Annotation

Decorating the genome sequence.

J. Craig Venter™ INSTITUTE

Where we are

- 13:30-14:00 – Primer Design to Amplify Microbial Genomes for Sequencing
- 14:00-14:15 – Primer Design Exercise
- 14:15-14:45 – Molecular Barcoding to Allow Multiplexed NGS
- 14:45-15:15 – Processing NGS Data – de novo and mapping assembly
- 15:15-15:30 – Break
- 15:30-15:45 – Assembly Exercise
- **15:45-16:15 – Annotation**
- 16:15-16:30 – Annotation Exercise
- 16:30-17:00 – Submitting Data to GenBank

J. Craig Venter™
INSTITUTE

What is Annotation?

- **Webster’s definition of “to annotate”:**
  - “to make or furnish critical or explanatory notes or comment”
- **Elements of the annotation process**
  - gene finding
  - gene model curation
  - functional assignment
- **What this includes for genomics**
  - gene product names/symbols
  - functional characteristics of gene products
  - physical characteristics of gene products
  - overall metabolic profile of the organism

J. Craig Venter™ INSTITUTE

Annotation – Bacterial Genome

- **Find ORF (Open Reading Frame)**
  - Defined by the absence of a translational “stop” codon
  - 3 “stops” in bacteria: TAA, TAG, TGA
  - Coding sequence goes from “stop” to “stop”

- **Find Protein Coding Gene:**
  - requires translational “start” codon
  - 3 “starts” in bacteria: ATG, GTG, TTG
  - coding sequence goes from “start” to “stop”

- **Assign function to genes**
  - based on homology to experimentally verified proteins, HMM, protein families etc

- **Find non-coding genes**

Differentiating between random ORFs and genes is the goal in the gene finding process.

