Streamlined Identification and Characterization of Viral Strains

Overcoming the top 9 bioinformatics challenges in the time of COVID-19

DNASTAR

ABOUT DNASTAR

DNASTAR is a pioneer in the field of bioinformatics, offering comprehensive software solutions for molecular biology, genomics, transcriptomics, and protein analysis. Our Lasergene Genomics software enables you to set up complex genomic sequencing projects in mere minutes and automates tasks that typically require extensive manual intervention in other software packages.

CONTENTS

Introduction 4

Chapter 1: General Procedure for Indentifying and Characterizing Viral Strains 5

- **Step 1:** Decide assembly approach 7
- **Step 2:** Choose technology and submit samples 7
- **Step 3:** Assemble raw reads 8
- **Step 4:** Compare draft genomes 9
- **Example:** Identifying the strain of an experimental virus sample using a phylogenetic tree 10

Chapter 2: Software Challenges Faced When Identifying Viral Strains 12

Chapter 3: Case Study: Rapid Analysis of Viral Strains During the COVID-19 Pandemic 20

Conclusion 22

Introduction

Viral genome information can be used to investigate viral pathogenesis and evolution as well as assisting in vaccine development and the study of vaccine resistance. Viral mutations can now be explored in detail by characterizing viral genomes from different hosts or even within a single individual. For example, thousands of SARS-CoV-2 variants have been discovered in the past two years, including several "variants of concern" that have been identified by the World Health Organization at the time of this ebook's publication.

The identification and characterization of viral genome strains is now a routine task thanks to advancements in read sequencing technology and sequence assembly and analysis software. Thousands of viral genomes have been published, with more being added daily.

But while the task of viral genome analysis may be routine, many researchers find it frustrating and time-consuming. DNASTAR recently surveyed over 100 virologists and genomics researchers about the challenges they face in identifying and characterizing viral samples. They pointed out several shortcomings in most assembly and analysis software as a significant barrier to easily performing this workflow.

This ebook will guide you through the process of viral genome sequencing and characterization: from choosing a sequencing strategy for templated or *de novo* assembly, through downstream analysis. It will also show the results from our survey, including a discussion of the top challenges faced by virologists performing this important work.

Chapter 1

General Procedure for Identifying and Characterizing Viral Strains

What is the general procedure for identifying and characterizing viral strains?

There are four main steps involved in identifying and characterizing viral strains. The first step happens even before you send your samples to be sequenced.

1. Decide whether you will be assembling your experimental sequence reads to a reference or assembling *de novo*
2. Choose the sequencing technology and submit your samples
3. Assemble raw reads into a set of draft genomes
4. Compare the experimental draft genomes to one another and/or to reference sequences

Step 1: Decide whether you will be assembling your experimental sequence reads to a reference or assembling *de novo*

Before deciding on a sequencing technology (Step 2), it is important to know whether you will be assembling your sequences *de novo* or against a reference sequence.

If you are working with a novel virus, or if your virus has a rapid mutation rate or large-scale changes (as seen with segmented reassortment), a closely matching reference genome will not be available. For these viruses, there is no choice but to do *de novo* assembly.

For more stable viral genomes, where the mutation rate is relatively slow, a genome reference sequence from the same (or closely-related) strain can be used to guide alignment of the sequencing reads. If a reference sequence is available for your virus of interest, we highly recommend that you use it. Templated assembly helps order contigs into larger scaffolds, giving you a more complete genomic sequence, and allows for both faster and more accurate assembly. These assemblies are ideal for monitoring the evolution of viral strains such as SARS-CoV-2. Downstream analysis typically focuses on identifying variations that define a viral strain or affect viral proteins.

Step 2: Choose the sequencing technology and submit your samples

If you will be doing a reference-guided assembly, Sanger or Illumina sequencing technologies are perfectly suitable for viral genome analysis. These technologies create shorter reads than some of the more recent technologies but are still highly accurate.

If you will be doing *de novo* assembly, the sequencing technology you choose will depend on the length of the viral genome you are studying. Excellent *de novo* assemblies for viral genomes up to 10Kbp in length can be produced with Sanger and Illumina 2X300bp MiSeq reads. However, accurate *de novo* assembly of long and/or more complex genomes like SARS-COV-2 (30Kbp) or influenza (segmented genome) is best accomplished using long reads.

Long-read sequence data provides many advantages to earlier sequencing types. At the time of this ebook's publication, there are three main long-read sequencing technologies: Oxford Nanopore Technologies (ONT/Nanopore), Pacific Biosciences (PacBio) CLR, and PacBio HiFi.

PacBio HiFi sequencing, in particular, produces long and accurate (>99.9%) sequence reads that work well for both reference-guided and *de novo* assembly. PacBio HiFi reads are produced via PacBio's Sequel HiFi sequencing platform using a hybrid short and long read sequencer mode called "circular consensus sequencing" (CSS). PacBio HiFi reads are shorter but more accurate than long-reads produced using the other two long read sequencing technologies. Due to this high level of accuracy, viral genome data sequenced using PacBio HiFi technology often assembles into a single contig.

Once you've decided on a sequencing technology, sequence your experimental samples in-house or submit them to a core facility. If you have chosen a long read technology, note that most small sequencing facilities do not have access to long-read sequencing platforms. To obtain reads in these formats, you may need to send your samples to a large outside facility.

Depending on the facility, you will get back either raw sequence reads or draft genomes.

If you will be starting with raw sequence reads, proceed to Step 3, where you will assemble those reads into draft genomes. If you already have draft genomes, instead go to Step 4, where you will analyze evolutionary relationships and find sequences and variants of interest.

Step 3: Assemble raw reads into a set of draft genomes

If you are starting with raw sequencing reads, you will need to use software to assemble them into a set of draft genomes. A couple of decades ago, this step could take up to an hour. Today, if you are using long reads that span the entire viral genome, modern software can complete your assembly in as little as a few seconds.

For *de novo* assemblies, you won't need anything other than the assembly software and sequence read file(s), typically in FASTQ format. The assembly software will produce one or more contigs that represent your viral draft genome. We recommend choosing software that allows you to visualize and edit these contigs in order to produce the most accurate draft genome. This draft genome can then be exported to FASTA format and compared to other viral strains in Step 4.

NCBI

If you have a closely related reference genome, you can use that to guide assembly of your raw sequencing data. The National Center for Biotechnology Information (NCBI) is an excellent source of reference genomes that can be downloaded for free. Click here to visit NCBI's nucleotide reference genome search page. Variants between the sequenced strain and reference genome can then be detected and written to a variant report or VCF file for further analysis in Step 4.

Step 4: Compare the experimental draft genomes to one another and/or to reference sequences

If your experimental draft genomes were assembled using a reference, you will likely be interested in finding important variants. Look for software that supports filtering so you can limit found variants to only the SNPs and indels of interest. The ideal software should accommodate common file formats for variant calls, like VCF, and should also allow you to easily compare variants across multiple samples using specific filtering criteria.

Multiple sequence alignment can be used to compare draft genomes to one another or to related reference sequences. When aligning large numbers of viral genomes, there are many algorithms available. However, MAFFT7 is considered to be superior for large numbers of viral genomes due to its accuracy and high capacity.

