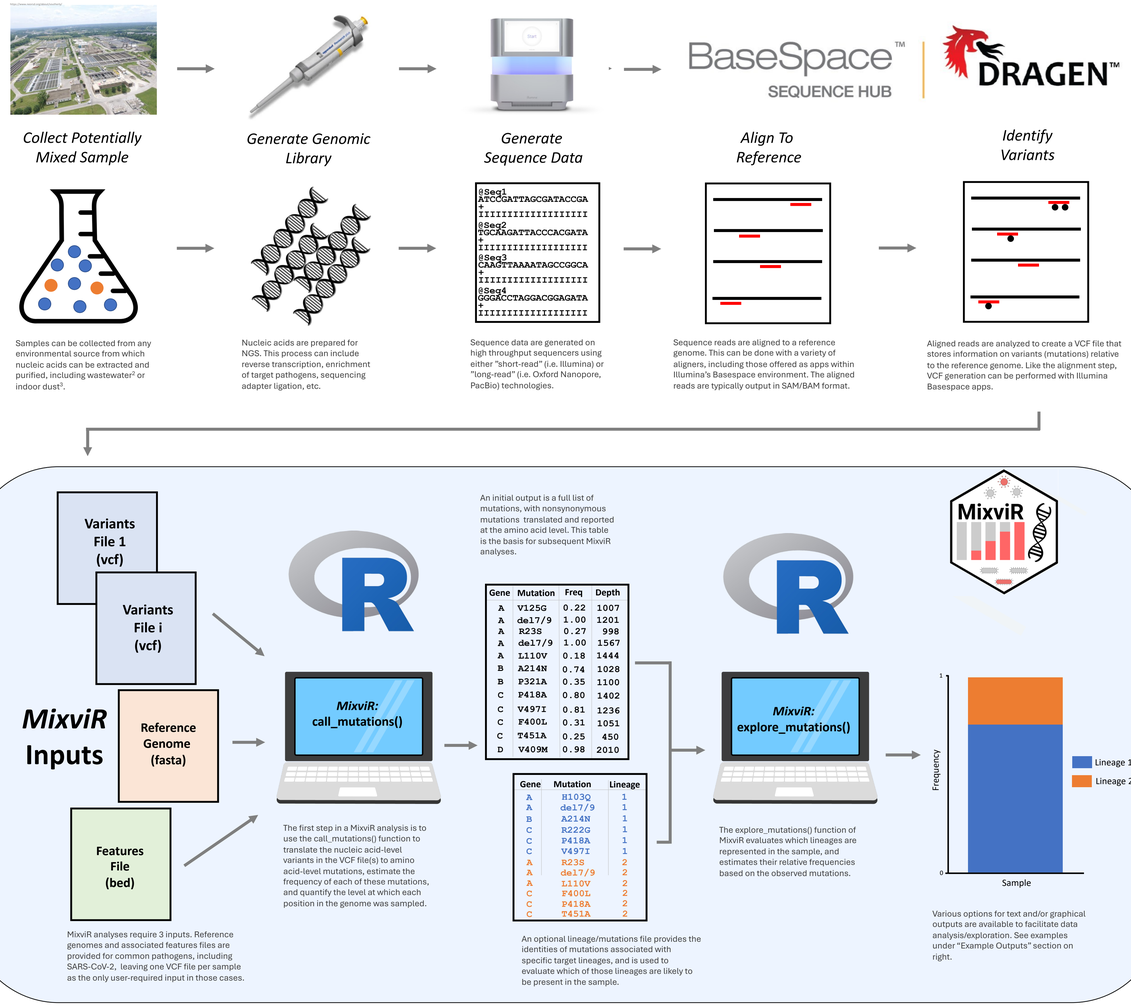
MixviR: A User-Friendly Computational Tool For Exploring Genomic Data From Environmental Samples Containing Mixed Pathogen Lineages

Michael G. Sovic, PhD; Freedom Genomics LLC, Pickerington, OH; www.freegenllc.com
Zuzana Bohrer, PhD; Ohio Department of Health, Bureau of Public Health Laboratory, Ohio Department of Health Laboratory (ODHL), Reynoldsburg, OH