J. Craig Venter™ INSTITUTE

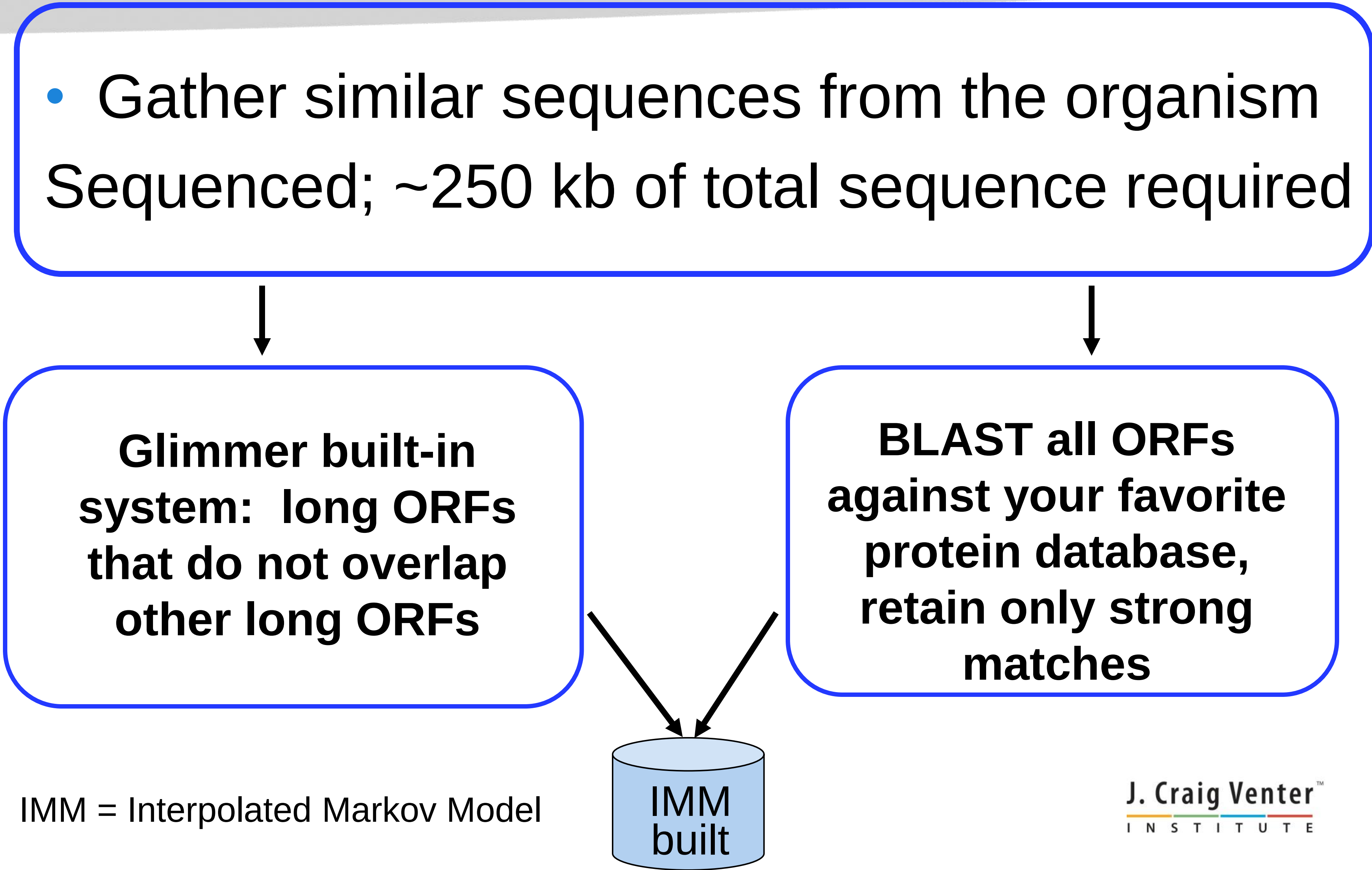
Gene Finding – Bacterial Genome

- Glimmer (**G**ene **L**ocator and **I**nterpolated **M**arkov **M**odeler) software tool – most commonly used.

- Glimmer uses Interpolated Markov Models (IMMs) to predict which ORFs in a genome contain real genes.

- Glimmer compares the nucleotide patterns it finds in a training set of “known real genes” to the nucleotide patterns of the ORFs in the whole genome. ORFs with patterns similar to the patterns in the training genes are considered real.