		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
A	HPV1_L1 - 1,509 bases		56.69	56.85	59.92	59.35	58.47	58.15	58.47	60.40	57.74	59.11	59.92	57.58	60.81	61.05	60.73	59.27	58.31	60.73	60.40	60.73	59.84	59.44	60.08	60.40	58.71	55.81
B	HPV2_L1 - 1,485 bases	0.43		68.63	57.10	57.58	67.34	64.52	59.03	57.42	68.47	67.10	57.42	65.24	58.55	56.69	66.61	57.82	68.15	58.23	57.98	56.77	56.29	56.94	56.94	58.71	64.03	90.16
C	HPV3_L1 - 1,515 bases	0.43	0.31		58.31	58.79	67.18	66.37	58.95	59.84	82.66	68.79	59.60	67.58	59.44	58.15	68.31	58.31	70.97	58.79	59.84	58.15	57.50	58.79	58.63	59.60	69.03	68.47
D	HPV4_L1 - 1,551 bases	0.40	0.43	0.42		65.08	61.45	61.21	62.10	63.06	58.79	60.89	63.06	60.65	65.97	62.82	62.34	61.85	62.34	65.56	64.92	65.32	61.13	61.94	64.19	65.16	60.08	57.58
E	HPV5_L1 - 1,551 bases	0.41	0.42	0.41	0.35		60.89	61.29	80.48	72.74	59.35	60.00	82.34	60.16	77.66	71.61	60.56	71.05	61.29	78.23	77.74	78.15	69.52	69.35	77.02	76.37	60.65	57.50
F	HPV6_L1 - 1,503 bases	0.42	0.33	0.33	0.39	0.39		68.79	60.48	59.19	66.85	85.73	60.00	77.58	60.00	60.08	70.81	60.00	68.47	59.84	60.32	59.35	58.63	59.03	59.19	60.73	69.35	67.18
G	HPV7_L1 - 1,518 bases	0.42	0.35	0.34	0.39	0.39	0.31		59.52	60.97	66.45	69.92	61.21	70.08	61.69	61.69	71.53	61.61	69.35	60.81	61.85	59.60	60.32	60.24	60.89	61.21	68.31	64.03
H	HPV8_L1 - 1,545 bases	0.42	0.41	0.41	0.38	0.20	0.40	0.40		70.24	60.97	59.68	81.77	60.08	77.74	70.81	61.61	70.48	59.84	77.58	76.21	76.85	68.31	69.84	76.85	76.21	60.40	59.11
I	HPV9_L1 - 1,524 bases	0.40	0.43	0.40	0.37	0.27	0.41	0.39	0.30		59.11	59.68	72.02	60.65	72.74	78.95	61.45	78.95	61.29	72.26	72.82	73.31	73.55	74.84	71.85	73.15	60.48	56.21
J	HPV10_L1 - 1,512 bases	0.42	0.32	0.17	0.41	0.41	0.33	0.34	0.39	0.41		67.42	59.11	67.50	60.65	57.02	68.95	57.82	70.48	59.35	61.13	60.00	57.90	58.95	59.19	59.84	69.84	68.15
K	HPV11_L1 - 1,506 bases	0.41	0.33	0.31	0.39	0.40	0.14	0.30	0.40	0.40	0.33		61.21	77.82	61.05	60.08	71.61	60.48	69.19	60.40	61.61	59.27	59.68	60.16	59.76	60.56	67.58	66.85
L	HPV12_L1 - 1,554 bases	0.40	0.43	0.40	0.37	0.18	0.40	0.39	0.18	0.28	0.41	0.39		61.13	78.31	70.81	61.85	70.32	61.53	79.11	78.23	76.77	67.82	68.87	76.94	77.34	59.76	58.79
M	HPV13_L1 - 1,500 bases	0.42	0.35	0.32	0.39	0.40	0.22	0.30	0.40	0.39	0.32	0.22	0.39		61.29	61.13	71.77	61.77	69.60	59.84	61.21	61.29	60.97	60.89	59.52	59.19	68.63	64.92
N	HPV14_L1 - 1,557 bases	0.39	0.41	0.41	0.34	0.22	0.40	0.38	0.22	0.27	0.39	0.39	0.22	0.39		72.02	64.11	72.58	62.82	84.27	86.05	86.85	70.65	71.94	77.98	84.19	61.21	58.79
O	HPV15_L1 - 1,524 bases	0.39	0.43	0.42	0.37	0.28	0.40	0.38	0.29	0.21	0.43	0.40	0.29	0.39	0.28		63.06	80.81	60.65	72.10	71.53	70.97	75.32	75.32	70.97	71.21	60.16	56.53
P	HPV16_L1 - 1,518 bases	0.39	0.33	0.32	0.38	0.39	0.29	0.28	0.38	0.39	0.31	0.28	0.38	0.28	0.36	0.37		61.37	70.24	62.82	63.39	60.97	61.45	61.45	61.69	63.87	70.16	67.02
Q	HPV17_L1 - 1,524 bases	0.41	0.42	0.42	0.38	0.29	0.40	0.38	0.30	0.21	0.42	0.40	0.30	0.38	0.27	0.19	0.39		61.85	71.77	71.13	71.29	75.24	75.08	70.73	70.73	59.03	57.98
R	HPV18_L1 - 1,524 bases	0.42	0.32	0.29	0.38	0.39	0.32	0.31	0.40	0.39	0.30	0.31	0.38	0.30	0.37	0.39	0.30	0.38		61.85	62.74	60.40	60.40	60.97	61.53	62.42	71.29	67.26
S	HPV19_L1 - 1,554 bases	0.39	0.42	0.41	0.34	0.22	0.40	0.39	0.22	0.28	0.41	0.40	0.21	0.40	0.16	0.28	0.37	0.28	0.38		83.79	83.95	69.92	69.92	77.50	86.37	60.65	59.19
T	HPV20_L1 - 1,551 bases	0.40	0.42	0.40	0.35	0.22	0.40	0.38	0.24	0.27	0.39	0.38	0.22	0.39	0.14	0.28	0.37	0.29	0.37	0.16		86.21	70.40	70.40	79.27	83.71	61.77	57.90
U	HPV21_L1 - 1,557 bases	0.39	0.43	0.42	0.35	0.22	0.41	0.40	0.23	0.27	0.40	0.41	0.23	0.39	0.13	0.29	0.39	0.29	0.40	0.16	0.14		69.68	71.05	78.39	83.06	60.48	57.66
V	HPV22_L1 - 1,533 bases	0.40	0.44	0.43	0.39	0.30	0.41	0.40	0.32	0.26	0.42	0.40	0.32	0.39	0.29	0.25	0.39	0.25	0.40	0.30	0.30	0.30		78.79	70.16	69.27	60.81	56.21
W	HPV23_L1 - 1,521 bases	0.41	0.43	0.41	0.38	0.31	0.41	0.40	0.30	0.25	0.41	0.40	0.31	0.39	0.28	0.25	0.39	0.25	0.39	0.30	0.30	0.29	0.21		70.48	69.92	60.48	55.97
X	HPV24_L1 - 1,539 bases	0.40	0.43	0.41	0.36	0.23	0.41	0.39	0.23	0.28	0.41	0.40	0.23	0.40	0.22	0.29	0.38	0.29	0.38	0.22	0.21	0.22	0.30	0.30		77.18	60.24	57.82
Y	HPV25_L1 - 1,554 bases	0.40	0.41	0.40	0.35	0.24	0.39	0.39	0.24	0.27	0.40	0.39	0.23	0.41	0.16	0.29	0.36	0.29	0.38	0.14	0.16	0.17	0.31	0.30	0.23		61.13	59.27
Z	HPV26_L1 - 1,512 bases	0.41	0.36	0.31	0.40	0.39	0.31	0.32	0.40	0.40	0.30	0.32	0.40	0.31	0.39	0.40	0.30	0.41	0.29	0.39	0.38	0.40	0.39	0.40	0.40	0.39		64.84
AA	HPV27_L1 - 1,458 bases	0.44	0.10	0.32	0.42	0.43	0.33	0.36	0.41	0.44	0.32	0.33	0.41	0.35	0.41	0.43	0.33	0.42	0.33	0.41	0.42	0.42	0.44	0.44	0.42	0.41	0.35	
AB	HPV28_L1 - 1,509 bases	0.43	0.33	0.13	0.42	0.42	0.34	0.34	0.42	0.42	0.16	0.33	0.41	0.33	0.40	0.43	0.31	0.41	0.30	0.42	0.41	0.43	0.42	0.42	0.43	0.42	0.31	0.32
AC	HPV29_L1 - 1,512 bases	0.41	0.31	0.22	0.44	0.41	0.33	0.33	0.40	0.38	0.19	0.32	0.40	0.32	0.40	0.42	0.32	0.40	0.30	0.40	0.40	0.41	0.42	0.42	0.41	0.40	0.30	0.32
AD	HPV30_L1 - 1,506 bases	0.41	0.33	0.30	0.38	0.39	0.34	0.33	0.39	0.39	0.30	0.32	0.38	0.32	0.39	0.38	0.30	0.39	0.29	0.40	0.38	0.40	0.40	0.39	0.39	0.39	0.31	0.35
AE	HPV31_L1 - 1,455 bases	0.39	0.33	0.30	0.39	0.38	0.27	0.28	0.39	0.37	0.31	0.28	0.38	0.28	0.37	0.38	0.21	0.38	0.30	0.39	0.37	0.38	0.38	0.39	0.38	0.38	0.29	0.32

