Developing a Pipeline for Viral Genome Closure

Jennifer M. Zaborsky, Jeffrey D. Sparenborg, Jessica B. Hostetler, Jeff N. Sitz, Naomi Sengamalay, Vik Subbu, Torrey L. Gallagher, Larry J. Overton, Xinyue Liu, Martin Shumway, Kristine Jones, Luke J. Tallon, David Spiro.

The J. Craig Venter Institute, Rockville, MD, USA

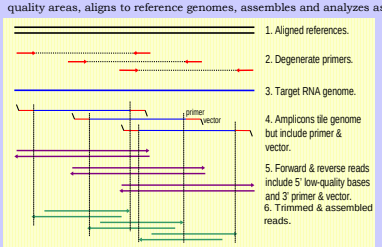
FF0077

Introduction

The Viral Genomics and Closure group at JCVI have developed a high throughput pipeline to sequence the Influenza A genome. The extensive genetic variation of Influenza and the large number of viral samples sequenced has led the Closure and Software team to develop new finishing techniques and software to match the unique needs of this pipeline. Closure uses a software suite, Elvira, to build contigs from amplicon-based PCR reads. Samples are tracked through the pipeline in the Closure Task Manager, where laboratory procedures tasks assigned and evaluated until the sample is finished. A combination of standard and custom primers is used in the finishing process to handle the highly varied Influenza genome and maintain the efficiency of the pipeline. To date, JCVI's pipeline has published over 2000 complete Influenza genomes in Genbank. The pipeline has recently been adapted to sequence and finish Influenza B, Coronavirus, and Rhinovirus samples. The evolution of the pipeline reflects the growth of interest in viral genomics and the need for finishing techniques that efficiently close highly varied genomes.

Amplicon Based Sequencing and Assembly

JCVI viral sequencing projects target closely related strains in large numbers. The projects rely on a reference genome sequence which is used to generate PCR primers for the related strains. Primers are used to tile the target genome by short amplicons. Once sequenced, we turn to Elvira for assembly. Elvira is a software suite developed for the unique challenge of assembling large numbers of PCR-amplified viral genomes. It tracks samples, trims primer sequence and low quality areas, aligns to reference genomes, assembles and analyzes assembly



1. Aligned references.
2. Degenerate primers.
3. Target RNA genome.
4. Amplicons tile genome but include primer & vector.
5. Forward & reverse reads include 5' low-quality bases and 3' primer & vector.
6. Trimmed & assembled reads.

Sequencing strategy. Use closely related reference genome sequences (1) to generate short degenerate primers (2). Perform RT-PCR on target genome (3) to produce amplicons (4). Each amplicon is shorter than a sequencing read, and collectively they tile the genome. Sequence each amplicon once in each direction to produce overlapping mate pairs (5). Use software to trim every read, removing 5' low-quality sequence and 3' primer & vector sequence and assemble the full set of mated reads (6). Generally, the genome end sequence is not recovered.

By giving Elvira a list of sample identifiers, we can assemble multiple samples in tandem. An assembly output is produced for each viral genome with useful information about assembly quality. This can then be added to the CTM log for each sample.

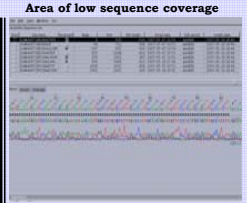
Elvira Assembly Output for Sample 19234

```
Assembly began: 2007-05-21 07:15:54
Assembly completed: 2007-05-21 07:20:53
Elapsed time: 4:59 (299 seconds)
Assembly generated 16 exceptions
-------------------------------------------------------------
[ 19234-PB1 ] - Missing 5' end coverage (23 bases).
[ 19234-PB1 ] - Missing 3' end coverage (21 bases).
[ 19234-NS ] - Missing 5' end coverage (16 bases).
[ 19234-NS ] - Missing 3' end coverage (21 bases).
[ 19234-NP ] - Missing 5' end coverage (14 bases).
```

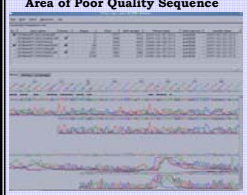
Initial Sample Analysis

Good Coverage and Quality

Area of low sequence coverage



Area of Poor Quality Sequence

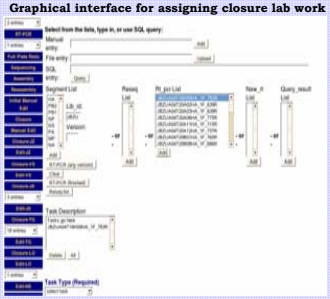


Once the samples have been assembled, each one is manually inspected by a member of the closure team. In order to be considered finished, all regions of the genome must have at least 2x high quality sequence coverage as well as 2x amplicon coverage. Samples that have areas of poor quality sequence or areas of low coverage require additional laboratory work to be performed. Samples that are finished are put through and validation procedure and approved for submission to Genbank.

Closure Task Manager

Sample homepage with tracking information

Graphical interface for assigning closure lab work



All work done for a sample is documented by the team members

The Closure Task Manager is essential for tracking the status of viral samples within the pipeline. Samples are tracked from the time they arrive at JCVI through publication in GenBank.