Gene Finding – With Glimmer



- Gather similar sequences from the organism

Sequenced; ~250 kb of total sequence required

Glimmer built-in system: long ORFs that do not overlap other long ORFs

BLAST all ORFs against your favorite protein database, retain only strong matches

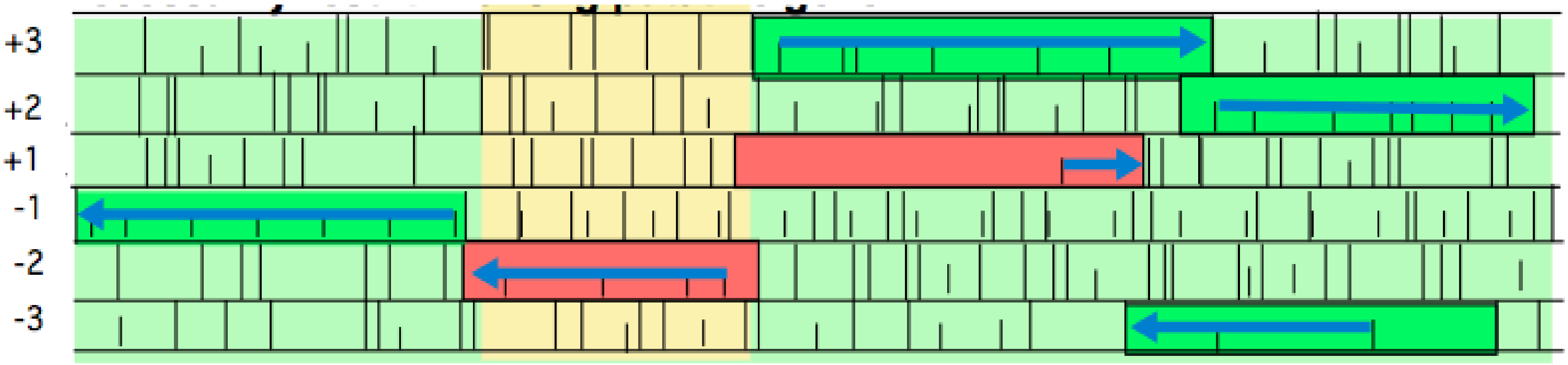
IMM built

IMM = Interpolated Markov Model

J. Craig Venter™ INSTITUTE

Candidate Vs. Final ORF – With Glimmer

Long vertical lines represent stops (TAA, TAG, TGA) Short vertical lines represent starts (ATG, GTG, TTG)