Figure 1. Section of a MegAlign Pro distance table for 10,000 viral samples aligned with MAFFT.

After multiple sequence alignment, the results can be viewed in the form of a distance table (**Figure 1**) or a phylogenetic tree. If you want to home in on the differences between two strains, some software lets you align just those two sequences using an algorithm optimized for pairwise comparison.

See the box on the following page for one example of how to use multiple sequence alignment to identify a viral strain of SARS-CoV-2 in an experimental subject.

Example: Identifying the strain of an experimental virus sample using a phylogenetic tree

This example is taken from a DNASTAR tutorial using MegAlign Pro multiple sequence assembly software.

First, an experimental viral sequence (*SRR13380669_NC_045512.2*) from a subject with an unknown strain of SARS-CoV-2 was uploaded to a multiple sequence alignment application. Reference sequences from NCBI for four known strains of SARS-CoV-2 were also uploaded (see table).

SARS-CoV-2 Reference Group	Strain Name
B.1.1.7	Alpha
B.1.351	Beta
B.1.617.2	Delta
C.37	Lambda

Next, the MAFFT algorithm was chosen, and a multiple sequence alignment was performed using all the uploaded sequences.

A phylogenetic tree was automatically generated from the resulting alignment (**Figure 2**). In the image below, the experimental sample *SRR13380669_NC_045512.2* is shown as part of the clade with a green background, B-1-351, or the “Beta” strain or SARS-CoV-2.

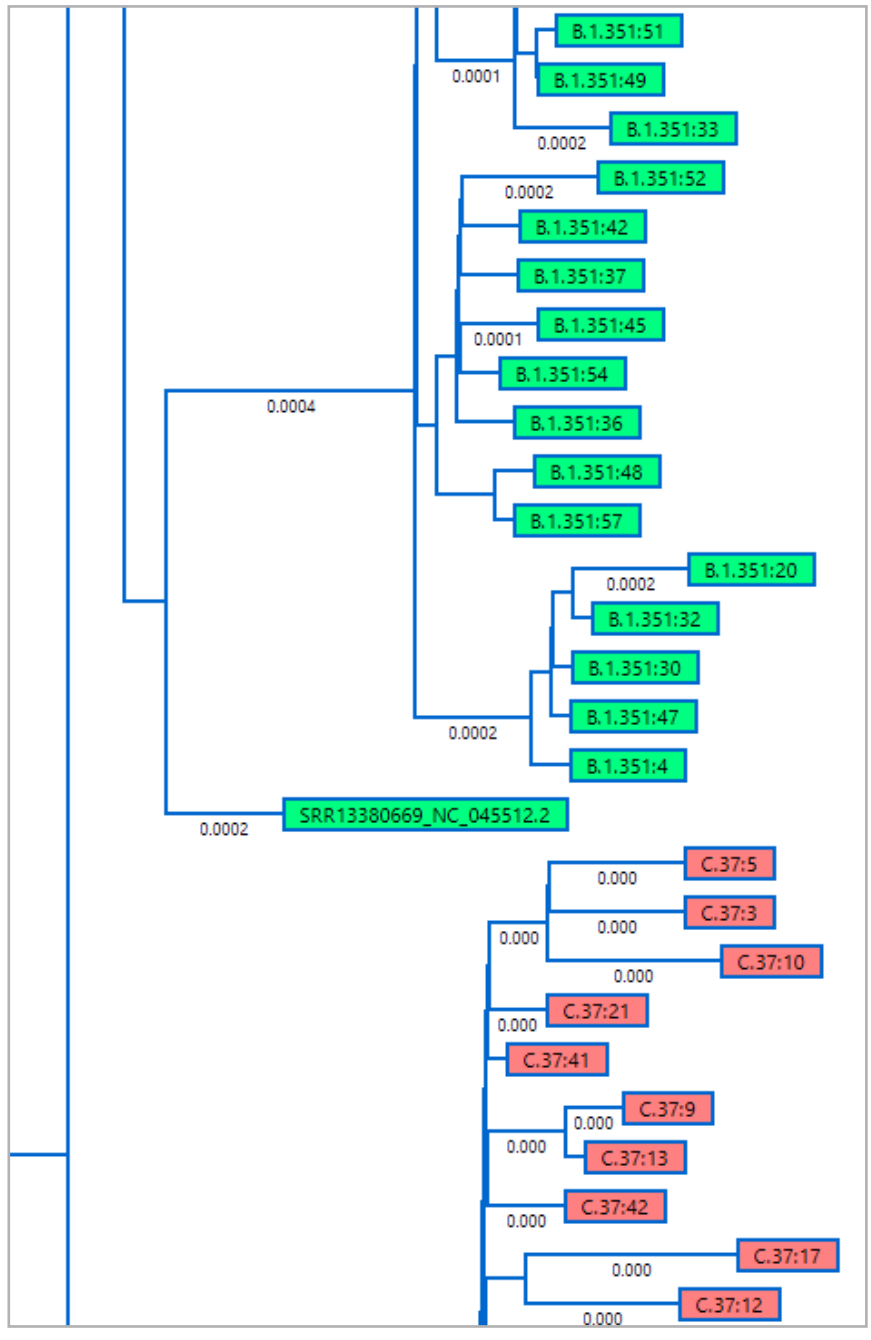
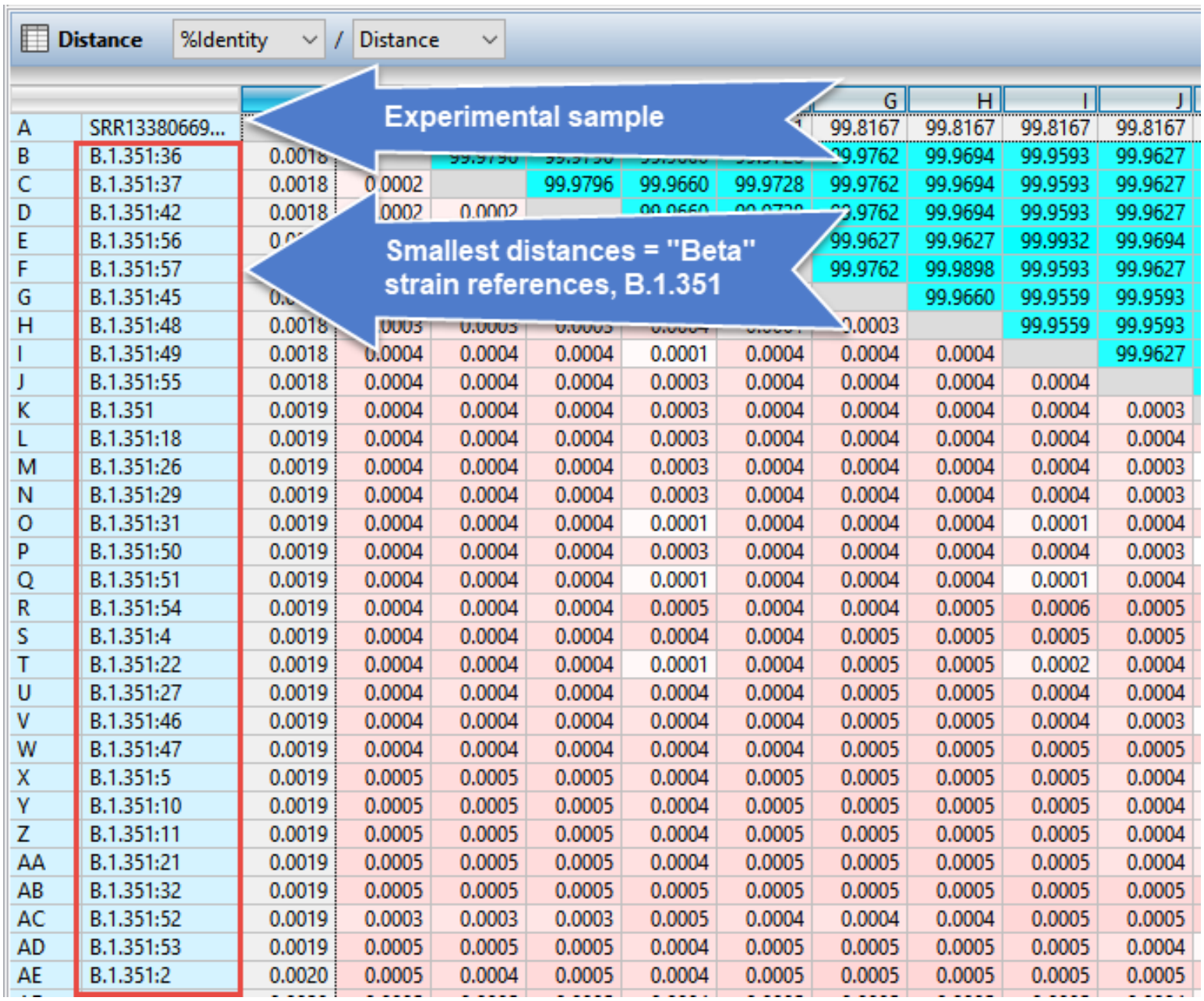