Abstract

Advancing technologies for high throughput nucleic acid sequencing provide powerful tools to monitor the presence and evolution of environmental pathogens. Application of such tools in wastewater systems holds great promise for efficient surveillance of pathogen dynamics at the community level, as was demonstrated during the SARS-CoV-2 pandemic. Ongoing methods development that underlies both the growing scale of datasets and the expanding number of specific pathogens targeted for study necessitates parallel development of new and diverse computational tools to analyze the data. As part of SARS-CoV-2 response efforts, we developed MixviR[1], an open-source, user-friendly bioinformatic analysis package written in R for rapid analysis, visualization, and exploration of genome-scale sequence data from environmental sources, including wastewater, where a mixture of various lineages may be present. MixviR provides methods for identifying such lineages and estimating their relative proportions from the sample. The program was written with a goal of making it readily accessible and usable by public health professionals, and since the original publication, we continue to update MixviR, emphasizing ease of use and an expanded scope of pathogens it can be applied to. This presentation provides an overview of the process of analyzing data with MixviR, and includes highlights of recent updates to the program.

MixviR Analysis Workflow



Genomic and Bioinformatics Jargon

Variant: A position that differs when comparing two or more aligned DNA sequences. Variants represent the outcome of mutation, but the term is also often used synonymously with "mutation". Note relationship to "Lineage" below.
SNP: Single nucleotide polymorphism; a common type of variant in which two individual nucleotides differ from each other (i.e. A vs T).
Indel: Short for insertion/deletion; a type of variant in which one or more nucleotides have been inserted or deleted from one DNA sequence relative to another.
Lineage: A group of individuals connected through a continuous line of descent. Individuals within a lineage share a unique set of genetic variants as a result of their common ancestry, and typically represent a large number of generations. Note that in practice, the term "variant" is also sometimes used synonymously with this definition of "lineage".
Genomics: Study of the complete, or nearly complete, set of an individual's hereditary information. Genomics differs from genetics largely with regards to the scale of the data generated, and has been facilitated by evolving NGS technologies.
NGS: Also referred to as high-throughput sequencing, second-generation sequencing, massively parallel sequencing, or "Illumina" sequencing (Illumina makes the instruments the majority of NGS studies are currently performed on). NGS technologies have revolutionized the natural sciences, allowing generation of genome-scale datasets that consist of up to billions of individual sequence reads in just days.
Basespace: A user-friendly, online (cloud-based) platform for bioinformatic analysis and data storage associated with Illumina sequencing instruments.
VCF: Variant Call Format; a specified format for text files that store genetic variants identified in samples.
BED: A specified format for text files that store information on relative positions of genes, or other features of interest, within a reference genome. The features are specified by a combination of a sequence/chromosome name and the beginning and ending coordinates/positions within the reference sequence.
SAM: Specified file format for text files storing sequence data after alignment to a reference. Compressed version is BAM.

Example Outputs

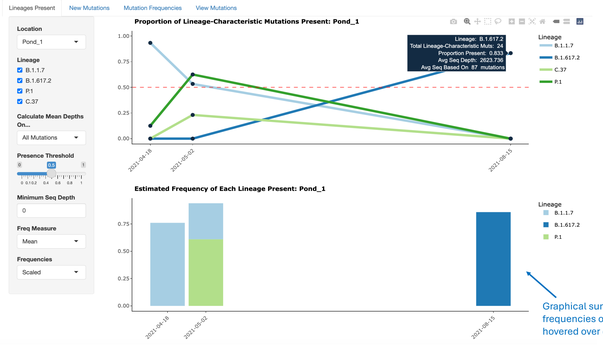
Lineages Table

The lineages table is written to a text file (csv format) that contains lineages and sublineages identified for each sample, along with estimated frequencies and information on mutations present and absent for each.