- Green ORFs scored well to the model, and was chosen by Glimmer as genes.
- Red ORFs scored less; therefore rejected.
- ORFs in the area of lateral transfer, although real genes, often will not be chosen since the nucleotide pattern doesn’t match that of the model

J. Craig Venter™ INSTITUTE

Assigning Function to Proteins

- **Homology searching**
  - Compare sequence of unknown function to those of known function for similarity
  - Relies on the assumption that shared sequence implies shared function
- **Pairwise alignments**
- **Multiple alignments**
- **Match to protein families (Pfam, NCBI's protein Clusters, SCOP)**
- **Experimental evidence (CharProt)**
- **Other sources (Uniprot, SCOP, PDB etc.)**

J. Craig Venter™ INSTITUTE

Assigning Function to Proteins

- **Protein name:** phosphomethylpyrimidine kinase
- **Gene symbol:** thiD
- **EC number (From Enzyme Commission):** 2.7.4.7
- **GO Terms (From Genome Ontology):**

Process: GO:0009228

Function: GO:0008972

Componant: GO:0005737

Prediction Of RNA

- **RNAs involved in protein synthesis**
  - Ribosomal RNA (rRNA)
  - Transfer RNA (tRNA)
  - Transfer-messenger RNA (tmRNA)
- **RNAs involved in post-transcriptional modification/DNA replication**
  - Small nucleolar RNA (snoRNA)
  - Ribonuclease P (Rnase P)
- **Regulatory RNAs**
  - Cis-regulartory elements (cis-reg)
  - Riboswitches

J. Craig Venter™ INSTITUTE

Annotation Of Viral Genomes Using VIGOR

- **Find ORF (Open Reading Frame)**
  - Using homology based **Vi**ral **G**enome **O**RF **R**eader
  - VIGOR is developed by JCVI and is customized to each virus type
  - Freely available to use

- **Find Protein Coding Gene:**
  - Polyprotein and mature peptides are annotated.
  - Structural annotation takes into account exon and intron sizes, overlapping genes and mutually exclusive genes
  - Exceptions are handles well such as translation exception, non-canonical splicing, RNA editing, ribosomal slippage, stop codon read-through etc.

- **Assign function to coding genes**
  - Vigor assigns gene symbol and protein names as part of annotation

J. Craig Venter™ INSTITUTE

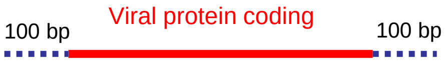
Algorithm of VIGOR to Predict Protein Coding Sequences

load sequences in FASTA format.
Divide input sequences into groups with maximal length of 5 Mb.

BlastX searching against custom DB consisting of viral protein sequences.

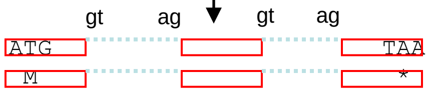
```
Query: 1667 VPHNLPDSQKNCFVTSLFRCSLGNEQRTGLRISLLEMINGYCMLAIPGKKSGSRSRKPR 1491
            VPHNL DSQK+CFVTSLFRCSLGNEQRTGLRISLLEMINGYCMLAIPG KSGSRSRKPR
Sbjct: 121  VPHNLSDSQKSCFVTSLFRCSLGNEQRTGLRISLLEMINGYCMLAIPGIKSGSRSRKPR 179
```

Define the regions encoding viral proteins with 100 bases flanking regions.



Detect ATG, stop codon and exons of ORFs in protein coding regions.

Identify the alternative splicing, i.e. in NS2, and M2 in flu genome.



Define the ribosome slippage site, i.e., in coronavirus genome and SARS genome

Generate cDNA sequences.

Extract the coding sequences and translate them into peptide sequences.

cleave the polypep into mature peptides, i.e., in Rhinovirus genome, orf1a and orf1ab in SARS coronavirus genome.

Functional Annotation of the predicted proteins based on sequence similarity

Sequence alignment between predicted protein and homolog in DB.