Figure 2. Portion of the MegAlign Pro phylogenetic tree showing placement of the experimental sample in the “Beta” strain clade.

As verification, the distance table for the alignment was automatically ordered from most- to least-related to the experimental sample (**Figure 3**). The most-related samples were all from the “Beta” strain, offering further confirmation that SRR13380669_NC-045512.2 was indeed from that strain.



							G	H	I	J	
A	SRR13380669...						99.8167	99.8167	99.8167	99.8167	
B	B.1.351:36	0.0018					9.9762	99.9694	99.9593	99.9627	
C	B.1.351:37	0.0018	0.0002		99.9796	99.9660	99.9728	99.9762	99.9694	99.9593	99.9627
D	B.1.351:42	0.0018	0.0002	0.0002				9.9762	99.9694	99.9593	99.9627
E	B.1.351:56	0.0						99.9627	99.9627	99.9932	99.9694
F	B.1.351:57							99.9762	99.9898	99.9593	99.9627
G	B.1.351:45	0.0							99.9660	99.9559	99.9593
H	B.1.351:48	0.0018	0.0003	0.0003				0.0003		99.9559	99.9593
I	B.1.351:49	0.0018	0.0004	0.0004	0.0004	0.0001	0.0004	0.0004	0.0004		99.9627
J	B.1.351:55	0.0018	0.0004	0.0004	0.0004	0.0003	0.0004	0.0004	0.0004	0.0004	
K	B.1.351	0.0019	0.0004	0.0004	0.0004	0.0003	0.0004	0.0004	0.0004	0.0004	0.0003
L	B.1.351:18	0.0019	0.0004	0.0004	0.0004	0.0003	0.0004	0.0004	0.0004	0.0004	0.0004
M	B.1.351:26	0.0019	0.0004	0.0004	0.0004	0.0003	0.0004	0.0004	0.0004	0.0004	0.0003
N	B.1.351:29	0.0019	0.0004	0.0004	0.0004	0.0003	0.0004	0.0004	0.0004	0.0004	0.0003
O	B.1.351:31	0.0019	0.0004	0.0004	0.0004	0.0001	0.0004	0.0004	0.0004	0.0001	0.0004
P	B.1.351:50	0.0019	0.0004	0.0004	0.0004	0.0003	0.0004	0.0004	0.0004	0.0004	0.0003
Q	B.1.351:51	0.0019	0.0004	0.0004	0.0004	0.0001	0.0004	0.0004	0.0004	0.0001	0.0004
R	B.1.351:54	0.0019	0.0004	0.0004	0.0004	0.0005	0.0004	0.0004	0.0005	0.0006	0.0005
S	B.1.351:4	0.0019	0.0004	0.0004	0.0004	0.0004	0.0004	0.0005	0.0005	0.0005	0.0005
T	B.1.351:22	0.0019	0.0004	0.0004	0.0004	0.0001	0.0004	0.0005	0.0005	0.0002	0.0004
U	B.1.351:27	0.0019	0.0004	0.0004	0.0004	0.0004	0.0004	0.0005	0.0005	0.0004	0.0004
V	B.1.351:46	0.0019	0.0004	0.0004	0.0004	0.0004	0.0004	0.0005	0.0005	0.0004	0.0003
W	B.1.351:47	0.0019	0.0004	0.0004	0.0004	0.0004	0.0004	0.0005	0.0005	0.0005	0.0005
X	B.1.351:5	0.0019	0.0005	0.0005	0.0005	0.0004	0.0005	0.0005	0.0005	0.0005	0.0004
Y	B.1.351:10	0.0019	0.0005	0.0005	0.0005	0.0004	0.0005	0.0005	0.0005	0.0005	0.0004
Z	B.1.351:11	0.0019	0.0005	0.0005	0.0005	0.0004	0.0005	0.0005	0.0005	0.0005	0.0004
AA	B.1.351:21	0.0019	0.0005	0.0005	0.0005	0.0004	0.0005	0.0005	0.0005	0.0005	0.0004
AB	B.1.351:32	0.0019	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005
AC	B.1.351:52	0.0019	0.0003	0.0003	0.0003	0.0005	0.0004	0.0004	0.0004	0.0005	0.0005
AD	B.1.351:53	0.0019	0.0005	0.0005	0.0005	0.0004	0.0005	0.0005	0.0005	0.0005	0.0004
AE	B.1.351:2	0.0020	0.0005	0.0004	0.0005	0.0004	0.0005	0.0005	0.0005	0.0005	0.0005

Figure 3. Portion of the MegAlign Pro distance table confirming placement of the experimental sample in the “Beta” strain clade.

Chapter 2

Software Challenges Faced When Identifying Viral Strains

What software challenges do virologists face when identifying viral strains?

