C-X3-C motif) receptor 1
***CXCL1*:** chemokine (C-X-C motif) ligand 1
***CXCL2*:** chemokine (C-X-C motif) ligand 2
DC: dendritic cell
DEG: differentially expressed gene
***DR3*:** death-domain receptor 3 (tumor necrosis factor receptor superfamily)
EAE: autoimmune encephalomyelitis
***EP300*:** E1A binding protein p300
***FOXO1*:** forkhead box O1
***FOXP2*:** forkhead box P2
***FOXP3*:** forkhead box P3

***GATA3*:** GATA binding protein 3
GEO: Gene Expression Omnibus
GO: The Gene Ontology
GREIN: GEO RNA-seq Experiments Interactive Navigator
***HIF*:** hypoxia-inducible factor
IBD: inflammatory bowel disease
IF1: NDV-induced circulating interferon
IFN: interferon
***IFNAR1*:** interferon (alpha and beta) receptor 1
***IFNAR2*:** interferon (alpha and beta) receptor 2
***IFNGR1*:** interferon gamma receptor 1
***IFNGR2*:** interferon gamma receptor 2
IL: interleukin
***IL-1R1*:** interleukin 1 receptor, type I
***IL-1RL1*:** interleukin 1 receptor-like 1
***IL-1RL1*:** interleukin 1 receptor-like 1
***IL-2R*:** interleukin 2 receptor, alpha chain
***IL-4R*:** interleukin 4 receptor, alpha
***IL-4RA*:** interleukin 4 receptor, alpha
***IL-6R*:** interleukin 6 receptor, alpha
***IL-7R*:** interleukin 7 receptor
***IL-9R*:** interleukin 9 receptor
***IL-10R2*:** interleukin 10 receptor, beta
***IL-12RB1*:** interleukin 12 receptor, beta 1
***IL-12RB1*:** interleukin 12 receptor, beta 1
***IL-17R*:** interleukin 17 receptor A
***IL-17RB*:** interleukin 17 receptor B
***IL-21R*:** interleukin 21 receptor
***IL-23R*:** interleukin 23 receptor
IRAK: interleukin-1 receptor-associated kinase 1
***IRF1*:** interferon regulatory factor
***JAK2*:** Janus kinase 2
***JUN*:** Jun proto-oncogene
KEGG: Kyoto Encyclopedia of Genes and Genomes
***MAF*:** avian musculoaponeurotic fibrosarcoma oncogene homolog
***MAPK*:** mitogen-activated protein kinase
MS: multiple sclerosis
***MYD88*:** myeloid differentiation primary response gene 88
***NFATC1*:** nuclear factor of activated T cells, cytoplasmic, calcineurin dependent 1
***NFATC2*:** nuclear factor of activated T cells, cytoplasmic, calcineurin dependent 2
NF-κB: nuclear factor kappa B
***NICD1*:** notch1 intracellular domain 1
***NOTCH1*:** neurogenic locus notch homolog protein 1
***p53*:** transformation-related protein 53
***PD1*:** programmed cell death protein 1
***PDL2*:** programmed cell death 1 ligand 2
***PPARG*:** peroxisome proliferator–activated receptor gamma
PPI: protein-protein interaction
***PU.1*:** spleen focus forming virus (SFFV) proviral integration oncogene
***R2*:** ribonucleotide reductase M2
***RORA*:** RAR-related orphan receptor alpha
***SATB1*:** special AT-rich sequence binding protein 1
***SGK1*:** serum/glucocorticoid regulated kinase 1
***SHP2*:** protein tyrosine phosphatase, non-receptor type 11
***SIRT1*:** sirtuin 1
SLE: systemic lupus erythematosus
***SMAD3*:** SMAD family member 3
***SMAD4*:** SMAD family member 4
***SMAD6*:** SMAD family member 6

***SOCS1*:** suppressor of cytokine signaling 1
***STAT*:** signal transducer and activator of transcription
TGF: transforming growth factor
Th: T helper
***TL1A*:** tumor necrosis factor (ligand) superfamily, member 15
TLR: Toll-like receptor
TNF: tumor necrosis factor
***TRAF6*:** TNF receptor–associated factor 6
***TSLP*:** thymic stromal lymphopoietin
***TSLPR*:** thymic stromal lymphopoietin receptor

Edited by T Leung; submitted 03.09.22; peer-reviewed by H Mohammed, R Pillai, O Rahaman; comments to author 08.11.22; revised version received 18.01.23; accepted 25.01.23; published 23.02.23.

Please cite as:
Khokhar M, Purohit P, Gadwal A, Tomo S, Bajpai NK, Shukla R
The Differentially Expressed Genes Responsible for the Development of T Helper 9 Cells From T Helper 2 Cells in Various Disease States: Immuno-Interactomics Study
JMIR Bioinform Biotech 2023;4:e42421
URL: https://bioinform.jmir.org/2023/1/e42421
doi:10.2196/42421
PMID:

Original Paper

SARS-CoV-2 Omicron Variant Genomic Sequences and Their Epidemiological Correlates Regarding the End of the Pandemic: In Silico Analysis

Ashutosh Kumar[1,2], MD; Adil Asghar[1,2], MD; Himanshu N Singh[2,3], PhD; Muneeb A Faiq[2,4], PhD; Sujeet Kumar[2,5], PhD; Ravi K Narayan[2,6], MD; Gopichand Kumar[1,2], MBBS; Prakhar Dwivedi[1,2], MBBS; Chetan Sahni[2,7], MD; Rakesh K Jha[1,2], MD; Maheswari Kulandhasamy[2,8], MD; Pranav Prasoon[2,9], PhD; Kishore Sesham[2,10], MD; Kamla Kant[2,11], MD; Sada N Pandey[2,12], PhD

[1]Department of Anatomy, All India Institute of Medical Sciences-Patna, Patna, India

[2]Etiologically Elusive Disorders Research Network, New Delhi, India

[3]Department of Systems Biology, Columbia University Irving Medical Center, New York, NY, United States

[4]New York University Langone Health Center, Robert I Grossman School of Medicine, New York University, New York, NY, United States

[5]Center for Proteomics and Drug Discovery, Amity Institute of Biotechnology, Amity University, Maharashtra, Mumbai, India

[6]Dr BC Roy Multi-speciality Medical Research Centre, Indian Institute of Technology, Kharagpur, India

[7]Department of Anatomy, Institute of Medical Sciences, Banaras Hindu University, Varanasi, India

[8]Department of Biochemistry, Maulana Azad Medical College, New Delhi, India

[9]School of Medicine, University of Pittsburgh, Pittsburgh, PA, United States

[10]Department of Anatomy, All India Institute of Medical Sciences-Mangalagiri, Mangalagiri, India

[11]Department of Microbiology, All India Institute of Medical Sciences-Bathinda, Bathinda, India

[12]Department of Zoology, Banaras Hindu University, Varanasi, India

Corresponding Author:
Ashutosh Kumar, MD
Department of Anatomy
All India Institute of Medical Sciences-Patna
Phulwari Sharif
Patna, 801507
India
Phone: 91 0612245 ext 1335
Email: drashutoshkumar@aiimspatna.org

Abstract

Background: Emergence of the new SARS-CoV-2 variant B.1.1.529 worried health policy makers worldwide due to a large number of mutations in its genomic sequence, especially in the spike protein region. The World Health Organization (WHO) designated this variant as a global variant of concern (VOC), which was named “Omicron.” Following Omicron’s emergence, a surge of new COVID-19 cases was reported globally, primarily in South Africa.

Objective: The aim of this study was to understand whether Omicron had an epidemiological advantage over existing variants.

Methods: We performed an in silico analysis of the complete genomic sequences of Omicron available on the Global Initiative on Sharing Avian Influenza Data (GISAID) database to analyze the functional impact of the mutations present in this variant on virus-host interactions in terms of viral transmissibility, virulence/lethality, and immune escape. In addition, we performed a correlation analysis of the relative proportion of the genomic sequences of specific SARS-CoV-2 variants (in the period from October 1 to November 29, 2021) with matched epidemiological data (new COVID-19 cases and deaths) from South Africa.

Results: Compared with the current list of global VOCs/variants of interest (VOIs), as per the WHO, Omicron bears more sequence variation, specifically in the spike protein and host receptor-binding motif (RBM). Omicron showed the closest nucleotide and protein sequence homology with the Alpha variant for the complete sequence and the RBM. The mutations were found to be primarily condensed in the spike region (n=28-48) of the virus. Further mutational analysis showed enrichment for the mutations decreasing binding affinity to angiotensin-converting enzyme 2 receptor and receptor-binding domain protein expression, and for increasing the propensity of immune escape. An inverse correlation of Omicron with the Delta variant was noted (r=–0.99,

P<.001; 95% CI –0.99 to –0.97) in the sequences reported from South Africa postemergence of the new variant, subsequently showing a decrease. There was a steep rise in new COVID-19 cases in parallel with the increase in the proportion of Omicron isolates since the report of the first case (74%-100%). By contrast, the incidence of new deaths did not increase (r=–0.04, *P*>.05; 95% CI –0.52 to 0.58).

Conclusions: In silico analysis of viral genomic sequences suggests that the Omicron variant has more remarkable immune-escape ability than existing VOCs/VOIs, including Delta, but reduced virulence/lethality than other reported variants. The higher power for immune escape for Omicron was a likely reason for the resurgence in COVID-19 cases and its rapid rise as the globally dominant strain. Being more infectious but less lethal than the existing variants, Omicron could have plausibly led to widespread unnoticed new, repeated, and vaccine breakthrough infections, raising the population-level immunity barrier against the emergence of new lethal variants. The Omicron variant could have thus paved the way for the end of the pandemic.

(JMIR Bioinform Biotech 2023;4:e42700) doi:10.2196/42700

KEYWORDS

COVID-19; pandemic; variants; immune escape; transmissibility; virulence; policy; mutations; epidemiology; data; Omicron; virus; transmission; genomic

Introduction

Background

A new variant of SARS-CoV-2 (lineage B.1.1.529) was reported from Botswana, South Africa, and multiple other countries [1], which the World Health Organization (WHO) designated as a global variant of concern (VOC) named “Omicron” [2]. The new variant was classified in the PANGO (Phylogenetic Assignment of Named Global Outbreak) lineage as BA.1. The presence of a large number of mutations in its genomic sequence—especially in the spike protein region, including in the host receptor-binding domain (RBD)—raised speculations that Omicron can prove to be a serious epidemiological threat and contributor to subsequent COVID-19 waves globally [3]. Multiple sublineages of Omicron were then identified with a slightly varying set of mutations [4]. These Omicron subvariants differentially affected the global population, leading to burst waves in various parts of the world [5]. Omicron is currently the predominant strain causing most of the new COVID-19 cases globally [5].

Significance of the Study

Owing to the heterogeneity of previous infections and vaccination coverage across the global population, there has been significant ambiguity in reports on the epidemiological properties of Omicron [6-9]. Specifically, it remains unclear whether the Omicron variant has an epidemiological advantage over existing variants [8]. Many researchers have proposed that Omicron’s emergence has changed the pandemic’s evolutionary course and speculated its end [10-12]. However, contradictory views are also being presented, suggesting against any sooner end of the pandemic and the possibility of the emergence of more lethal variants as the immunity that the global population gained from previous infections and vaccines fades [13]. Therefore, we aimed to resolve the existing ambiguity over the epidemiological properties of the Omicron variant using an integrated approach combining viral genomic sequence analysis and epidemiological data. Integrating viral genomic analysis with epidemiological data is a relatively novel approach; however, its success in predicting the epidemiological properties of SARS-CoV-2 variants and the future course of the COVID-19 pandemic has been validated in recent bioinformatic studies [14,15]. The findings of this study will thus provide concrete insights into the origin and epidemiological attributes of this variant to pave the way for the end of the pandemic.

Objectives

We performed an in silico analysis of the complete genomic sequences of the Omicron BA.1 variant available on the Global Initiative on Sharing Avian Influenza Data (GISAID) platform [16] with the primary objective of predicting the functional impact of the mutations present in this variant on virus-host interactions in terms of viral transmissibility, virulence, and immune-escape capabilities. Moreover, we assessed the relative proportion of the genomic sequences of existing SARS-CoV-2 variants, which was correlated with the rise in new COVID-19 cases in the global geographical location most affected by Omicron to understand whether the new variant had an epidemiological advantage in terms of transmissibility and virulence/lethality over existing variants.

Methods

Data Collection

The SARS-CoV-2 genomic sequence for the Omicron variant and other global VOCs/variants of interest (VOIs) were downloaded from the EpiCoV database of GISAID [16] using the automatic search function feeding information for geographical location, SARS-CoV-2 lineage, sample collection, and sequence reporting dates (up to December 10, 2021). The optimum length and coverage of the downloaded sequences (used for variant comparisons) were obtained by selecting the “complete sequence” and “high coverage” options in the search function.

Data Analysis

Mutational analysis on the genomic sequences was performed, and the 3D structure of the spike protein with amino acid changes in Omicron was generated using the CoVsurver app provided by GISAID [16], employing hCoV-19/Wuhan/WIV04/2019 as the reference strain. Further, a comparative mutational analysis of Omicron with existing global VOCs/VOIs (as per the WHO) [17] was generated using

the "compare lineages" function at outbreak.info [18] with GISAID as the source of genomic sequence data. The Expasy Swiss Bioinformatics portal [19] was used for protein sequence translation from the viral genomic sequences. A comparative assessment of the Omicron nucleotide and protein sequences with existing global VOCs/VOIs was performed using the National Center for Biotechnology Information (NCBI) Blast tool [20].

Furthermore, the functional impact of the mutations present at the RBD of the variants was assessed using an open analysis pipeline developed by Starr et al [21], which integrates a yeast-display platform with deep mutational scanning to determine how all possible RBD amino acid mutations affect angiotensin-converting enzyme 2 (ACE2)-binding affinity and protein expression (a correlate of protein folding stability) as compared to the wild-type SARS-CoV-2 strain [22].

The epidemiological correlates of the Omicron variant were assessed based on the comparative analysis of the genomic sequences from GISAID [16] and current epidemiological data (daily new cases and deaths) made available at Worldometer for South Africa [23], as one of the regions most strongly affected by the variant (last date of collection: December 10, 2021). The number of sequences for each SARS-CoV-2 variant was retrieved using an automatic search function feeding information for the lineage and collection dates in the EpiCoV database of GISAID for the period of October 1, 2021, to December 10, 2021. A 3-day sum of the total number of sequences was noted for each variant and their relative proportions were calculated (in percentages). Data were tabulated and the distribution of each variant was charted against the COVID-19 epidemiological data (3-day sum of new cases and deaths). Statistical analysis was performed to appreciate the changes in the relationship between the variables before and after the emergence of Omicron.

Statistical Analysis

An expected (*E*) value ≤0 was considered significant for the sequence homology match through NCBI Blast. An E value close to 0 or below and a higher Max score indicate a higher sequence homology ranking (see [24] for further details of the statistical methods in predicting significance in similarity scores). For the mutational analysis, only the mutations present in at least 75% of sequenced samples were considered for functional characterization.

For the analysis of epidemiological data, statistical tests were performed to evaluate intergroup differences among SARS-CoV-2 variants in Microsoft Excel 2019 and the R statistical package version 4.2.2. The normality of the data was examined using the Shapiro-Wilk test. Pearson (*r*) and Spearman (ρ) correlation tests were performed for the normally distributed and skewed data, respectively. A correlation matrix was generated and linear regression analysis was performed between the comparing variables (presented as *r* values, ranging from 0 to 1, and 95% CIs). Results were considered statistically significant at $P \leq .05$. Graphs were plotted to visualize the data trends.

Ethical Considerations

Approval from the institutional ethics committee was precluded as publicly available/open access databases were used for this study.

Results

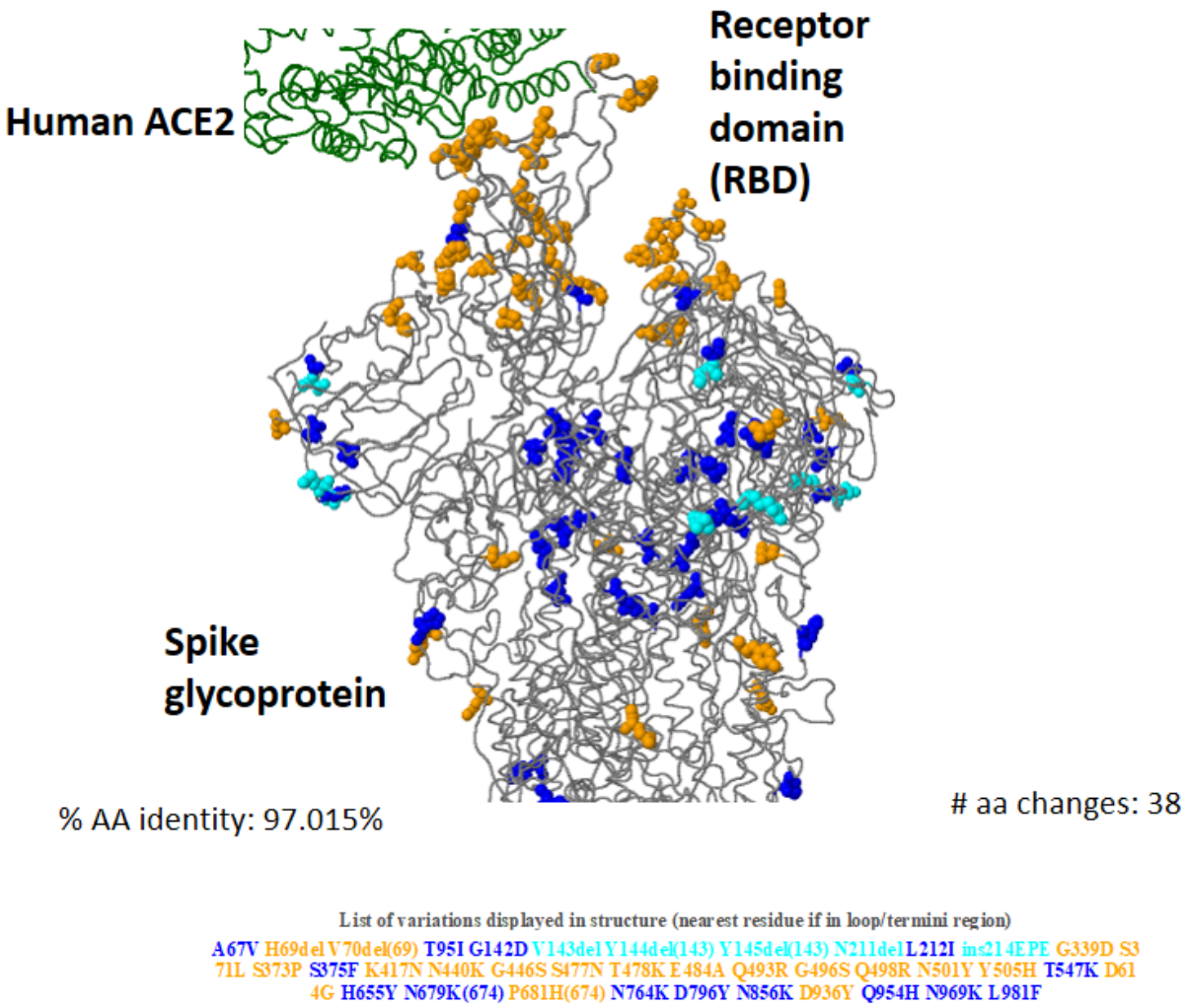
Data Summary

A total of 3604 genomic sequences of Omicron from 54 countries were uploaded on GISAID up to December 10, 2021 (see Figure S1 in Multimedia Appendix 1), which were analyzed for mutational characteristics. The mutations found were primarily condensed in the spike protein region (n=28-48) of the virus; however, frequent nonspike mutations were also noted (n=20-26). In this study, we focused on analyzing the genomic sequences of Omicron's initially most prevalent sublineage (BA.1).

Sequence Homology of Omicron (BA.1) With Wild-Type Strains

Compared to the current list of global VOCs/VOIs (as per the WHO), Omicron showed more sequence variation, specifically in the spike protein (nucleotides 21,563-25,384; amino acids 1-1273), including the receptor-binding motif (RBM; nucleotides 22,869-23,089 and amino acids 438-508), where the riffs were most prominent (Table S1 in Multimedia Appendix 1). The homology of Omicron to the reference strain (hCoV-19/Wuhan/WIV04/2019) for the spike protein sequence varied from 96.23% to 97.8% (28-48 mutations) in the analyzed sequences (Figure 1).

Figure 1. Three-dimensional structure of Omicron (BA.1) spike glycoprotein in the interaction of human angiotensin-converting enzyme 2 (ACE2), showing key amino acid substitutions. (Data source: CoVsurver app from GISAID [16]). AA/aa: amino acid.



Sequence Homology of Omicron (BA.1) With Existing SARS-CoV-2 VOCs/VOIs

The analysis of Omicron's genomic and protein sequence homology with the reference strain and current global VOCs/VOIs (as per the WHO) showed the highest similarity of Omicron with the Alpha variant for the complete sequence as well as for the RBM. However, the highest similarity for the complete nucleotide and protein sequences for the spike protein were noted with the Beta and Delta variant, respectively (see Table S1 in Multimedia Appendix 1).

Mutational Analysis

Multiple clusters of closely spaced mutations were noted across the sequence, which were most densely located in the spike protein region, particularly in its S1 subunit, including the host RBM (Figure 2, Table 1). Many of the mutations in Omicron are shared with the current global VOCs/VOIs (Figure 3).

Tables 2-6 summarize the reported mutations in Omicron (BA.1) (present in at least 75% of sequences) and their functional characteristics based on the existing literature [16,21,25-41]. According to the available evidence, these mutations in PANGO lineage BA.1 can be broadly categorized into four major groups: immune escape (n=20) (Table 2), host receptor binding (n=10) (Table 3), virus replication (n=18) (Table 4), and host adaptability (n=3) (Table 5). Mutations outside of the spike protein are summarized in Table 6.

Figure 2. The mutational landscape in SARS-CoV-2 variant B.1.1.529 (Omicron, sublineage: BA.1). The analysis of the mutations present at the RBD using a deep mutational scanning pipeline by Starr et al [21] reflected prominent ACE2-binding affinity and protein expression changes (see Table 1). Notably, mutations decreasing the ACE2-binding affinity and protein expression were significantly greater in number. ACE2: angiotensin-converting enzyme 2; ORF: open reading frame; RBD: receptor-binding domain; RBM: receptor-binding motif.