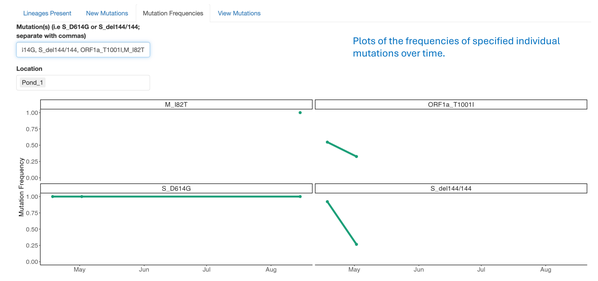
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Sample	Location	Date	Lineage	Num_Target_Muts	Proportion_Targets_Present	Estimated_Freq	Mean_Depth	Target_Muts_Present	Target_Muts_Absent	Sublineages_Identified	Sub_Proportion_Targets_Present	Sub_Estimated_Freq	Sub_Target_Muts_Present	
2	Sample1	Pond_1	4/18/21	B.1.1.7	15	0.933333333	0.760033188	2403.857143	N_D3L;ORF1a_T1001I;S_A570D;S_D1118H;(ORF1a_I2230T		NA	NA	NA	NA	
3	Sample2	Pond_1	5/2/21	B.1.1.7	15	0.533333333	0.329557729	2683.973684	ORF1a_T1001I;S_A570D;ORF8_Y73C;ORF8_ N_D3L;S_D1118H;ORF1a_A1708D		NA	NA	NA	NA	
4	Sample2	Pond_1	5/2/21	P.1	16	0.625	0.608661805	2683.973684	ORF1a_K1795Q,N_P80R;ORF3a_S253P;S_HI ORF1b_E1264D;S_P26S;S_T20N;S		NA	NA	NA	NA	
5	Sample3	Pond_1	8/15/21	B.1.617.2	24	0.833333333	0.857675897	2871.2	ORF3a_S26L;M_I82T;ORF1b_P1000L;N_R20 S_E156G;S_P681R;ORF1a_T3646A	AY.4.2		0.5	0.024	S_A222V	
6															

Lineage identified from the sample
# of total characteristic mutations for the target lineage
Proportion of the characteristic mutations identified in the sample
Estimated relative frequency of the lineage in the sample
Identities of characteristic mutations observed in the sample
Identities of characteristic mutations not observed in the sample
Sublineages identified in the sample
Proportion of characteristic sublineage mutations observed in the sample
Estimated relative frequency of the sublineage in the sample
Identities of characteristic sublineage mutations observed in the sample