DNASTAR recently conducted a survey of virologists about the challenges researchers face in analyzing viral genomes. Though a common concern was finding good quality template sequences with up-to-date annotations, inadequate or hard-to-use software was cited as the most significant hurdle to obtaining characterized viral genomes. Our respondents reported that the software they had tried:

1. Was command-line based and difficult to learn
2. Couldn’t automatically remove host DNA prior to assembly
3. Didn’t integrate with existing pipelines or work with multiple types of read technology (e.g., Sanger, Illumina, Nanopore)
4. Required manual data manipulation
5. Didn’t offer useful guidance on setting up or analyzing the project
6. Didn’t allow them to visualize variants once the samples were assembled against a template
7. Didn’t support automated batch extraction of the sequence portion of interest (e.g., the part of viral samples known to affect potential vaccine efficiency) and multiple alignment of only that portion of the sequence
8. Couldn’t align viral genomes over 200kb in length
9. Couldn’t calculate mutation rates

How does Lasergene software address these challenges?

Lasergene software eliminates many of the common challenges faced by researchers identifying viral strains by offering tools that are easy to use, automated, and accurate.

1. Lasergene has an intuitive interface and is easy to learn

All Lasergene applications involved in the viral analysis workflow feature easy project setup with intuitive graphic interfaces and wizards. For instance, the SeqMan NGen assembly application guides you step-by-step through choosing a workflow and uploading your read sequence and reference files. Yellow highlighting (**Figure 4**) shows you exactly where your input is needed.

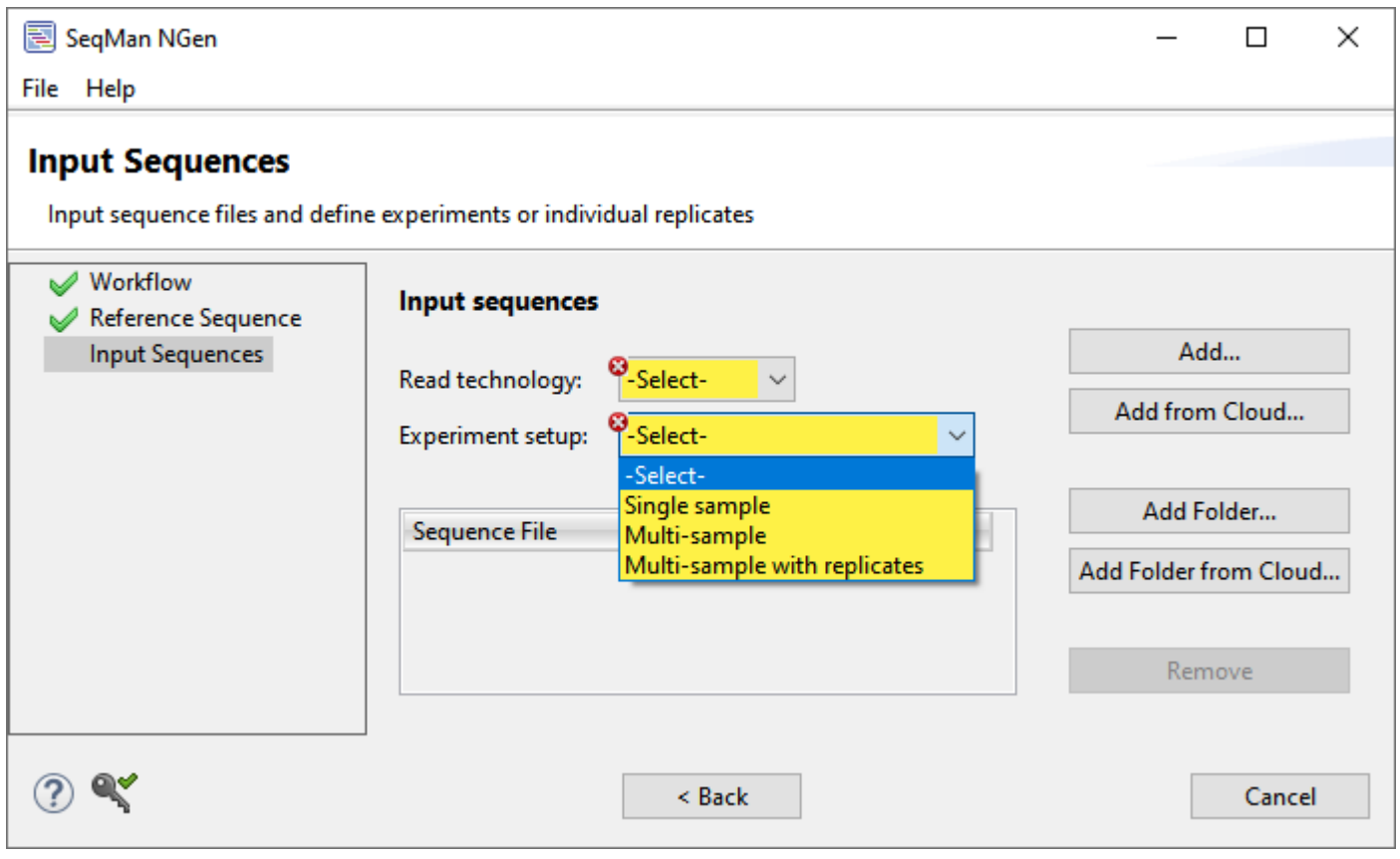


Figure 4. SeqMan NGen's "Input Sequences" screen prompts you to specify the read technology and experiment setup.

2. **Lasergene can automatically remove host DNA prior to assembly**

   SeqMan NGen's templated assembly set-up wizard lets you check a box and add one vector file or a whole folder of them (**Figure 5**). *De novo* assembly workflows allow you to specify automatic removal of common adapters, specific vectors, poor quality data, contaminants and/or repeats prior to assembly.