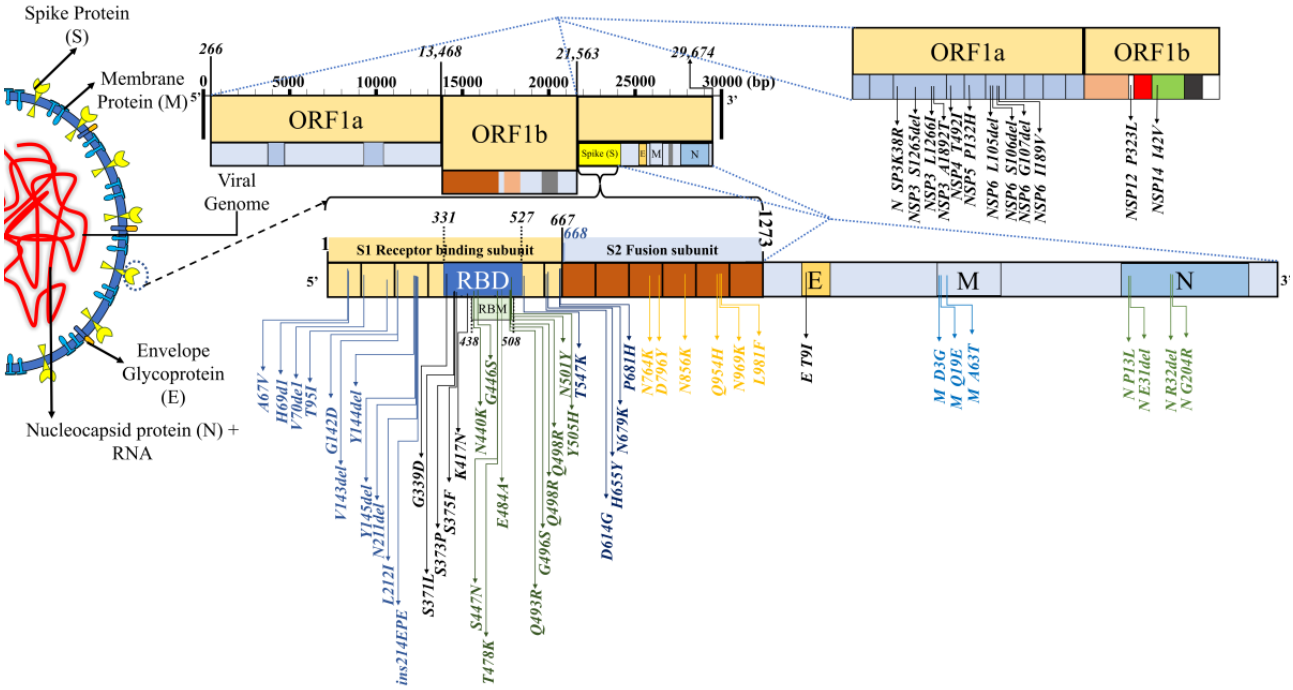


Table 1. Predicted impact of receptor-binding motif variations in the SARS-CoV-2 variant B.1.1.529 (Omicron, sublineage: BA.1) on interactions with the host.[a]

ACE2[b] binding site mutations	ACE2 binding (Δlog10 KD app[c,d])	Protein expression (Δlog mean MFI[e,f])	ACE2 contact with SARS-CoV-2	RSA[g] bound	SARS-CoV-1 amino acid	RaTG13 amino acid	GD Pangolin-CoV amino acid
G339D	0.06	0.30	false	0.47	G	G	G
S371L	–0.14	–0.61	false	0.46	S	S	S
S373P	–0.08	–0.22	false	0.48	F	S	S
S375F	–0.55	–1.81	false	0.48	S	S	S
K417N	–0.45	0.10	true	0.19	V	K	R
N440K	0.07	–0.12	false	0.68	N	H	N
G446S	–0.20	–0.40	true	0.55	T	G	G
S477N	0.06	0.06	false	0.76	G	S	S
T478K	0.02	0.02	false	0.48	K	K	T
E484A	–0.07	–0.23	false	0.50	P	T	E
Q493R	–0.09	–0.06	true	0.10	N	Y	Q
G496S	–0.63	0.12	true	0.04	G	G	G
Q498R	–0.06	0.10	true	0.00	Y	Y	H
N501Y	0.24	–0.14	true	0.03	T	D	N
Y505H	–0.71	0.16	true	0.12	Y	H	Y

[a]Based on the study of Starr et al [21].

[b]ACE2: angiotensin-converting enzyme 2.

[c]KD app: apparent dissociation constant.

[d]A positive Δlog10 KD app value relative to the unmutated SARS-CoV-2 receptor-binding domain (3.9×10^{-11} M) indicates stronger binding.

[e]MFI: mean fluorescence intensity.

[f]Positive Δlog MFI values relative to the unmutated SARS-CoV-2 receptor-binding domain indicate increased expression.

[g]RSA: relative solvent accessibility.

Figure 3. Lineage comparison between Omicron and other global variants of concerns/interest. Only mutations with >75% prevalence in at least one lineage are shown. (Data source: outbreak.info, based on the SARS-CoV-2 genomic sequences uploaded in GISAID until December 6, 2021).

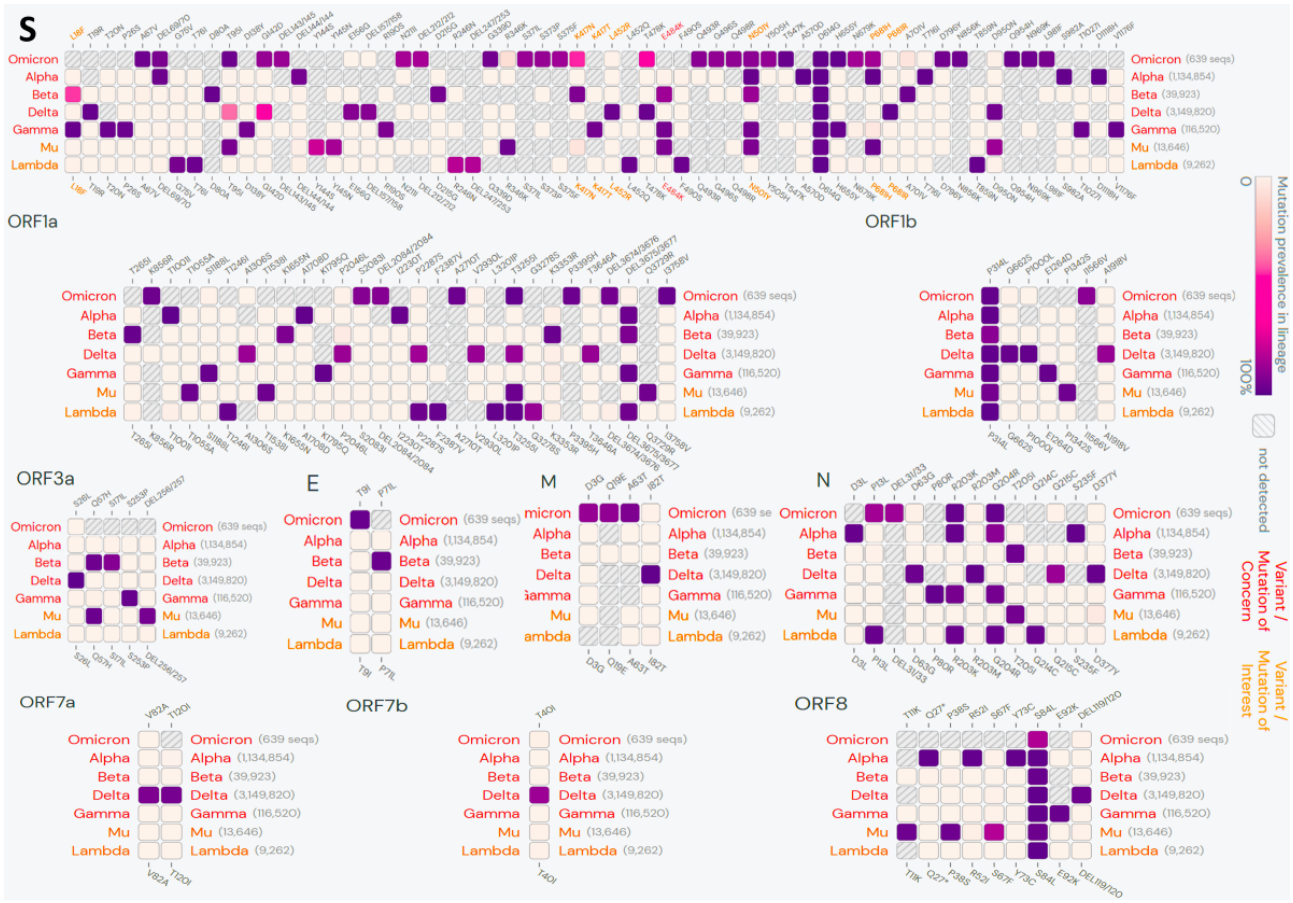


Table 2. Mutations in SARS-CoV-2 variant B.1.1.529 (Omicron, sublineage BA.1) spike protein influencing immune escape via antibody recognition sites and/or antigenic drift.[a]

Mutation	Frequency (%)[b]	Remarks	Reference
H69del	20.35	H69del+V70del have 2-fold higher infectivity compared to the wild type. H69del+V70del-containing viruses showed reduced neutralization sensitivity to mAb[c] COVA1-21, targeting an as-yet-undefined epitope outside the RBD[d]	[16]
V70del	20.37	H69del+V70del have 2-fold higher infectivity compared to the wild type. H69del+V70del-containing viruses showed reduced neutralization sensitivity to mAb COVA1-21, targeting an as-yet-undefined epitope outside the RBD	[16]
V143del	0.12	N/A[e]	[16]
Y144del	20.94	Decreased sensitivity to convalescent sera	[25,26]
Y145del	2.33	Decreased sensitivity to convalescent sera	[16,25,26]
G339D	0.01	N/A	[16]
S371L	0.00	N/A	[16,21]
S373P	0.01	N/A	[16,21]
S375F	0.00	N/A	[16,27]
K417N	0.83	N/A	[16,21,28]
N440K	0.17	N/A	[16,21]
G446S	0.01	N/A	[16,28]
S477N	1.31	S477N was also resistant to neutralization by the human convalescent sera tested in this study, but not to vaccine-elicited sera	[16,21,29]
E484A	0.02	N/A	[27]
Q493R	0.01	N/A	[30,31]
G496S	0.01	N/A	[21]
Q498R	0.00	N/A	[16,27]
N501Y	24.11	Associated with increased transmissibility and increased affinity for human ACE2[f] receptor	[16,21,28,32]
H655Y	2.25	N/A	[34-36]

[a]Based on the genomic sequences of Omicron uploaded on GISAID [16] (last date of collection December 10, 2021).

[b]Among all SARS-CoV-2 genomic sequences uploaded on GISAID [16].

[c]mAB: monoclonal antibody.

[d]RBD: receptor-binding domain.

[e]N/A: not applicable.

[f]ACE2: angiotensin-converting enzyme 2.

Table 3. Mutations in SARS-CoV-2 variant B.1.1.529 (Omicron, sublineage BA.1) spike protein influencing receptor binding.[a]

Mutation	Frequency (%)[b]	Effect on virus-host interactions	Remarks	Reference
G339D	0.01	Increased RBD[c] expression	N/A[d]	[16,21]
S371L	0.00	Increased ACE2[e] binding	N/A	[16,21]
S373P	0.01	Increased RBD expression	N/A	[16,21]
K417N	0.83	Increased RBD expression	N/A	[16,21,28]
N440K	0.17	Increased ACE2 binding	N/A	[16,21]
S477N	1.31	Increased ACE2 binding/ increased RBD expression	S477N was also resistant to neutralization by the human convalescent sera tested in this study, but not to vaccine-elicited sera	[16,21,29]
T478K	52.56	Increased ACE2 binding/increased RBD expression	Decreased sensitivity to convalescent sera	[21,25]
Q493R	0.01	Host change	N/A	[30,31]
G496S	0.01	Increased RBD expression	N/A	[21]
N501Y	24.11	Increased ACE2 binding/host change	Associated with increased transmissibility and increased affinity for human ACE2 receptor	[16,21,28,32]
Y505H	0.00	Increased RBD expression	N/A	[16,21]
D614G	98.51	Increased infectivity	Lower cycle threshold values were observed in G614 infections, indicating a higher viral load	[16,25,33]

[a]Based on the genomic sequences of Omicron uploaded on GISAID [16] (last date of collection December 10, 2021).

[b]Among all SARS-CoV-2 genomic sequences uploaded on GISAID [16].

[c]RBD: receptor-binding domain.

[d]N/A: not applicable.

[e]ACE2: angiotensin-converting enzyme 2.

Table 4. Mutations in SARS-CoV-2 variant B.1.1.529 (Omicron, sublineage BA.1) spike protein influencing viral oligomerization interfaces.[a]

Mutations	Frequency (%)[b]	Remarks	Reference
S371L	0.00	N/A[c]	[16,21]
S373P	0.01	N/A	[16,21]
S375F	0.00	N/A	[16,27]
K417N	0.83	N/A	[16,21,28]
S477N	1.31	S477N was also resistant to neutralization by the human convalescent sera tested in this study, but not to vaccine-elicited sera	[16,21,29]
Q493R	0.01	N/A	[30,31]
N501Y	24.11	Associated with increased transmissibility and increased affinity for human ACE2[d] receptor	[16,21,28,32]
Y505H	0.00	N/A	[16,21]
N764K	0.01	N/A	[16]
D796Y	0.08	N/A	[16]
N856K	0.00	N/A	[16]
Q954H	0.00	N/A	[16,40]
N969K	0.00	N/A	[16]
L981F	0.00	N/A	[16]

[a]Based on the genomic sequences of Omicron uploaded on GISAID [16] (last date of collection December 10, 2021).

[b]Among all SARS-CoV-2 genomic sequences uploaded on GISAID [16].

[c]N/A: not applicable.

[d]ACE2: angiotensin-converting enzyme 2.

Table 5. Mutations in SARS-CoV-2 variant B.1.1.529 (Omicron, sublineage BA.1) spike protein influencing host adaptation and other mechanisms.[a]

Mutations	Frequency (%)[b]	Effect on virus-host interactions	Remarks	Reference
A67V	0.36	Unknown	N/A[c]	[16]
T95I	21.32	Unknown	N/A	[16]
G142D	33.40	Unknown	N/A	[16]
Q954H	0.00	Host adaptation (cell culture)	N/A	[16,40]
N211del	0.02	Unknown	N/A	[16]
L212I	0.01	Unknown	N/A	[16]
ins214EPE	0.00	Unknown	N/A	[16]
H655Y	2.25	Host adaptation (cats); spike glycoprotein fusion efficiency	N/A	[32–34]
N679K	0.09	Unknown	N/A	[16,37]
P681H	22.73	Unknown	P681H mutation at the S1/S2 site of the SARS-CoV-2 spike protein may increase its cleavability by furin-like proteases, but this does not translate into increased virus entry or membrane fusion	[16,38,39]
T547K	0.00	Unknown	N/A	[16]
N856K	0.00	Ligand binding	N/A	[16]

[a]Based on the genomic sequences of Omicron uploaded on GISAID [16] (last date of collection December 10, 2021).

[b]Among all SARS-CoV-2 genomic sequences uploaded on GISAID [16].

[c]N/A: not applicable.

Table 6. Mutations in SARS-CoV-2 variant B.1.1.529 (Omicron, sublineage BA.1) outside of the spike protein.[a]

Mutations	Frequency (%)[b]	Effect on virus-host interactions	Remarks	References
Envelope (E) T9I	0.09	Viral oligomerization interfaces	N/A[c]	[16]
Membrane (M)				
M D3G	0.08	Unknown	N/A	[16]
M Q19E	0.00	Unknown	N/A	[16]
M A63T	0.01	Unknown	N/A	[16]
Nucleocapsid (N)				
N P13L	0.63	Antigenic drift	P13L variant in B*27:05-restricted CD8+ nucleocapsid epitope, showing complete loss of responsiveness to the T-cell lines evaluated	[16,41]
N E31del	0.00	Unknown	N/A	[16]
N R32del	0.00	Unknown	N/A	[16]
N G204R	26.20	Unknown	N/A	[16]
Nonstructural protein (NSP)				
NSP3 K38R	0.01	Unknown	N/A	[16]
NSP3 S1265del	0.02	Unknown	N/A	[16]
NSP3 L1266I	0.02	Unknown	N/A	[16]
NSP3 A1892T	0.00	Unknown	N/A	[16]
NSP4 T492I	47.76	Viral oligomerization interfaces	N/A	[16]
NSP5 P132H	0.01	Unknown	N/A	[16]
NSP6 L105del	0.02	Unknown	N/A	[16]
NSP6 S106del	24.74	Unknown	N/A	[16]
NSP6 G107del	24.74	Unknown	N/A	[16]
NSP6 I189V	0.03	Unknown	N/A	[16]
NSP12 P323L	96.69	Viral oligomerization interfaces	N/A	[16]
NSP14 I42V	0.00	Viral oligomerization interfaces	N/A	[16]

[a]Based on the genomic sequences of Omicron uploaded on GISAID [16] (last date of collection December 10, 2021).

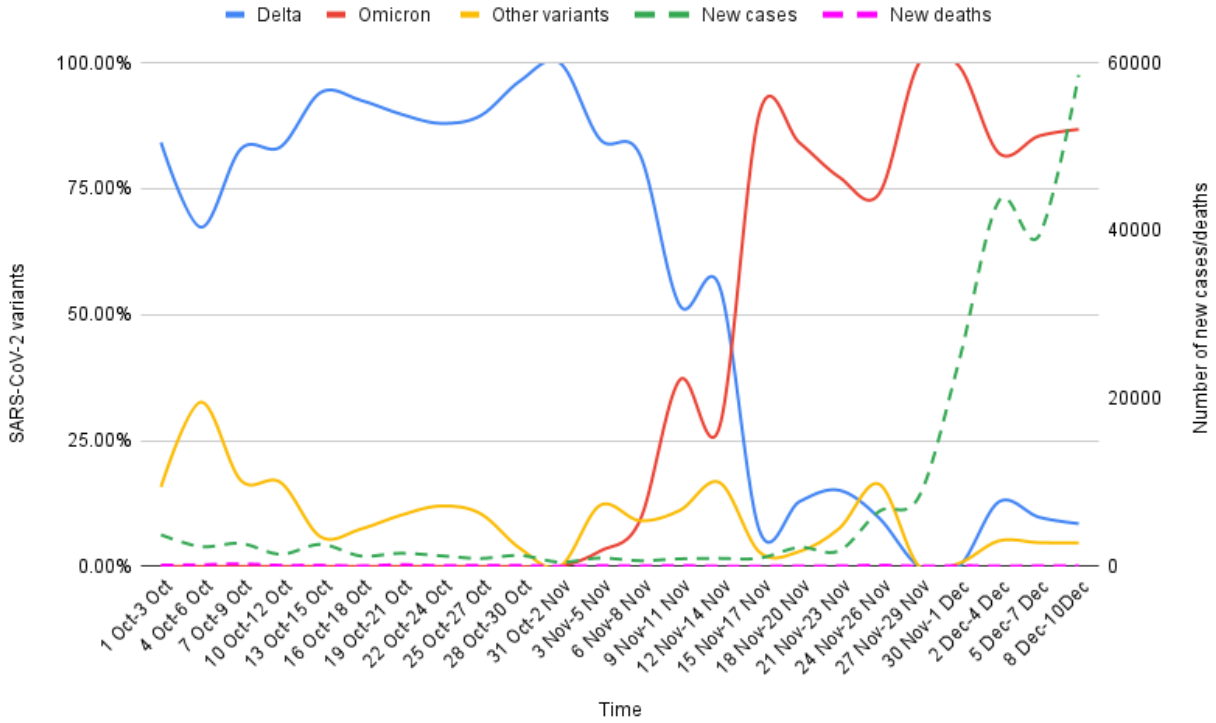
[b]Among all SARS-CoV-2 genomic sequences uploaded on GISAID [16].

[c]N/A: not applicable.

Epidemiological Correlates

A total of 4224 SARS-CoV-2 genomic sequences (Delta, n=999; Omicron, n= 2937; and others, n= 288) were uploaded on GISAID from South Africa in the period of study. For the complete duration of the study, Delta correlated negatively with the number of new COVID-19 cases (r=–0.567, P=.004; 95% CI –0.79 to –0.21) but correlated positively with the number of new deaths (r=0.38, P=.07; 95% CI –0.025 to 0.68). The differential analysis of the SARS-CoV-2 genomic sequences from South Africa before and after the emergence of the first case of Omicron (dated November 5, 2021, EPI_ISL_7456440) reflected a sharp change in the dominance of the variant from Delta to Omicron (Figure 4). An inverse correlation of Omicron with Delta variants was noted (r=–0.99, P<.001; 95% CI –0.99 to –0.97) in the period of study. There has been a steep rise in the number of new COVID-19 cases in parallel with the increase in the proportion of Omicron since the first case of Omicron (74%-100% of total genomic sequences after November 15-17, 2021). However, no parallel increase was observed in the death cases, which otherwise showed a reverse trend (r=–0.04, P=.02; 95% CI –0.52 to 0.58) (Figure 4).

Figure 4. Epidemiological correlates of Omicron and Delta variants genomic sequences reported on GISAID from South Africa for the period of October 1 to December 10, 2021. The proportion of Delta and Omicron variants among the total SARS-CoV-2 genomic sequences were correlated with the new COVID-19 cases and deaths in the study period (3-day sum of each variable). A sharp change in the dominance from Delta to Omicron was observable since the report of the first Omicron case (November 5, 2021). The rise of Omicron cases paralleled the increase in the new COVID-19 cases. In comparison, the Delta variant showed a fall in the same period. Notably, there has been no increase in the number of deaths postemergence of Omicron. (Data sources: GISAID and Worldometer).



Discussion

Principal Findings

Our analysis of the SARS-CoV-2 genomic sequences and epidemiological data from South Africa unravels multiple observations regarding host-virus interactions, which may help to predict the further epidemiological potential of the Omicron variant. We found that compared to the current list of global VOCs/VOIs (as per the WHO), Omicron bears more sequence variation, specifically in the spike protein and RBM. Omicron showed the closest nucleotide and protein sequence homology with the Alpha variant. Further, the mutational analysis showed enrichment for the mutations decreasing ACE2-binding affinity and RBD protein expression, but increased propensity of immune escape. The analysis of the viral genomic sequences and epidemiological data from South Africa reflected an inverse correlation of Omicron with Delta variant infections, with a subsequent decrease. There was a steep rise in the number of new COVID-19 cases in parallel with the increase in the proportion of Omicron since the report of the first case; however, the incidence of deaths did not increase.

Sequence Homology With Wild-Type Strains and Existing SARS-CoV-2 VOCs/VOIs

Our analyses showed that among the existing VOCs and VOIs, Omicron bears the highest homology of the complete sequence and RBM (nucleotide and protein sequences) with the Alpha variant (Table S1 in Multimedia Appendix 1). Interestingly, similar to Alpha variant spike gene target failure, polymerase chain reaction (PCR)-based detection is a sensitive method for detecting Omicron in clinical samples [42].

As Omicron bears key mutations from multiple existing VOCs/VOIs, with approximate sequence homology variation rather than a direct descendance, the numerous recombination events between the variants inside hosts can be a more plausible explanation for its origin.

It will be pertinent to explore the evolutionary mechanisms involved in accumulating such a large number of mutations in Omicron. Speculations were raised that the long-term persistence of SARS-CoV-2 infection in an immunocompromised host could be a probable mechanism behind the origin of Omicron [43-46]. Avanzato et al [43] and Choi et al [45] reported case studies of the persistence of infection and accumulation of novel mutations in the SARS-CoV-2 spike gene and RBD in chronically ill and immunocompromised COVID-19 patients. Another such case was reported by Karim et al [44]. The authors documented the long persistence of SARS-CoV-2 infection (for more than 6 months) in a patient with advanced HIV and antiretroviral treatment failure. Through whole-genome sequencing for SARS-CoV-2 performed at multiple time points from patient samples, the authors demonstrated the early emergence of the E484K substitution, followed by N501Y, K417T, and many other mutations (including some novel mutations) in the spike gene and RBD. An increase in the genomic diversity reflecting the intrahost evolution of SARS-CoV-2 during prolonged infection was also noted in a recent cohort study by Voloch et al [46].