Input

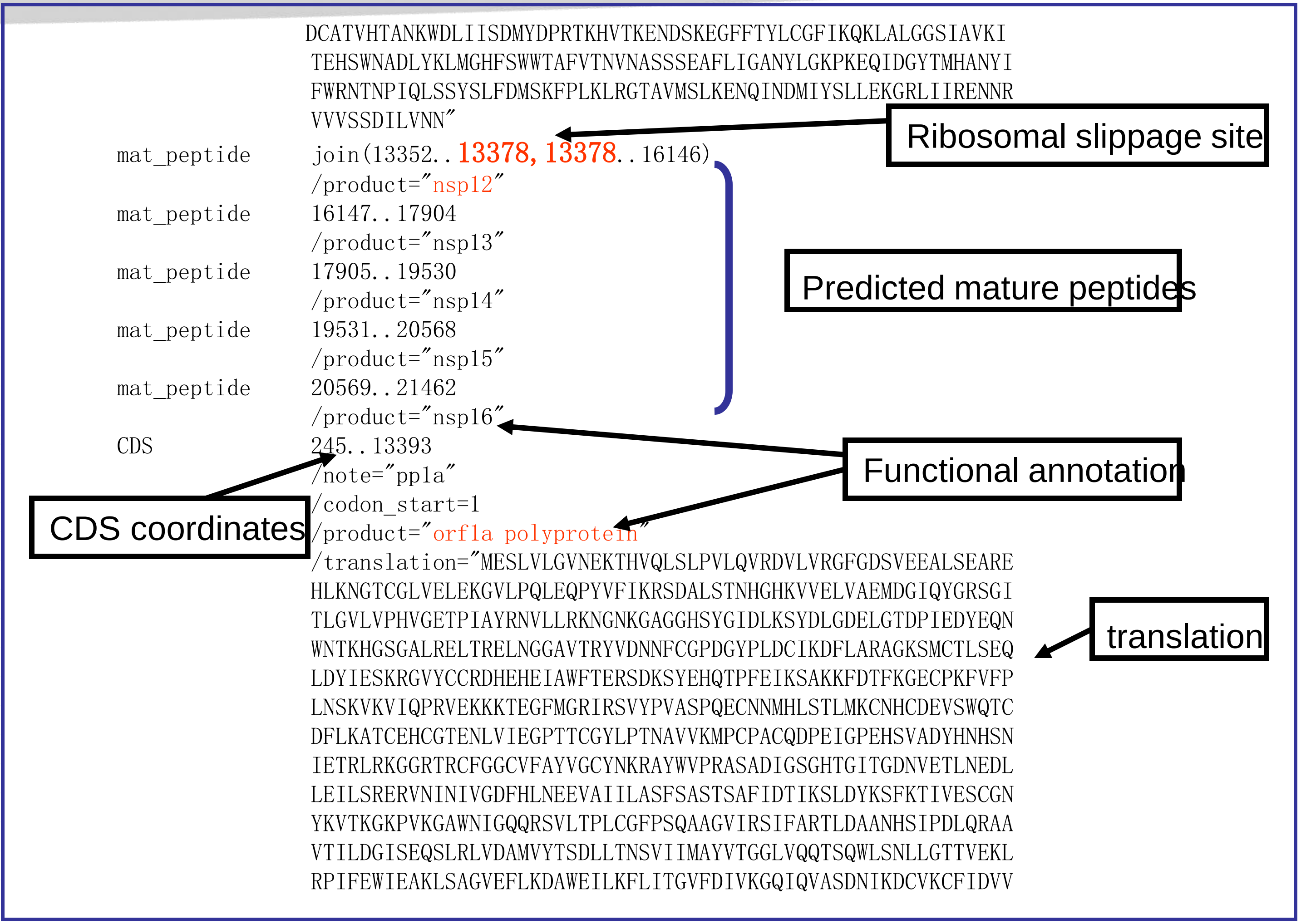
Genomic seq

- Predicted peptide
- Predicted cDNA
- Protein alignment
- .tbl file

Output

J. Craig Venter™
INSTITUTE

VIGOR Prediction of a SARS Coronavirus Genome



```
                     DCATVHTANKWDLIISDMYDPRTKHVTKENDSKEGFFTYLCGFIKQKLALGGSIAVKI
                     TEHSWNADLYKLMGHFSWWTAFVTNVNASSSEAFLIGANYLGKPKEQIDGYTMHANYI
                     FWRNTNPIQLSSYSLFDMSKFPLKLRGTAVMSLKENQINDMIYSLLEKGRLIIRENNR
                     VVVSSDILVNN″
     mat_peptide     join(13352..13378,13378..16146)
                     /product=″nsp12″
     mat_peptide     16147..17904
                     /product=″nsp13″
     mat_peptide     17905..19530
                     /product=″nsp14″
     mat_peptide     19531..20568
                     /product=″nsp15″
     mat_peptide     20569..21462
                     /product=″nsp16″
     CDS             245..13393
                     /note=″pp1a″
                     /codon_start=1
                     /product=″orf1a polyprotein″
                     /translation=″MESLVLGVNEKTHVQLSLPVLQVRDVLVRGFGDSVEEALSEARE
                     HLKNGTCGLVELEKGVLPQLEQPYVFIKRSDALSTNHGHKVVELVAEMDGIQYGRSGI
                     TLGVLVPHVGETPIAYRNVLLRKNGNKGAGGHSYGIDLKSYDLGDELGTDPIEDYEQN
                     WNTKHGSGALRELTRELNGGAVTRYVDNNFCGPDGYPLDCIKDFLARAGKSMCTLSEQ
                     LDYIESKRGVYCCRDHEHEIAWFTERSDKSYEHQTPFEIKSAKKFDTFKGECPKFVFP
                     LNSKVKVIQPRVEKKKTEGFMGRIRSVYPVASPQECNNMHLSTLMKCNHCDEVSWQTC
                     DFLKATCEHCGTENLVIEGPTTCGYLPTNAVVKMPCPACQDPEIGPEHSVADYHNHSN
                     IETRLRKGGRTRCFGGCVFAYVGCYNKRAYWVPRASADIGSGHTGITGDNVETLNEDL
                     LEILSRERVNINIVGDFHLNEEVAIILASFSASTSAFIDTIKSLDYKSFKTIVESCGN
                     YKVTKGKPVKGAWNIGQQRSVLTPLCGFPSQAAGVIRSIFARTLDAANHSIPDLQRAA
                     VTILDGISEQSLRLVDAMVYTSDLLTNSVIIMAYVTGGLVQQTSQWLSNLLGTTVEKL
                     RPIFEWIEAKLSAGVEFLKDAWEILKFLITGVFDIVKGQIQVASDNIKDCVKCFIDVV
```

Ribosomal slippage site, functional annotation are highlighted. The prediction was converted into gbf format.