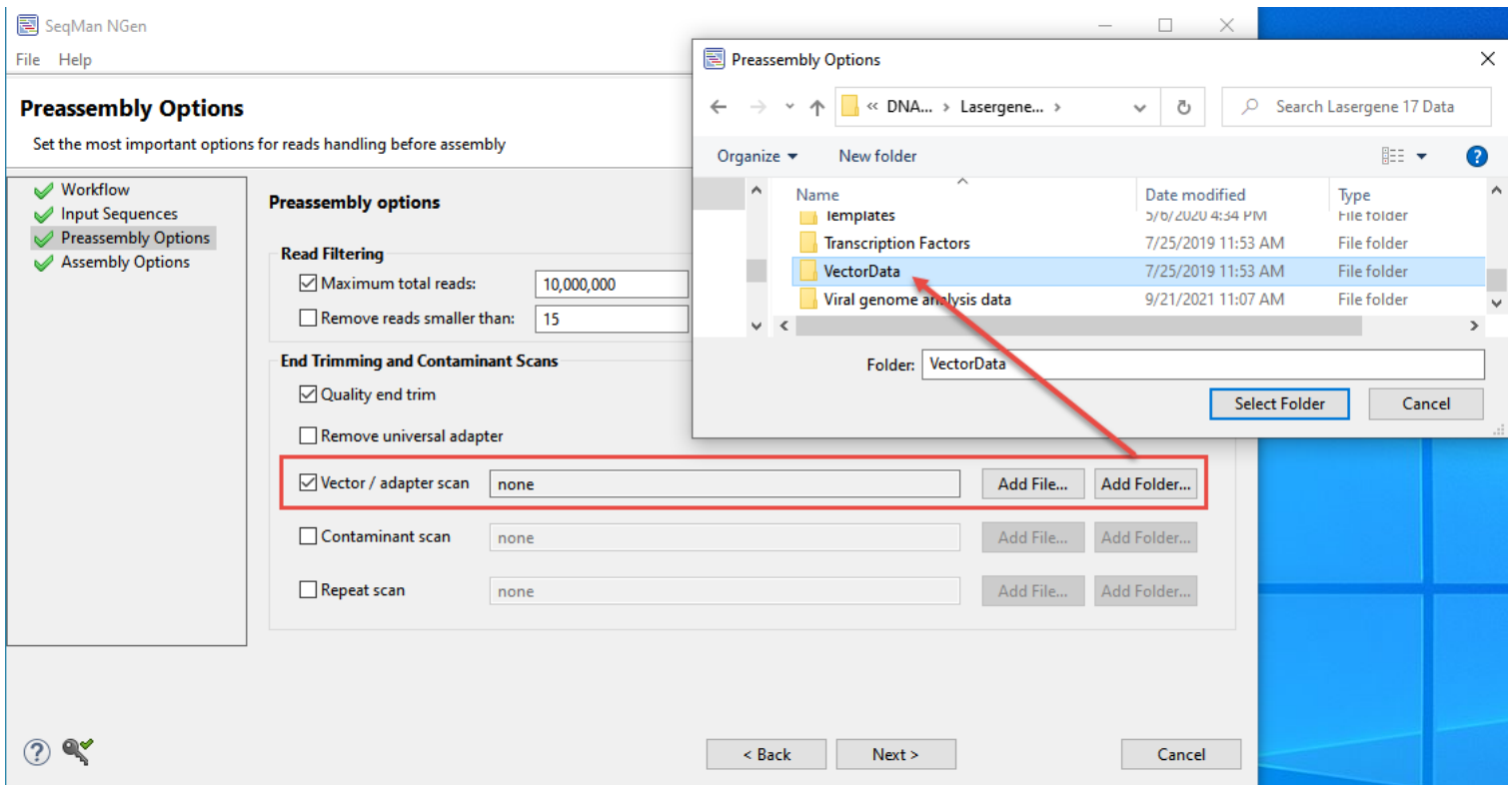


Figure 5. Specifying vector sequence removal in SeqMan NGen.

3. **Lasergene integrates with existing pipelines and works with multiple types of read technology (e.g., Sanger, Illumina, Nanopore)**

   All Lasergene applications can open and export most commonly-used sequence formats.

   Lasergene's viral genome analysis workflow supports assembly of Sanger, Illumina, Ion Torrent, PacBio and Nanopore reads. In fact, SeqMan NGen can import and assemble sequence reads with any of two dozen file extensions ranging from .abi to .zip.

   If you are starting with draft genomes, MegAlign Pro can import and multiply-align or pairwise-align draft genome sequences with any of nearly three dozen different file extensions (**Figure 6**). The application can also export a wide range of alignment and tree file formats, meaning it integrates easily with other phylogenetic analysis programs.

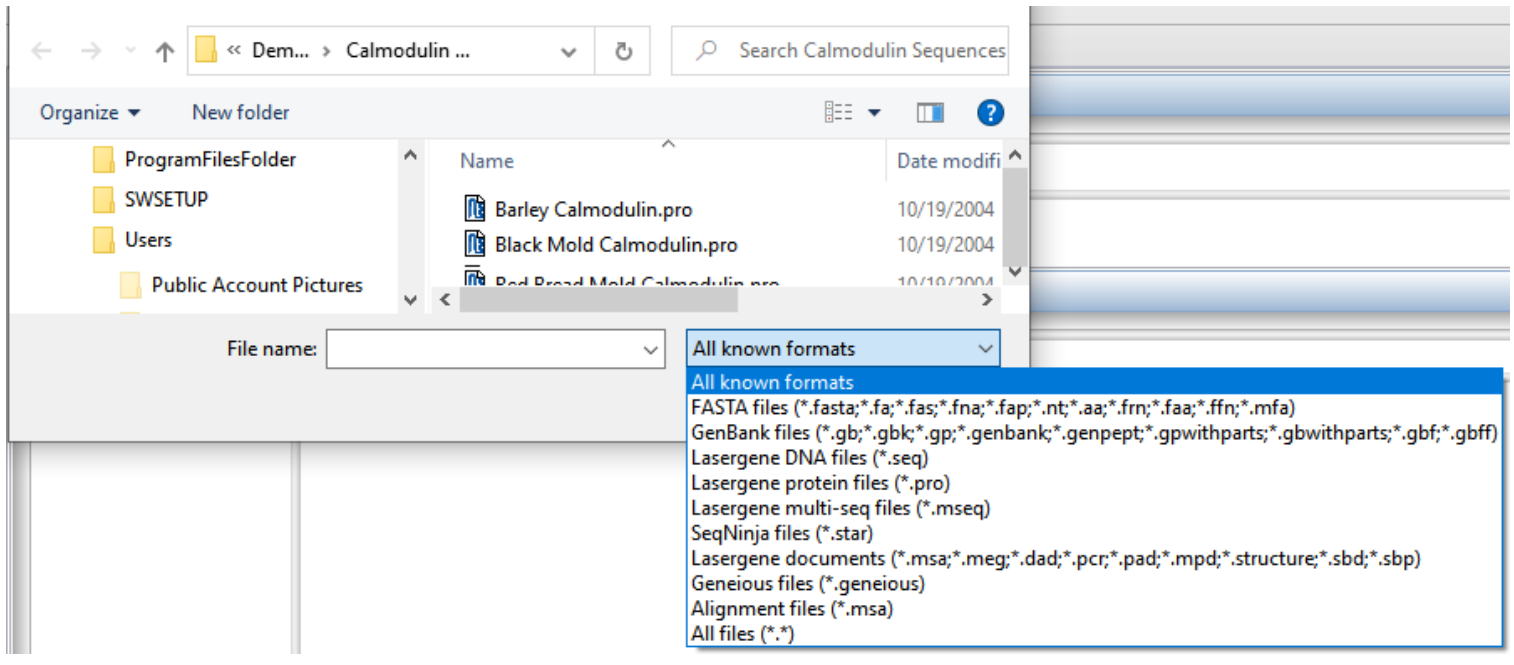


Figure 6. When adding sequences or draft genomes to MegAlign Pro for pairwise or multiple sequence alignment, numerous file formats can be uploaded.

4. Lasergene does not require manual data manipulation

Because it can recognize and use so many file types, Lasergene's viral analysis workflow does not require manual data manipulation or file reformatting.

On the other hand, several Lasergene applications in the viral analysis workflow do support editing when *you* want to edit.

- ✓ SeqMan Ultra can automatically remove of poor-quality Sanger reads prior to assembly and allows you to manually edit contigs after assembly.
- ✓ MegAlign Pro supports semi-automated end trimming prior to multiple sequence alignment.
- ✓ SeqMan Ultra and MegAlign Pro both let you export editable images to Microsoft PowerPoint.

5. After assembling reads against a template, Lasergene lets you easily visualize variants

Lasergene applications work together to create a seamless workflow. For example, once SeqMan NGen finishes assembling raw data reads, you simply click a button to begin variant analysis in SeqMan Ultra (**Figure 7**) and/or multi-sample comparison in ArrayStar.

Each application lets you specify what information you'd like to see in tables and which colors, fonts and other styles should be used in graphical displays. You can even display SNPs with specific color coding that can help you focus on variants of interest.