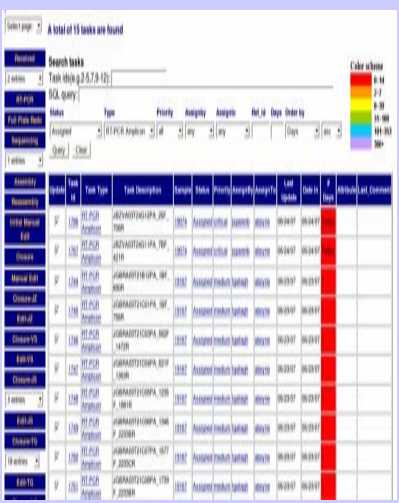
Every member of the project team can view the history of a sample, including past statuses, and work performed by other members of the project.

Any work that is performed on a sample is documented in this reference log. This allows every member of the project team the ability to work on any sample in the pipeline, regardless of it's history.

Closure work is assigned and tracked through the closure task manager. Specific lab tasks can be selected from a drop down menu containing all possible options.

RT-PCR

A list of all RT-PCR closure tasks that are currently assigned

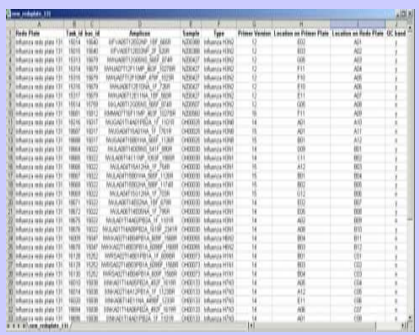


Ambiguous regions and gaps in the genome are often resolved by performing RT-PCR of specific, individual amplicons that span the problem region. Tasks are assigned by a member of the closure team.

When a sufficient number of closure tasks exist, the viral lab group runs a script designed to generate an automatic pick list of these sample specific RT-PCR tasks, based on the information stored in the project database. The script outputs information linking the assigned amplicons to the correct Influenza samples, providing improved tracking and increased efficiency in performing laboratory tasks.

After the lab group completes the tasks, the excel sheet is updated and the samples are prepared for sequencing. The closure group monitors the status of the sequencing and assembles the new data as it arrives.

The file used by the viral lab group to complete closure tasks



Unique Solutions for Finishing Viral Genomes

Task attributes allow valuable information to be tracked and viewed by all project members

Task details for a sample are easily communicated between groups

Customized primers are rapidly designed by closure personnel when primers from the pre-designed primer blocks fail to finish the genome. Primers are targeted for problem regions of the particular genome.

Once the primers are designed, closure tasks are entered into the CTM. These tasks require that specific attributes, or descriptors, be entered as a part of the task. This method ensures that the appropriate information is conveyed to the viral lab group, who will complete the RT PCR with these primers. Using this system, the closure group can track the success or failure of a primer as an attribute that is stored in the project database.

All custom primers that successfully amplify regions of the genome are added to a primer block and incorporated permanently into the finishing pipeline. Once included in a primer block, these custom primers play the same role as the pre-designed primers do to quickly finish influenza genomes. Custom primers add a flexibility component but also support the consistent stream-lined philosophy of the sequencing pipeline.

A new primer set was developed from the successful closure primers

A	H2 1-626	H2 1-681	H2 626-1131	H2 626-1197	H2 1091-1773	H2 1146-1780	H3 1-400	H3 1-600	H3 924-1769	H3 939-1769	H3 980-1464	H3 1032-1480
B	H3 1100-1765	H3 1400-1765	H3 1400-1765	H5 19-214	H6 366-863	H6 440-896	H6 450-880	H6 766-1167	H10 1-279	H10 1-639	H10 1B-827	H10 682-1245
C	H10 1061-1728	H11 1-618	H11 1-1122	H11 1-1144	H11 1082-1780	H11 1384-1759	H12 1-823	H12 1046-1737	N1 1-510	N2 1-488	N2 1-450	N2 1-517
D	N2 960-1466	N2 962-1460	N2 980-1466	N3 468-1233	N6 412-1042	N6 501-983	N6 915B-1401	N7 1-955	N7 2-103	N7 220-680	N7 453-1248	N8 1-235
E	N8 1-545	N8 23-372	N8 480-971	N8 500-1122	N8 512-1030	N8 523-1015	N8 1104-1460	N9 1-378	N9 1-644	N9 1-637	N9 120-644	N9 368-909
F	N9 457-998	N9 468-988	N9 499-988	N9 499-1049	N9 740-1390	PB2 1-600	PB2 2-254	PB2 500-1208	PB2 532-1400	PB2 593-1210B	PB2 640-1200	PB2 630-1230
G	PB2 1375-1800	PB2 1393-2000	PB2 1415-2000	PB1 535-1087	PB1 600-1385	PB1 980-1615	PB1 1028-1528	PB1 1033-1529	PB1 1670-2325	PB1 1710-2341	PA 7B-421	PA 25-706
H	PA 613-1409	PA 1410-1775	PA 1450-1775	PA 1500-2233	PA 1730-2233	PA 1790-2233	MP 1-890	MP 500-1000	MP 620-1080	MP 1-468	MP 280-790	MP 776-1027

Acknowledgements

The authors would like to express their gratitude to Naomi Sengamalay and the other members of the Viral genomics group at JCVI for their help with research and development of many of the aspects of the Viral closure pipeline. In addition, we would like to thank Martin Shumway, Jeff Sitz, and the other members of the Software group at JCVI, who have continually worked to develop novel software tools that meet the unique needs of the Viral genomics projects.