Effect on Virus-Host Interactions

Our analysis shows that Omicron accumulated multiple closely spaced mutations at the RBM with ACE2 (Figure 2). Notably, this variant has many of the mutations common with the earlier VOCs (Figure 3), many of which have been shown to enhance RBD-ACE2 binding in comparison to the wild-type strain [47] (Tables 1 and 3). The selective mutations present at or near the vicinity of the RBM (N440K, S477N, T478K, and N501Y) in most of the Omicron sequences are believed to stabilize binding with ACE2 (Tables 2-3). D614G, a critical mutation in all B.1 descendants [47], is known to stabilize the trimeric structure and create a more open conformation of the RBD, allowing stronger binding with ACE2 [47]. Paradoxically, our analysis suggests that the majority of the novel or rare spike mutations (<0.2% prevalence in the total sequenced samples, Tables 2-6) in Omicron may have a deleterious effect on host interactions owing to their presence at the constrained RBD regions in terms of ACE2 binding (10/15) and/or RBD expression (8/15) (Table 3). Notably, most of the spike mutations that predicted a favorable effect on ACE2 binding, RBD expression, or both are present in current VOCs, primarily the Delta (T478K), Alpha (N501Y), and Beta (K417N) variants. Further, a set of mutations in Omicron that are present inside (P681H) or in the vicinity (D614G, H655Y) of the furin cleavage site of SARS-CoV-2 spike protein—a small stretch of peptide (PRRAR) inserted at the intersection of spike segments S1 and S2 (amino acid residues 681-685)—can enhance proteolytic cleavage of spike protein by a host protease (furin), which is considered to improve its fusion to the host cell membrane [48]. P681H is characteristically present in multiple VOCs/VOIs such as B.1.1.7, P.1, Q.1, and B.1.621 lineage variants [49]. A mutation at the exact location, P681R, has been present in the Delta variant and its emerging sublineages [50]. Characterizing the individual mutations on RBM specifies that Omicron may not have more efficient interactions with the host than existing VOCs/VOIs, specifically Delta. Further assessment of the allosteric influence and dynamic interactions of the mutations present at the RBD and other regions of spike protein and in situ/in vivo studies will be necessary to understand their exact impact on host-receptor binding and its clinical correlates. The clinical data on the severity of the disease indicated a milder illness in Omicron infection than in the existing VOCs [51,52].

Viral Replication

Many of the mutations, especially in the nonspike regions, are linked with viral oligomerization, synthesis, and packaging of the ribonucleic acid core (Tables 4 and 6). These mutations likely have a role in virus replication inside the host cells [53]. The NSP12 P323L mutation located in the RNA-dependent RNA polymerase coding region is of particular interest (Figure 1, Table 6), as this has been a frequently observed mutation in the earlier variants (96.69%) (Table S1 in Multimedia Appendix 1). However, whether these mutations will have a positive or negative impact on viral replication remains unclear. Interestingly, the results of a comparative study [54] that employed ex vivo cultures of SARS-CoV-2 strains isolated from the respiratory tract of infected patients indicated higher replication rates for the Omicron variant. The authors observed that after 24 hours of incubation, Omicron replicated 70 times faster than wild-type and Delta variant strains in the human bronchus. In contrast, it replicated less efficiently (>10 times lower) in the human lung tissue than the wild-type strain and the replication rate was also lower than that of the Delta variant.

Immune Escape

Most spike mutations (18/32) in Omicron have occurred at the known antibody recognition sites (Table 2). Existing studies have established the role of these mutations in immune escape against convalescent sera, vaccine-acquired antibodies, and therapeutically used monoclonal antibodies (Table 2). The evidence from in situ studies indicates potential immune escape by Omicron against convalescent sera, vaccine-acquired antibodies, and therapeutically used monoclonal antibodies [42,55,56]. Interestingly, Omicron contains the K417N and E484A mutations, which are present in multiple existing variants and are believed to contribute to immune escape [47]. Of note, the K417 locus is a known epitope for CB26, a therapeutically used monoclonal antibody in COVID-19 [47]. A more significant number of mutations in Omicron spike protein, specifically in the RBD, may be an evolutionary gain in this variant, providing it with higher immune-escape ability. Support for this notion comes from a study by Nabel et al [57], who demonstrated that SARS-CoV-2 pseudotypes containing up to seven mutations, as opposed to the one to three found in earlier VOCs, were more resistant to neutralization by therapeutic antibodies and serum from vaccine recipients [57].

A nonspike mutation in the nucleocapsid (N) protein (P13L) present in Omicron (Table 6) was shown to cause complete loss of recognition by epitope-specific (B∗27:05-restricted CD8+ nucleocapsid epitope QRNAPRITF$_{9\text{-}17}$) T cells in a cell line–based in situ study [41]. However, no such evidence in human samples is currently available. In another study, Redd et al [58] examined peripheral blood mononuclear cell samples from PCR-confirmed, recovered/convalescent COVID-19 cases (N=30) for their anti-SARS-CoV-2 CD8+ T-cell responses with Omicron. The authors noted that only one low-prevalence (found in 7%) epitope (GVYFASTEK, restricted to HLA*A03:01 and HLA*A11:01) from the spike protein (T95I) region was mutated in Omicron [58]. The presence of these mutations raises concerns about escaping T cell immunity by Omicron [59] and hence should be explored in further detail.

The overall evidence supports Omicron's very high immune-escape ability [42,55,56,60]. Cele et al [42] tested the ability of plasma from 14 BNT162b2-vaccinated study participants to neutralize Omicron versus the wild-type D614G virus in a live virus neutralization assay. The authors observed that Omicron showed a 41-fold decline in the 50% focus reduction neutralization test geometric mean titer compared to the wild-type D614G virus in subjects without previous infection (6/14). Interestingly, earlier, those with the infection showed relatively higher neutralization titers with Omicron (6/14), which indicated that the last infection, followed by vaccination or booster, might increase the neutralization levels and confer protection from severe disease in cases of Omicron infection.

Epidemiological Correlates: Omicron Versus Delta Variants

The analysis of the SARS-CoV-2 genomic sequences from South Africa indicates that Omicron gained an advantage in terms of transmissibility over the Delta variant (Figure 4). A third COVID-19 wave driven by the Delta variant occurred in South Africa [61]; hence, the epidemiological characteristics of the Delta and Omicron variants in the local population should be analyzed in this backdrop. We observed that before the arrival of Omicron, the Delta variant was dominant locally; by contrast, at present, the majority of new sequences are from Omicron (Omicron vs Delta r=–0.99, P<.001; 95% CI –0.99 to –0.97) (Figure 4). The steep rise in the new COVID-19 cases in South Africa seems to be driven by Omicron, whereas Delta variant–linked cases are seeing a decline (Figure 4). The rapid rise in new COVID-19 cases connected with the emergence of a new SARS-CoV-2 variant strongly indicated the commencement of a new COVID-19 wave in South Africa [14].

Further, death, which is considered a strong indicator of virulence/lethality, showed a negative correlation (r=–0.04, P=.02; 95% CI –0.52 to 0.58) (Figure 4) with the rise in Omicron. However, death correlated positively with the Delta variant in the period postemergence (r=0.38, P=.07; 95% CI –0.025 to 0.68) over the complete study period. This pattern indicates that the reported incidences of death were primarily linked with Delta rather than with Omicron. The significantly reduced lethality of Omicron compared to Delta has been confirmed through recent epidemiological studies [62-64].

An approximately 2.4 (2.0-2.7) times higher transmissibility was suggested with Omicron compared to the Delta variant in the South African population [65]. An estimate from the United Kingdom indicated that Omicron's risk of spreading the infection to members of a household is 3 times higher than that of the Delta variant [66]. A significantly shorter incubation period and early reaching of the peak have been reported for the Omicron variant [67]. Based on the epidemiological patterns observed in South Africa in our analysis, an epidemiological advantage to Omicron in comparison to Delta can be inferred in terms of transmissibility [66]. However, we found no indications of increased lethality with Omicron compared to Delta and other variants circulating in the South African population.

Notably, the presence of an immunological barrier in the population imparted by the recent COVID-19 wave mediated by the Delta variant could be a likely reason for this variant's fall in new cases [7,68]. A continuous fall in Delta cases was also noticeable in the period before the emergence of Omicron (Figure 4), further substantiating this notion. The data records showed that a significant proportion of the local population in South Africa was fully vaccinated at the time of Omicron's emergence (25.2%) [69]. Notably, the high number of immune escape–related mutations in Delta could have contributed to lowered efficacy of the vaccines, immunity from natural infections, and therapeutically used antibodies [47]. As Omicron contains a much higher number of immune escape–related mutations, including many shared with Delta (Figure 3), Omicron might have added potential for vaccine breakthrough infections and reinfections. Similar speculations were presented by other authors and global health regulatory bodies [2,70,71]. A higher risk of reinfections with Omicron was indicated by Pulliam et al [72] based on a retrospective analysis of routine epidemiological surveillance data to examine whether SARS-CoV-2 reinfection risk has changed over time in South Africa in the context of the emergence of the consecutive variants: Beta, Delta, and Omicron. The authors noted that as compared to the first wave driven by wild-type strains, subsequent waves by Beta and Delta variants had a lower estimated hazard ratio for reinfection versus primary infection (relative hazard ratio for wave 2 versus wave 1: 0.75, 95% CI 0.59-0.97; for wave 3 versus wave 1: 0.71, 95% CI 0.56-0.92) in comparison to Omicron (Omicron surge for the period of November 1-27, 2021, versus wave 1: 2.39, 95% CI 1.88-3.11).

Study Limitations

We analyzed a limited number of genomic sequences and epidemiological data from specific geographical regions affected by Omicron. Further, the relative frequency of specific lineage-characterizing mutations in the Omicron variant may have varied since the study's inception. Both of these limitations may have an impact on the quality of the results.

Conclusion

In silico analysis of viral genomic sequences suggests that the Omicron variant has more remarkable immune-escape ability than the existing VOCs/VOIs, including Delta, but reduced virulence/lethality than other reported variants. The higher power for immune escape for Omicron was a likely reason for the resurgence in COVID-19 cases and its soon becoming a globally dominant strain. Being more infectious but less lethal than the existing variants, Omicron could have plausibly led to widespread unnoticed new, repeated, and vaccine breakthrough infections, raising the population-level immunity barrier against the emergence of new lethal variants. The Omicron variant could have thus paved the way for the end of the pandemic.

Data Sharing

Primary data used for this study are publicly available on the GISAID database [16] for SARS-CoV-2 genomic sequences and Worldometer [23] for epidemiological data. The categorized data for the study period can be availed from the corresponding author upon reasonable request.

Authors' Contributions

AK, GK, and PD collected samples and analyzed data. AK wrote the first draft. AA and HNS performed the statistical analysis. MAF, SK, RKN, RKJ, CS, MK, PP, KS, KK, and SNP reviewed and edited the paper. All authors consented to submit the final draft.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Global spread of the SARS-CoV-2 Omicron variant (Figure S1). Nucleotide and protein sequence homology of Omicron (BA.1) with wild-type SARS-CoV-2 and other global variants of concern/interest (Table S1).
[PDF File (Adobe PDF File), 343 KB - bioinform_v4i1e42700_app1.pdf]

References

1. Tracking of hCoV-19 variants. GISAID. URL: https://www.gisaid.org/hcov19-variants/ [accessed 2021-07-20]
2. Update on Omicron. World Health Organization. 2021 Nov 28. URL: https://www.who.int/news/item/28-11-2021-update-on-omicron [accessed 2021-12-04]
3. Callaway E. Heavily mutated Omicron variant puts scientists on alert. Nature 2021 Dec 25;600(7887):21. [doi: 10.1038/d41586-021-03552-w] [Medline: 34824381]
4. Omicron variant report. outbreak.info. URL: https://outbreak.info/situation-reports/omicron [accessed 2022-11-26]
5. One year since the emergence of COVID-19 virus variant Omicron. World Health Organization. 2022 Nov 25. URL: https://www.who.int/news-room/feature-stories/detail/one-year-since-the-emergence-of-omicron [accessed 2022-11-26]
6. Nealon J, Cowling BJ. Omicron severity: milder but not mild. Lancet 2022 Jan 29;399(10323):412-413 [FREE Full text] [doi: 10.1016/S0140-6736(22)00056-3] [Medline: 35065007]
7. Madhi SA, Kwatra G, Myers JE, Jassat W, Dhar N, Mukendi CK, et al. Population immunity and Covid-19 severity with Omicron variant in South Africa. N Engl J Med 2022 Apr 07;386(14):1314-1326 [FREE Full text] [doi: 10.1056/NEJMoa2119658] [Medline: 35196424]
8. Bhattacharyya RP, Hanage WP. Challenges in inferring intrinsic severity of the SARS-CoV-2 Omicron variant. N Engl J Med 2022 Feb 17;386(7):e14. [doi: 10.1056/NEJMp2119682] [Medline: 35108465]
9. Strasser Z, Hadavand A, Murphy S, Strasser Z, Hadavand PA, Murphy S. SARS-CoV-2 Omicron variant is as deadly as previous waves after adjusting for vaccinations, demographics, and comorbidities. Research Square Preprints. 2022 May 02. URL: https://assets.researchsquare.com/files/rs-1601788/v1_covered.pdf?c=1651763984 [accessed 2022-12-19]
10. Daria S, Islam MR. The SARS-CoV-2 omicron wave is indicating the end of the pandemic phase but the COVID-19 will continue. J Med Virol 2022 Jun 04;94(6):2343-2345 [FREE Full text] [doi: 10.1002/jmv.27635] [Medline: 35098543]
11. Murray CJL. COVID-19 will continue but the end of the pandemic is near. Lancet 2022 Jan 29;399(10323):417-419 [FREE Full text] [doi: 10.1016/S0140-6736(22)00100-3] [Medline: 35065006]
12. Robertson D, Doshi P. The end of the pandemic will not be televised. BMJ 2021 Dec 14;375:e068094. [doi: 10.1136/bmj-2021-068094] [Medline: 34906970]
13. Das S, Samanta S, Banerjee J, Pal A, Giri B, Kar SS, et al. Is Omicron the end of pandemic or start of a new innings? Travel Med Infect Dis 2022 Jul;48:102332 [FREE Full text] [doi: 10.1016/j.tmaid.2022.102332] [Medline: 35472451]
14. Kumar A, Asghar A, Dwivedi P, Kumar G, Narayan RK, Jha RK, et al. A bioinformatics tool for predicting future COVID-19 waves based on a retrospective analysis of the second wave in India: model development study. JMIR Bioinform Biotech 2022 Sep 22;3(1):e36860 [FREE Full text] [doi: 10.2196/36860] [Medline: 36193192]
15. de Hoffer A, Vatani S, Cot C, Cacciapaglia G, Chiusano ML, Cimarelli A, et al. Variant-driven early warning via unsupervised machine learning analysis of spike protein mutations for COVID-19. Sci Rep 2022 Jun 03;12(1):9275. [doi: 10.1038/s41598-022-12442-8] [Medline: 35661750]
16. GISAID. URL: https://gisaid.org/ [accessed 2022-11-26]
17. Tracking SARS-CoV-2 variants. World Health Organization. URL: https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/ [accessed 2021-06-17]
18. SARS-CoV-2 data explorer. outbreak.info. URL: https://outbreak.info/ [accessed 2022-11-26]
19. Expasy. Swiss Institute of Bioinformatics. URL: https://www.expasy.org/ [accessed 2022-12-19]
20. Basic Local Alignment Search Tool. National Library of Medicine. URL: https://blast.ncbi.nlm.nih.gov/Blast.cgi [accessed 2022-11-26]
21. Starr TN, Greaney AJ, Hilton SK, Ellis D, Crawford KH, Dingens AS, et al. Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. Cell 2020 Sep 03;182(5):1295-1310 [FREE Full text] [doi: 10.1016/j.cell.2020.08.012] [Medline: 32841599]
22. SARS-CoV-2 RBD DMS. JBloom Lab. 2022 Mar 03. URL: https://jbloomlab.github.io/SARS-CoV-2-RBD_DMS/ [accessed 2022-12-19]

23. South Africa COVID - Coronavirus Statistics. Worldometer. URL: https://www.worldometers.info/coronavirus/country/south-africa [accessed 2022-11-26]
24. The statistics of similarity scores. National Center for Biotechnology Information (NCBI). URL: https://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html [accessed 2022-11-26]
25. Li Q, Wu J, Nie J, Zhang L, Hao H, Liu S, et al. The impact of mutations in SARS-CoV-2 spike on viral infectivity and antigenicity. Cell 2020 Sep 03;182(5):1284-1294 [FREE Full text] [doi: 10.1016/j.cell.2020.07.012] [Medline: 32730807]
26. Shen L, Bard JD, Triche TJ, Judkins AR, Biegel JA, Gai X. Rapidly emerging SARS-CoV-2 B.1.1.7 sub-lineage in the United States of America with spike protein D178H and membrane protein V70L mutations. Emerg Microbes Infect 2021 Dec 26;10(1):1293-1299 [FREE Full text] [doi: 10.1080/22221751.2021.1943540] [Medline: 34125658]
27. Greaney AJ, Starr TN, Gilchuk P, Zost SJ, Binshtein E, Loes AN, et al. Complete mapping of mutations to the SARS-CoV-2 spike receptor-binding domain that escape antibody recognition. Cell Host Microbe 2021 Jan 13;29(1):44-57 [FREE Full text] [doi: 10.1016/j.chom.2020.11.007] [Medline: 33259788]
28. Wang P, Nair MS, Liu L, Iketani S, Luo Y, Guo Y, et al. Antibody resistance of SARS-CoV-2 variants B.1.351 and B.1.1.7. Nature 2021 May 08;593(7857):130-135. [doi: 10.1038/s41586-021-03398-2] [Medline: 33684923]
29. Wu J, Zhang L, Zhang Y, Wang H, Ding R, Nie J, et al. The antigenicity of epidemic SARS-CoV-2 variants in the United Kingdom. Front Immunol 2021 Jun 17;12:687869 [FREE Full text] [doi: 10.3389/fimmu.2021.687869] [Medline: 34220844]
30. Greaney AJ, Loes AN, Crawford KH, Starr TN, Malone KD, Chu HY, et al. Comprehensive mapping of mutations in the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human plasma antibodies. Cell Host Microbe 2021 Mar 10;29(3):463-476 [FREE Full text] [doi: 10.1016/j.chom.2021.02.003] [Medline: 33592168]
31. Wu K, Peng G, Wilken M, Geraghty RJ, Li F. Mechanisms of host receptor adaptation by severe acute respiratory syndrome coronavirus. J Biol Chem 2012 Mar 16;287(12):8904-8911 [FREE Full text] [doi: 10.1074/jbc.M111.325803] [Medline: 22291007]
32. Wibmer CK, Ayres F, Hermanus T, Madzivhandila M, Kgagudi P, Oosthuysen B, et al. SARS-CoV-2 501Y.V2 escapes neutralization by South African COVID-19 donor plasma. Nat Med 2021 Apr 02;27(4):622-625. [doi: 10.1038/s41591-021-01285-x] [Medline: 33654292]
33. Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, Sheffield COVID-19 Genomics Group, et al. Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. Cell 2020 Aug 20;182(4):812-827 [FREE Full text] [doi: 10.1016/j.cell.2020.06.043] [Medline: 32697968]
34. Braun KM, Moreno GK, Halfmann PJ, Hodcroft EB, Baker DA, Boehm EC, et al. Transmission of SARS-CoV-2 in domestic cats imposes a narrow bottleneck. PLoS Pathog 2021 Feb 26;17(2):e1009373 [FREE Full text] [doi: 10.1371/journal.ppat.1009373] [Medline: 33635912]
35. Dieterle ME, Haslwanter D, Bortz RH, Wirchnianski AS, Lasso G, Vergnolle O, et al. A replication-competent vesicular stomatitis virus for studies of SARS-CoV-2 spike-mediated cell entry and its inhibition. Cell Host Microbe 2020 Sep 09;28(3):486-496 [FREE Full text] [doi: 10.1016/j.chom.2020.06.020] [Medline: 32738193]
36. Baum A, Fulton BO, Wloga E, Copin R, Pascal KE, Russo V, et al. Antibody cocktail to SARS-CoV-2 spike protein prevents rapid mutational escape seen with individual antibodies. Science 2020 Aug 21;369(6506):1014-1018 [FREE Full text] [doi: 10.1126/science.abd0831] [Medline: 32540904]
37. Sabir DK. Analysis of SARS-COV2 spike protein variants among Iraqi isolates. Gene Rep 2022 Mar;26:101420 [FREE Full text] [doi: 10.1016/j.genrep.2021.101420] [Medline: 34754982]
38. Voss C, Esmail S, Liu X, Knauer M, Ackloo S, Kaneko T, et al. Epitope-specific antibody responses differentiate COVID-19 outcomes and variants of concern. JCI Insight 2021 Jul 08;6(13):e148855. [doi: 10.1172/jci.insight.148855] [Medline: 34081630]
39. Lubinski B, Fernandes MH, Frazier L, Tang T, Daniel S, Diel DG, et al. Functional evaluation of the P681H mutation on the proteolytic activation of the SARS-CoV-2 variant B.1.1.7 (Alpha) spike. iScience 2022 Jan 21;25(1):103589 [FREE Full text] [doi: 10.1016/j.isci.2021.103589] [Medline: 34909610]
40. Ramirez S, Fernandez-Antunez C, Galli A, Underwood A, Pham LV, Ryberg LA, et al. Overcoming culture restriction for SARS-CoV-2 in human cells facilitates the screening of compounds inhibiting viral replication. Antimicrob Agents Chemother 2021 Jun 17;65(7):e0009721 [FREE Full text] [doi: 10.1128/AAC.00097-21] [Medline: 33903110]
41. de Silva TI, Liu G, Lindsey B, Dong D, Moore S, Hsu N, COVID-19 Genomics UK (COG-UK) Consortium, ISARIC4C Investigators, et al. The impact of viral mutations on recognition by SARS-CoV-2 specific T cells. iScience 2021 Nov 19;24(11):103353 [FREE Full text] [doi: 10.1016/j.isci.2021.103353] [Medline: 34729465]
42. Cele S, Jackson L, Khoury DS, Khan K, Moyo-Gwete T, Tegally H, NGS-SA, COMMIT-KZN Team, et al. Omicron extensively but incompletely escapes Pfizer BNT162b2 neutralization. Nature 2022 Feb;602(7898):654-656 [FREE Full text] [doi: 10.1038/s41586-021-04387-1] [Medline: 35016196]
43. Avanzato VA, Matson MJ, Seifert SN, Pryce R, Williamson BN, Anzick SL, et al. Case study: Prolonged infectious SARS-CoV-2 shedding from an asymptomatic immunocompromised individual with cancer. Cell 2020 Dec 23;183(7):1901-1912 [FREE Full text] [doi: 10.1016/j.cell.2020.10.049] [Medline: 33248470]
44. Karim F, Moosa M, Gosnell B, Cele S, Giandhari J, Pillay S. Persistent SARS-CoV-2 infection and intra-host evolution in association with advanced HIV infection. medRxiv. 2021 Jun 04. URL: https://www.medrxiv.org/content/10.1101/2021.