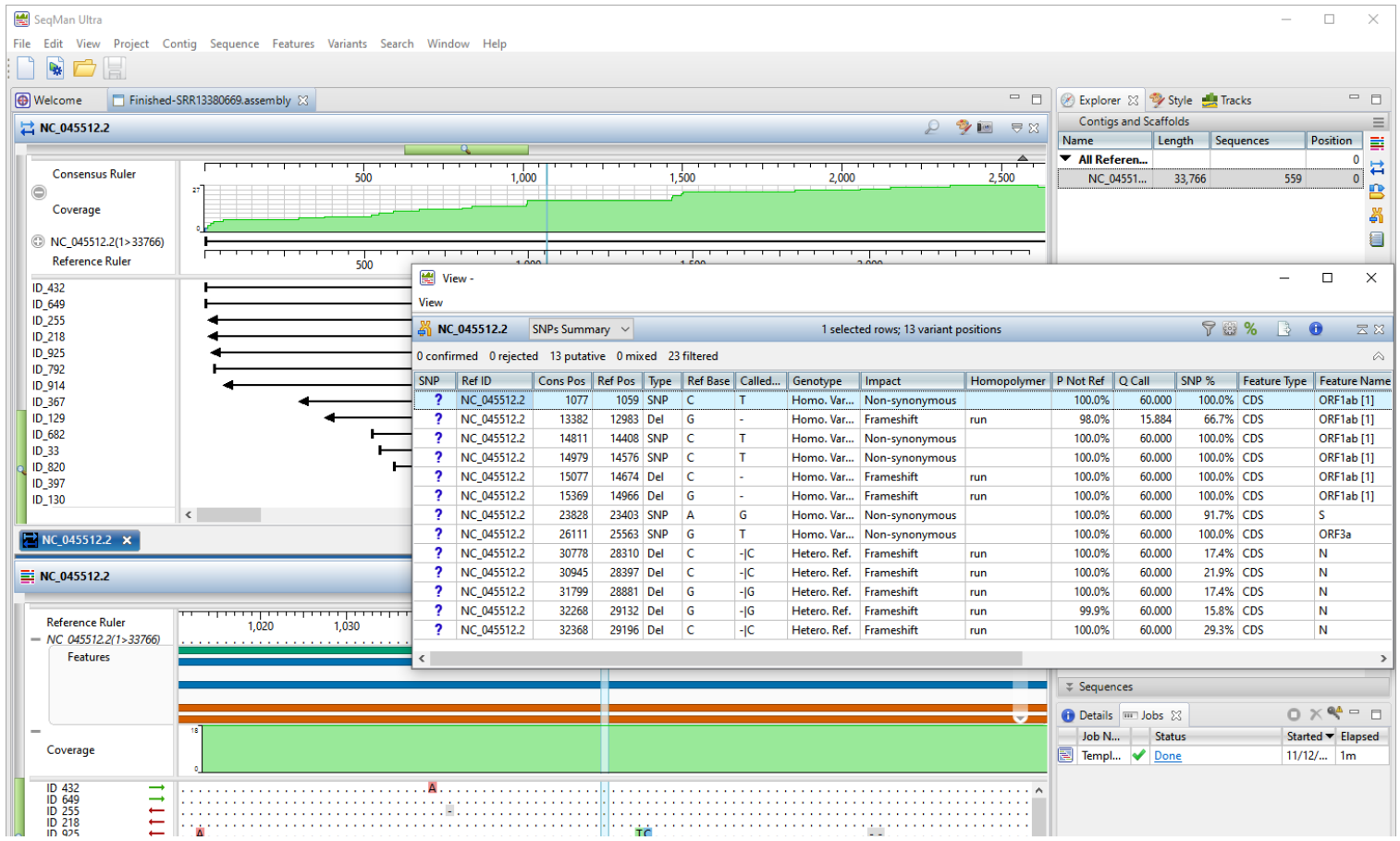


Figure 7. Viral genome data set displayed in SeqMan Ultra. The table in the middle of the image shows variants that met specific criteria. The lower portion of the image shows the location of all variants, which have been automatically color-coded according to their chemistry.

6. Lasergene offers useful guidance about how to set up and analyze the project

All Lasergene applications involved in the viral genome workflow start with a screen designed to effortlessly get you started.

Within the application, you can access the User Guide at any time by choosing the first command in the Help menu. Each User Guide has comprehensive information about every screen, view and workflow in the application, as well as how-to videos and step-by-step tutorials with downloadable data.

Don’t want to search for the User Guide topic you need? Many application screens, views, dialogs and popups have a blue “info” button that takes you straight to the relevant page of the User Guide. You can also hover over tools, table cells and other parts of the application window for helpful tool tips (**Figure 8**).

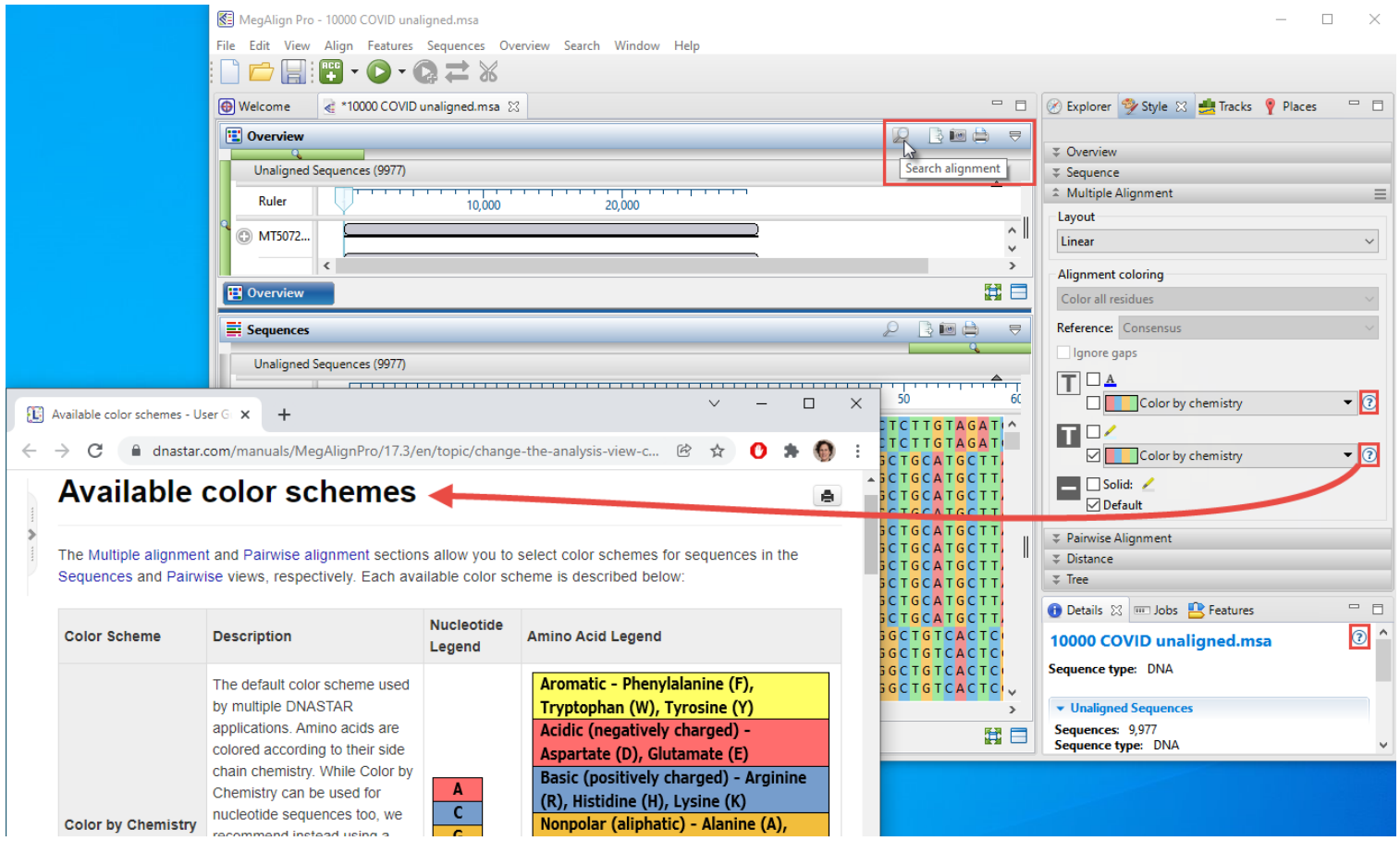


Figure 8. The MegAlign Pro window showing a tooltip (boxed), three info buttons and the user guide page accessed by clicking one of those buttons.