J. Craig Venter™ INSTITUTE

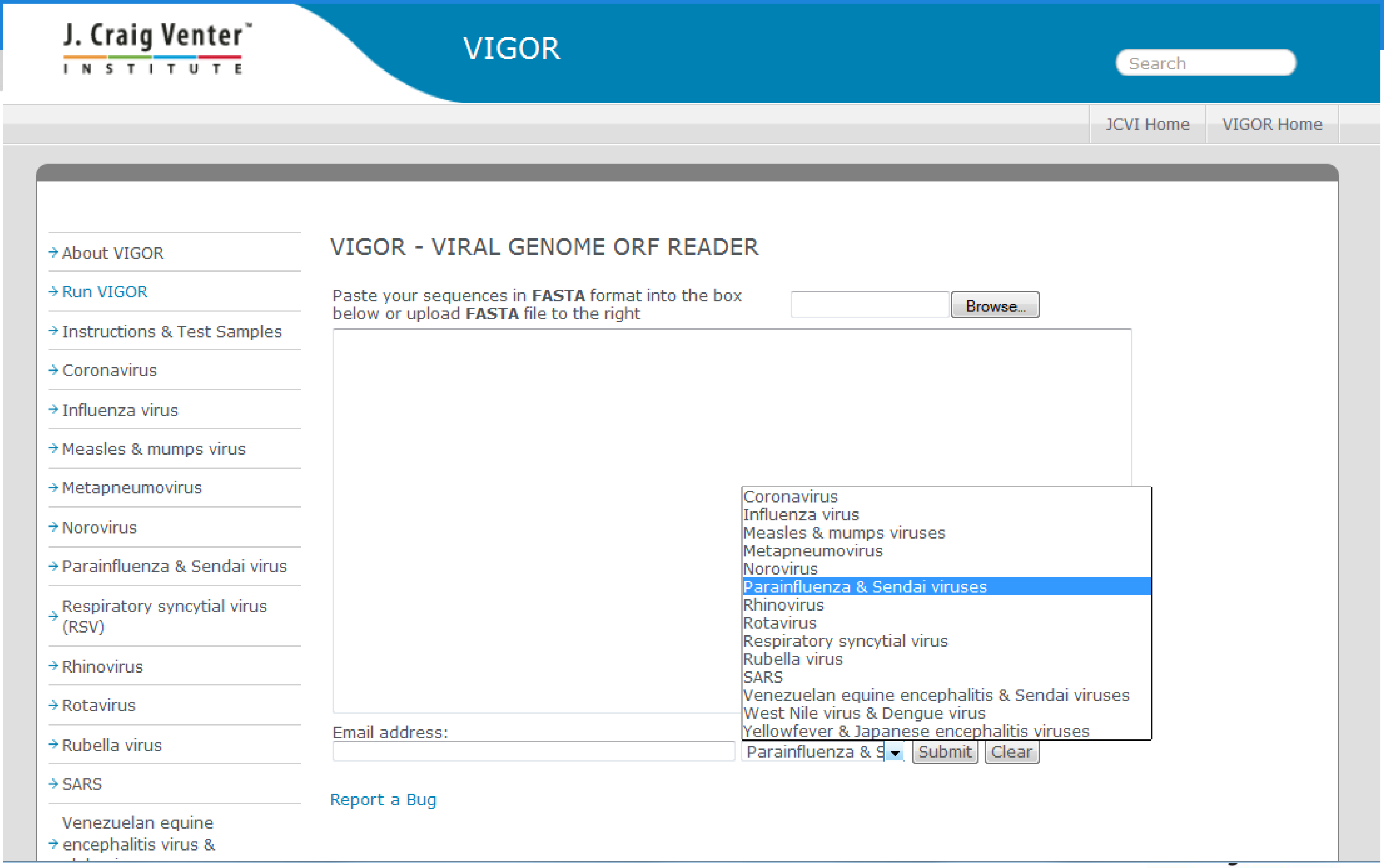
List Of Virus VIGOR can Annotate

Virus Name	Special features	Mature peptide prediction
Influenza	Splicing	N/A[1]
Rotavirus	Gene overlapping	N/A
Coronavirus	Ribosomal slippage	
SARS	Ribosomal slippage	Yes
Rhinovirus	No	Yes
MPV	Gene overlapping	N/A
Measles Virus& Mumps Viruses	RNA editing	N/A
Norovirus	No	Yes
Parainfluenza & Sendai Virus	RNA editing, Ribosomal shunting	N/A
RSV	Gene overlapping	N/A
Rubella Virus	No	Yes
VEEV & Alphavirus	Stop codon leakage,	Yes
YFV and JEV	No	Yes
West Nile Virus & Dengue Virus	Frame-shift	Yes

[1] Not Applicable

J. Craig Venter™
INSTITUTE

VIGOR Web Interface (www.jcvi.org/vigor)



VIGOR - VIRAL GENOME ORF READER

Paste your sequences in **FASTA** format into the box below or upload **FASTA** file to the right

Email address:

Parainfluenza & S

Report a Bug

Annotation Resources

- **Free Microbial Annotation for the community:** Submit your genome fasta file and you will receive your annotated genome
- JCVI annotation Service is now deprecated but continuing efforts are being carried out by the Institute of Genome Sciences (IGS).

http://ae.igs.umaryland.edu/cgi/index.cgi

- **Free Viral Annotation for the community:**

Upload your genome fasta file and you will receive output file with annotation

http://www.jcvi.org/vigor/submission.php

More General Resources

- **Swiss-Prot**
  - All entries manually curated based on experimental evidence
  - Annotation includes links to references, coordinates of protein features, links to cross-referenced databases
- **TrEMBL**
  - Entries have not been manually curated
- **UniProt** http://www.uniprot.org/

  Swiss-Prot + TrEMBL
- **Pfam**: http://pfam.sanger.ac.uk/
- **Enzyme Commission**: http://enzyme.expasy.org/
- **NCBI**: http://www.ncbi.nlm.nih.gov/
- **KEGG Pathway database:** http://www.genome.jp/kegg/kegg1.html
- **NCBI Protein Clusters**: http://www.ncbi.nlm.nih.gov/proteinclusters
- **Gene Ontology (Amigo)**: http://amigo.geneontology.org/cgi-bin/amigo/go.cgi

J. Craig Venter™ INSTITUTE

Resources For RNA Prediction

RNAs are found using a combination of the following programs/methods:

- tRNAscan (Lowe/Eddy, 1997)
- http://lowelab.ucsc.edu/tRNAscan-SE/

- Aragorn (Laslett/Canback 2004)
- http://mbio-serv2.mbioekol.lu.se/ARAGORN/

- BLAST searches
- http://blast.ncbi.nlm.nih.gov/Blast.cgi

- Rfam (Sanger, WashU)
- http://rfam.sanger.ac.uk/

J. Craig Venter™ INSTITUTE