06.03.21258228v1#:~:text=While%20most%20people%20effectively%20clear,HIV%20and%20antiretroviral%20treatment%20failure [accessed 2022-12-19]
45. Choi B, Choudhary MC, Regan J, Sparks JA, Padera RF, Qiu X, et al. Persistence and evolution of SARS-CoV-2 in an immunocompromised host. N Engl J Med 2020 Dec 03;383(23):2291-2293 [FREE Full text] [doi: 10.1056/NEJMc2031364] [Medline: 33176080]
46. Voloch C, da Silva Francisco R, de Almeida LGP, Brustolini OJ, Cardoso CC, Gerber AL, et al. Intra-host evolution during SARS-CoV-2 prolonged infection. Virus Evol 2021 Sep 29;7(2):veab078 [FREE Full text] [doi: 10.1093/ve/veab078] [Medline: 34642605]
47. Kumar A, Parashar R, Kumar S, Faiq MA, Kumari C, Kulandhasamy M, et al. Emerging SARS-CoV-2 variants can potentially break set epidemiological barriers in COVID-19. J Med Virol 2022 Apr;94(4):1300-1314 [FREE Full text] [doi: 10.1002/jmv.27467] [Medline: 34811761]
48. Kumar A, Prasoon P, Kumari C, Pareek V, Faiq MA, Narayan RK, et al. SARS-CoV-2-specific virulence factors in COVID-19. J Med Virol 2021 Mar;93(3):1343-1350. [doi: 10.1002/jmv.26615] [Medline: 33085084]
49. S:P681H mutation report. outbreak.info. URL: https://outbreak.info/situation-reports?muts=S%3AP681H [accessed 2021-12-08]
50. S:P681R mutation report. outbreak.info. URL: https://outbreak.info/situation-reports?muts=S%3AP681R [accessed 2021-12-08]
51. Esper F, Adhikari T, Tu Z, Cheng Y, El-Haddad K, Farkas D, et al. Alpha to Omicron: disease severity and clinical outcomes of major SARS-CoV-2 variants. J Infect Dis 2022 Oct 10:jiac411 [FREE Full text] [doi: 10.1093/infdis/jiac411] [Medline: 36214810]
52. Consolazio D, Murtas R, Tunesi S, Lamberti A, Senatore S, Faccini M, et al. A comparison between Omicron and earlier COVID-19 variants' disease severity in the Milan area, Italy. Front Epidemiol 2022 Jun 28;2:891162. [doi: 10.3389/fepid.2022.891162]
53. Naqvi A, Fatima K, Mohammad T, Fatima U, Singh I, Singh A, et al. Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: Structural genomics approach. Biochim Biophys Acta Mol Basis Dis 2020 Oct 01;1866(10):165878 [FREE Full text] [doi: 10.1016/j.bbadis.2020.165878] [Medline: 32544429]
54. Hui KPY, Ho JCW, Cheung M, Ng K, Ching RHH, Lai K, et al. SARS-CoV-2 Omicron variant replication in human bronchus and lung ex vivo. Nature 2022 Mar 01;603(7902):715-720. [doi: 10.1038/s41586-022-04479-6] [Medline: 35104836]
55. Wilhelm A, Widera M, Grikscheit K, Toptan T, Schenk B, Pallas C, et al. Limited neutralisation of the SARS-CoV-2 Omicron subvariants BA.1 and BA.2 by convalescent and vaccine serum and monoclonal antibodies. EBioMedicine 2022 Aug;82:104158 [FREE Full text] [doi: 10.1016/j.ebiom.2022.104158] [Medline: 35834885]
56. Hu J, Peng P, Cao X, Wu K, Chen J, Wang K, et al. Increased immune escape of the new SARS-CoV-2 variant of concern Omicron. Cell Mol Immunol 2022 Feb 11;19(2):293-295 [FREE Full text] [doi: 10.1038/s41423-021-00836-z] [Medline: 35017716]
57. Nabel KG, Clark SA, Shankar S, Pan J, Clark LE, Yang P, et al. Structural basis for continued antibody evasion by the SARS-CoV-2 receptor binding domain. Science 2022 Jan 21;375(6578):eabl6251 [FREE Full text] [doi: 10.1126/science.abl6251] [Medline: 34855508]
58. Redd A, Nardin A, Kared H, Bloch E, Abel B, Pekosz A, et al. Minimal cross-over between mutations associated with Omicron variant of SARS-CoV-2 and CD8+ T cell epitopes identified in COVID-19 convalescent individuals. bioRxiv. 2021 Dec 09. URL: https://www.biorxiv.org/content/10.1101/2021.12.06.471446v1 [accessed 2022-12-19]
59. Grifoni A, Sidney J, Vita R, Peters B, Crotty S, Weiskopf D, et al. SARS-CoV-2 human T cell epitopes: adaptive immune response against COVID-19. Cell Host Microbe 2021 Jul 14;29(7):1076-1092 [FREE Full text] [doi: 10.1016/j.chom.2021.05.010] [Medline: 34237248]
60. SARS-CoV-2 variants of concern and variants under investigation in England. Variant of concern: Omicron, VOC21NOV-01 (B.1.1.529). Technical briefing 30. UK Health Security Agency. 2021 Dec 03. URL: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1038404/Technical_Briefing_30.pdf [accessed 2022-12-19]
61. Tegally H, Wilkinson E, Althaus C, Giovanetti M, San J, Giandhari J. Rapid replacement of the Beta variant by the Delta variant in South Africa. medRxiv. 2021 Sep 27. URL: https://www.medrxiv.org/content/10.1101/2021.09.23.21264018v1 [accessed 2022-12-19]
62. Sigal A, Milo R, Jassat W. Estimating disease severity of Omicron and Delta SARS-CoV-2 infections. Nat Rev Immunol 2022 May 12;22(5):267-269 [FREE Full text] [doi: 10.1038/s41577-022-00720-5] [Medline: 35414124]
63. Ward IL, Bermingham C, Ayoubkhani D, Gethings OJ, Pouwels KB, Yates T, et al. Risk of covid-19 related deaths for SARS-CoV-2 omicron (B.1.1.529) compared with delta (B.1.617.2): retrospective cohort study. BMJ 2022 Aug 02;378:e070695 [FREE Full text] [doi: 10.1136/bmj-2022-070695] [Medline: 35918098]
64. Nyberg T, Ferguson NM, Nash SG, Webster HH, Flaxman S, Andrews N, COVID-19 Genomics UK (COG-UK) consortium, et al. Comparative analysis of the risks of hospitalisation and death associated with SARS-CoV-2 omicron (B.1.1.529) and

delta (B.1.617.2) variants in England: a cohort study. Lancet 2022 Apr 02;399(10332):1303-1312 [FREE Full text] [doi: 10.1016/S0140-6736(22)00462-7] [Medline: 35305296]
65. Pearson C, Silal S, Dushoff J, Abbott S, van Schalkwyk C, Bingham J, on behalf of the South African COVID-19 Modelling Consortium. Google Drive. URL: https://drive.google.com/file/d/1hA6Mec2Gq3LGqTEOj35RqSeAb_SmXpbI/view [accessed 2021-12-04]
66. SARS-CoV-2 variants of concern and variants under investigation in England. Technical briefing 31. UK Health Security Agency. 2021 Dec 10. URL: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1040076/Technical_Briefing_31.pdf [accessed 2022-12-19]
67. Lundberg AL, Lorenzo-Redondo R, Ozer EA, Hawkins CA, Hultquist JF, Welch SB, et al. Has Omicron changed the evolution of the pandemic? JMIR Public Health Surveill 2022 Jan 31;8(1):e35763 [FREE Full text] [doi: 10.2196/35763] [Medline: 35072638]
68. Lundberg AL, Lorenzo-Redondo R, Hultquist JF, Hawkins CA, Ozer EA, Welch SB, et al. Overlapping Delta and Omicron outbreaks during the COVID-19 pandemic: dynamic panel data estimates. JMIR Public Health Surveill 2022 Jun 03;8(6):e37377 [FREE Full text] [doi: 10.2196/37377] [Medline: 35500140]
69. Coronavirus (COVID-19) vaccinations. Our World in Data. URL: https://ourworldindata.org/covid-vaccinations?country=~ZAF) [accessed 2021-12-06]
70. Threat Assessment Brief: Implications of the further emergence and spread of the SARS-CoV-2 B.1.1.529 variant of concern (Omicron) for the EU/EEA - first update. European Centre for Disease Prevention and Control. 2021 Dec 02. URL: https://www.ecdc.europa.eu/en/publications-data/covid-19-threat-assessment-spread-omicron-first-update [accessed 2021-12-05]
71. Science Brief: Omicron (B.1.1.529) Variant. Centers for Disease Control and Prevention. 2021 Dec 02. URL: https://www.cdc.gov/coronavirus/2019-ncov/science/science-briefs/scientific-brief-omicron-variant.html#6 [accessed 2021-12-05]
72. Pulliam J, van Schalkwyk C, Govender N, von Gottberg A, Cohen C, Groome M, et al. Increased risk of SARS-CoV-2 reinfection associated with emergence of Omicron in South Africa. Science 2022 May 06;376(6593):eabn4947 [FREE Full text] [doi: 10.1126/science.abn4947] [Medline: 35289632]

Abbreviations

ACE2: angiotensin-converting enzyme 2
GISAID: Global Initiative on Sharing Avian Influenza Data
NCBI: National Center for Biotechnology Information
PANGO: Phylogenetic Assignment of Named Global Outbreak
PCR: polymerase chain reaction
RBD: receptor-binding domain
RBM: receptor-binding motif
VOC: variant of concern
VOI: variant of interest
WHO: World Health Organization

Edited by A Mavragani; submitted 15.09.22; peer-reviewed by A Prabahar, A Krishnan; comments to author 11.11.22; revised version received 29.11.22; accepted 16.12.22; published 10.01.23.

Please cite as:
Kumar A, Asghar A, Singh HN, Faiq MA, Kumar S, Narayan RK, Kumar G, Dwivedi P, Sahni C, Jha RK, Kulandhasamy M, Prasoon P, Sesham K, Kant K, Pandey SN
SARS-CoV-2 Omicron Variant Genomic Sequences and Their Epidemiological Correlates Regarding the End of the Pandemic: In Silico Analysis
JMIR Bioinform Biotech 2023;4:e42700
URL: https://bioinform.jmir.org/2023/1/e42700
doi:10.2196/42700
PMID:36688013

Original Paper

Mutations of SARS-CoV-2 Structural Proteins in the Alpha, Beta, Gamma, and Delta Variants: Bioinformatics Analysis

Saima Rehman Khetran[1*], MPhil; Roma Mustafa[1*], DPhil

Department of Life Sciences, Sardar Bahadur Khan Women's University, Quetta, Pakistan

[*]all authors contributed equally

Corresponding Author:
Saima Rehman Khetran, MPhil
Department of Life Sciences
Sardar Bahadur Khan Women's University
Bawrery Road near Kidney Hospital Quetta
Quetta, 87300
Pakistan
Email: aspirantcss2022@gmail.com

Related Article:

This is a corrected version. See correction statement: https://bioinform.jmir.org/2024/1/e64915

Abstract

Background: COVID-19 and Middle East Respiratory Syndrome are two pandemic respiratory diseases caused by coronavirus species. The novel disease COVID-19 caused by SARS-CoV-2 was first reported in Wuhan, Hubei Province, China, in December 2019, and became a pandemic within 2-3 months, affecting social and economic platforms worldwide. Despite the rapid development of vaccines, there have been obstacles to their distribution, including a lack of fundamental resources, poor immunization, and manual vaccine replication. Several variants of the original Wuhan strain have emerged in the last 3 years, which can pose a further challenge for control and vaccine development.

Objective: The aim of this study was to comprehensively analyze mutations in SARS-CoV-2 variants of concern (VoCs) using a bioinformatics approach toward identifying novel mutations that may be helpful in developing new vaccines by targeting these sites.

Methods: Reference sequences of the SARS-CoV-2 spike (YP_009724390) and nucleocapsid (YP_009724397) proteins were compared to retrieved sequences of isolates of four VoCs from 14 countries for mutational and evolutionary analyses. Multiple sequence alignment was performed and phylogenetic trees were constructed by the neighbor-joining method with 1000 bootstrap replicates using MEGA (version 6). Mutations in amino acid sequences were analyzed using the MultAlin online tool (version 5.4.1).

Results: Among the four VoCs, a total of 143 nonsynonymous mutations and 8 deletions were identified in the spike and nucleocapsid proteins. Multiple sequence alignment and amino acid substitution analysis revealed new mutations, including G72W, M2101I, L139F, 209-211 deletion, G212S, P199L, P67S, I292T, and substitutions with unknown amino acid replacement, reported in Egypt (MW533289), the United Kingdom (MT906649), and other regions. The variants B.1.1.7 (Alpha variant) and B.1.617.2 (Delta variant), characterized by higher transmissibility and lethality, harbored the amino acid substitutions D614G, R203K, and G204R with higher prevalence rates in most sequences. Phylogenetic analysis among the novel SARS-CoV-2 variant proteins and some previously reported β-coronavirus proteins indicated that either the evolutionary clade was weakly supported or not supported at all by the β-coronavirus species.

Conclusions: This study could contribute toward gaining a better understanding of the basic nature of SARS-CoV-2 and its four major variants. The numerous novel mutations detected could also provide a better understanding of VoCs and help in identifying suitable mutations for vaccine targets. Moreover, these data offer evidence for new types of mutations in VoCs, which will provide insight into the epidemiology of SARS-CoV-2.

(JMIR Bioinform Biotech 2023;4:e43906) doi:10.2196/43906

KEYWORDS

virus evolution; influenza and other respiratory viruses; advances in virus research; COVID-19; protein; mutation; genomic; vaccine development; phylogenetic analysis; biochemistry

Introduction

The emergence of SARS-CoV-2 during the early months of 2020 made headlines worldwide. Since then, several new variants of SARS-CoV-2 have emerged and are classified based on their ability to cause a threat to public health in two groups: variants of concern (VoCs) and variants of interest (VoIs) [1]. VoIs are defined as variants with specific genetic markers causing mutations that facilitate virus transmissibility, reduce the accuracy of diagnostic results, and reduce antibody neutralization acquired through natural infection or vaccination [2]. VoCs are associated with the level of virus transmissibility, infection, reduced effectiveness of vaccines and treatment, failure in virus detection, and reduced levels of neutralizing antibodies generated during previous vaccination or infection. The main SARS-CoV-2 VoCs that emerged include the Alpha (B.1.1.7), Beta (B.1.351), Gamma (P.1), Delta (B.1.617.2), and the most recent Omicron variants [2]

The first reported VoC, B.1.1.7 (Alpha variant), was isolated in the United Kingdom in December of 2020 [3,4], which contained a total of 23 mutations [5]. These mutations directly affect the open reading frame (ORF)1ab and ORF8 regions of the spike (S) protein as well as the nucleocapsid (N) protein [6]. The Alpha variant was characterized by substantially higher levels of infectivity and transmissibility. A total of seven mutations were reported in the S protein of this variant, including N501Y, A570D, D614G, P681H, T716I, S982A, and D1118H, along with two deletions (Δ69-70 and Δ145) [7]. In addition, several mutations were identified in the N protein sequence of the Alpha variant, including D3L, P13L, D103Y, S197L, S188L, S93I, I292T, R203K, G204R, S190I, S194L, S202N, S235F, D348H, and D401Y [8].

The second reported VoC of SARS-CoV-2, B.1.351 (Beta variant), was identified during the second wave in Africa in October 2020 [9]. This variant included a total of nine point mutations (L18F, D80A, D215G, R246I, K417N, E484K, N501Y, D614G, and A701V) and three deletions (Δ242-Δ244) in the S protein, but only one mutation in the N protein (T205I) [10,11]. Subsequently, another VoC, P.1 (Gamma variant), was first identified in Brazil at the end of January 2021 [12] harboring 10 mutations in the S protein (L18F, T20N, P26S, D138Y, R190S, K417T, E484K, N501Y, H655Y, T1027I) [2,12] and 3 mutations in the N protein (R203K, G204R/X, P80R) [13,14].

Among the other reported VoCs, B.1.617.2 (Delta variant), the most dominant variant of SARS-CoV-2, was first detected in India during the second wave of the COVID-19 outbreak in April 2021 [15]. The Delta variant harbored 10 point mutations (T19R, G142D, R158G, L452R, T478K, D614G, P681R, and D950N) [2] and two deletions (Δ156, Δ157) in S protein, but the most lethal among them were L452R and P681R [15,16], as isolates carrying these mutations were responsible for most of the deaths in India and other countries. Moreover, Delta acquired several mutations in the N protein, including G18C, D63G, L139F, R203M, G215C, A252S, S327L, D377Y, and R385K [17,18].

In addition to the S and N proteins, all of these reported VoCs comprise the other main structural proteins of SARS-CoV-2, membrane protein (M) and envelope protein (E). Each protein of the virus plays a vital role and also takes part in the replication cycle. The S protein is required for the attachment and amalgamation of the virus to host cell surface receptors, which enables the virus to enter the host cell [19]. The main function of the N protein is to form the nucleocapsid by binding to the RNA genome of the virus, which has unique properties compared to the other proteins. The shape of the virus envelope is developed with the help of M protein, which is present in abundance inside the virus and also facilitates interactions with other viral proteins as well as in organizing the assembly of proteins. E protein is the smallest SARS-CoV-2 protein, whose function remains somewhat mysterious. During replication, E protein is abundantly expressed in the host cell, whereas only a small portion of the protein is incorporated into the virus envelope [20]. Almost all of the major structural proteins possess mutations at the receptor-binding domain (RBD) and N-terminal domain sites [21], and they have one mutation in common (N501Y) except for the Delta variant [21].

Recent studies have shown that several mutations are responsible for the spread and lethality of SARS-CoV-2, with more than 10 SARS-CoV-2 variants reported to date that are categorized as either VoCs or VoIs [3]. However, it remains unclear how these sequences are mutating during transfer of the virus from person to person. To answer this question, a total of 127 full-length amino acid sequences of SARS-CoV-2 isolates from 14 countries submitted to NCBI up to July 15, 2021, were retrieved to investigate and identify amino acid substitutions in SARS-CoV-2 lineages and their mutational pattern in major structural proteins.

In this study, we used bioinformatics methods to identify nonsynonymous mutations in the S and N proteins of the four main VoCs of SARS-CoV-2 and to determine how they affect the structure and functional dynamics of the virus. This analysis will help to better understand the epidemiology of SARS-CoV-2 and its emerging VoCs, which might ultimately identify suitable mutations as new vaccine targets.

Methods

Data Source

On the basis of a high predominance rate, the data were collected from isolates reported in 14 countries (Pakistan, Turkey, China, Iran, Morocco, United States, United Kingdom, France, Italy, Spain, India, Japan, Egypt, and Russia). A total of 127 nucleotide sequences of SARS-CoV-2 were retrieved from the National Center for Biotechnology Information (NCBI) Virus

SARS-CoV-2 Data Hub [22] along with their major structural proteins (Table 1).

SARS-CoV-2 sequences for the four respective proteins (S, N, M, and E) were downloaded in FASTA format. Reference sequences were also considered for data comparison. A total of 127 amino acid sequences were obtained for analysis, which were converted into nucleotides using the online reverse translation tool Sequence Manipulation Suite.

Table 1. Number of isolates corresponding to the sequences retrieved from various nations.

Variant of SARS-CoV-2	Countries	Number of isolates
Reference sequence	China	1
β-Coronavirus isolates	United States	5
Alpha variant	Pakistan, Turkey, China, Iran, Morocco, United States, United Kingdom, France, Italy, Spain, India, Japan, Egypt, Russia	78
Beta variant	United States, Italy, Spain, France, India	20
Gamma variant	United States, Italy, Pakistan, Spain, Egypt, India	15
Delta variant	India, Egypt, United States, Spain	10

Quality Profiling for Sequence Selection and Phylogenetic Analysis

Quality profiling for sequence selection was performed to differentiate between countries according to the epidemic record. This study included four types of SARS-CoV-2 variants taking into account their S and N proteins. Data were compared by constructing phylogenetic trees for each protein. All other types of SARS-CoV-2 variants and their respective proteins were excluded from the analysis.

The phylogenetic tree was constructed from 127 sequences of the major proteins along with the corresponding reference protein sequences (S protein: YP_009724392; N protein: YP_009724393; M protein: YP_009724397; and E protein: YP_009724390). Protein sequences of β-coronavirus strains were selected as outgroups: human coronavirus (hCoV)-NL63 (YP_003767), hCoV-229E (NP_073551), hCoV-OC43 (YP_009555241), Middle East Respiratory Syndrome (MERS; YP_009047204), and SARS (NC_004718).

Multiple sequence alignment was performed and phylogenetic trees were constructed by the neighbor-joining method with 1000 bootstrap replicates using MEGA (version 6) [23]. The FASTA file was computed with a gap-opening penalty of 15 and gap-extension penalty of 6.66, maintaining a delay divergent cutoff of 30%. Amino acid substitutions that were unique to SARS-CoV-2 were identified by visual inspection of the alignments.

Mutation Identification With MultAlin

For the detection of widespread nonsynonymous mutations in the S, N, M, and E proteins, amino acid sequences were analyzed using MultAlin (version 5.4.1) [20] and each mutation was recorded separately.

This tool enabled identifying the exact location of the mutation in the genome sequence of each strain by providing the position of the mutated site.

Ethics Considerations

This study was based on analysis of secondary data that are publicly available at NCBI [22] and did not require any ethical approval.

Results and Discussion

Mutation Hotspots in the S Protein of SARS-CoV-2 Variants

The main VoCs of SARS-CoV-2 all contain the four major structural proteins S, N, M, and E, and numerous studies have elucidated similarities and differences among the viral genomes and their proteins using different types of bioinformatics tools [23]. Among the isolates of the 14 countries considered in this study, strong evidence was found for occurrence of the D614G mutation (see Multimedia Appendix 1), indicating replacement of the amino acid aspartic acid (D) with glycine (G) at position 614 in the sequence. The D614G mutation affects the interaction with the host receptor angiotensin-converting enzyme 2 (ACE2), resulting in greater stability and the ability to transmit more efficiently, although binding of the mutant was not as competent as compared to the normal binding of the viral protein [24]. The majority of the Pakistan isolates (MW421982–92) also carried the D614G mutation along with some unreported additional mutations, including P26L, D80Y, S813N, Q1207H, D1163Y, and T1117I (see Multimedia Appendix 1). Two of the Egyptian isolates (MW533286, MW533289) displayed unique mutations (Q23X, S12X, Q677X, and P681X), where the amino acids glutamine (Q), serine (S), and proline (P) were replaced with an unknown amino acid (X) at different positions. These mutations might be important for the future study of the mechanism underlying virus lethality. A much higher rate of mutations along with D614G was observed in the UK isolate MT906649, with a series of novel mutations (T22X, P25X, G142X, Y144-5X, S735X, K1191X) identified in which the existing amino acids tyrosine (Y), threonine (T), proline (P), lysine (K), and glycine (G) were substituted to result in a change of the conformation of S protein (see Multimedia Appendices 1 and 2).