Need personalized help? Licensed Lasergene users can get live support by email, phone, or a one-on-one online training session.

7. Lasergene can batch-extract and align only the important parts of viral sequences

With Lasergene, you can automatically select a portion of each sequence in the project (**Figure 9**) and align just the selected portions of the sequences.

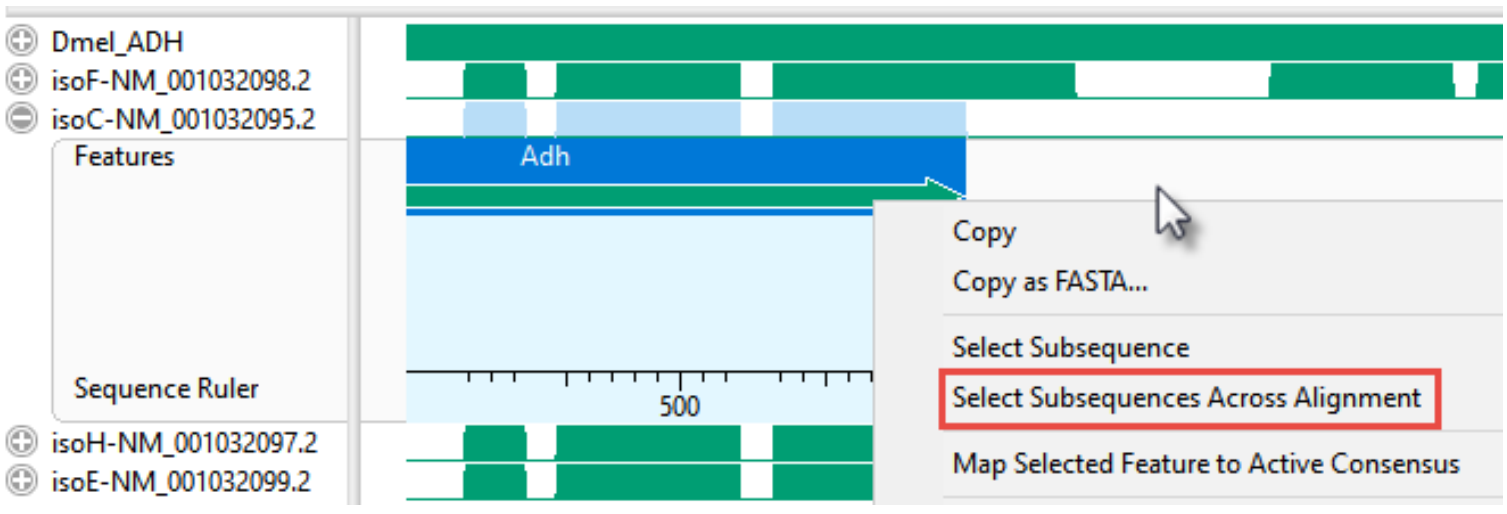


Figure 9. Selecting the portion of sequence corresponding to the Adh gene using MegAlign Pro.

8. Lasergene can align viral genomes over 200kb in length

Lasergene supports multiple alignment with the MAFFT v.7 algorithm, which can align approximately twenty sequences of one million bp in length. As sequences get shorter, MAFFT supports aligning more of them. For instance, by selecting the MAFFT algorithm within MegAlign Pro, we have successfully aligned 10,000 SARS CoV-2 genomes (30kb) in just two minutes.

9. Lasergene can calculate mutation rates

Lasergene provides an easy way to measure the mutation rate of a viral strain. Collect samples over time (e.g., SARS-CoV-2 strains from various stages of the pandemic) and use the MegAlign Pro application to perform a multiple sequence alignment. In the Distance table, specify that Distance values be displayed. These values can be used to calculate the rate of mutation.

Chapter 3

Case Study: Rapid Analysis of Viral Strains During the COVID-19 Pandemic

Case Study: Rapid analysis of viral strains during the COVID-19 pandemic

While many researchers can find commercial or open-source software solutions that are sufficient to perform their viral genome analysis, sometimes the sequencing protocols evolve more quickly than the software. This has been especially true during the COVID-19 pandemic, when innovative technologies were developed to support rapid sequencing of a high volume of patient samples. One example is the ARTIC Network, which employs MinION sequencing to support the creation of real-time epidemiological data.

DNASTAR recently partnered with Impetus Bioscience in Germany to develop software to analyze Oxford Nanopore reads generated through the ARTIC protocol. The group was using this protocol to sequence and identify SARS-coV-2 strains in the German population but needed a comprehensive software solution to get from sequencing data to useful results. The team at Impetus didn't have the resources to develop or modify bioinformatics command line tools, and time was of the essence.

Working closely with their team, DNASTAR developed a software solution that enabled researchers to input sequencing data, perform assembly and visualize results without any special bioinformatics expertise. With the help of DNASTAR software, the team was able to quickly and easily analyze thousands of samples.

"In our Lab we are analysing COVID 19 samples by PCR. Since Feb 2021 5% of COVID 19 positive samples must be sequenced in Germany. Therefore, we used the Oxford Nanopore technic. The sequencing technic runs very well, but for the data analysis you cannot get a commercial software package from Oxford Nanopore. In the community you can find some protocols, but these protocols are not plug and play, if you are not familiar with bioinformatics (I mean writing programs). So, we looked for an easy-to-handle analysis tool. With the DNASTAR we can analyse the data set for 1 Sample in 5-10 min.

I believe many people have a problem with handling the Oxford Nanopore data, because the sequencing itself is easy to handle in every lab, which is a big advantage (time and money). But many small labs like use have no bioinformatics for data analysis. With DNASTAR this problem can be solved."

-Dr. Olav Grundmann, Manager R&D at Impetus Bioscience

Conclusion

Human health depends upon our ability to control or eradicate viruses that affect us and the foods we eat. Characterization of viral genomes is critical to the investigation of how viruses spread and mutate. In the face of the COVID-19 pandemic, virology research has gained both appreciation and increased governmental support around the world.

Improvements in read sequencing technology and analysis software have made the identification and characterization of viruses an everyday event. Nevertheless, our recent survey showed that researchers found several areas in which they would like to see analysis software improve. From non-existent user interfaces to missing features—such as the inability to remove host DNA or align long sequences—poor software can negatively impact the ability to correctly identify and analyze viral strains.

DNASTAR has addressed each of these issues with a fully integrated workflow for viral genome analysis and an intuitive interface that provides guidance each step of the way. We believe Lasergene provides the easiest, fastest and most accurate software solutions for identifying and characterizing viral strains.

Want to try this workflow for yourself?

If you'd like to try Lasergene's integrated viral genome assembly and analysis workflow, we invite you to download a fully-functional 14-day free trial.

Download Free Trial

Once you're installed the free trial, we recommend starting with our illustrated step-by-step tutorial, Templated long-read workflow (ARTIC), which has downloadable data. This is an expanded version of the example used earlier in this ebook and starts with sequence assembly.

See the Tutorial

CONTACT US

To learn more about how Lasergene can fit into your virology research, email or call us at one of the numbers listed below.

www.dnastar.com
Email: sales@dnastar.com
Phone: 608.258.7420
Toll Free in the U.S. and Canada: 866.511.5090
Call free in the UK: 0.808.271.1041
Call free in Germany: 0.800.182.4747