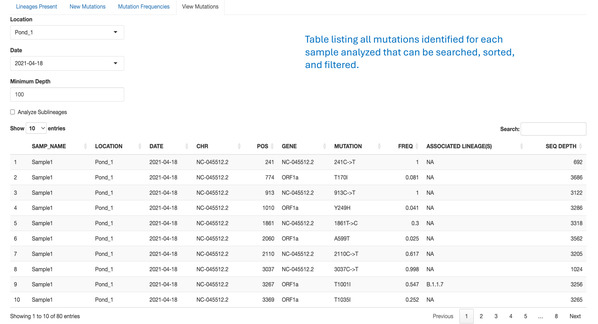
Lineages Present Tab



Mutation Frequencies Tab



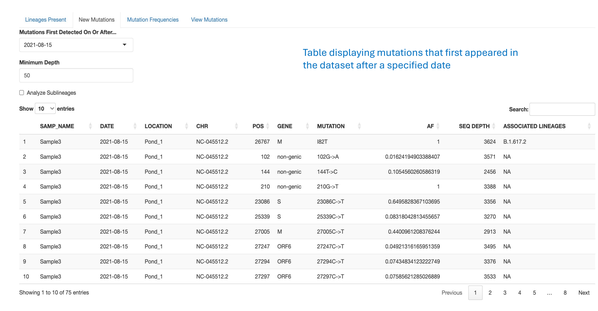
View Mutations Tab



	SAMP_NAME	LOCATION	DATE	CHR	POS	GENE	MUTATION	FREQ	ASSOCIATED LINEAGE(S)	SEQ DEPTH
1	Sample1	Pond_1	2021-04-18	NC-045512.2	241	NC-045512.2	241C->T	1	NA	692
2	Sample1	Pond_1	2021-04-18	NC-045512.2	774	ORF1a	T170I	0.081	NA	3686
3	Sample1	Pond_1	2021-04-18	NC-045512.2	913	NC-045512.2	913C->T	1	NA	3122
4	Sample1	Pond_1	2021-04-18	NC-045512.2	1010	ORF1a	Y249H	0.041	NA	3286
5	Sample1	Pond_1	2021-04-18	NC-045512.2	1861	NC-045512.2	1861T->C	0.3	NA	3318
6	Sample1	Pond_1	2021-04-18	NC-045512.2	2060	ORF1a	A599T	0.025	NA	3562
7	Sample1	Pond_1	2021-04-18	NC-045512.2	2110	NC-045512.2	2110C->T	0.617	NA	3205
8	Sample1	Pond_1	2021-04-18	NC-045512.2	3037	NC-045512.2	3037C->T	0.998	NA	1024
9	Sample1	Pond_1	2021-04-18	NC-045512.2	3267	ORF1a	T1001I	0.547	B.1.1.7	3256
10	Sample1	Pond_1	2021-04-18	NC-045512.2	3369	ORF1a	T1035I	0.252	NA	3265

Showing 1 to 10 of 80 entries

New Mutations Tab



	SAMP_NAME	DATE	LOCATION	CHR	POS	GENE	MUTATION	AF	SEQ DEPTH	ASSOCIATED LINEAGES
1	Sample3	2021-08-15	Pond_1	NC-045512.2	26767	M	I82T	1	3624	B.1.617.2
2	Sample3	2021-08-15	Pond_1	NC-045512.2	102	non-genic	102G->A	0.016241949033884O7	3571	NA
3	Sample3	2021-08-15	Pond_1	NC-045512.2	144	non-genic	144T>C	0.10545602805863 19	2456	NA
4	Sample3	2021-08-15	Pond_1	NC-045512.2	210	non-genic	210G->T	1	3388	NA
5	Sample3	2021-08-15	Pond_1	NC-045512.2	23086	S	23086C->T	0.6495828367103695	3356	NA
6	Sample3	2021-08-15	Pond_1	NC-045512.2	25339	S	25339C->T	0.0631804281345565 7	3270	NA
7	Sample3	2021-08-15	Pond_1	NC-045512.2	27005	M	27005C->T	0.4400961208376244	2913	NA
8	Sample3	2021-08-15	Pond_1	NC-045512.2	27247	ORF6	27247C->T	0.0492131616991359	3495	NA
9	Sample3	2021-08-15	Pond_1	NC-045512.2	27294	ORF6	27294C->T	0.0743483412322274 9	3376	NA
10	Sample3	2021-08-15	Pond_1	NC-045512.2	27297	ORF6	27297C->T	0.0758562128502688 9	3533	NA

Showing 1 to 10 of 75 entries

What's New/What's Coming

- Option to explicitly analyze sublineages in addition to main lineages available as of MixviR v. 3.5.0 (released late 2022)
- More flexibility in VCF input file format requirements coming in next version release, including option to analyze datasets based only on the identity of mutations (no depth or mutation frequency information)
- Updates and validation for pathogens with segmented genomes in progress
- Potential integration into Illumina Basespace for seamless workflow from sequencing to results.
- **Input welcome for additional features/updates!!**

Check out an online tutorial/demo for MixviR!

References

[1]Sovic, Michael G., Francesca Savona, Zuzana Bohrerova, and Seth A. Faith. "MixviR: an R package for exploring variation associated with genomic sequence data from environmental SARS-CoV-2 and other mixed microbial samples." *Applied and Environmental Microbiology* 88, no. 22 (2022): e00874-22.
[2]Bohrerova, Zuzana, Nichole E. Brinkman, Ritu Chakravarti, Saurabh Chattopadhyay, Seth A. Faith, Jay Garland, James Herrin et al. "Ohio Coronavirus Wastewater Monitoring Network: implementation of statewide monitoring for protecting public health." *Journal of Public Health Management and Practice* 29, no. 6 (2023): 845-853.
[3]Van Dusen, John, Haley LeBlanc, Nicholas Nastasi, Jenny Panescu, Austin Shamblin, Jacob W. Smith, Michael G. Sovic et al. "Identification of SARS-CoV-2 variants in indoor dust." *Plos one* 19, no. 2 (2024): e0297172.

Contact Info

Mike Sovic
Freedom Genomics
www.freegenllc.com
labservices@freegenllc.com
614-282-8600