Another mutation of concern identified in S protein was A570D, in which alanine (A) was replaced by aspartic acid (D) at position 570, which was found to co-occur with a Δ145 deletion and three other mutations: T716I, where tryptophan (T) was replaced by isoleucine (I) at position 716; S982A, where serine (S) was replaced by alanine (A) at position 982; and D1118H, where aspartic acid (D) was replaced by histidine (H) at position 1118 (Multimedia Appendix 1). The mutations A570D, T716I, S982A, and D1118H were a result of a series of accumulated mutations, which collectively increased the lethality and transmissibility of the virus [25]. These mutations were observed in isolates from the United States, India, Italy, and Spain simultaneously; nevertheless, the US isolates (MW725912, MW725900, MW725904, MW725907, MW712865, MW712862, MW712864, MW725917, and MW725924) also harbored the mutation G72W, in which glycine (G) was replaced by tryptophan (W) at position 72W, along with D614G although the effect of this mutation remains unknown. The co-occurrence of the mutation A570D with D614G and S982A was also observed in some isolates of Italy (MW491232, MW711159), the United States (MZ311101 and MW725906), and India (MW600456), with no other novel mutations identified in these cases (Multimedia Appendices 1 and 2). The mutations A570D, D614G, and S982A correspondingly help in minimizing contact between the individual trimeric spike promoter chains, thereby promoting increased cleavage between the S1 and S2 domains of S protein to consequently enhance the host fusion capability while rearranging the overall dynamic structure of the virus [26].

The third most prominent mutation identified was N501Y, in which asparagine (N) was replaced by tyrosine (Y) at position 501 of the S protein (Multimedia Appendix 1). The transmissibility of the virus harboring the N501Y mutation (located at the receptor-binding motif) increased by 70%-80% and this mutation also improved the binding affinity of the virus onto host cells [27]. This mutation in combination with 7 other mutations (A570D, P681H, T716I, S982A, D1118H, and Δ69-70, Δ145) were termed to be "mutations of major concern" [27,28] and were consistently detected in isolates from the United States, India, Italy, and Spain. The deletion of histidine at position 69 (Δ69) and valine at position 70 (Δ70) also evolved in other variants (Multimedia Appendix 1) and are considered to be responsible for increasing the transmissibility as well as infectivity of the virus, along with causing S gene target failure, resulting in nondetection of the virus [7,29]. Another deletion of tyrosine at position 144 (Δ144) was considered to be responsible for changing the conformation of the S protein's surface, thereby facilitating evasion of host immunity and increasing infection [30]. Apart from these mutations, deletions at position 85-89 (Δ85-Δ89) in a Spanish isolate (MW715071) along with other unique mutations of S protein, such as V90T (in which valine is replaced by threonine at position 90) [31], A93Y (in which alanine is replaced by tyrosine at position 93), and D138H (in which aspartic acid is replaced by histidine at position 138), were also observed (Multimedia Appendices 1 and 2). Although the specific function of these mutations remains unknown, their identification and further analysis may help to better understand virus structure and lethality.

Additionally, the trio mutations A220V (alanine replaced by valine at position 220), ORF10 V30L, and Spike A222V, were identified in the S and N proteins of Spanish isolates (MW715068-MW715080). These mutations formed different types of clades when combined with other mutations [32], although the A220V mutation was identified with no additional mutations from the reported data. Furthermore, some of the main mutations included in South African variants were L18F (leucine replaced by phenylalanine at position 18), D80A (aspartic acid replaced by alanine at position 80), D215G (aspartic acid replaced by glycine at position 215), R246I (arginine replaced by isoleucine at position 246), K417N (lysine replaced by asparagine at position 417), and E484K (glutamic acid replaced by lysine at position 484), along with N501Y, D614G, and A222V. The mutations K417N, E484K, and N501Y located in the RBD help the virus in binding to the ACE2 receptors of host cells [9,10] A recent study also reported that the E484K mutation might alter the conformation of S protein, thereby affecting the neutralizing capability of the antibody response in host cells, as cases of reinfection were also increased in patients with isolates harboring the E48K mutation at the peak (ie, the majority of the isolates possessed the E484K mutation) during mid-2021 [27]. Several studies have also reported the E484K mutation as a major cause of decreased effectiveness of current vaccines [27,33]. The mutations N501Y and E484K along with L18F and K417T/N are considered to decrease ACE2 binding affinity [34] and were reported in isolates from Italy (MW642250 and MW642248) and the United States (MZ320527). Some of the mutations of the Alpha and Gamma variants, such as N501Y, D614G, E484K, A701V, and N501Y, were also observed in isolates from Italy and the United States (Multimedia Appendices 1 and 2).

As the variants continued to spread across different regions, another VoC emerged in Spain toward the end of 2020. This variant possessed an exceptional mutation, A222V (alanine replaced by valine at position 222), in the S protein (Multimedia Appendix 1). The mutation A222V alone had no direct impact on transmissibility of the virus, in contrast to the effect of D614G [35]; however, in combination with other Beta variant mutations such as L18F, D80A, K417N, E484K, N501Y, A701V, D215G, and deletions at position 242-244 (Δ242-Δ244), A222V causes a severe hindrance in antibody binding [29]. We found these mutations combined with D614G in Spanish isolates (MW715072 and MW715075). A new type of deletion (Δ139-Δ144) was also observed in two isolates from Spain (MW715068 and MW715078) along with the L18F, A222V, and D614G mutations Multimedia Appendices 1 and 2).

In addition to the Alpha and Beta variants, another VoC was the Brazil variant, which consists of mutations almost identical to S protein mutations of the Beta variant (N501Y, E484K, A701Y) except for the K417T mutation, where lysine (K) was replaced by threonine (T) at position 417, also causing a decrease in ACE2 binding affinity [34]. The dominance of these mutations in many VoCs that play an important role in ACE2 binding affinity during viral attachment [34] might also increase the chances of reinfection [36]. These mutations occurred in isolates from Italy (MW642250, MW642248, MW711159, and

MW491232), the United States (MZ320527), and France (MW580244) (Multimedia Appendices 1 and 3).

As compared to other VoCs, the Delta variant was of major concern, which consists of four types of signature mutations: L452R, T478K, D614G, and P681R. The P681R mutation, in which proline (P) was replaced by arginine (R) at position 681, increased the rate of the cleavage process in S1 and S2 subunits (at the furin cleavage site), facilitating virus transmissibility [34,37]. A famous virologist at Cornell University in New York stated that "This little insert (P681R) sticks out and hits you in the face" [37]. The P681R mutation was considered to be responsible for the rapid spread of SARS-CoV-2 around the globe [37]. These signature mutations (L452R, T478K, D614G) were observed in isolates from Egypt (MW533290), India (MZ310590 and MZ310591), and Spain (MW715070) (Multimedia Appendix 4). L452R is the only S protein mutation that clasps the virus with the host cell surface, facilitating injection of the viral genetic material into host cells [4]. The L452R mutation was identified in isolates from the United States (MW725963) and Spain (MW715074) along with D614G, covering more than 90% of variants that emerged since 2020, conferring the virus with increased replication and infectivity abilities [24,37] (Multimedia Appendices 1 and 4).

These data demonstrated that the UK variant 20I/N501Y.V1 derived from lineage B.1.1.7 and the Brazil variant 20J/501Y.V2 derived from lineage B.1.351 (termed P.1) consisted of several mutations at specific points of the nucleic acid sequence, causing several physical changes as well as functional changes affecting virus lethality.

Mutation Hotspots in N protein of SARS-CoV-2 Variants

Among the other structural proteins of SARS-CoV-2, N protein, which is known to be more stable and conserved than other proteins, consists of three domains: the N-terminal domain, serine/arginine-rich linker region, and C-terminal domain [38]. The function of N protein is to make the nucleocapsid for the virus by binding to its RNA genome [39]. A few mutations have been observed in the N protein of the Alpha variant, including D3L (aspartic acid replaced by leucine at position 3), R203K (arginine replaced by lysine at position 203), G204R (glycine replaced by arginine at position 204), S194L (serine replaced by leucine at position 194), and S235F (serine replaced by phenylalanine at position 235), along with a single mutation of the Beta variant (T205I, in which threonine is replaced by isoleucine at position 205). Few mutations were observed in the Gamma variant, including R203K and G204R/X. Moreover, the N protein of the Delta variant exhibited the mutations R203M (arginine replaced by methionine at position 203), G204R, and D377Y (aspartic acid replaced by tyrosine at position 377).

The tetrad mutations D3L, R203K, G204R, and S235F were observed in isolates from the United States (MW712861-64, MZ311101, and MW725900-24), Italy (MW711159, MW491232), India (MW600456-58), and Spain (MW715071). The extraordinary sequence of Spain (MW715071) also possesses a unique deletion at position 209-211 whose function has not yet been reported. This sequence (MW715071) is termed extraordinary because of its unique genotypic characteristics along with the presence of a high number of unique mutations that had not been previously identified, including V90T, A93Y, D138H, and G212S (glycine replaced by serine at position 212), and deletions at positions 85-89 of S protein as well as at positions 209-211 of N protein (Multimedia Appendics 2 and 5).

Mutually, R203K-G204R mutations were observed in the serine/arginine-rich linker region (responsible for cellular processes such as the cell cycle and characterized by high flexibility) of N protein, which also affect virus assembly [11,40]. Additionally, R203K-G204R mutations belong to the Alpha (along with D3L and S235F) and Gamma variants of SARS-CoV-2 (also expressed as G204R/X) [40,41]. The function of the G204X mutation can be considered ambiguous at present, because the specific amino acid replacing glycine at position 204 is unknown. Additionally, the amino acid substitutions R203K-G204R were identified in various isolates from 11 regions. The Egypt isolates (MW533286, MW533289) also possess the peculiar mutations G212X and G25X, in which glycine is replaced by an unknown residue at positions 25 and 212 (Multimedia Appendix 2), whereas a Pakistan isolate (MW422070) harbors the mutations R203K, G204R, and D614G of the Alpha variant along with an additional mutation A152X, where alanine (A) is replaced by an unknown amino acid residue (X) at position 152 (Multimedia Appendix 3). The mutations R203K, G204R, and D614G also increase viral infectivity due to a higher replication rate; thus, the presence of the dual mutation R203K/G204R in N protein along with the D614G and N501Y mutations of S protein result in an overall increase in the severity of disease and viral infectivity in the host body [14].

The R203K/G204R and N501Y mutations were also associated with disease severity, infectivity of the virus, and an increase in the mortality rate of host cells [42,43]. The combinations of R203K/G204R and N501Y along with the P80R, K417T, and E484K mutations were observed in isolates from Italy (MW642250, MW642248), the United States (MZ320527), and France (MW580244) (Multimedia Appendices 3 and 5). Conversely, the Delta variant possesses the R203M, G204R, and D377Y mutations that might cause a functional disruption in viral efficiency [14]. The trio mutations R203M, G204R, and D377Y were only observed in isolates of India (MZ702716, MZ310590, MZ310591) (Multimedia Appendices 4 and 5).

Furthermore, one of the mutations of interest in N protein was S194L, which is in a region responsible for protein oligomerization [44] (formation of hetero oligomers), and these hetero oligomers form an N-M protein complex that is critical for virus assembly [44,45]. The mutation S194L was identified with no other co-occurring mutations in isolates from India (MZ310512, MW600461-63), the United States (MW725958), and Iran (MT889692) (Multimedia Appendices 2 and 5). The S194L mutation was also identified during the SARS outbreak in 2003 [40]. In addition, another mutation, T205I, was frequently identified in the majority of the global variants evaluated, including isolates from Spain (MW715082, MW715069), France (MW580244), the United States (MW725963), and India (MW595912, MW595915, MW595914, MZ310507) (Multimedia Appendix 6).

Mutation Hotspots in M and E Proteins of SARS-CoV-2 Variants

M protein interacts with the S and E proteins to establish the traditional shape of the virus envelope, and also helps in connecting as well as organizing other proteins of the virus [46]. We identified only five mutations in M protein in our sequence analysis: V70L (valine replaced by leucine at position 70), F28X (phenylalanine replaced by an unknown amino acid at position 28), E12X (glutamic acid replaced by an unknown amino acid at position 12), I82T (isoleucine replaced by threonine at position 82), and deletion at position 72 (Δ72) (Multimedia Appendices 7 and 8). The Δ72 deletion was observed in an isolate from Spain (MW375731), which also contains the S protein mutation D614G (Multimedia Appendix 1). The E12X and F28X mutations were observed in a UK isolate (MT906649), which also possesses the mutation D614G of S protein and the T30I and L51X mutations of E protein (Multimedia Appendix 7). The I82T mutation was present in an Indian isolate (MZ702716) that also harbored the T182I mutation of E protein (Multimedia Appendix 7); L452R, T478K, D614G, P681R mutations of S protein (Multimedia Appendix 1); and the N protein mutations R203M and D377Y from the Delta variant (Multimedia Appendix 5). The last mutation V70L was observed in an isolate from Egypt (MW533290), which stands out from all other sequences because it consists of top controversial mutations (as these mutations were present in almost every variant of SARS-CoV-2) from the S protein of the Alpha (D614G) and Delta (P681R) variants, as well as N protein mutations from the Gamma variant (R203K, G204X) (Multimedia Appendices 5, 7, and 8).

E protein of SARS-CoV-2 plays a significant role in the assembly, pathogenesis, envelope formation, and budding of the virus [7]. As the smallest of the major structural proteins, the expression of E protein is abundant inside the host cell, but only a small portion of this protein is incorporated into the virus envelope [47]. We identified five mutations in E protein: L28P (leucine replaced by proline at position 28), T30I (threonine replaced by isoleucine at position 30), L51X (leucine replaced by unknown amino acid at position 51), V58F (valine replaced by phenylalanine at position 58), and P71L (proline replaced by leucine at position 71) (Multimedia Appendices 7 and 8). The mutation V58F was present in an isolate from India (MW595915), in addition to the D614G mutation of S protein (Multimedia Appendix 1) and T205I mutation of N protein (Multimedia Appendix 5). The L28P mutation was observed only in an Iran isolate (MT994881) with no other major mutations present. The third mutation, P71L, was present in US isolates (MW725914 and MW725923), along with the D614G, R203K, and G204R mutations. The mutation P71L was also observed in an isolate from France (MW580244), along with the N501Y and E484K mutations and the A701V mutation from S protein of the Gamma variant. In contrast, mutations T30I and L51X were observed in a UK isolate (MT906649) along with D614G (from S protein), E12X, and F28X (from M protein) (Multimedia Appendices 1,5, 7, and 8).

According to the predicted functions of these major mutations, it was concluded that four mutations from M protein and five mutations from E protein of SARS-CoV-2 variants along with other mutations of S and N proteins might increase the transmissibility, susceptibility, and lethality of the virus [8]. Additionally, analysis of the mutational patterns showed that the SARS-CoV-2 variants displayed unique mutations in isolates from different countries (Multimedia Appendix 9).

Phylogenetic Analysis

Along with an overall visual investigation of the relevant mutations, phylogenetic analysis was performed to analyze the evolutionary relationships among different strains of SARS-CoV-2.

Analysis of the nodes of the tree constructed with S protein sequences showed that hCoV-NL63 (YP_003767) and hCoV-229E (NP_073551) displayed a strong association (100%), while SARS-CoV (NC_004718) and MERS-CoV (YP_009047204) exhibited strongly associated clades (76%). Moreover, the reference sequence of S protein (YP_009724392) presented a weak association (68%) and was distantly related to S protein sequences of other β-coronaviruses. In addition, the majority of the SARS-CoV-2 variants displayed no support to the reference sequence clades with some being only distantly related. Therefore, the level of observed clades in each strain differed, providing a set of contradictory nodes during cladogram comparison among S protein variants (Figure 1).

The cladogram of N protein showed a different pattern than that constructed for S protein. All four β-coronavirus sequences of N protein exhibited strong associations among each other (100%), but there was no support for an association to the reference sequence of N protein (YP_009724393). The clades of India (QQY49667, QQY679) and the United States (QSU75744, QSU75637) were well-associated (79%-84%). The repeatability of bootstrap values below 50% was high, whereas few clades possessed weak or strong associations. Overall, no evolutionary relationship was observed among the clades of the reference sequence and retrieved nucleotide sequences (Figure 2).

The clades of M and E proteins of the examined isolates along with their reference sequences (YP_009724397 and YP_009724390) showed no association with the β-coronavirus species, whereas the β-coronavirus species displayed strong associations among themselves (100%) (Multimedia Appendix 10). Overall, the neighbor-joining trees for the four major proteins indicated total divergence among β-coronavirus species and retrieved sequences of SARS-CoV-2, and there was only weak or no support between the SARS-CoV-2 clades. The length of the branches of the neighbor-joining tree represents the genetic distance between species (Figures 1-2, Multimedia Appendix 10). Moreover, all the alternative and noncontradictory nodes as well as the repeatability of bootstrap values were rejected in this analysis.

Figure 1. Neighbor-joining phylogenetic tree of SARS-CoV-2 spike (S) protein. The tree is divided into seven groups: Group 1 (B.1.1.7) showed 9 mutations of the Alpha variant (A501Y, A570D, D614G, P681H, T716I, S982A, D1118H, and deletion 69-70, 144 in red); Group 2 (B.1.1.7) showed the most prominent mutation (D614G in green); Group 3 (P.1 and B.1.351) of the Beta and Gamma variants comprises two mutations (E484K and N501Y in blue); Group 4 (B.1.617.2) of the Delta variant showed two mutations (D614G and L452R in pink); Group 5 (B.1.617.2) of the Delta variant showed three mutations (D614G, P681R, and L452R in brown); Group 6 (B.1.351) of the Beta variant has three mutations (D614G, L18F, and A222V in orange); and Group 7 represents the β-coronavirus (BCOV) strains used as outgroups for data comparison.

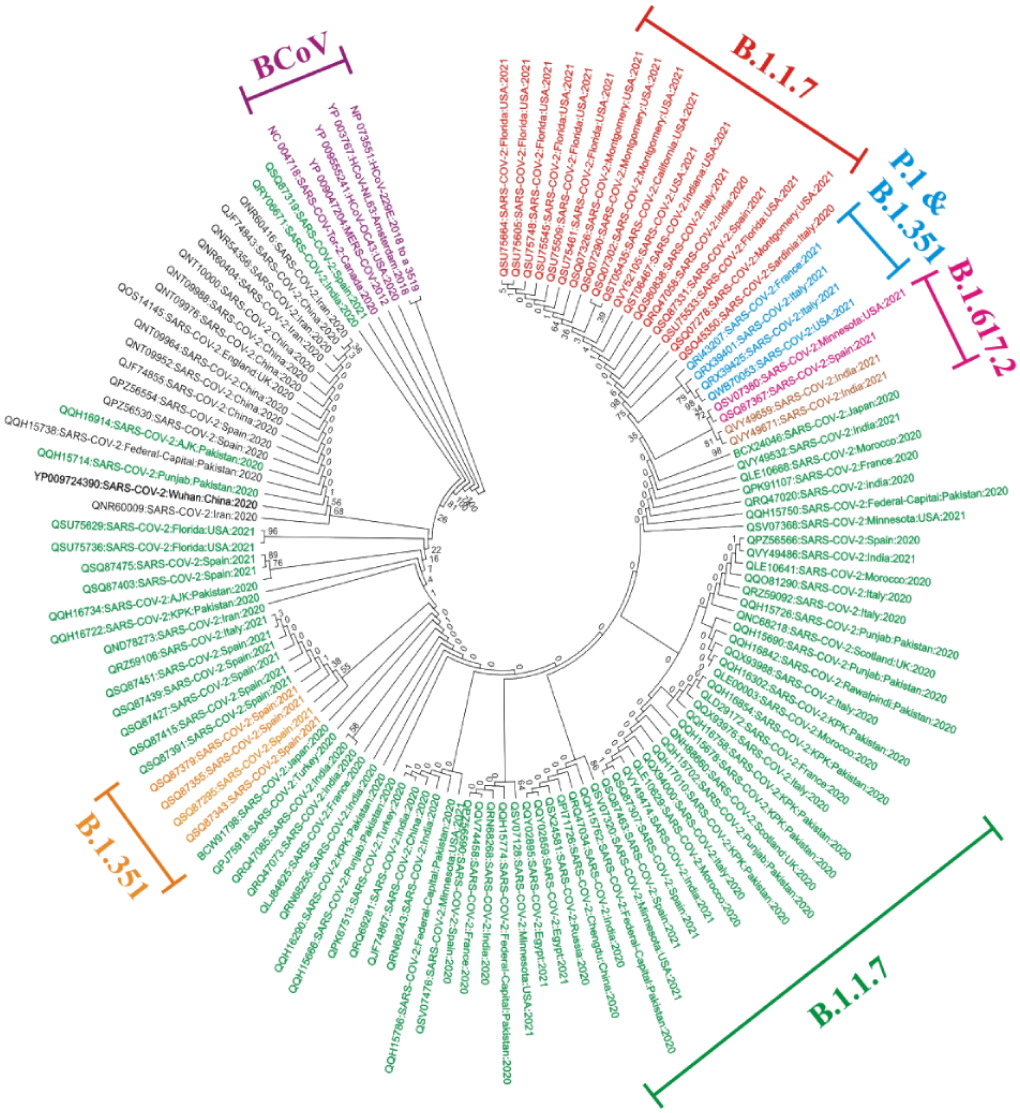
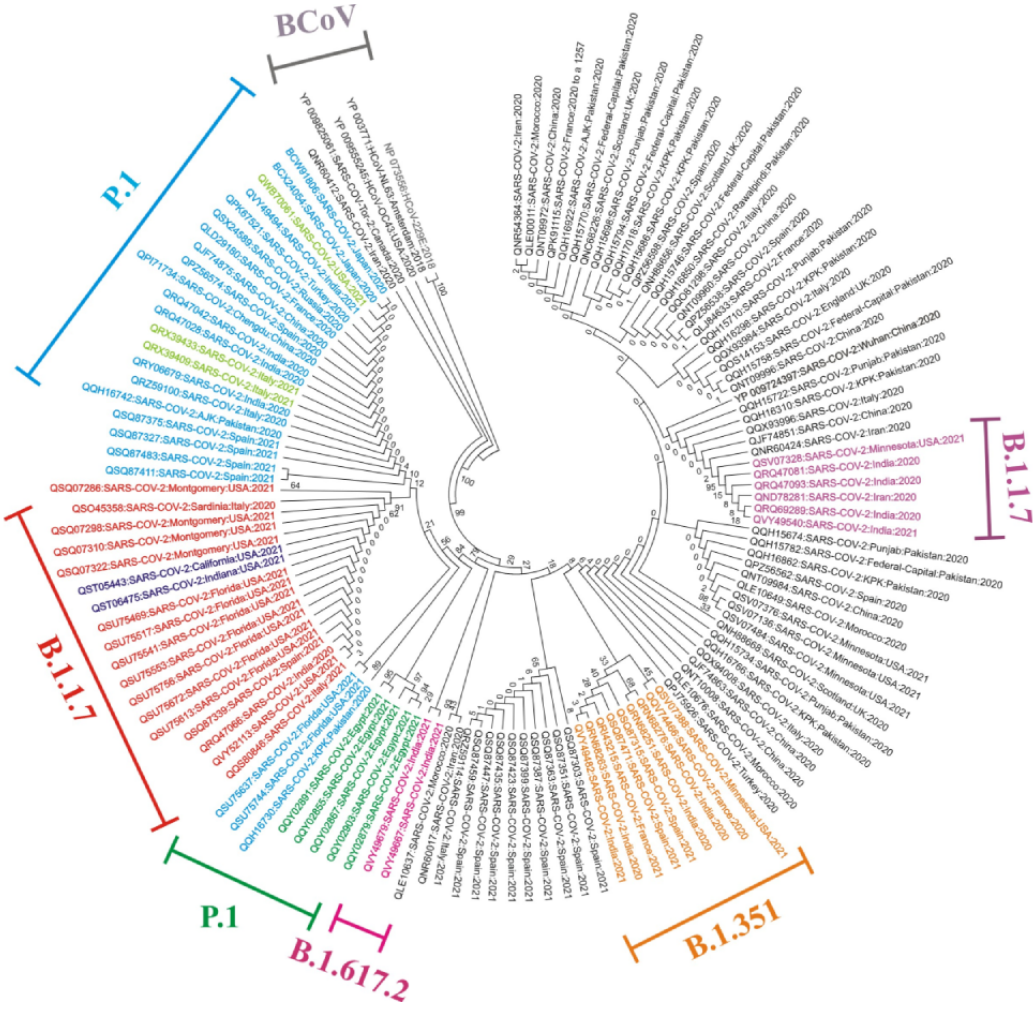


Figure 2. Neighbor-joining phylogenetic tree of SARS-CoV-2 nucleocapsid (N) protein. The tree is divided into eight groups: Group 1 (B.1.1.7, in red) showed four mutations of the Alpha variant (D3L, S235F, R203K, and G204R); Group 2 (B.1.1.7, in dark blue) showing three mutations (S235F, R203K, and G204R); Group 3 (B.1.1.7, in indigo) showing one mutation (S194L); Group 4 (P.1, in green) of the Gamma variant showing two mutations (R203K and G204X); Group 5 (P.1) of the Gamma variant showing two mutations (R203K and G204R, in sky blue), and P80R, R203K, G204R (in parrot green); Group 6 (B.1.351, in orange) of the Beta variant harboring the T205I mutation; Group 7 (B.1.617.2, in pink) of the Delta variant showing three mutations (D377Y, R203M, and G204R); and Group 8 representing the β-coronavirus (BCOV) strains used as outgroups for the comparison of mutations (grey).



Conclusion

The world has witnessed a global pandemic during the 21st century and the majority of nations have contributed to the development of vaccines. Nevertheless, there have been obstacles in the distribution of the vaccines, including a lack of fundamental resources, poor immunization, and manual vaccine replication. Overall, this study can offer a better understanding of the main VoCs of SARS-CoV-2. Several new mutations were detected in this study (see Multimedia Appendix 11), which may contribute to gaining a better understanding of the VoCs as well as in identifying suitable mutations for vaccine targets. These data can further provide evidence for new types of mutations in VoCs, which will help in gaining a better understanding of the epidemiology of SARS-CoV-2 and its dynamic mutational patterns.

Acknowledgments

We appreciate the National Center for Biotechnology Information online portal for providing free access to full-length genomes of SARS-CoV-2 variants. We also gratefully acknowledge the various originating and submitting laboratories for providing the full viral genome sequences and the metadata that were included in this study. We extend our appreciation to the various software developer programmers, including the developers of MEGA 6, coreldraw12, and MultAlin portal. The authors received no specific funding to support this work.

Data Availability

The data sets generated during this study are available in the National Center for Biotechnology Information SARS-COV-2 resources repository [22].

Authors' Contributions

SRK performed the main analyses and wrote the first draft of the manuscript. IA and RM coordinated the research and edited the paper. All authors have read and approved the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Mutations in the S protein of SARS-CoV-2 variants.
[DOCX File , 1130 KB - bioinform_v4i1e43906_app1.docx]

Multimedia Appendix 2
Isolates of SARS-CoV-2 Alpha variant from different countries and their mutations on S and N proteins.
[DOCX File , 36 KB - bioinform_v4i1e43906_app2.docx]

Multimedia Appendix 3
Nonsynonymous mutations observed on S and N proteins of the SARS-CoV-2 Gamma variant.
[DOCX File , 19 KB - bioinform_v4i1e43906_app3.docx]

Multimedia Appendix 4
Mutations in S and N proteins of SARS-CoV-2 Delta variant.
[DOCX File , 18 KB - bioinform_v4i1e43906_app4.docx]

Multimedia Appendix 5
Mutations in the N protein of SARS-CoV-2 variants.
[DOCX File , 1090 KB - bioinform_v4i1e43906_app5.docx]

Multimedia Appendix 6
Mutations identified on S and N proteins of the Beta variant.
[DOCX File , 22 KB - bioinform_v4i1e43906_app6.docx]

Multimedia Appendix 7
Mutations in the M and E proteins of SARS-CoV-2 variants.
[DOCX File , 511 KB - bioinform_v4i1e43906_app7.docx]

Multimedia Appendix 8
Mutations identified from M and E proteins of SARS-CoV-2 variants.
[DOCX File , 17 KB - bioinform_v4i1e43906_app8.docx]

Multimedia Appendix 9
Shift in emergence of SARS-CoV-2 variants worldwide.
[DOCX File , 159 KB - bioinform_v4i1e43906_app9.docx]

Multimedia Appendix 10
Neighbor-joining phylogenetic trees of M and E proteins of SARS-CoV-2.
[DOCX File , 626 KB - bioinform_v4i1e43906_app10.docx]

Multimedia Appendix 11
Novel mutations in S and N proteins of SARS-CoV-2.
[DOCX File , 20 KB - bioinform_v4i1e43906_app11.docx]

References

1. Dougherty K, Mannell M, Naqvi O, Matson D, Stone J. SARS-CoV-2 B.1.617.2 (Delta) variant COVID-19 outbreak associated with a gymnastics facility - Oklahoma, April-May 2021. MMWR Morb Mortal Wkly Rep 2021 Jul 16;70(28):1004-1007. [doi: 10.15585/mmwr.mm7028e2] [Medline: 34264910]

2. Aleem A, Samad ABA, Vaqar S. Emerging variants of SARS-CoV-2 and novel therapeutics against coronavirus (COVID-19). Treasure Island, FL: StatPearls Publishing; May 08, 2023.
3. Galloway SE, Paul P, MacCannell DR, Johansson MA, Brooks JT, MacNeil A, et al. Emergence of SARS-CoV-2 B.1.1.7 lineage - United States, December 29, 2020-January 12, 2021. MMWR Morb Mortal Wkly Rep 2021 Jan 22;70(3):95-99. [doi: 10.15585/mmwr.mm7003e2] [Medline: 33476315]
4. Volz E, Mishra S, Chand M, Barrett JC, Johnson R, Geidelberg L, COVID-19 Genomics UK (COG-UK) consortium, et al. Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. Nature 2021 May;593(7858):266-269 [FREE Full text] [doi: 10.1038/s41586-021-03470-x] [Medline: 33767447]
5. Yang T, Yu P, Chang Y, Liang K, Tso H, Ho M, et al. Effect of SARS-CoV-2 B.1.1.7 mutations on spike protein structure and function. Nat Struct Mol Biol 2021 Sep;28(9):731-739. [doi: 10.1038/s41594-021-00652-z] [Medline: 34385690]
6. Janik E, Niemcewicz M, Podogrocki M, Majsterek I, Bijak M. The emerging concern and interest SARS-CoV-2 variants. Pathogens 2021 May 21;10(6):633 [FREE Full text] [doi: 10.3390/pathogens10060633] [Medline: 34064143]
7. Meng B, Kemp SA, Papa G, Datir R, Ferreira IATM, Marelli S, COVID-19 Genomics UK (COG-UK) Consortium, et al. Recurrent emergence of SARS-CoV-2 spike deletion H69/V70 and its role in the Alpha variant B.1.1.7. Cell Rep 2021 Jun 29;35(13):109-292 [FREE Full text] [doi: 10.1016/j.celrep.2021.109292] [Medline: 34166617]
8. Rahman MS, Islam MR, Alam ASMRU, Islam I, Hoque MN, Akter S, et al. Evolutionary dynamics of SARS-CoV-2 nucleocapsid protein and its consequences. J Med Virol 2021 Apr;93(4):2177-2195. [doi: 10.1002/jmv.26626] [Medline: 33095454]
9. Tegally H, Wilkinson E, Giovanetti M, Iranzadeh A, Fonseca V, Giandhari J, et al. Detection of a SARS-CoV-2 variant of concern in South Africa. Nature 2021 Apr;592(7854):438-443. [doi: 10.1038/s41586-021-03402-9] [Medline: 33690265]
10. Mwenda M, Saasa N, Sinyange N, Busby G, Chipimo PJ, Hendry J, et al. Detection of B.1.351 SARS-CoV-2 variant strain - Zambia, December 2020. MMWR Morb Mortal Wkly Rep 2021 Feb 26;70(8):280-282. [doi: 10.15585/mmwr.mm7008e2] [Medline: 33630820]
11. Wu K, Werner AP, Moliva JI, Koch M, Choi A, Stewart-Jones GBE, et al. mRNA-1273 vaccine induces neutralizing antibodies against spike mutants from global SARS-CoV-2 variants. bioRxiv 2021 Jan 25:7948. [doi: 10.1101/2021.01.25.427948] [Medline: 33501442]
12. Faria NR, Mellan TA, Whittaker C, Claro IM, Candido DDS, Mishra S, et al. Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. Science 2021 May 21;372(6544):815-821 [FREE Full text] [doi: 10.1126/science.abh2644] [Medline: 33853970]
13. Washington NL, Gangavarapu K, Zeller M, Bolze A, Cirulli ET, Schiabor Barrett KM, et al. Emergence and rapid transmission of SARS-CoV-2 B.1.1.7 in the United States. Cell 2021 May 13;184(10):2587-2594 [FREE Full text] [doi: 10.1016/j.cell.2021.03.052] [Medline: 33861950]
14. Zhu Z, Liu G, Meng K, Yang L, Liu D, Meng G. Rapid spread of mutant alleles in worldwide SARS-CoV-2 strains revealed by genome-wide single nucleotide polymorphism and variation analysis. Genome Biol Evol 2021 Feb 03;13(2):evab015 [FREE Full text] [doi: 10.1093/gbe/evab015] [Medline: 33512495]
15. Shiehzadegan S, Alaghemand N, Fox M, Venketaraman V. Analysis of the Delta variant B.1.617.2 COVID-19. Clin Pract 2021 Oct 21;11(4):778-784 [FREE Full text] [doi: 10.3390/clinpract11040093] [Medline: 34698149]
16. Cherian S, Potdar V, Jadhav S, Yadav P, Gupta N, Das M, et al. SARS-CoV-2 spike mutations, L452R, T478K, E484Q and P681R, in the second wave of COVID-19 in Maharashtra, India. Microorganisms 2021 Jul 20;9(7):1542 [FREE Full text] [doi: 10.3390/microorganisms9071542] [Medline: 34361977]
17. Suratekar R, Ghosh P, Niesen MJM, Donadio G, Anand P, Soundararajan V, et al. High diversity in Delta variant across countries revealed by genome-wide analysis of SARS-CoV-2 beyond the Spike protein. Mol Syst Biol 2022 Feb;18(2):e10673 [FREE Full text] [doi: 10.15252/msb.202110673] [Medline: 35156767]
18. Syed AM, Taha TY, Tabata T, Chen IP, Ciling A, Khalid MM, et al. Rapid assessment of SARS-CoV-2-evolved variants using virus-like particles. Science 2021 Dec 24;374(6575):1626-1632 [FREE Full text] [doi: 10.1126/science.abl6184] [Medline: 34735219]
19. Jackson CB, Farzan M, Chen B, Choe H. Mechanisms of SARS-CoV-2 entry into cells. Nat Rev Mol Cell Biol 2022 Jan;23(1):3-20 [FREE Full text] [doi: 10.1038/s41580-021-00418-x] [Medline: 34611326]
20. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol Biol Evol 2013 Dec;30(12):2725-2729 [FREE Full text] [doi: 10.1093/molbev/mst197] [Medline: 24132122]
21. Chi X, Yan R, Zhang J, Zhang G, Zhang Y, Hao M, et al. A neutralizing human antibody binds to the N-terminal domain of the Spike protein of SARS-CoV-2. Science 2020 Aug 07;369(6504):650-655 [FREE Full text] [doi: 10.1126/science.abc6952] [Medline: 32571838]
22. SARS-CoV-2 Data. National Library of Medicine. URL: https://www.ncbi.nlm.nih.gov/sars-cov-2/ [accessed 2023-06-23]
23. Corpet F. Multiple sequence alignment with hierarchical clustering. Nucleic Acids Res 1988 Nov 25;16(22):10881-10890 [FREE Full text] [doi: 10.1093/nar/16.22.10881] [Medline: 2849754]
24. Zhang L, Jackson CB, Mou H, Ojha A, Peng H, Quinlan BD, et al. SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity. Nat Commun 2020 Nov 26;11(1):6013. [doi: 10.1038/s41467-020-19808-4] [Medline: 33243994]

25. Zhou P, Yang X, Wang X, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 2020 Mar;579(7798):270-273 [FREE Full text] [doi: 10.1038/s41586-020-2012-7] [Medline: 32015507]
26. Ostrov DA. Structural consequences of variation in SARS-CoV-2 B.1.1.7. J Cell Immunol 2021;3(2):103-108 [FREE Full text] [doi: 10.33696/immunology.3.085] [Medline: 33969357]
27. Liu H, Yuan M, Huang D, Bangaru S, Zhao F, Lee CD, et al. A combination of cross-neutralizing antibodies synergizes to prevent SARS-CoV-2 and SARS-CoV pseudovirus infection. Cell Host Microbe 2021 May 12;29(5):806-818 [FREE Full text] [doi: 10.1016/j.chom.2021.04.005] [Medline: 33894127]
28. Liu Y, Liu J, Plante KS, Plante JA, Xie X, Zhang X, et al. The N501Y spike substitution enhances SARS-CoV-2 infection and transmission. Nature 2022 Feb;602(7896):294-299 [FREE Full text] [doi: 10.1038/s41586-021-04245-0] [Medline: 34818667]
29. McCarthy KR, Rennick LJ, Nambulli S, Robinson-McCarthy LR, Bain WG, Haidar G, et al. Recurrent deletions in the SARS-CoV-2 spike glycoprotein drive antibody escape. Science 2021 Mar 12;371(6534):1139-1142 [FREE Full text] [doi: 10.1126/science.abf6950] [Medline: 33536258]
30. La Rosa G, Mancini P, Bonanno Ferraro G, Veneri C, Iaconelli M, Lucentini L, et al. Rapid screening for SARS-CoV-2 variants of concern in clinical and environmental samples using nested RT-PCR assays targeting key mutations of the spike protein. Water Res 2021 Jun 01;197:117104 [FREE Full text] [doi: 10.1016/j.watres.2021.117104] [Medline: 33857895]
31. Stojanov D. Phylogenicity of B.1.1.7 surface glycoprotein, novel distance function and first report of V90T missense mutation in SARS-CoV-2 surface glycoprotein. Meta Gene 2021;30:100967. [doi: 10.1016/j.mgene.2021.100967]
32. Patro LPP, Sathyaseelan C, Uttamrao PP, Rathinavelan T. The evolving proteome of SARS-CoV-2 predominantly uses mutation combination strategy for survival. Comput Struct Biotechnol J 2021;19:3864-3875 [FREE Full text] [doi: 10.1016/j.csbj.2021.05.054] [Medline: 34109017]
33. Wise J. Covid-19: The E484K mutation and the risks it poses. BMJ 2021 Feb 05;372:n359. [doi: 10.1136/bmj.n359] [Medline: 33547053]
34. Starr TN, Greaney AJ, Hilton SK, Ellis D, Crawford KHD, Dingens AS, et al. Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. Cell 2020 Sep 03;182(5):1295-1310 [FREE Full text] [doi: 10.1016/j.cell.2020.08.012] [Medline: 32841599]
35. Hodcroft EB, Zuber M, Nadeau S, Vaughan TG, Crawford KHD, Althaus CL, SeqCOVID-SPAIN Consortium, et al. Emergence and spread of a SARS-CoV-2 variant through Europe in the summer of 2020. medRxiv. 2021 Mar 24. URL: https://www.medrxiv.org/content/10.1101/2020.10.25.20219063v1 [accessed 2023-06-23]
36. Nonaka CKV, Franco MM, Gräf T, de Lorenzo Barcia CA, de Ávila Mendonça RN, de Sousa KAF, et al. Genomic evidence of SARS-CoV-2 reinfection involving E484K spike mutation, Brazil. Emerg Infect Dis 2021 May;27(5):1522-1524. [doi: 10.3201/eid2705.210191] [Medline: 33605869]
37. Peacock TP, Sheppard CM, Brown JC, Goonawardane N, Zhou J, Whiteley M, PHE Virology Consortium, et al. The SARS-CoV-2 variants associated with infections in India, B. 1.617, show enhanced spike cleavage by furin. BioRxiv. 2021. URL: https://www.biorxiv.org/content/10.1101/2021.05.28.446163v1 [accessed 2023-06-23]
38. Zhao S, Lou J, Chong MKC, Cao L, Zheng H, Chen Z, et al. Inferring the association between the risk of COVID-19 case fatality and N501Y substitution in SARS-CoV-2. Viruses 2021 Apr 08;13(4):638 [FREE Full text] [doi: 10.3390/v13040638] [Medline: 33918060]
39. Plante JA, Liu Y, Liu J, Xia H, Johnson BA, Lokugamage KG, et al. Spike mutation D614G alters SARS-CoV-2 fitness. Nature 2021 Apr;592(7852):116-121 [FREE Full text] [doi: 10.1038/s41586-020-2895-3] [Medline: 33106671]
40. Wu S, Tian C, Liu P, Guo D, Zheng W, Huang X, et al. Effects of SARS-CoV-2 mutations on protein structures and intraviral protein-protein interactions. J Med Virol 2021 Apr;93(4):2132-2140 [FREE Full text] [doi: 10.1002/jmv.26597] [Medline: 33090512]
41. Zhou P, Shi Z. SARS-CoV-2 spillover events. Science 2021 Jan 08;371(6525):120-122. [doi: 10.1126/science.abf6097] [Medline: 33414206]
42. Funk T, Pharris A, Spiteri G, Bundle N, Melidou A, Carr M, COVID study groups. Characteristics of SARS-CoV-2 variants of concern B.1.1.7, B.1.351 or P.1: data from seven EU/EEA countries, weeks 38/2020 to 10/2021. Euro Surveill 2021 Apr;26(16):2100348 [FREE Full text] [doi: 10.2807/1560-7917.ES.2021.26.16.2100348] [Medline: 33890566]
43. Martins AF, Zavascki AP, Wink PL, Volpato FCZ, Monteiro FL, Rosset C, Ramos, et al. Detection of SARS-CoV-2 lineage P.1 in patients from a region with exponentially increasing hospitalisation rate, February 2021, Rio Grande do Sul, Southern Brazil. Euro Surveill 2021 Mar;26(12):2100276 [FREE Full text] [doi: 10.2807/1560-7917.ES.2021.26.12.2100276] [Medline: 33769251]
44. Yu I, Oldham ML, Zhang J, Chen J. Crystal structure of the severe acute respiratory syndrome (SARS) coronavirus nucleocapsid protein dimerization domain reveals evolutionary linkage between corona- and arteriviridae. J Biol Chem 2006 Jun 23;281(25):17134-17139 [FREE Full text] [doi: 10.1074/jbc.M602107200] [Medline: 16627473]
45. He R, Leeson A, Ballantine M, Andonov A, Baker L, Dobie F, et al. Characterization of protein-protein interactions between the nucleocapsid protein and membrane protein of the SARS coronavirus. Virus Res 2004 Oct;105(2):121-125 [FREE Full text] [doi: 10.1016/j.virusres.2004.05.002] [Medline: 15351485]

46. Siu YL, Teoh KT, Lo J, Chan CM, Kien F, Escriou N, et al. The M, E, and N structural proteins of the severe acute respiratory syndrome coronavirus are required for efficient assembly, trafficking, and release of virus-like particles. J Virol 2008 Nov;82(22):11318-11330 [FREE Full text] [doi: 10.1128/JVI.01052-08] [Medline: 18753196]
47. Schoeman D, Fielding BC. Coronavirus envelope protein: current knowledge. Virol J 2019 May 27;16(1):69 [FREE Full text] [doi: 10.1186/s12985-019-1182-0] [Medline: 31133031]

Abbreviations

ACE2: angiotensin-converting enzyme 2
E protein: envelope protein
hCoV: human coronavirus
M protein: membrane glycoprotein
MERS: Middle East Respiratory Syndrome
NCBI: National Center for Biotechnology Information
N protein: nucleocapsid protein
ORF: open reading frame
RBD: receptor-binding domain
S protein: surface glycoprotein (spike protein)
VoC: variant of concern
VoI: variant of interest

Edited by A Uzun; submitted 29.10.22; peer-reviewed by J Vandana, I Shaker; comments to author 19.12.22; revised version received 02.03.23; accepted 08.06.23; published 14.07.23.

Please cite as:
Khetran SR, Mustafa R
Mutations of SARS-CoV-2 Structural Proteins in the Alpha, Beta, Gamma, and Delta Variants: Bioinformatics Analysis
JMIR Bioinform Biotech 2023;4:e43906
URL: https://bioinform.jmir.org/2023/1/e43906
doi:10.2196/43906
PMID:37485046

Original Paper

The Identification of Potential Drugs for Dengue Hemorrhagic Fever: Network-Based Drug Reprofiling Study

Praveenkumar Kochuthakidiyel Suresh[1*], MSc; Gnanasoundari Sekar[2*], MTech; Kavya Mallady[3], MSc; Wan Suriana Wan Ab Rahman[4], MD; Wan Nazatul Shima Shahidan[4], PhD; Gokulakannan Venkatesan[4*], MTech

[1]Central Research Facility, Sri Ramachandra Institute of Higher Education and Research, Chennai, India

[2]Department of Biotechnology, Bharathidasan Institute of Technology Campus, Anna University, Tiruchirappalli, India

[3]Centre for Toxicology and Developmental Research, Sri Ramachandra Institute of Higher Education and Research, Chennai, India

[4]School of Dental Sciences, Universiti Sains Malaysia, Health Campus, Kelantan, Malaysia

[*]these authors contributed equally

Corresponding Author:
Gokulakannan Venkatesan, MTech
School of Dental Sciences
Universiti Sains Malaysia
Health Campus
16150 Kubang Kerian
Kota Bharu, Kelantan
Kelantan, 16150
Malaysia
Phone: 60 162543854
Email: gokulkannancmr@gmail.com

Abstract

Background: Dengue fever can progress to dengue hemorrhagic fever (DHF), a more serious and occasionally fatal form of the disease. Indicators of serious disease arise about the time the fever begins to reduce (typically 3 to 7 days following symptom onset). There are currently no effective antivirals available. Drug repurposing is an emerging drug discovery process for rapidly developing effective DHF therapies. Through network pharmacology modeling, several US Food and Drug Administration (FDA)-approved medications have already been researched for various viral outbreaks.

Objective: We aimed to identify potentially repurposable drugs for DHF among existing FDA-approved drugs for viral attacks, symptoms of viral fevers, and DHF.

Methods: Using target identification databases (GeneCards and DrugBank), we identified human–DHF virus interacting genes and drug targets against these genes. We determined hub genes and potential drugs with a network-based analysis. We performed functional enrichment and network analyses to identify pathways, protein-protein interactions, tissues where the gene expression was high, and disease-gene associations.

Results: Analyzing virus-host interactions and therapeutic targets in the human genome network revealed 45 repurposable medicines. Hub network analysis of host-virus-drug associations suggested that aspirin, captopril, and rilonacept might efficiently treat DHF. Gene enrichment analysis supported these findings. According to a Mayo Clinic report, using aspirin in the treatment of dengue fever may increase the risk of bleeding complications, but several studies from around the world suggest that thrombosis is associated with DHF. The human interactome contains the genes *prostaglandin-endoperoxide synthase 2* (*PTGS2*), *angiotensin converting enzyme (ACE)*, and *coagulation factor II, thrombin* (*F2)*, which have been documented to have a role in the pathogenesis of disease progression in DHF, and our analysis of most of the drugs targeting these genes showed that the hub gene module (human-virus-drug) was highly enriched in tissues associated with the immune system (P=7.29 × 10^{-24}) and human umbilical vein endothelial cells (P=1.83 × 10^{-20}); this group of tissues acts as an anticoagulant barrier between the vessel walls and blood. Kegg analysis showed an association with genes linked to cancer (P=1.13 × 10^{-14}) and the advanced glycation end products–receptor for advanced glycation end products signaling pathway in diabetic complications (P=3.52 × 10^{-14}), which indicates that DHF patients with diabetes and cancer are at risk of higher pathogenicity. Thus, gene-targeting medications may play a significant part in limiting or worsening the condition of DHF patients.

Conclusions: Aspirin is not usually prescribed for dengue fever because of bleeding complications, but it has been reported that using aspirin in lower doses is beneficial in the management of diseases with thrombosis. Drug repurposing is an emerging field in which clinical validation and dosage identification are required before the drug is prescribed. Further retrospective and collaborative international trials are essential for understanding the pathogenesis of this condition.

(JMIR Bioinform Biotech 2023;4:e37306) doi:10.2196/37306

KEYWORDS

dengue hemorrhagic fever; drug reprofiling; network pharmacology; network medicine; DHF; repurposable drugs; viral fevers; drug repurposing

Introduction

Dengue fever, also known as "breakbone fever," is characterized by acute, severe fever in patients 3 to 14 days after they are bitten by an infected mosquito. Migraine, retro-orbital pain, myalgia, muscle ache, signs of hemolytic anemia, rash, and a low white blood cell count are only a few of the symptoms [1]. Dengue hemorrhagic fever (DHF), a severe and sometimes fatal manifestation of the disease, affects certain patients with dengue fever. These patients may show warning signs of serious disease close to the period the fever begins to diminish (typically 3 to 7 days following symptom onset). Severe abdominal discomfort, continuous vomiting, a significant change in temperature (from fever to hypothermia), hemorrhagic manifestations, or a change in mental status (eg, irritability, confusion, or obtundation) are also warning indicators [2]. Restlessness, chilly, clammy skin, a rapid, weak pulse, and a narrowing of pulse pressure (both systolic blood pressure and diastolic blood pressure) are all early indications of shock. Patients with dengue fever should be advised to return to the hospital if any of these symptoms appear.

According to one estimate, 390 million dengue virus infections occur each year (95% credible interval [CI] 284 million to 528 million), with 96 million (95% CI 67 million to 136 million) showing clinical symptoms of any severity [3]. According to the World Health Organization (WHO), 3.9 billion individuals are at risk of contracting the dengue virus. Although there is a risk of infection in 129 nations, Asia bears 70% of the actual burden. Over the last 2 decades, the number of dengue cases reported to the WHO has increased more than 8 times, from 505,430 cases in 2000 to over 2.4 million in 2010 and 5.2 million in 2019. Between 2000 and 2015, the number of deaths reported grew from 960 to 2032. This worrying rise in case numbers can be explained in part by a shift in national practices for recording and reporting dengue fever to health ministries and the WHO. However, it also symbolizes the government's acknowledgment of the problem, and hence the need to disclose the prevalence of dengue fever [4]. As a result, while the complete global burden of the disease remains unknown, the observed growth only takes us closer to a more precise estimate of the full extent of the burden.

In 2021, dengue fever increased in Bangladesh, Brazil, the Cook Islands, Ecuador, India, Indonesia, the Maldives, Mauritania, Mayotte (an overseas department of France), Nepal, Singapore, Sri Lanka, Sudan, Thailand, Timor-Leste, and Yemen. Dengue fever was also still a problem in Brazil, the Cook Islands, Colombia, Fiji, Kenya, Paraguay, Peru, and Reunion Island in 2021 [5]. The COVID-19 epidemic is placing an enormous strain on health care and management systems all across the world. During this critical period, the WHO has stressed the significance of maintaining efforts to prevent, identify, and treat vector-borne diseases such as dengue fever and other arboviral infections as case numbers rise in various countries, putting urban people at risk for both diseases [6].

Recent systems biology developments suggest a unique testable hypothesis for systematic drug repurposing [7,8]. This can greatly decrease the time spent on research and development compared to traditional drug development programs. The typical strategy takes 10 to 16 years to develop a new treatment. A medication repurposing plan costs $1.6 billion to create, whereas a typical strategy costs $12 billion [9]. The identification of new targets and illness proteins has been made possible by rapid advances in genomic, proteomic, structural, functional, and systems investigations of existing targets and other disease proteins [10].

In this study, we provide an embedded medicine platform that uses a network-based method to quantify the association of DHF with human-host interactions, we examine the efficacy of existing US Food and Drug Administration (FDA)-approved medications as potentially repurposable drugs, and we also examine their associations with DHF-host genes. To discover and prioritize existing pharmacological targets in the DHF pathway, we chose FDA-approved medications from a clinical registry database.

Methods

Building the DHF-Human Interactome

We performed an extensive electronic-literature similarity search from January 2000 to January 2022 for keywords related to DHF and human interactions, including "dengue hemorrhagic fever," "dengue hemorrhagic fever and human gene interaction," and "dengue hemorrhagic fever human interactome," with a focus on reviews, editorials, commentary, letters, case reports, and original research manuscripts published on PubMed, Google Scholar, and other widely used databases. We manually removed duplicates based on variables such as author, nationality, and collaborations, and we finally extracted 31 articles. We performed a database similarity search using GeneCards and found 588 DHF-targeting human genes. A total of 59 host-interacting DHF genes were sorted based on a hit score of 50. Experimental evidence for interactions between human proteins and dengue virus proteins was obtained with high-throughput yeast 2-hybrid screening methods [11].

Recently, Dey and Mukhopadhyay [12] reported the development of DenvInt, a database of manually curated experimental data of dengue protein and host protein interactions. We merged the data from published references to DenvInt and used them in our analysis along with the dengue-host interactome data from recent investigations. Infectious diseases are the result of molecular crosstalk between hosts and their pathogens. This crosstalk is in part mediated by host-pathogen (HP) protein-protein interactions (PPIs). HP-PPIs play crucial roles in infections [13]. The best way of unveiling their mechanisms is to investigate the HP-PPI network [14].

Human (Host) Gene–DHF Gene Interactome

The key host genes involved in DHF were identified from the GeneCards database using the search terms "dengue hemorrhagic fever" and "dengue hemorrhagic fever interacting human genes." GeneCards is a searchable, integrative database that provides comprehensive, user-friendly information on all annotated and predicted human genes. The knowledge base automatically integrates gene-centric data from approximately 150 web sources, including genomic, transcriptomic, proteomic, genetic, clinical, and functional information. As of January 13, 2022, GeneCards comprises 326,787 genes, including 18,870 disease genes and 500 host genes [15]. The functional genes identified from GeneCards and related literature were collected and are presented in Multimedia Appendix 1. The PPI network was built with Cytoscape (version 3.9.0; Cytoscape Consortium) and Gephi (version 0.9.2; Gephi Association) [16].

Drug Target (Human Gene) Interactome

We collected 87 FDA-approved antiviral and 137 anti-DHF drugs from the Therapeutic Target Database, compared them with the results from the DrugBank database [17,18], identified drug targets, and formulated them as a data set (Multimedia Appendix 1). Human-drug interactions are based on drug targets (ie, drug targeting genes); these were visualized using Cytoscape [19]. The nodes in a network represent antiviral drugs or anti-DHF drugs and the edges of the network represent drugs targeting human genes [20].

Building the Drug-to-Human Interactome

A network pharmacology–based host–DHF–antiviral–anti-DHF drug interactome was constructed by assembling the host-DHF interacting proteins with or without antivirals and anti-DHF drugs. The PPI network was built with Gephi [21] and Cytoscape. Each node in the constructed PPI network indicates a host gene and an edge indicates an interacting drug target.

Network Hub Gene Identification

Highly connected nodes (hubs) in biological networks are topologically important to the structure of the network and have also been shown to be preferentially associated with a range of phenotypes of interest. Hub genes can be identified using the Contextual Hub Analysis Tool plug-in of Cytoscape [22], which enables users to easily construct and visualize a network of interactions from a gene list of interest.

Functional Enrichment Analysis of Genes and Drugs

Functional enrichment analysis is a method to determine classes of genes or drugs that are overrepresented in a large group of genes or drugs and may have relations with disease phenotypes. This approach uses statistical methods to determine significantly enriched groups of genes. The biological relevance and functional pathways of our data sets were revealed by enriching the semantic similarities of the pathways and tissue. All functional enrichment analyses were performed using the Enrichr enrichment platform (Icahn School of Medicine) [23] as additional evidence for drug repurposing. Enrichr is a comprehensive gene enrichment analysis platform that comprises 382,208 terms from 192 libraries.

Results

Human (Host)-DHF (Viral) Gene Interactome

We constructed a host-DHF interactome consisting of 59 interacting genes with 60 nodes and 59 edges (Multimedia Appendix 2, Figure S1A). A Kegg pathway enrichment analysis indicated that genes involved in the advanced glycation end products–receptor for advanced glycation end products (AGE-RAGE) signaling pathway in diabetic complications were most enriched (P=3.01^{32}), which indicates that patients with DHF have a higher chance of poor blood sugar management; meanwhile, AGE-RAGE signaling has been shown to increase oxidative stress and promote diabetes-mediated vascular calcification through activation of nicotinamide adenine dinucleotide phosphate oxidase-1 and decreased expression of superoxide dismutase type 1 [24]. Chagas disease (P=4.65 × 10^{-32}) and influenza A (P=6.17 × 10^{-31}) pathways were also typically enriched. Compared to the Kegg pathway analysis, a reactome pathway analysis indicated that the immune system (P=3.93 × 10^{-28}) and cytokine signaling in the immune system (P=7.06 × 10^{-27}) were enriched, which indicates that DHF most strongly hijacks the human immune system–associated gene pathways; a gene set was also identified in which the immune system (P=1.12 × 10^{-61}) and bronchoalveolar lavage (P=2.85 × 10^{-50}) tissues were most enriched (Multimedia Appendix 3, Figure S1B).

Host–Viral–Antiviral Drug Target Interactome

A host–DHF–antiviral drug interactome was built with 298 nodes and 370 edges from 237 interacting genes (Multimedia Appendix 4, Figure S2A). Kegg pathway gene enrichment analysis revealed that upon antiviral drug administration, the most enriched gene was a neuroactive ligand–receptor interaction (P=3.22 × 10^{-45}), that is, a collection of genes associated with intracellular and extracellular signaling pathways in the plasma membrane and mitogen-activated protein kinase pathways (P=9.57 × 10^{-45}), which relay, amplify, and integrate signals from a diverse range of stimuli and elicit appropriate physiological responses in mammalian cells, including cellular proliferation, differentiation, and development; an inflammatory response; and apoptosis (Multimedia Appendix 5, Figure S2B). The reactome pathway analysis indicated that pathway genes were enriched in phase 2, the plateau phase (P=2.85 × 10^{-34}), which sustains cardiac action potential muscle contraction [25,26] and transmission across chemical synapses (ie, neurotransmitters; P=6.03 × 10^{-34}) after antiviral drug

administration for DHF, as were genes in adult (P=2.52 × 10^{-69}) and immune-system (P=1.77 × 10^{-49}) tissue types.

Host–Viral–Anti-DHF Drug Target Interactome

A host–DHF interactome–anti-DHF drug interactome was built with 558 nodes and 861 edges from 419 interacting genes (Multimedia Appendix 6, Figure S3A). Neuroactive ligand–receptor interaction (P=3.37 × 10^{-75}) and the cAMP signaling pathway (P=2.29 × 10^{-54}), which is also known as the adenylyl cyclase pathway and is a G protein–coupled receptor-triggered signaling cascade used in cell communication, were the most enriched gene pathways according to Kegg pathway analysis (Multimedia Appendix 7, Figure S3B). Amine ligand–binding receptors (P=1.76 × 10^{-47}), which act as neurotransmitters in humans, and signal transduction (P=1.40 × 10^{-46}), which involves the binding of extracellular signaling molecules and ligands to receptors located on the cell surface, were highly enriched. Adult (P=1.74 × 10^{-59}) and immune-system (P=3.35 × 10^{-54}) tissue types were the most prominent after anti-DHF drug administration in COVID-19 patients.

Host–Viral–Antiviral–Anti-DHF Drug Target Interactome

Based on all the interactomic data sets, we combined all the data sets to frame a network-based drug reprofiling approach to testing their robustness, which involved a network with 717 nodes and 1175 edges from 487 interacting genes (Figure 1). Gene functional enrichment analysis of the Kegg pathway revealed that gene sets involved in neurotransmitter pathways (P=1.13 × 10^{-84}) and calcium-signaling pathways (P=1.78 × 10^{-66}) were highly enriched, similarly to previous drug-related host-virus interactomes in humans; these also provide a stable outcome when combined drug administration (eg, an antiviral and an anti-DHF drug) is used for patients with systemic DHF (Figure 2). The majority of the gene set was enriched in adult (P=1.53 × 10^{-77}) and immune-system (P=3.07 × 10^{-63}) tissues. Genes related to signal transduction (P=6.76 × 10^{-56}) and signaling by G protein–coupled receptors (P=1.96 × 10^{-49}) were the prevalent reactome pathways enriched in the patients with DHF and combined drug administration.

Figure 1. Human-anti–dengue hemorrhagic fever–antiviral drug–anti–dengue hemorrhagic fever interaction network. ADHF: anti–dengue hemorrhagic fever; AV: antiviral; DB: DrugBank; DHF: dengue hemorrhagic fever. Higher-resolution version of this figure is available in Multimedia Appendix 8.

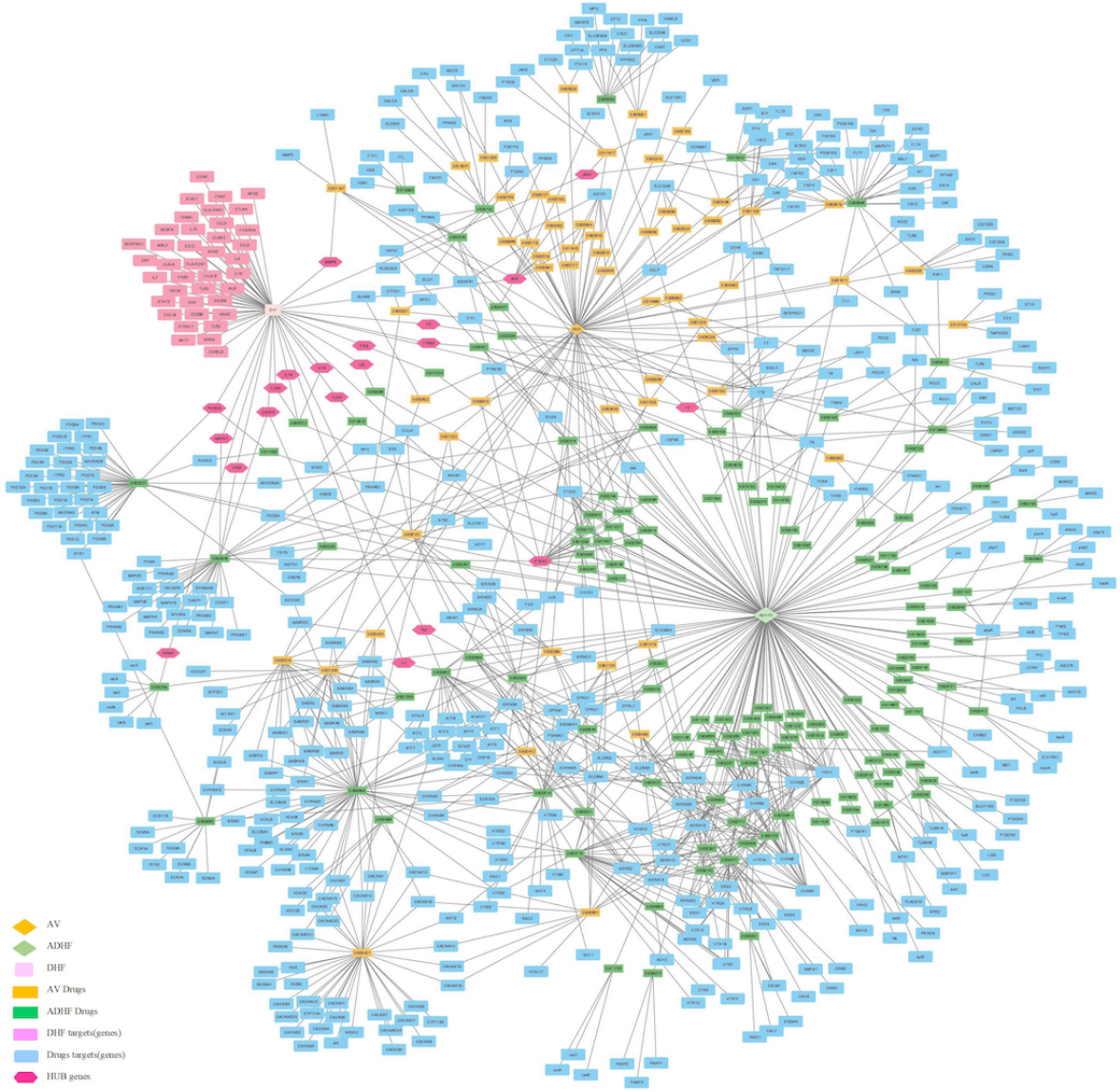


Figure 2. Enrichment analysis of human-anti–dengue hemorrhagic fever–antiviral drug–anti–dengue hemorrhagic fever. (A) Kegg pathways, (B), reactome pathways, and (C) tissues. Higher-resolution version of this figure is available in Multimedia Appendix 9.

Network-Based Drug Repurposing Based on Hub Gene Analysis

We predicted a hub gene module containing 20 interacting genes (66 nodes and 113 edges) from the above interactome of the host-virus-drugs systems framework (Figure 3). A total of 45 drugs were repurposed from the hub gene module, of which 13 were antiviral drugs and 32 were anti-DHF drugs. From the hub gene–drug association network, we determined that 3 major drugs bound efficiently with DHF-targeting human genes: aspirin, captopril, and rilonacept. Thus, these are efficient FDA-approved drugs that can be used in the treatment of DHF (Figure 4). We identified 18 *PTGS2* genes, 10 *ACE* genes, and 4 *F2* genes targeting drugs in hub genes in the network (Figure 2A). Interestingly, 18 of 17 *PTGS2*-targeting drugs were anti-DHF drugs and 10 of 9 *ACE*-targeting genes were antiviral drugs. Moreover, *F2*-targeting drugs had equal numbers of these

2 types of drugs: they included 2 antiviral drugs and 2 anti-DHF drugs (Figure 5).

Gene enrichment analysis showed that the hub gene module was highly enriched in tissues associated with the immune system (P=7.29 × 10^{-24}) and human umbilical vein endothelial cells (P=1.83 × 10^{-20}). This group of tissues acts as an anticoagulant barrier between the vessel walls and blood. Kegg analysis showed that genes associated with cancer (P=1.13 × 10^{-14}) and the AGE-RAGE signaling pathway in diabetic complications (P=3.52 × 10^{-14}) were enriched, which indicates that DHF patients with diabetes and cancer are at risk of higher pathogenicity. Reactome pathway gene enrichment analysis provided evidence that immune system–associated pathways, including signaling by interleukins (P=2.04^{-14}) and cytokine signaling in the immune system (P=7.12^{-14}), were most enriched (Figure 6).

Figure 3. Representation of human-interacting anti-dengue hemorrhagic fever–antiviral drug–anti-dengue hemorrhagic fever hub network. ADHFD: anti–dengue hemorrhagic fever drugs; AVD: antiviral drugs; DB: DrugBank; DHF: dengue hemorrhagic fever.

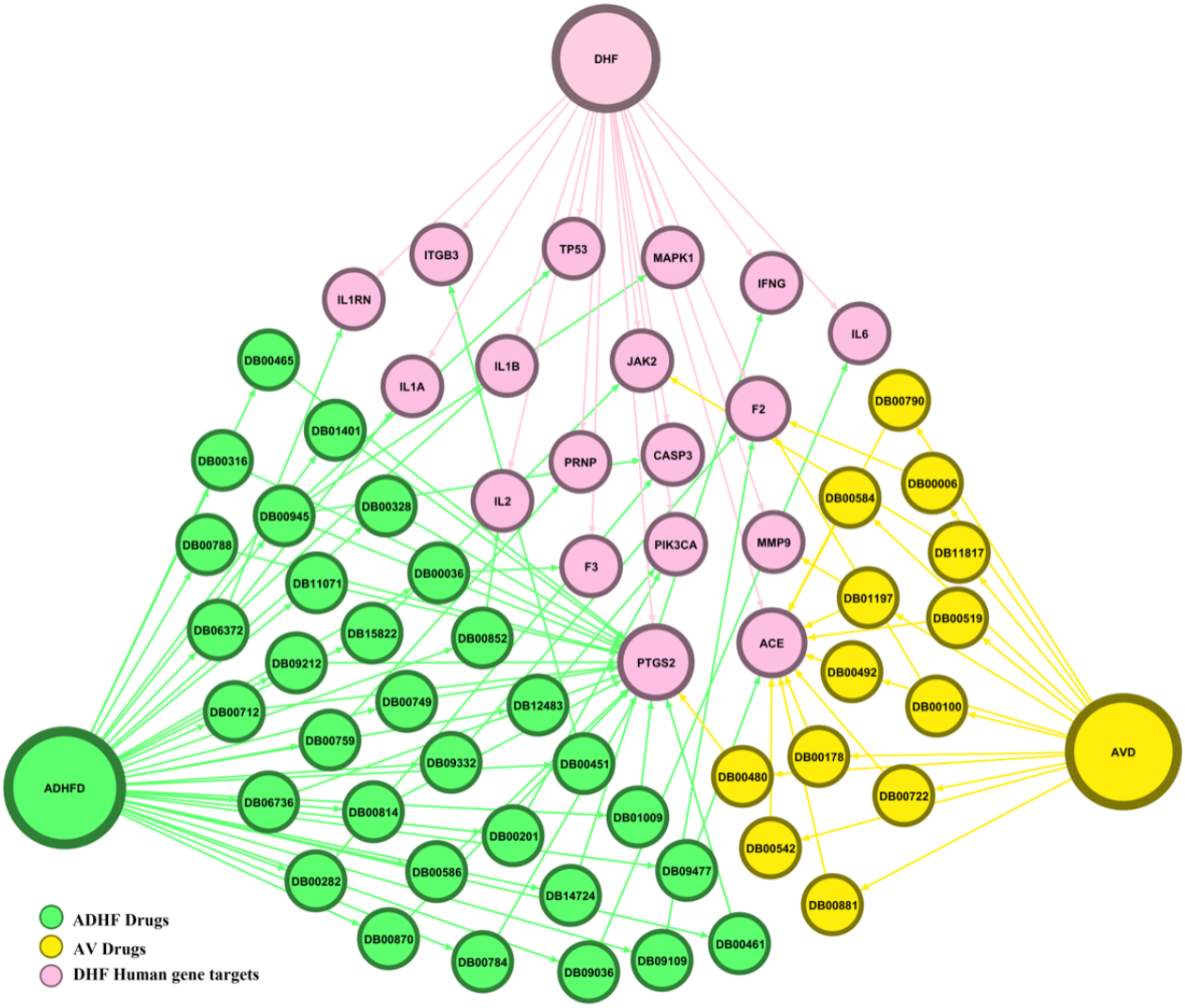


Figure 4. Gene interactions in hub network for (A) *PTGS2*, (B), *F2*, and (C) *ACE*. ADHFD: anti–dengue hemorrhagic fever drugs; AVD: antiviral drugs; DB: DrugBank; DHF: dengue hemorrhagic fever.

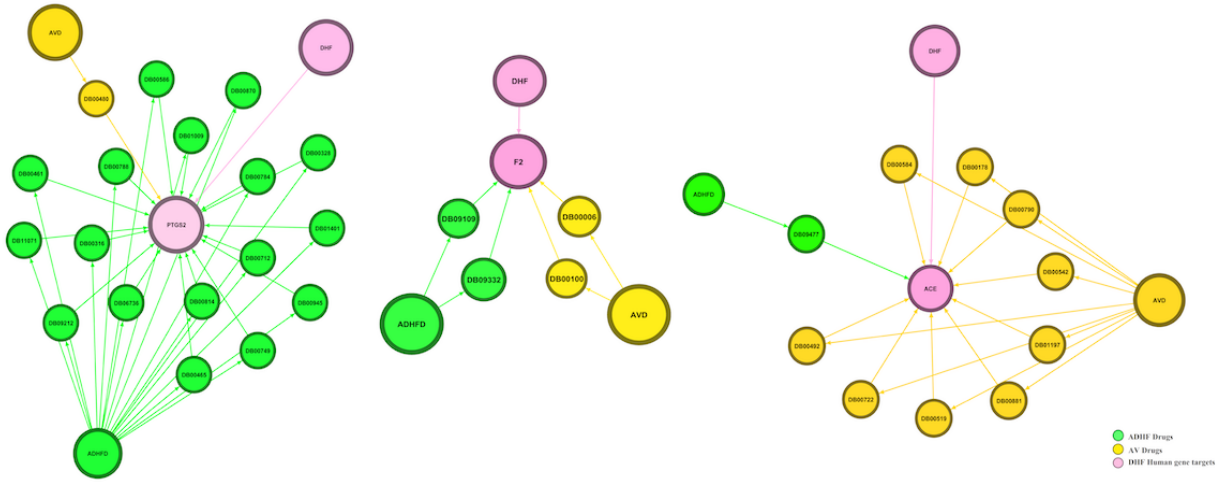


Figure 5. Repurposable drugs identified through hub network analysis: (A) aspirin, (B) rilonacept, and (C) capropril. ADHFD: anti–dengue hemorrhagic fever drugs; AVD: antiviral drugs; DHF: dengue hemorrhagic fever.

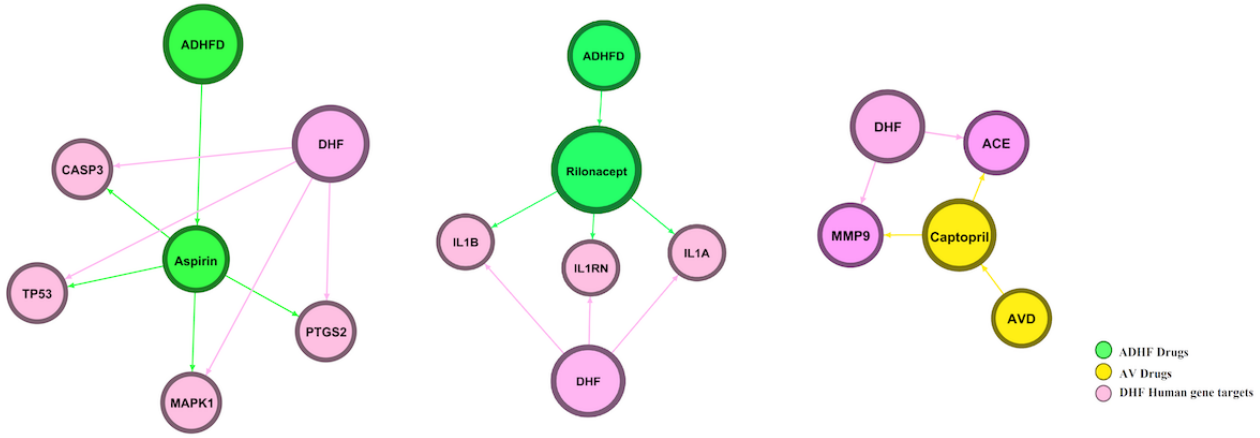
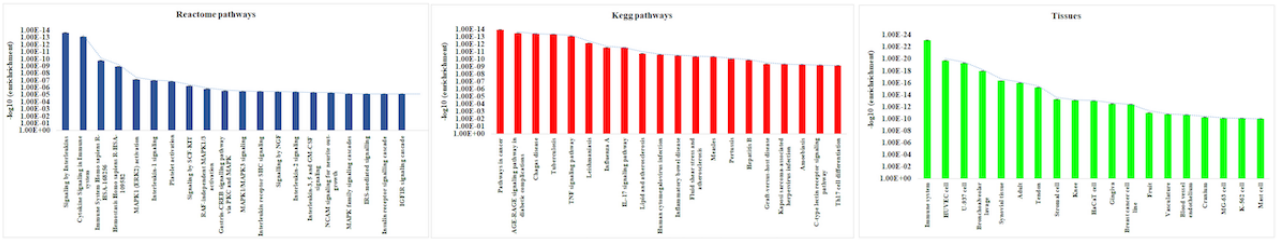


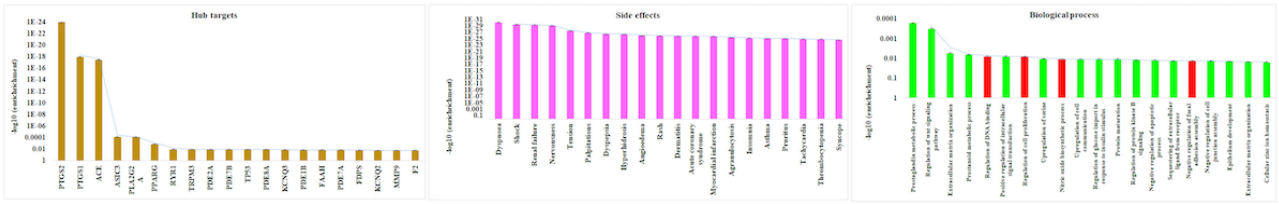
Figure 6. Functional hub gene enrichment analyses: (A) reactome pathways, (B), Kegg pathways, and (C) tissues. Higher-resolution version of this figure is available in Multimedia Appendix 10.



Functional Enrichment Analysis of Drugs Based on Hub Gene Prediction

Hub gene analysis showed a total of 45 repurposable drugs in which the human pathogen gene interacted with the drug target. The hub gene mechanism also showed where the genes were expressed in biological systems and side effects (Figure 7). Flurbiprofen, mefenamic acid, acetylsalicylic acid, indomethacin, naproxen, ketoprofen, acetaminophen, ketorolac, aceclofenac, lenalidomide, diclofenac, suprofen, loxoprofen, and nabumetone targeted the hub gene *PTGS2*. Module dyspnea, shock, renal failure, nervousness, and tension were prominent side effects of these drugs. The prostaglandin metabolic process (P=.00016) and regulation of the Wingless-related integration site signaling pathway (P=.00031 were prominent upregulated gene expression pathways after administration of the above drugs.

Figure 7. Functional hub gene–drug enrichment analyses: (A) hub targets, (B) side effects, and (C) biological process. Higher-resolution version of this figure is available in Multimedia Appendix 11.



Discussion

Principal Findings

We systematically studied the association of dengue viral interactions with the human genome through network-based association analysis. We hypothesized that a host protein that is functionally associated with this virus is localized in a corresponding subnetwork within the comprehensive human interactome network. The host dependency factors mediating virus infection and effective molecular targets should be identified for developing broad-spectrum antiviral drugs and anti-DHF drugs. In our network-based analysis, we identified 45 repurposable drug candidates against DHF, including 13 antivirals targeting human genes and 32 anti-DHF drugs targeting 20 human genes. The most prevalent side effects identified in repurposed drug enrichment were dyspnea and shock. The *PTGS2*, *F2*, and *ACE* genes were highly targeted by the repurposed drugs.

Comparison to Prior Work

The pathogenicity of the *PTGS2* and *COX-2* gene pathways in the progression of DHF has already been reported [27]. Most importantly, the *PTGS2* gene has a direct relationship with severe dengue, in which the blood vessels become damaged and leaky, and the number of clot-forming cells (platelets) in the bloodstream drops. This can lead to shock, internal bleeding, organ failure, and even death [28]. Inhibiting this will help further prevent heart disease and improve the management of DHF. For that, we identified several effective targeted drugs with our repurposing approach, including aceclofenac, acetaminophen, aspirin, choline magnesium trisalicylate, diclofenac, etodolac, epinephrine, indomethacin, ketoprofen, ketorolac, loxoprofen, mefenamic acid, meloxicam, nabumetone, naproxen, phenyl salicylate, suprofen, and lenalidomide. Controversially, aspirin is an antiplatelet drug that prevents clotting of the blood in dengue patients, and it has been noted that high-dose (>1 gram) aspirin was linked to increased bleeding risk, probably because of its permanent antiplatelet effects [29], making it important to accurately decide the dosage while treating the condition. Despite the wide range of increased procoagulant activity during sickness, thrombotic events have not been extensively recorded in dengue, even though hemorrhagic events of various degrees have often been described [30]. Between January and March 2011, there was a localized dengue fever outbreak brought on by dengue virus type 1 and dengue virus type 2 in Brazil; individuals with dengue fever experienced multiple incidences of thrombotic events that impacted large veins [31]. These cases were similar to others reported in different parts of the world. Enrichment analysis emphasizes that most pathways of the immune system are highly enriched, and adult and immune system–associated tissues were associated with an enriched viral and drug response throughout the study. The network-based analysis suggested that 3 drugs have repurposable properties: aspirin, captopril, and rilonacept; however, because of their anticoagulant qualities, patients should be encouraged to be properly hydrated and avoid aspirin (ie, acetylsalicylic acid), aspirin-containing medicines, and other nonsteroidal anti-inflammatory drugs, such as ibuprofen [30]. On the other hand, several studies have reported that thrombosis is associated with dengue fever [32,33]; hence, further studies are essential to prove the efficiency of aspirin and the other repurposed drugs identified in this study for the treatment of DHF.

Limitations of This Study

This study took into account all targets of DHF and sorted them based on scores from databases, but the data set represents a large population size, so it may or may not be generalizable to groups within the population. In the real world, the expressed genes may vary based on ethnicity, heredity, and drug use, so it is essential to test the most expressed genes and their associations with the pathogenesis of DHF. A network-based drug reprofiling approach will be helpful to enable effective personalized medicine, but whole-genome sequencing is still not a cost-effective method compared to traditional treatment methods.

Conclusion

Aspirin is not usually prescribed for dengue fever because of bleeding complications, but it has been reported that using aspirin in lower doses is beneficial in the management of diseases with thrombosis. Drug repurposing is an emerging field in which clinical validation and dosage identification are required before drugs are prescribed. Further retrospective and collaborative international trials are essential for understanding the pathogenesis of this condition.

Acknowledgments

We acknowledge funding from the Ministry of Higher Education of Malaysia as part of the Fundamental Research Grant Scheme (FRGS/1/2020/SKK0/USM/03/16).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Functional genes identified from GeneCards and related literature.
[XLSX File (Microsoft Excel File), 71 KB - bioinform_v4i1e37306_app1.xlsx]

Multimedia Appendix 2
Host-DHF interactome.
[PNG File , 551 KB - bioinform_v4i1e37306_app2.png]

Multimedia Appendix 3
Results of Kegg pathway analysis and reactome pathway analysis.
[PNG File , 152 KB - bioinform_v4i1e37306_app3.png]

Multimedia Appendix 4
Host–DHF–antiviral drug interactome.
[PNG File , 2405 KB - bioinform_v4i1e37306_app4.png]

Multimedia Appendix 5
Kegg pathway gene enrichment analysis.
[PNG File , 150 KB - bioinform_v4i1e37306_app5.png]

Multimedia Appendix 6
Host–DHF interactome–anti–dengue hemorrhagic fever drug interactome.
[PNG File , 2958 KB - bioinform_v4i1e37306_app6.png]

Multimedia Appendix 7
Most enriched gene pathways according to Kegg pathway analysis.
[PNG File , 136 KB - bioinform_v4i1e37306_app7.png]

Multimedia Appendix 8
Higher resolution version of Figure 1. Human-anti–dengue hemorrhagic fever–antiviral drug–anti–dengue hemorrhagic fever interaction network. ADHF: anti–dengue hemorrhagic fever; AV: antiviral; DB: DrugBank; DHF: dengue hemorrhagic fever.
[PNG File , 5069 KB - bioinform_v4i1e37306_app8.png]

Multimedia Appendix 9
Higher resolution version of Figure 2. Enrichment analysis of human-anti–dengue hemorrhagic fever–antiviral drug–anti–dengue hemorrhagic fever. (A) Kegg pathways, (B), reactome pathways, and (C) tissues.
[PNG File , 143 KB - bioinform_v4i1e37306_app9.png]

Multimedia Appendix 10
Higher resolution version of Figure 6. Functional hub gene enrichment analyses: (A) reactome pathways, (B), Kegg pathways, and (C) tissues.
[PNG File , 124 KB - bioinform_v4i1e37306_app10.png]

Multimedia Appendix 11
Higher resolution version of Figure 7. Functional hub gene–drug enrichment analyses: (A) hub targets, (B) side effects, and (C) biological process.
[PNG File , 105 KB - bioinform_v4i1e37306_app11.png]

References

1. Heilman J, De Wolff J, Beards GM, Basden BJ. Dengue fever: a Wikipedia clinical review. Open Med 2014;8(4):e105-e115 [FREE Full text] [Medline: 25426178]

2. Simmons CP, Farrar JJ, Nguyen VVC, Wills B. Dengue. N Engl J Med 2012 Apr 12;366(15):1423-1432. [doi: 10.1056/NEJMra1110265] [Medline: 22494122]
3. Dengue and severe dengue. World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue [accessed 2023-04-05]
4. Din M, Asghar M, Ali M. COVID-19 and dengue coepidemics: A double trouble for overburdened health systems in developing countries. J Med Virol 2021 Feb;93(2):601-602 [FREE Full text] [doi: 10.1002/jmv.26348] [Medline: 32706408]
5. Dengue worldwide overview. European Centre for Disease Prevention and Control. URL: https://www.ecdc.europa.eu/en/dengue-monthly [accessed 2023-04-05]
6. Sutherst RW. Global change and human vulnerability to vector-borne diseases. Clin Microbiol Rev 2004 Jan;17(1):136-173 [FREE Full text] [doi: 10.1128/CMR.17.1.136-173.2004] [Medline: 14726459]
7. Zhou Y, Hou Y, Shen J, Huang Y, Martin W, Cheng F. Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. Cell Discov 2020;6:14 [FREE Full text] [doi: 10.1038/s41421-020-0153-3] [Medline: 32194980]
8. Mohammadi E, Benfeitas R, Turkez H, Boren J, Nielsen J, Uhlen M, et al. Applications of genome-wide screening and systems biology approaches in drug repositioning. Cancers (Basel) 2020 Sep 21;12(9):2694 [FREE Full text] [doi: 10.3390/cancers12092694] [Medline: 32967266]
9. Mohs RC, Greig NH. Drug discovery and development: Role of basic biological research. Alzheimers Dement (N Y) 2017 Nov 10;3(4):651-657 [FREE Full text] [doi: 10.1016/j.trci.2017.10.005] [Medline: 29255791]
10. Roessler HI, Knoers NVAM, van Haelst MM, van Haaften G. Drug repurposing for rare diseases. Trends Pharmacol Sci 2021 Apr;42(4):255-267. [doi: 10.1016/j.tips.2021.01.003] [Medline: 33563480]
11. Mairiang D, Zhang H, Sodja A, Murali T, Suriyaphol P, Malasit P, et al. Identification of new protein interactions between dengue fever virus and its hosts, human and mosquito. PLoS One 2013;8(1):e53535 [FREE Full text] [doi: 10.1371/journal.pone.0053535] [Medline: 23326450]
12. Dey L, Mukhopadhyay A. DenvInt: A database of protein-protein interactions between dengue virus and its hosts. PLoS Negl Trop Dis 2017 Oct;11(10):e0005879 [FREE Full text] [doi: 10.1371/journal.pntd.0005879] [Medline: 29049286]
13. Nicod C, Banaei-Esfahani A, Collins BC. Elucidation of host-pathogen protein-protein interactions to uncover mechanisms of host cell rewiring. Curr Opin Microbiol 2017 Oct;39:7-15 [FREE Full text] [doi: 10.1016/j.mib.2017.07.005] [Medline: 28806587]
14. Khorsand B, Savadi A, Naghibzadeh M. Comprehensive host-pathogen protein-protein interaction network analysis. BMC Bioinformatics 2020 Sep 10;21(1):400 [FREE Full text] [doi: 10.1186/s12859-020-03706-z] [Medline: 32912135]
15. Stelzer G, Rosen N, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, et al. The GeneCards suite: from gene data mining to disease genome sequence analyses. Curr Protoc Bioinformatics 2016 Jun 20;54:1.30.1-1.30.33. [doi: 10.1002/cpbi.5] [Medline: 27322403]
16. Ramos P, Arge LWP, Lima NCB, Fukutani KF, de Queiroz ATL. Leveraging user-friendly network approaches to extract knowledge from high-throughput omics datasets. Front Genet 2019;10:1120 [FREE Full text] [doi: 10.3389/fgene.2019.01120] [Medline: 31798629]
17. Zhou Y, Zhang Y, Lian X, Li F, Wang C, Zhu F, et al. Therapeutic target database update 2022: facilitating drug discovery with enriched comparative data of targeted agents. Nucleic Acids Res 2022 Jan 07;50(D1):D1398-D1407 [FREE Full text] [doi: 10.1093/nar/gkab953] [Medline: 34718717]
18. Wishart D, Knox C, Guo A, Shrivastava S, Hassanali M, Stothard P, et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res 2006 Jan 01;34(Database issue):D668-D672 [FREE Full text] [doi: 10.1093/nar/gkj067] [Medline: 16381955]
19. Shannon P, Markiel A, Ozier O, Baliga N, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003 Nov;13(11):2498-2504 [FREE Full text] [doi: 10.1101/gr.1239303] [Medline: 14597658]
20. Huang L, Law JN, Murali TM. Automating the PathLinker app for Cytoscape. F1000Res 2018;7:727 [FREE Full text] [doi: 10.12688/f1000research.14616.1] [Medline: 30057757]
21. Groshek J, de Mees V, Eschmann R. Modeling influence and community in social media data using the digital methods initiative-twitter capture and analysis toolkit (DMI-TCAT) and Gephi. MethodsX 2020;7:101164 [FREE Full text] [doi: 10.1016/j.mex.2020.101164] [Medline: 33665150]
22. Muetze T, Goenawan IH, Wiencko HL, Bernal-Llinares M, Bryan K, Lynn DJ. Contextual Hub Analysis Tool (CHAT): A Cytoscape app for identifying contextually relevant hubs in biological networks. F1000Res 2016;5:1745 [FREE Full text] [doi: 10.12688/f1000research.9118.2] [Medline: 27853512]
23. Xie Z, Bailey A, Kuleshov MV, Clarke DJB, Evangelista JE, Jenkins SL, et al. Gene set knowledge discovery with Enrichr. Curr Protoc 2021 Mar;1(3):e90 [FREE Full text] [doi: 10.1002/cpz1.90] [Medline: 33780170]
24. Kay A, Simpson CL, Stewart JA. The role of AGE/RAGE signaling in diabetes-mediated vascular calcification. J Diabetes Res 2016;2016:6809703 [FREE Full text] [doi: 10.1155/2016/6809703] [Medline: 27547766]
25. Park DS, Fishman GI. The cardiac conduction system. Circulation 2011 Mar 01;123(8):904-915 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.110.942284] [Medline: 21357845]

26. Grant AO. Cardiac ion channels. Circ Arrhythm Electrophysiol 2009 Apr;2(2):185-194. [doi: 10.1161/CIRCEP.108.789081] [Medline: 19808464]
27. Lin C, Tseng C, Wu Y, Liaw C, Lin C, Huang C, et al. Cyclooxygenase-2 facilitates dengue virus replication and serves as a potential target for developing antiviral agents. Sci Rep 2017 Mar 20;7:44701 [FREE Full text] [doi: 10.1038/srep44701] [Medline: 28317866]
28. Ruan C, So S, Ruan K. Inducible COX-2 dominates over COX-1 in prostacyclin biosynthesis: mechanisms of COX-2 inhibitor risk to heart disease. Life Sci 2011 Jan 03;88(1-2):24-30 [FREE Full text] [doi: 10.1016/j.lfs.2010.10.017] [Medline: 21035466]
29. Kellstein D, Fernandes L. Symptomatic treatment of dengue: should the NSAID contraindication be reconsidered? Postgrad Med 2019 Mar;131(2):109-116. [doi: 10.1080/00325481.2019.1561916] [Medline: 30575425]
30. Dengue: Treatment. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/dengue/healthcare-providers/treatment.html [accessed 2023-04-05]
31. Cavalcanti LPDG, Coelho ICB, Vilar DCLF, Holanda SGS, Escóssia KNFD, Souza-Santos R. Clinical and epidemiological characterization of dengue hemorrhagic fever cases in northeastern, Brazil. Rev Soc Bras Med Trop 2010;43(4):355-358 [FREE Full text] [doi: 10.1590/s0037-86822010000400003] [Medline: 20802929]
32. da Costa PSG, Ribeiro GM, Junior CS, da Costa Campos L. Severe thrombotic events associated with dengue fever, Brazil. Am J Trop Med Hyg 2012 Oct;87(4):741-742 [FREE Full text] [doi: 10.4269/ajtmh.2012.11-0692] [Medline: 22949517]
33. Ranasinghe K, Dissanayaka D, Thirumavalavan K, Seneviratne M. An unusual case of dengue shock syndrome complicated by ilio-femoral deep vein thrombosis; a case report. BMC Infect Dis 2020 May 12;20(1):335 [FREE Full text] [doi: 10.1186/s12879-020-05062-y] [Medline: 32398134]

Abbreviations

***ACE*:** angiotensin converting enzyme
AGE-RAGE: advanced glycation end products–receptor for advanced glycation end products
CI: credible interval
DHF: dengue hemorrhagic fever
***F2*:** coagulation factor II, thrombin
FDA: Food and Drug Administration
HP: host-pathogen
PPI: protein-protein interaction
***PTGS2*:** prostaglandin-endoperoxide synthase 2
WHO: World Health Organization

Edited by T Leung; submitted 15.02.22; peer-reviewed by M Khokhar, D Pessoa, B Foroutan; comments to author 01.05.22; revised version received 30.09.22; accepted 28.03.23; published 09.05.23.

Please cite as:
Kochuthakidiyel Suresh P, Sekar G, Mallady K, Wan Ab Rahman WS, Shima Shahidan WN, Venkatesan G
The Identification of Potential Drugs for Dengue Hemorrhagic Fever: Network-Based Drug Reprofiling Study
JMIR Bioinform Biotech 2023;4:e37306
URL: https://bioinform.jmir.org/2023/1/e37306
doi:10.2196/37306
PMID:

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org